LECTURE NOTES FOR MA20217 (ALGEBRA 2B)

Abstract. This course introduces abstract ring theory in order to establish in full the structure theorem for linear operators on a finite dimensional vector space.

Contents

What's Algebra 2B about?

An $n \times n$ matrix A with coefficients in $\mathbb{C}$ defines a linear map $f_A \colon \mathbb{C}^n \to \mathbb{C}^n$ by sending each column vector $v \in \mathbb{C}^n$ to the column vector $Av \in \mathbb{C}^n$ obtained by multiplying the matrices together. Matrices arise throughout mathematics, but they do so precisely because they act on vectors, and this action is what the linear map describes.

Given an invertible $n \times n$ matrix P, the $n \times n$ matrix

$$B := P^{-1}AP$$

represents the same linear map as A, but expresses it in a different choice of basis on $\mathbb{C}^n$. It can be helpful to think in terms of a commutative diagram of linear maps as shown:

(0.1)
$$\begin{array}{ccc} \mathbb{C}^n & \xrightarrow{f_B} & \mathbb{C}^n \\ {\scriptstyle f_P}\downarrow & & \uparrow{\scriptstyle f_{P^{-1}}} \\ \mathbb{C}^n & \xrightarrow{f_A} & \mathbb{C}^n \end{array}$$

Therefore we may replace the matrix A by any such matrix B because we only care about how matrices *act* on vectors; that is, we care only about the linear map f_A. This expresses the key conclusion of Algebra 1B for a square matrix A. We now ask:

Question: How best to choose the matrix P to give an especially simple matrix B? And if this can be done, is the resulting matrix unique?

Ideally, we would choose B to be diagonal, but this isn't always possible. For the general case we take inspiration from the following:

Theorem (**Fundamental Theorem of Arithmetic**) *Every integer $n \geq 2$ can be expressed as the product of finitely many prime numbers, and this expression is unique up to a change of order of the prime factors.*

To state the analogous result for matrices, a *Jordan block matrix* is an $m \times m$ matrix

$$J(\lambda, m) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}$$

where $\lambda \in \mathbb{C}$ are the entries in the diagonal, 1's lie immediately above the diagonal, and there are zeroes everywhere else. Given an $m \times m$ matrix A and an $n \times n$ matrix B, the $(m+n) \times (m+n)$ block matrix $A \oplus B$ is the matrix with A in the top-left $m \times m$ block, with B in the bottom-right $n \times n$ block, and zeroes everywhere else.

The goal of Algebra 2B is to prove the following result:

Theorem (**Jordan Normal Form**) *Every $n \times n$ matrix A is similar to a matrix B (this means there exists invertible P such that $B = P^{-1}AP$) which can be expressed as the direct sum of finitely many Jordan block matrices, and this expression is unique up to a change of order of the Jordan blocks.*

The idea that ties these theorems together is the existence of a 'unique decomposition', more typically called *unique factorisation*. The Jordan Normal Form theorem provides the answer to the question that we asked above; it tells us that we can get pretty close to diagonalising any square matrix with entries in $\mathbb{C}$.

Our proof of the Jordan Normal Form theorem uses many fundamental properties of the polynomial ring $\mathbb{C}[t]$, and for this reason, we begin with an introduction to rings. We'll also spend some time investigating a class of rings that exhibit nice unique factorisation properties, called Unique Factorisation Domains (UFDs). We also present several applications to various branches of algebra.

1. Rings

1.1. **A reminder on groups.** Informally, a ring is simply a set equipped with 'sensible' notions of addition and multiplication that are compatible. We would like the definition to be broad enough to include examples like the integers, the set of $n \times n$ complex-valued matrices under the usual matrix addition and multiplication, and the set of complex-valued polynomials under the usual polynomial addition and multiplication. At the same time, we don't want the definition to be so broad that we're unable to prove any interesting theorems.

Before introducing the formal definition of a ring (and recalling that of a group), recall that a *binary operation* on a set S is a function

$$f \colon S \times S \to S.$$

The binary operations that crop up here are typically addition, denoted $+$, or multiplication, denoted $\cdot$. We write $a + b$ rather than $+(a, b)$, and $a \cdot b$ rather than $\cdot(a, b)$.

Definition 1.1 (**Group**)**.** A *group* is a pair $(G, *)$, where G is a set, $*$ is a binary operation on G and the following axioms hold:

- (The associative law)
$$(a * b) * c = a * (b * c) \text{ for all } a, b, c \in G.$$
- (Existence of an identity) There exists an element $e \in G$ with the property that
$$e * a = a \text{ and } a * e = a \text{ for all } a \in G.$$
- (The existence of an inverse) For each $a \in G$ there exists $b \in G$ such that
$$a * b = b * a = e.$$

If it is clear from the context what the group operation $*$ is, one often simply refers to the group G rather than to the pair $(G, *)$.

Remarks 1.2. Both the identity element and the inverse of a given element are unique:

(1) if $e, f \in G$ are two elements satisfying the identity property from (b) above, then

$$f = e * f = e,$$

where the first identity follows from the fact that e satisfies the property and the latter from the fact that f satisfies the property.

(2) Given $a \in G$, if $b, c \in G$ are both elements satisfying (c) above, then

$$b = b * e = b * (a * c) = (b * a) * c = e * c = c.$$

This unique element b is called the *inverse* of a. It is often denoted a^{-1}.

Definition 1.3 (**Abelian group**). A group $(G, *)$ is *abelian* if $a*b = b*a$ for all $a, b \in G$.

The binary operation in an abelian group is often written as $+$, in which case the identity element is denoted 0, and the inverse of an element $a \in G$ is denoted $-a \in G$.

Definition 1.4 (**Subgroup**). A nonempty subset H of a group G is a *subgroup* of G iff

$$\forall\ a, b \in H,\ \text{we have } a * b^{-1} \in H. \tag{1.1}$$

This version of the definition is great when you want to show that a subset is a subgroup, because there's so little to check. Despite this, we have (see Algebra 1A):

Lemma 1.5. *A nonempty subset H of a group $(G, *)$ is a subgroup if and only if $(H, *)$ is a group.*

Proof. Let H be a subgroup of $(G, *)$. Since H is nonempty, there exists $a \in H$ and hence $e = a * a^{-1} \in H$ by equation (1.1). For $a \in H$, apply condition (1.1) to the elements $e, a \in H$ to see that $a^{-1} = e*a^{-1} \in H$. Also, for $a, b \in H$, we've just shown that $b^{-1} \in H$, so applying condition (1.1) to the elements $a, b^{-1} \in H$ gives $a * b = a * (b^{-1})^{-1} \in H$. In particular, $*$ is a binary operation on H, and since $(G, *)$ is a group, the operation $*$ on H is associative. For the converse, let H be a subset of G such that $(H, *)$ is a group. Then the identity element $e \in H$, so H is nonempty. Let $a, b \in H$. Then b^{-1} lies in H since H is a group, and since $*$ is a binary operation on H we have $a * b^{-1} \in H$ as required. □

1.2. **Definitions and basic properties of rings.** We now move on to rings.

Definition 1.6 (**Ring**). A *ring* is a triple $(R, +, \cdot)$, where R is a set with binary operations

$$+ \colon R \times R \to R \quad (a, b) \mapsto a + b \qquad \text{and} \qquad \cdot \colon R \times R \to R \quad (a, b) \mapsto a \cdot b$$

such that the following axioms hold:

- $(R, +)$ is an abelian group. Write 0 for the (unique) additive identity, and $-a$ for the (unique) additive inverse of $a \in R$, so

$$\begin{aligned} (a+b)+c &= a+(b+c) && \text{for all } a,b,c \in R; \\ a+0 &= a && \text{for all } a \in R; \\ a+b &= b+a && \text{for all } a,b \in R; \\ a+(-a) &= 0 && \text{for all } a \in R. \end{aligned}$$

- (The associative law under multiplication)

$$(a \cdot b) \cdot c = a \cdot (b \cdot c) \qquad \text{for all } a,b,c \in R;$$

- (The distributive laws hold)

$$\begin{aligned} a \cdot (b+c) &= (a \cdot b) + (a \cdot c) && \text{for all } a,b,c \in R; \\ (b+c) \cdot a &= (b \cdot a) + (c \cdot a) && \text{for all } a,b,c \in R. \end{aligned}$$

Notation 1.7. We often omit $\cdot$ and write ab instead of $a \cdot b$. For simplicity we often avoid brackets when there is no ambiguity. Here the same conventions hold as for real numbers, i.e., that $\cdot$ has priority over $+$. For example $ab + ac$ stands for $(a \cdot b) + (a \cdot c)$ and not $(a \cdot (b + a)) \cdot c$. One also writes a^2 for $a \cdot a$ and $2a$ for $a + a$ and so on.

Lemma 1.8. *In any ring $(R, +, \cdot)$, we have*

(1) $a \cdot 0 = 0$ *and* $0 = 0 \cdot a$ *for all $a \in R$; and*
(2) $a \cdot (-b) = -(a \cdot b)$ *and* $-(a \cdot b) = (-a) \cdot b$ *for all $a, b \in R$.*

Proof. For (1), let $a \in R$. Since 0 is an additive identity, one of the distributive laws gives

$$a \cdot 0 = a \cdot (0 + 0) = a \cdot 0 + a \cdot 0.$$

Adding $-(a \cdot 0)$ on the left on both sides gives

$$-(a \cdot 0) + a \cdot 0 = -(a \cdot 0) + a \cdot 0 + a \cdot 0.$$

The left hand side is zero, and the associativity law gives that the right hand side is

$$(-(a \cdot 0) + a \cdot 0) + a \cdot 0 = 0 + a \cdot 0 = a \cdot 0$$

as required. The second identity is similar. To prove (2), note that

$$a \cdot b + a \cdot (-b) = a \cdot (b + (-b)) = a \cdot 0 = 0.$$

This means that $a \cdot (-b)$ is the additive inverse of ab, that is, $a \cdot (-b) = -(a \cdot b)$. The second identity is similar. □

Definition 1.9 (**Units in a ring with 1**). A ring $(R, +, \cdot)$ is called a *ring with* 1 (also called a *unital ring*) if there is a multiplicative identity, i.e., an element $1 \in R$ satisfying

$$a \cdot 1 = 1 \cdot a = a \text{ for all } a \in R.$$

An element $a \in R$ in a ring with 1 is a *unit* if it has a multiplicative inverse, i.e., if there exists $b \in R$ such that $a \cdot b = b \cdot a = 1$.

Remarks 1.10. Let R be a ring with 1. Then:

(1) the multiplicative identity is unique. The same argument as before works, i.e., if $1, \bar{1}$ are both multiplicative identity elements, then $\bar{1} = \bar{1} \cdot 1 = 1$.
(2) The multiplicative inverse of a unit is unique, see Remark 1.2(2) for the argument. We denote the multiplicative inverse by a^{-1}.

Definition 1.11 (**Other common types of ring**)**.** Let $(R, +, \cdot)$ be a ring. Then:

(1) R is a *commutative ring* if $a \cdot b = b \cdot a$ for all $a, b \in R$.
(2) R is an *integral domain* if it is a commutative ring with 1 in which $0 \neq 1$, such that if $a, b \in R$ satisfy $ab = 0$, then $a = 0$ or $b = 0$.
(3) R a *division ring* if it is a ring with 1 in which $0 \neq 1$, such that every non-zero element is a unit, i.e.,
$$\text{for all } a \in R \setminus \{0\}, \text{ there exists } b \in R \text{ such that } ab = 1 = ba.$$
(4) R is a *field* if it is a commutative division ring.

Remark 1.12. Every field $\Bbbk$ is an integral domain. Indeed, if $a, b \in \Bbbk$ satisfy $ab = 0$ and if $a \neq 0$, then $b = 1 \cdot b = a^{-1}ab = a^{-1} \cdot 0 = 0$.

1.3. **Examples of rings.** We'll start with a few familiar examples.

Examples 1.13. (1) Every field is a commutative ring and hence so are $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ with respect to the usual addition and multiplication.
(2) Division rings need not be commutative, so division rings need not be fields. We'll see an example of a noncommutative division ring (hence not a field) in section 4.
(3) The ring $\mathbb{Z}$ is an integral domain, but it's not a division ring, so it's not a field.
(4) The commutative ring $\mathbb{Z}_4 = \{[0], [1], [2], [3]\}$ satisfies $[2] \cdot [2] = [4] = [0]$ and yet $[2] \neq [0]$, so $\mathbb{Z}_4$ is not an integral domain.

Example 1.14 (**The ring of $n \times n$ matrices over R**)**.** For any ring R, let $M_n(R)$ denote the set of all $n \times n$ matrices with coefficients in the ring R. Then $M_n(R)$ is a ring with respect to usual addition and multiplication of square matrices. If R is a ring with 1 then so is $M_n(R)$, but this ring is not commutative in general even if R is commutative (ask yourself: what goes wrong?).

Example 1.15 (**The ring of formal power series with coefficients in R**)**.** Let R be a ring and let x be a variable. A *formal power series f over R* is a formal expression
$$f = \sum_{k=0}^{\infty} a_k x^k = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots$$
with $a_k \in R$ for $k \geq 0$ (we don't worry about convergence: R is any ring, so we have no notion of 'distance' between two elements). Let
$$R[[x]] := \left\{ \sum_{k=0}^{\infty} a_k x^k \mid a_k \in R \text{ for all } k \geq 0 \right\}$$

be the set of all formal power series over R, where addition and multiplication on $R[[x]]$ are defined as follows:

$$\begin{aligned}\sum_{k=0}^{\infty} a_k x^k + \sum_{k=0}^{\infty} b_k x^k &:= \sum_{k=0}^{\infty} (a_k + b_k) x^k \\ \left(\sum_{k=0}^{\infty} a_k x^k\right) \cdot \left(\sum_{k=0}^{\infty} b_k x^k\right) &:= a_0 b_0 + (a_1 b_0 + a_0 b_1) x + (a_2 b_0 + a_1 b_1 + a_0 b_2) x^2 + \cdots \\ &= \sum_{k=0}^{\infty} \left(\sum_{i+j=k} a_i b_j\right) x^k.\end{aligned}$$

As R is an abelian group with respect to the ring addition it follows readily that $(R[[x]], +)$ is an abelian group in which the power series $0 = 0 + 0x + 0x^2 + \cdots$ is the zero element. To see that $(R[[x]], +, \cdot)$ is a ring, it remains to see that the multiplication is associative and that the distributive laws hold. For this, let

$$f = \sum_{k=0}^{\infty} a_k x^k, \quad g = \sum_{k=0}^{\infty} b_k x^k, \quad h = \sum_{k=0}^{\infty} c_k x^k$$

be formal power series. The coefficent of x^n in the product $(fg)h$ is

$$\sum_{i+j+k=n} (a_i b_j) c_k$$

which (as multiplication in R is associative) is the same as

$$\sum_{i+j+k=n} a_i (b_j c_k),$$

the coefficient of x^n in $f(gh)$. It follows that $(fg)h = f(gh)$, so multiplication in $R[[x]]$ is associative. Finally we check the distributive laws. The coefficent of x^n in $f(g+h)$ is

$$\sum_{i+j=n} a_i (b_j + c_j) = \sum_{i+j=n} a_i b_j + \sum_{i+j=n} a_i c_j$$

which equals the coefficient of x^n in $fg + fh$, so $f(g+h) = fg + fh$. Similary one proves that $(g+h)f = gf + hf$. This completes the proof that $(R[[x]], +, \cdot)$ is a ring.

Remarks 1.16. (1) Two formal power series $\sum_{k=0}^{\infty} a_k x^k$ and $\sum_{k=0}^{\infty} b_k x^k$ coincide if and only if $(a_k) = (b_k)$, i.e., the variable x is superfluous.

(2) Many properties of the ring R carry over to $R[[x]]$; see Exercise sheet 1.

1.4. **Subrings.** We now introduce subrings of a ring.

Definition 1.17 (**Subring**)**.** A nonempty subset S of a ring R is a *subring* iff

$$\forall\, a, b \in S, \quad \text{we have} \quad a - b \in S.$$
$$\forall\, a, b \in S, \quad \text{we have} \quad a \cdot b \in S.$$

The sets of the form $r + S = \{r + s \mid s \in S\}$ for $r \in R$ are the *cosets* of S in R.

Lemma 1.18. *Let S be a subset of a ring $(R, +, \cdot)$. Then S is a subring of R if and only if $(S, +, \cdot)$ is a ring.*

Proof. This is an exercise. □

Examples 1.19. (1) For any ring R, both $\{0\}$ and R are subrings of R.

(2) The ring $\mathbb{Z}$ is a subring of $\mathbb{Q}$ which is a subring of $\mathbb{R}$ which is a subring of $\mathbb{C}$ under the usual operations of addition and multiplication.

(3) The even integers $\mathbb{Z}2$ are a subring of $\mathbb{Z}$, and hence they form a ring in their own right by Lemma 1.17. This ring is not a 'ring with 1'. In particular, a subring of a 'ring with 1' need not be a 'ring with 1' (!).

(4) The Gaussian integers $\mathbb{Z}[i] := \{a + bi \in \mathbb{C} \mid a, b \in \mathbb{Z}\}$ form a subring of the field $\mathbb{C}$, so $\mathbb{Z}[i]$ is a ring. The next result implies that $\mathbb{Z}[i]$ is an integral domain.

Lemma 1.20. *If a subring S of an integral domain R contains the element 1, then S is an integral domain.*

Proof. The only property of an integral domain R that is not necessarily inherited by every subring is the existence of 1, but this follows from the assumptions. □

Example 1.21 (**The ring of polynomials with coefficients in** R)**.** Let R be a ring and let $\sum_{k=0}^{\infty} a_k x^k \in R[[x]]$ be a formal power series. If only finitely many of the coefficients a_k are nonzero, we say that $\sum_{k=0}^{\infty} a_k x^k$ is a *polynomial* and we write

$$R[x] := \left\{ \sum_{k=0}^{\infty} a_k x^k \in R[[x]] \mid a_k \neq 0 \text{ for only finitely many } k \geq 0 \right\}$$

for the subset of polynomials. In particular, by ignoring the terms with coefficient equal to zero, any polynomial can be written as $a_0 + a_1 x + \cdots + a_n x^n$ for some $n \geq 0$. The *degree* of a nonzero polynomial is the largest n such that $a_n \neq 0$.

We claim that $R[x]$ is a subring of $R[[x]]$. Indeed, if $f = \sum_{k=0}^{\infty} a_k x^k, g = \sum_{k=0}^{\infty} b_k x^k$ are polynomials of degree m and n respectively, then

$$f - g = \sum_{k=0}^{\infty} a_k x^k - \sum_{k=0}^{\infty} b_k x^k = \sum_{k=0}^{\infty} (a_k - b_k) x^k$$

is a polynomial of degree at most $\max(m, n)$, and

$$\sum_{k=0}^{\infty} a_k x^k \cdot \sum_{k=0}^{\infty} b_k x^k = \sum_{k=0}^{\infty} \left(\sum_{i+j=k} a_i b_j \right) x^k.$$

is a polynomial of degree at most $m + n$. In particular, $R[x]$ is a ring by Lemma 1.17.

End of Week 1

1.5. **Quotient rings.** For the moment, let R be any set. Recall that a *relation* $\sim$ on R is a subset $S \subset R \times R$, in which case we write

$$a \sim b \iff (a, b) \in S.$$

An *equivalence* relation on R is a relation $\sim$ that is reflexive, symmetric and transitive, and the *equivalence class* of an element $a \in R$ is the (nonempty) set

$$[a] := \{b \in R \mid b \sim a\}$$

of elements that are equivalent to a. Every element lies in a unique equivalence class, and any two distinct equivalences classes are disjoint subsets of R; we say that the equivalence classes *partition* the set R (see Algebra 1A).

The key point for us is that an equivalence relation on a set R produces a new set, namely the set of equivalence classes

$$R/\sim := \big\{[a] \mid a \in R\big\}.$$

Question 1.22. *If R is a ring (not just a set), do we require extra conditions on an equivalence relation $\sim$ to ensure that the set $R/\sim$ of equivalence classes is a ring?*

You've already seen examples of this in Algebra 1A:

Example 1.23 (**The ring $\mathbb{Z}_n$ of integers mod n**)**.** For any $n \in \mathbb{Z}$, consider the subset $\mathbb{Z}n := \{mn \in \mathbb{Z} \mid m \in \mathbb{Z}\}$ of integers that are divisible by n (notice that $\mathbb{Z}n = \mathbb{Z}(-n)$, so we may as well assume $n \geq 0$). There is an equivalence relation $\sim$ on $\mathbb{Z}$ defined by

$$a \sim b \iff n|(b-a) \iff b - a \in \mathbb{Z}n.$$

Any integer m can be written in the form $m = qn + r$ for a unique $0 \leq r < n$, in which case $[m] = [r]$. Therefore the set of equivalence (or *congruence*) classes is simply

$$\mathbb{Z}_n := \mathbb{Z}/\sim = \big\{[a] \mid a \in \mathbb{Z}\big\} = \big\{[0], [1], \ldots, [n-1]\big\}.$$

The crucial point for us is that $\mathbb{Z}_n$ is more than a set: addition and multiplication can be defined as follows:

$$[a] + [b] := [a + b] \quad \text{and} \quad [a] \cdot [b] := [a \cdot b].$$

This says simply that we add and multiply the representatives a and b in $\mathbb{Z}$, and then take the equivalence class of the result using the fact that $[n] = [0]$. To be explicit, $\mathbb{Z}/\mathbb{Z}3$ has three elements $[0], [1]$ and $[2]$, and the addition and multiplication tables are

$+$	$[0]$	$[1]$	$[2]$
$[0]$	$[0]$	$[1]$	$[2]$
$[1]$	$[1]$	$[2]$	$[0]$
$[2]$	$[2]$	$[0]$	$[1]$

$\cdot$	$[0]$	$[1]$	$[2]$
$[0]$	$[0]$	$[0]$	$[0]$
$[1]$	$[0]$	$[1]$	$[2]$
$[2]$	$[0]$	$[2]$	$[1]$

In this case, notice that both $[1]$ and $[2]$ have a multiplicative inverse. This shouldn't be a surprise: you know that $\mathbb{Z}_n$ is a field if and only if n is a prime.

Definition 1.24 (**Congruence relation**). Let R be a ring and let $\sim$ be an equivalence relation on R. We say that $\sim$ is a *congruence* iff for all $a, b, a', b' \in R$, we have

$$a \sim a' \text{ and } b \sim b' \implies a + b \sim a' + b' \text{ and } a \cdot b \sim a' \cdot b'. \tag{1.2}$$

The equivalence classes of a congruence $\sim$ are called *congruence classes*.

Remark 1.25. This says simply that one can add or multiply any two equivalence classes $[a], [b] \in R/\sim$ by first adding or multiplying *any representative* of the equivalence classes in the ring R, and then taking the congruence class of the result.

Addition and multiplication in $\mathbb{Z}_n$ is possible precisely because the equivalence relation $\sim$ on $\mathbb{Z}$ defined in Example 1.23 is a congruence. More generally, we have the following:

Theorem 1.26 (**Quotient rings**). *Let $\sim$ be a congruence on a ring R. Define addition and multiplication on the set $R/\sim$ of equivalence classes as follows: for $a, b \in R$, define*

$$[a] + [b] := [a + b] \quad \text{and} \quad [a] \cdot [b] := [a \cdot b].$$

Then $(R/\sim, +, \cdot)$ is a ring with zero element $[0]$. Moreover:

1. *if R is a ring with 1, then the element $[1]$ makes $R/\sim$ into a ring with 1; and*
2. *if R is commutative then so is $R/\sim$.*

Proof. We first check that addition and multiplication are well-defined for equivalence classes. For this, consider alternative representatives of the equivalence classes $[a]$ and $[b]$, say $a' \in R$ satisfying $[a] = [a']$ and $b' \in R$ satisfying $[b] = [b']$. Then

$$\begin{aligned} [a'] + [b'] &= [a' + b'] && \text{by definition} \\ &= [a + b] && \text{by the congruence property} \\ &= [a] + [b] && \text{by definition,} \end{aligned}$$

and similarly

$$\begin{aligned} [a'] \cdot [b'] &= [a' \cdot b'] && \text{by definition} \\ &= [a \cdot b] && \text{by the congruence property} \\ &= [a] \cdot [b] && \text{by definition} \end{aligned}$$

as required. This means that addition and multiplication define binary operations on the set $R/\sim$ of equivalence classes. We now check that all the ring axioms hold:

1. To check that $(R/\sim, +)$ is an abelian group, (look at Exercise 1.1 or) note that for $a, b, c \in R$ we have

$$([a] + [b]) + [c] = [a + b] + [c] = [(a + b) + c] = [a + (b + c)] = [a] + [b + c] = [a] + ([b] + [c]),$$

$$[a] + [b] = [a + b] = [b + a] = [b] + [a].$$

Also, we have $[a] + [0] = [a + 0] = [a]$, so $[0]$ is the zero element. Moreover, $[a] + [-a] = [a + (-a)] = [0]$, so $[-a]$ is the additive identity of $[a]$.

(2) To check that $(R/\sim, \cdot)$ is associative, note that for $a, b, c \in R$ we have

$$([a] \cdot [b]) \cdot [c] = [ab] \cdot [c] = [(ab)c] = [a(bc)] = [a] \cdot [bc] = [a] \cdot ([b] \cdot [c]).$$

(3) To check that $R/\sim$ satisfies the distributive laws, note that for $a, b, c \in R$ we have

$$\begin{aligned} [c] \cdot ([a] + [b]) = [c] \cdot [a + b] &= [c(a + b)] \\ &= [ca + cb] \\ &= [ca] + [cb] \\ &= [c] \cdot [a] + [c] \cdot [b]. \end{aligned}$$

One proves that $([a] + [b]) \cdot [c] = [a] \cdot [c] + [b] \cdot [c]$ similarly.

This completes the proof that $(R/\sim, +, \cdot)$ is a ring with zero element $[0]$. To finish off, note first that if R is a ring with 1, then $[1] \in R/\sim$ is a multiplcative identity because

$$[a] \cdot [1] = [a \cdot 1] = [a] = [1 \cdot a] = [1] \cdot [a],$$

hence $R/\sim$ is a ring with 1. Finally, if R is commutative then

$$[a] \cdot [b] = [a \cdot b] = [b \cdot a] = [b] \cdot [a],$$

so $R/\sim$ is commutative. □

In order to produce many examples of congruences, we first establish a link between congrences and a very special class of subrings.

Definition 1.27 (Ideal). A nonempty subset I of a ring R is an *ideal* in R if and only if

$$\begin{aligned} \forall\, a, b \in I, &\quad \text{we have} \quad a - b \in I \\ \forall\, a \in I, r \in R, &\quad \text{we have} \quad r \cdot a, a \cdot r \in I. \end{aligned}$$

Remark 1.28. This simply means that an ideal is an additive subgroup that is closed under multiplication by *all elements of the ring*. Notice that every ideal I in R is a subring of R. In particular, Lemma 1.17 implies that every ideal contains 0_R.

Example 1.29 (Principal ideal). Let R be a commutative ring and let $a \in R$. The set

$$Ra := \{r \cdot a \in R \mid r \in R\}$$

(sometimes denoted $\langle a \rangle$ if the ring R is clear from the context) is an ideal in R; this is called the *ideal generated by* a, and every ideal of this form is called a *principal ideal*.

Lemma 1.30. *Let $\sim$ be a congruence relation on a ring R, and let $I := [0]$ denote the congruence class of 0. Then I is an ideal in the ring R. Moreover:*

(1) *for $a, b \in R$, we have $a \sim b \iff a - b \in [0]$; and*
(2) *the congruence classes of $\sim$ are the cosets of I, i.e., $[a] = a + [0]$ for all $a \in R$.*

Proof. See Exercise Sheet 2. □

Proposition 1.31. *Let I be an ideal in R, and define $\sim$ on R by setting*

$$a \sim b \text{ if and only if } a - b \in I.$$

Then $\sim$ is a congruence relation in which the equivalence classes are the cosets of I in R, i.e., we have $[a] = a + I$ for all $a \in R$. In particular, $[0] = I$.

Proof. We first show that $\sim$ is an equivalence relation. Let $a, b, c \in R$. Then $a - a = 0 \in I$ means $a \sim a$, so $\sim$ is reflexive. If $a \sim b$ then $a - b \in I$ and hence $b - a = -(a - b) \in I$ by Lemma 1.17. This gives $b \sim a$, so $\sim$ is symmetric. Finally if $a \sim b$ and $b \sim c$ then $a - b, b - c \in I$. As I is closed under addition, it follows that $(a - b) + (b - c) = a - c \in I$ and hence $a \sim c$. This shows that $\sim$ is transitive, so $\sim$ is an equivalence relation.

To prove that $\sim$ is a congruence, let $a, b, a', b' \in R$ and suppose that $a \sim a'$ and $b \sim b'$. Then $a - a', b - b' \in I$. Since I is an ideal, we have

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

by the first defining property of an ideal, so $a+b \sim a'+b'$. Finally, by adding $0 = -ab'+ab'$ below, we get

$$ab - a'b' = ab + \big[- ab' + ab' \big] - a'b' = a(b - b') + (a - a')b' \in I$$

by the second defining property of an ideal, so $ab \sim a'b'$ as required.

For $a \in R$, the equivalence class of a is

$$\begin{aligned} [a] := \{b \in R \mid b \sim a\} &= \{b \in R \mid b - a \in I\} \\ &= \{b \in R \mid \exists\, i \in I \text{ such that } b - a = i\} \\ &= \{a + i \mid i \in I\} \\ &= a + I \end{aligned}$$

as claimed. □

Proposition 1.31 says that ideals determine congruence relations, and it provides the converse to Lemma 1.30. These results together establish a one-to-one correspondence between congruences on a ring and ideals in that ring. We may therefore change our point-of-view when considering quotient rings: then next definition simply rewrites the definition of the quotient ring $R/\sim$ constructed in Theorem 1.26 directly in terms of the ideal I associated to the congruence class $\sim$.

Definition 1.32 (Quotient rings from ideals). Let I be an ideal in a ring R. The *quotient ring* R/I is the set

$$R/I = \{a + I : a \in R\}$$

of cosets of I in R, where we define addition and multiplication in the ring R/I by

$$\begin{aligned} (a + I) + (b + I) &= (a + b) + I \\ (a + I) \cdot (b + I) &= (a \cdot b) + I. \end{aligned}$$

Remark 1.33. Remember that these addition and multiplication formulas simply mean that we add and multiply the representatives a and b of each coset as if we're adding and multiplying in R, and then we take the coset of the resulting element of R.

Example 1.34. In Example 1.23, the subset $\mathbb{Z}n$ of $\mathbb{Z}$ is an ideal, so $\mathbb{Z}_n := \mathbb{Z}/\mathbb{Z}n$ is a ring. It's a commutative ring with 1 because $\mathbb{Z}$ is too (recall that we may assume $n \geq 1$).

Example 1.35. For a ring R, consider the polynomial ring $R[x]$. Let

$$\langle x^2 \rangle := \{ f \cdot x^2 \in R[x] \mid f \in R[x] \}$$

denote the ideal in $R[x]$ generated by x^2. This ideal determines the congruence relation $\sim$ on $R[x]$, where for $f, g \in R[x]$

$$f \sim g \iff f - g \in \langle x^2 \rangle \iff x^2 | f - g.$$

Any polynomial f can be written in the form $f = gx^2 + ax + b$ for unique $a, b \in R$, so $[f] = [ax + b]$ for some $a, b \in R$. Therefore

$$R[x]/\langle x^2 \rangle = \{ [ax + b] \mid a, b \in R \},$$

where addition and multiplication are given by

$$[ax + b] + [cx + d] = [(a + c)x + (b + d)]$$

and

$$[ax + b] \cdot [cx + d] = [acx^2 + (ad + bc)x + bd] = [(ad + bc)x + bd]$$

respectively. Notice that we add and multiply as if we're working with polynomials and then we modify the result using the fact that $[x^2] = [0]$.

End of Week 2.

---

2. Ring homomorphisms

2.1. **Definitions and examples.** We now introduce ring homomorphims which do for rings what maps do for sets, what linear maps do for vector spaces and what group homomorphisms do for groups.

Definition 2.1 (Ring homomorphism). Let R, S be rings. A map $\phi\colon R \to S$ is said to be a *ring homomorphism* if and only if for all $a, b \in R$, we have

$$\phi(a + b) = \phi(a) + \phi(b) \quad \text{and} \quad \phi(a \cdot b) = \phi(a) \cdot \phi(b).$$

Examples 2.2. Consider two maps from the integers involving the number 2:

(1) The function $\phi\colon \mathbb{Z} \to \mathbb{Z}_2$ defined by

$$\phi(n) = \begin{cases} 0 & \text{if } n \text{ is even} \\ 1 & \text{if } n \text{ is odd} \end{cases}$$

is a ring homomorphism. Indeed, if we compare the rules for adding and multiplying even and odd integers

$+$	even	odd
even	even	odd
odd	odd	even

$\cdot$	even	odd
even	even	even
odd	even	odd

with the addition and multiplication tables for $\mathbb{Z}_2$, we see that computing in $\mathbb{Z}$ and then applying ϕ is the same as applying ϕ and then computing in $\mathbb{Z}_2$.

(2) The function $\phi\colon \mathbb{Z} \to 2\mathbb{Z}$ defined by $\phi(n) = 2n$ is not a ring homomorphism, because $\phi(nm) = 2nm$ is typically not equal to $4nm = (2n)(2m) = \phi(n)\phi(m)$.

Lemma 2.3. *The composition of two ring homomorphisms is a ring homomorphism.*

Proof. This is an exercise. □

Lemma 2.4. *If $\phi : R \to S$ is a ring homomorphism then*

(1) *for $a, b \in R$, we have $\phi(b - a) = \phi(b) - \phi(a)$;*
(2) *$\phi(0_R) = 0_S$;*
(3) *for $a \in R$, we have $\phi(-a) = -\phi(a)$.*

Proof. For part (1), we have

$$\phi(b-a) + \phi(a) = \phi\big((b-a)+a\big) = \phi\big(b+(a-a)\big) = \phi(b+0) = \phi(b),$$

and add $-\phi(a)$ to both sides. For (2), substitute $b = a$ in (1) to obtain

$$\phi(0_R) = \phi(a-a) = \phi(a) - \phi(a) = 0_S.$$

For part (3), substitute $b = 0$ into part (1) and use part (2) to obtain

$$\phi(-a) = \phi(0_R - a) = \phi(0_R) - \phi(a) = 0_S - \phi(a) = -\phi(a)$$

as required. □

Example 2.5. For $n \in \mathbb{Z}_{\geq 0}$, the map $\phi\colon \mathbb{Z} \to \mathbb{Z}_n$ sending m to its equivalence class $[m]$ modulo n is a ring homomorphism. To see this, generalise the method from Example 2.2 above. For details, see Example 2.10 which provides a much broader generalisation.

Example 2.6 (**Evaluation map**)**.** Let R be a commutative ring and choose $r \in R$. Let S be a subring of R (the first time you read this example, assume $S = R$ for simplicity). Given a polynomial $f \in S[x]$ with coefficients in S, then $f(r) \in R$, so we obtain a map

$$\phi\colon S[x] \to R \ : \ f \mapsto f(r)$$

given by *evaluating each polynomial at $r \in R$*, i.e., substitute $r \in R$ into each polynomial.

We claim that this map is a ring homomorphism. Indeed, given any two polynomials $f = \sum_{k=0}^{m} a_k x^k$ and $g = \sum_{k=0}^{n} b_k x^k$, we have for $\ell = \max\{m, n\}$ that

$$\phi(f+g) = \phi\left(\sum_{k=0}^{\ell}(a_k + b_k)x^k\right) = \sum_{k=0}^{\ell}(a_k + b_k)r^k = \sum_{k=0}^{m} a_k r^k + \sum_{k=0}^{n} b_k r^k = \phi(f) + \phi(g),$$

where the third equals sign uses commutativity of addition and distributivity in R. Also, for $\ell = m + n$, we have that

$$\begin{aligned}
\phi(fg) &= \phi\left(\sum_{k=0}^{\ell}\left(\sum_{i+j=k} a_i b_j\right)x^k\right) && \text{by definition of multiplication in } R[x] \\
&= \sum_{k=0}^{\ell}\left(\sum_{i+j=k} a_i b_j\right) r^k \\
&= \sum_{i=0}^{m} a_i r^i \cdot \sum_{j=0}^{n} b_j r^j && \text{see below} \\
&= \phi\left(\sum_{i=0}^{m} a_i x^i\right) \cdot \phi\left(\sum_{j=0}^{n} b_j x^j\right) \\
&= \phi(f) \cdot \phi(g),
\end{aligned}$$

where the middle equals sign requires the distributive laws, commutativity of addition and associativity of both addition and multiplication in the ring R.

2.2. **Kernel and Image.** A ring homomorphism $\phi\colon R \to S$ defines a subset in R and a subset in S that play an important role in what follows:

Definition 2.7 (**Kernel and image**)**.** Let $\phi\colon R \to S$ be a ring homomorphism. The *kernel* of ϕ is the subset of R given by

$$\mathrm{Ker}(\phi) = \{a \in R \mid \phi(a) = 0\}$$

and the *image* of ϕ is the subset of S given by

$$\mathrm{Im}(\phi) = \{\phi(a) \in S \mid a \in R\}.$$

Lemma 2.8 (**Properties of the kernel**)**.** *Let $\phi\colon R \to S$ be a ring homomorphism. Then* $\mathrm{Ker}(\phi)$ *is an ideal of R. Moreover, ϕ is injective iff* $\mathrm{Ker}(\phi) = \{0\}$.

Proof. Since $\phi(0_R) = 0_S$ we have $0_R \in \mathrm{Ker}(\phi)$ and hence $\mathrm{Ker}(\phi) \neq \emptyset$. For $a, b \in \mathrm{Ker}(\phi)$,

$$\phi(a - b) = \phi(a) - \phi(b) = 0 - 0 = 0,$$

and for $r \in R$ and $a \in \mathrm{Ker}(\phi)$ we have

$$\phi(ra) = \phi(r)\phi(a) = \phi(r) \cdot 0 = 0 \quad \text{and} \quad \phi(ar) = \phi(a)\phi(r) = 0 \cdot \phi(r) = 0.$$

Thus $a - b, ra, ar \in \mathrm{Ker}(\phi)$, so $\mathrm{Ker}(\phi)$ is an ideal in R.

To prove the second statement, assume $\mathrm{Ker}(\phi) = \{0\}$ and suppose that $a, b \in R$ satisfy $\phi(a) = \phi(b)$. Then Lemma 2.4(1) implies that

$$\phi(b - a) = \phi(b) - \phi(a) = 0$$

so $b - a \in \mathrm{Ker}(\phi)$. This forces $a = b$, so ϕ is injective. Conversely, assume ϕ is injective and let $a \in \mathrm{Ker}(\phi)$. Lemma 2.4(2) gives $\phi(0) = 0 = \phi(a)$, and injectivity of ϕ forces $a = 0$, hence $\mathrm{Ker}(\phi) = \{0\}$ as required. □

Lemma 2.9 (**Properties of the image**). *Let $\phi\colon R \to S$ be a ring homomorphism. Then $\mathrm{Im}(\phi)$ is a subring of S. Moreover, ϕ is surjective iff $\mathrm{Im}(\phi) = S$.*

Proof. Again $\phi(0_R) = 0_S$, so $\mathrm{Im}(\phi)$ is nonempty. Let $a, b \in \mathrm{Im}(\phi)$, so there exists $c, d \in R$ such that $a = \phi(c)$ and $b = \phi(d)$. Then

$$a - b = \phi(c) - \phi(d) = \phi(c - d)$$

by Lemma 2.4(1), and $ab = \phi(c)\phi(d) = \phi(cd)$. This gives $a - b, ab \in \mathrm{Im}(\phi)$, so $\mathrm{Im}(\phi)$ is a subring of S. That ϕ is surjective if and only if $\mathrm{Im}(\phi) = S$ holds by definition. □

Example 2.10 (**Two fundamental maps**). Let I be an ideal in a ring R, and consider the map $\pi\colon R \to R/I$ defined by setting $\pi(a) = a + I$. This is a ring homomorphism, because

$$\pi(a + b) = (a + b) + I = (a + I) + (b + I) = \pi(a) + \pi(b),$$

and

$$\pi(ab) = ab + I = (a + I)(b + I) = \pi(a) \cdot \pi(b).$$

It's clearly surjective, and $\pi(a) = 0 \iff a \in I$. Therefore $\mathrm{Im}(\pi) = R/I$ and $\mathrm{Ker}(\pi) = I$.

Now let S be a subring of a ring R. Consider the map $\iota\colon S \to R$ defined by sending each element $s \in S$ to the same element considered as an element in R, i.e., $\iota(s) = s \in R$. This is a ring homomorphism because

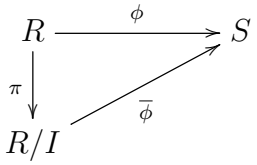
$$\iota(a + b) = a + b = \iota(a) + \iota(b)$$

and

$$\iota(a \cdot b) = a \cdot b = \iota(a) \cdot \iota(b).$$

It's clearly injective and it has image $S \subseteq R$, so $\mathrm{Ker}(\iota) = \{0\}$ and $\mathrm{Im}(\iota) = S$.

Theorem 2.11 (**Universal property of the quotient map**). *Let $\phi\colon R \to S$ is a ring homomorphism and let I be an ideal in R satisfying $I \subseteq \mathrm{Ker}(\phi)$. Then there exists a unique ring homomorphism $\overline{\phi}\colon R/I \to S$ such that the diagram*



commutes, i.e., $\overline{\phi} \circ \pi = \phi$ (here $\pi\colon R \to R/I$ is the quotient map from Example 2.10).

Proof. Consider the map $\overline{\phi}\colon R/I \to S$ defined by setting $\overline{\phi}\big(a+I\big) = \phi(a)$. To see that this map is well-defined independent of any choices, notice that

$$\begin{aligned} a+I = b+I \iff a-b \in I \implies a-b \in \operatorname{Ker}(\phi) \\ \iff 0 = \phi(a-b) = \phi(a) - \phi(b) \iff \phi(a) = \phi(b). \end{aligned}$$

In particular, we've shown that $a+I = b+I \implies \phi(a) = \phi(b)$, so $\overline{\phi}$ does not depend on the choice of representative a in the coset $a+I$.

To see that $\overline{\phi}$ is a ring homomorphism, notice that

$$\overline{\phi}((a+I)+(b+I)) = \overline{\phi}((a+b)+I) = \phi(a+b) = \phi(a)+\phi(b) = \overline{\phi}(a+I)+\overline{\phi}(b+I)$$

and

$$\overline{\phi}((a+I)\cdot(b+I)) = \overline{\phi}(ab+I) = \phi(ab) = \phi(a)\cdot\phi(b) = \overline{\phi}(a+I)\cdot\overline{\phi}(b+I).$$

This ring homomorphism satisfies $\overline{\phi}\circ\pi = \phi$, because for all $a \in R$ we have

$$\big(\overline{\phi}\circ\pi\big)(a) = \overline{\phi}(a+I) = \phi(a)$$

as required. Notice that it's the unique such ring homomorphism: the condition $\overline{\phi}\circ\pi = \phi$ forces us to have $(\overline{\phi}\circ\pi)(a) = \phi(a)$ for all $a \in R$, and since $\pi(a) = a+I$, this forces our map to satisfy $\overline{\phi}(a+I) = \phi(a)$ for all $a \in R$. □

2.3. **Isomorphisms of rings.** We now study isomorphisms of rings.

Definition 2.12 (**Ring isomorphism**)**.** Let R, S be rings. A homomorphism $\phi\colon R \to S$ is called an *isomorphism* if there is a ring homomorphism $\psi\colon S \to R$ such that $\psi(\phi(r)) = r$ for all $r \in R$ and $\phi(\psi(s)) = s$ for all $s \in S$. Given an isomorphism $\phi\colon R \to S$, we say that R is *isomorphic* to S and write $R \cong S$.

Remarks 2.13.
(1) Clearly the inverse of a ring isomorphism is a ring isomorphism. Indeed, forgetting for a moment the addition and multiplication, an isomorphism $\phi\colon R \to S$ is bijective as a map of sets, and the inverse is the map $\phi^{-1} = \psi$ from Definition 2.12. In particular, we're allowed to say that R and S are isomorphic without having to worry about whether we say R first or S first.
(2) If R is isomorphic to S then there is no structural difference between the two rings (see Exercise sheet 4).

Theorem 2.14 (**The first isomorphism theorem**)**.** *Let $\phi\colon R \to S$ be a ring homomorphism. Then there is a ring isomorphism*

$$\overline{\phi}\colon \big(R/\operatorname{Ker}(\phi)\big) \longrightarrow \operatorname{Im}(\phi).$$

Proof. Applying the universal property from Theorem 2.11 to the ideal $I := \operatorname{Ker}(\phi)$ gives a ring homomorphism $\overline{\phi}\colon R/\operatorname{Ker}(\phi) \to S$ given by $\overline{\phi}\big(a+I\big) = \phi(a)$. We may write this as a surjective ring homomorphism

$$\overline{\phi}\colon R/\operatorname{Ker}(\phi) \to \operatorname{Im}(\phi) \tag{2.1}$$

simply by changing the target of the morphism from S to the image of ϕ. To see that $\overline{\phi}$ is injective, suppose $\overline{\phi}(a+I) = \overline{\phi}(b+I)$, i.e, $\phi(a) = \phi(b)$. Then $\phi(a-b) = \phi(a) - \phi(b) = 0$, giving $a - b \in \mathrm{Ker}(\phi) = I$ and hence $a + I = b + I$ as required. Therefore the map $\overline{\phi}$ from (2.1) is a bijective ring homomorphism, so it's an isomorphism by Exercise Sheet 3. □

Remark 2.15. (1) **It is impossible to overstate how important Theorem 2.14 is**. We'll give several applications in the weeks ahead.

(2) Theorem 2.14 says in particular that every ring homomorphism can be written as the composition of a surjective ring homomorphism, then an isomorphism, and finally an injective ring homomorphism as shown below:

$$\begin{array}{ccc} R & \xrightarrow{\phi} & S \\ {\scriptstyle\pi}\downarrow & & \uparrow{\scriptstyle\iota} \\ R/I & \xrightarrow[\overline{\phi}]{} & \mathrm{Im}(\phi) \end{array}$$

2.4. **The characteristic of a ring with 1.** We use the following standard short hand notation for iterated sums in a ring R: for any positive integer n and for $a \in R$, we write

$$na = \underbrace{a + \cdots + a}_{n} \quad \text{and} \quad (-n)a = -(na).$$

In particular, zero copies of an element $a \in R$ is the zero element 0_R in the ring R (one might write this as $0a = 0_R$, where 0 is the zero element in $\mathbb{Z}$). This is just notation and *has nothing to do with the ring multiplication.* Notice that $0_R \cdot a = 0_R$ is a fact that we proved in Lemma 1.8, but $0a = 0_R$ is just a natural notation when 0 is the zero integer.

Definition 2.16 (**Characteristic of a ring with 1**)**.** Let R be a ring with 1. The *characteristic* of R, denoted $\mathrm{char}(R)$, is a non-negative integer defined as follows; if there is a positive integer m such that $m1_R = 0_R$, then $\mathrm{char}(R)$ is the smallest such positive integer; otherwise, there is no such positive integer and we say that $\mathrm{char}(R) = 0$.

Examples 2.17. (1) The zero ring $R = \{0\}$ is actually a ring with 1 (!!), and it's the only ring for which $\mathrm{char}(R) = 1$.

(2) For any positive integer n, we have that $\mathrm{char}(\mathbb{Z}_n) = n$.

(3) The field $\mathbb{C}$ has characteristic zero, and hence so do $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$.

Lemma 2.18. *Let R be a ring of characteristic $n > 0$. Then $na = 0$ for all $a \in R$.*

Proof. For $a \in R$, we have

$$na = \underbrace{a + \cdots + a}_{n} = (\underbrace{1_R \cdot a + \cdots + 1_R \cdot a}_{n}) = (\underbrace{1_R + \cdots + 1_R}_{n}) \cdot a = 0_R \cdot a = 0_R$$

as required. □

Let R be a ring with 1. It's easy to see that the following subset is a subring of R:

$$\mathbb{Z}1_R := \{n\,1_R \mid n \in \mathbb{Z}\} = \{\cdots, (-2)1_R, -1_R, 0_R, 1_R, (2)1_R, \cdots\}.$$

Lemma 2.19. *Let R be a ring with 1. Then either:*

(1) *$\mathrm{char}(R) = 0$, in which case $\mathbb{Z}1_R$ is isomorphic to $\mathbb{Z}$; or*
(2) *$\mathrm{char}(R) = n > 0$, in which case $\mathbb{Z}1_R$ is isomorphic to $\mathbb{Z}_n$.*

Proof. The map $\phi\colon \mathbb{Z} \to R$ given by $\phi(n) = n1_R$ is a ring homomorphism because

$$\phi(n+m) = (n+m)1_R = n1_R + m1_R = \phi(n) + \phi(m),$$

and the distributive law gives

$$\phi(nm) = nm1_R = n1_R \cdot m1_R = \phi(n) \cdot \phi(m).$$

Moreover, the image of ϕ is clearly $\mathbb{Z}1_R$.

Suppose first that $\mathrm{char}(R) = 0$. Then $\phi(n) = n\,1_R$ equals 0_R if and only if $n = 0$, so $\mathrm{Ker}(\phi) = \{0\}$. Applying the fundamental isomorphism theorem to ϕ gives $\mathbb{Z} \cong \mathbb{Z}1_R$ which proves part (1). Otherwise, $\mathrm{char}(R) = n > 0$. Then $\phi(m) = m1_R = 0$ if and only if $n|m$, therefore $\mathrm{Ker}(\phi) = \mathbb{Z}n$. Applying the fundamental isomorphism theorem to ϕ gives $\mathbb{Z}_n \cong \mathbb{Z}1_R$, so part (2) holds. □

Proposition 2.20. *The characteristic of an integral domain is either 0 or a prime.*

Proof. Let R be an integral domain. Notice first that since $R \neq \{0\}$, we have $\mathrm{char}(R) \neq 1$. Suppose that $n := \mathrm{char}(R)$ is neither 0 nor a prime, i.e., $n = rs$ for some $1 < r, s < n$. Then $0 = n\,1_R = (rs)1_R = (r\,1_R) \cdot (s\,1_R)$, but since R is an integral domain it follows that either $r\,1_R = 0$ or $s\,1_R = 0$. Either case is impossible in a ring of characteristic n because $r, s < n$. Thus, the characteristic must be zero or prime after all. □

End of Week 3.

2.5. **The Chinese remainder theorem.** In this section we revisit the fabulously named 'Chinese remainder theorem' that you met in Algebra 1A. We first introduce and study two new ideals that we can associate to a pair of ideals.

Definition 2.21 (**Sum and intersection of ideals**)**.** Let I and J be ideals of R. The *sum* of I and J is the subset

$$I + J := \{a + b \in R \mid a \in I, b \in J\},$$

and the *intersection* of I and J is the subet

$$I \cap J := \{a \in R \mid a \in I \text{ and } a \in J\}.$$

Lemma 2.22. *Both $I \cap J$ and $I + J$ are ideals of R.*

Proof. See Exercise Sheet 4.

□

Example 2.23. For $m, n \in \mathbb{Z}$, if we write $I = \mathbb{Z}m = \langle m \rangle$ and $J = \mathbb{Z}n = \langle n \rangle$, then

$$I + J = \langle \gcd(m, n) \rangle \quad \text{and} \quad I \cap J = \langle \operatorname{lcm}(m, n) \rangle.$$

Definition 2.24 (**Direct product of rings**)**.** Let R and S be rings. The *direct product* of R and S is the ring

$$R \times S = \big\{(r, s) \mid r \in R,\ s \in S\big\},$$

where the operations are $(a, b) + (c, d) = (a + c, b + d)$ and $(a, b) \cdot (c, d) = (ac, bd)$.

Theorem 2.25 (**Chinese remainder theorem**)**.** *Let R be a commutative ring with 1. Let I, J be ideals in R satisfying $I + J = R$. Then there is a ring isomorphism*

$$\frac{R}{I \cap J} \cong \frac{R}{I} \times \frac{R}{J}.$$

Proof. Consider the map $\phi\colon R \to R/I \times R/J$ defined by setting $\phi(a) = (a + I, a + J)$. It's a ring homomorphism because

$$\begin{aligned}
\phi(a + b) &= (a + b + I, a + b + J) \\
&= \big((a + I) + (b + I), (a + J) + (b + J)\big) && \text{by Definition 1.32} \\
&= (a + I, a + J) + (b + I, b + J) && \text{by Definition 2.24} \\
&= \phi(a) + \phi(b)
\end{aligned}$$

and

$$\begin{aligned}
\phi(a \cdot b) &= (a \cdot b + I, a \cdot b + J) \\
&= \big((a + I) \cdot (b + I), (a + J) \cdot (b + J)\big) && \text{by Definition 1.32} \\
&= (a + I, a + J) \cdot (b + I, b + J) && \text{by Definition 2.24} \\
&= \phi(a) \cdot \phi(b).
\end{aligned}$$

We now compute the kernel of ϕ. For this, notice that

$$a \in \operatorname{Ker}(\phi) \iff (a + I, a + J) = (0 + I, 0 + J) \iff a \in I \cap J,$$

so $\operatorname{Ker}(\phi) = I \cap J$. Apply the Fundamental Isomorphism Theorem 2.14 to ϕ to see that

$$\overline{\phi}\colon R/(I \cap J) \longrightarrow \operatorname{Im}(\phi)$$

is an isomorphism. It remains to show that the image of ϕ is equal to the ring $R/I \times R/J$. To see this, consider an arbitrary element $(a + I, b + J) \in R/I \times R/J$. Since $R = I + J$, there exists $x \in I$ and $y \in J$ such that $1 = x + y$. Define $r := ay + bx \in R$. Then

$$\begin{aligned}
\phi(r) &= (ay + bx + I, ay + bx + J) \\
&= (ay + I, bx + J) && \text{as } bx \in I \text{ and } ay \in J \\
&= \big(a(1 - x) + I, b(1 - y) + J\big) && \text{as } 1 = x + y \\
&= \big(a - ax + I, b - by + J\big) \\
&= (a + I, b + J) && \text{as } x \in I \text{ and } y \in J.
\end{aligned}$$

Since $(a + I, b + J) \in R/I \times R/J$ was arbitrary, it follows that ϕ is surjective. $\square$

Example 2.26. Let $m, n \in \mathbb{Z}$ be coprime natural numbers. This means there exists $\lambda, \mu \in \mathbb{Z}$ such that $1 = \lambda m + \mu n$, i.e., $\mathbb{Z} = \mathbb{Z}m + \mathbb{Z}n$. Theorem 2.25 gives an isomorphism $\overline{\phi}\colon \mathbb{Z}_{mn} \to \mathbb{Z}_m \times \mathbb{Z}_n$; this is the Chinese Remainder Theorem from Algebra 1A.

2.6. **Field of fractions of an integral domain.** We now construct from every integral domain R, an injective homomorphism to a field $F(R)$, called the *field of fractions* of R.

Consider the set $T = \{(a, b) \in R \times R \mid b \neq 0\}$ together with two binary operations $T \times T \to T$ given by

$$(a, b) + (c, d) := (ad + bc, bd) \quad \text{and} \quad (a, b) \cdot (c, d) := (ac, bd).$$

These operations are well defined - that is, the formulas each define a map from $T \times T$ to T - precisely because R is an integral domain. Indeed, suppose otherwise, i.e., suppose that $bd = 0$. The fact that R is an integral domain forces either $b = 0$ or $d = 0$, but then either $(a, b) \notin T$ or $(c, d) \notin T$ which is absurd.

Lemma 2.27. *Define a relation $\sim$ on T by setting*

$$(a, b) \sim (c, d) \iff ad = bc.$$

Then for all $a, a', b, b', c, c', d, d' \in R$ with $b, b', d, d' \neq 0$, we have that

$$(a, b) \sim (a', b') \text{ and } (c, d) \sim (c', d') \implies \begin{cases} (a, b) + (c, d) \sim (a', b') + (c', d') \\ (a, b) \cdot (c, d) \sim (a', b') \cdot (c', d') \end{cases}$$

In other words, $\sim$ satisfies the conditions of being a congruence relation on T.

Proof. We're allowed to use

$$(a, b) \sim (a', b') \text{ i.e., that } ab' = a'b \tag{2.2}$$

and that

$$(c, d) \sim (c', d') \text{ i.e., that } cd' = c'd. \tag{2.3}$$

Notice that

$$\begin{aligned} (ad + bc)b'd' &= ab'dd' + bb'cd' \\ &= a'bdd' + bb'c'd && \text{using both (2.2) and (2.3)} \\ &= (a'd' + b'c')bd. \end{aligned}$$

This is equivalent to having

$$(a, b) + (c, d) = (ad + bc, bd) \sim (a'd' + b'c', b'd') = (a', b') + (c', d')$$

Similarly for the product formula, notice that (2.2) and (2.3) give

$$ab'cd' = a'bcd' = a'bc'd$$

which is equivalent to having $(a, b) \cdot (c, d) = (ac, bd) \sim (a'c', b'd') = (a', b') \cdot (c', d')$ as required. □

Just as with a congruence relation on a ring (see Definition 1.24), Lemma 2.27 shows that taking equivalence classes commutes with both of our binary operations on T, so we get a pair of binary operations on the set of equivalence classes

$$F(R) := T/\sim .$$

Following the standard convention in $\mathbb{Q}$, we write equivalence classes as $\frac{a}{b}$ rather than as $[(a,b)]$, so our operations become the familiar operations $F(R) \times F(R) \to F(R)$ given by

$$\frac{a}{b} + \frac{c}{d} := \frac{ad+bc}{bd} \quad \text{and} \quad \frac{a}{b} \cdot \frac{c}{d} := \frac{ac}{bd}. \tag{2.4}$$

Remark 2.28. It might have been nice to apply Theorem 1.26 directly to show that $T/\sim$ is a ring, but this is impossible because T isn't a ring: an element $(a,b) \in T$ does not have an additive inverse if b is not a unit. Despite this, we can show that $T/\sim$ is a ring; in fact, it's a field.

Theorem 2.29. *Let R be an integral domain. The set $F(R)$ with the binary operations from (2.4) above is a field; this is called the* field of fractions *of R. Moreover, the map $R \to F(R)$ defined by sending a to $\frac{a}{1}$ is an injective homomorphism.*

Proof. To check that $F(R)$ is an abelian group under addition, notice that for $\frac{a}{b}, \frac{c}{d}, \frac{e}{f} \in F(R)$, we have

$$\left(\frac{a}{b} + \frac{c}{d}\right) + \frac{e}{f} = \frac{ad+bc}{bd} + \frac{e}{f} = \frac{adf+bcf+bde}{bdf} = \frac{a}{b} + \frac{cf+de}{df} = \frac{a}{b} + \left(\frac{c}{d} + \frac{e}{f}\right)$$

so addition is associative. Addition is commutative in $F(R)$ because multiplication in the integral domain R is commutative (and addition is commutative) and hence

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd} = \frac{cb+da}{db} = \frac{c}{d} + \frac{a}{b}$$

The zero element is $\frac{0}{1}$ because

$$\frac{a}{b} + \frac{0}{1} = \frac{a \cdot 1 + b \cdot 0}{b \cdot 1} = \frac{a}{b} = \frac{0 \cdot b + 1 \cdot a}{1 \cdot b} = \frac{0}{1} + \frac{a}{b},$$

and the additive inverse of $\frac{a}{b}$ is $\frac{-a}{b}$ because $0 \cdot 1 = 0 = b^2 \cdot 0$ and hence in $F(R)$ we have

$$\frac{a}{b} + \frac{-a}{b} = \frac{ab + (-a)b}{b^2} = \frac{0}{b^2} = \frac{0}{1} = \frac{-ab+ab}{b^2} = \frac{-a}{b} + \frac{a}{b}.$$

Associativity of multiplication is much easier: multiplication in R is associative, so

$$\frac{a}{b} \cdot \left(\frac{c}{d} \cdot \frac{e}{f}\right) = \frac{a}{b} \cdot \frac{ce}{df} = \frac{a(ce)}{b(df)} = \frac{(ac)e}{(bd)f} = \frac{ac}{bd} \cdot \frac{e}{f} = \left(\frac{a}{b} \cdot \frac{c}{d}\right) \cdot \frac{e}{f}.$$

For the distributive laws, $b^2df(acf + ade) = bdf(abcf + abde)$, so in $F(R)$ we have

$$\begin{aligned}\frac{a}{b}\cdot\left(\frac{c}{d}+\frac{e}{f}\right) = \frac{a}{b}\cdot\frac{cf+de}{df} &= \frac{a(cf+de)}{bdf}\\ &= \frac{acf+ade}{bdf}\\ &= \frac{abcf+abde}{b^2df}\\ &= \frac{ac}{bd}+\frac{ae}{bf} = \frac{a}{b}\cdot\frac{c}{d}+\frac{a}{b}\cdot\frac{e}{f}.\end{aligned}$$

The other distributive law is similar. This proves that $F(R)$ with the given operations is a ring. Since R is a commutative ring with 1, the ring $F(R)$ is commutative (easy!) and $\frac{1}{1}$ makes it a ring with 1. It's not the zero ring, because $\frac{0}{1} \neq \frac{1}{1}$ (otherwise $0 = 1$ in R which is absurd). It remains to show that every nonzero element has a multiplicative inverse. For this, let $\frac{a}{b} \in F(R)$ be a nonzero equivalence class. Then $\frac{b}{a} \in F(R)$ satisfies

$$\frac{a}{b}\cdot\frac{b}{a} = \frac{ab}{ba} = \frac{1}{1} = \frac{ba}{ab} = \frac{b}{a}\cdot\frac{a}{b}$$

in $F(R)$ as required. Proving the statement about the homomorphism is easy. □

Examples 2.30. The two best known examples of this construction are the field $F(\mathbb{Z}) = \mathbb{Q}$ of rational numbers and, for any field $\Bbbk$, the field $F(\Bbbk[x]) = \Bbbk(x)$ of *rational functions*.

End of Week 4.

---

3. Factorisation in integral domains

Throughout this section we let R be an integral domain. We introduce several special classes of such rings and study factorisation properties.

3.1. **Primes and irreducibles in integral domains.** We've already seen many examples of integral domains:

(1) any field $\Bbbk$ is an integral domain by Remark 1.12;
(2) the ring of integers $\mathbb{Z}$ and the ring of Gaussian integers $\mathbb{Z}[i]$ are both integral domains by Lemma 1.20 (because they're both subrings of $\mathbb{C}$).
(3) the rings $R[x]$ and $R[[x]]$ associated to an integral domain R are integral domains by Exercise Sheets 1 and 2.

We first establish a property that characterises integral domains.

Lemma 3.1 (**Cancellation property**). *Let R be a commutative ring with 1 such that $0 \neq 1$. Then R is an integral domain if and only if for all $a, b, c \in R$, we have*

$$ab = ac \text{ and } a \neq 0 \implies b = c.$$

Proof. First, let R be an integral domain, and suppose $ab = ac$ and $a \neq 0$. Then

$$0 = ab + (-ac) = ab + a(-c) = a(b + (-c)).$$

Since R is an integral domain and $a \neq 0$, we have $b + (-c) = 0$, that is $b = c$. For the opposite implication, let R be a commutative ring with 1 such that $0 \neq 1$, and assume the cancellation property. Suppose $a, b \in R$ satisfies $ab = 0$ and $a \neq 0$. Then $ab = 0 = a \cdot 0$, and since $a \neq 0$ the cancellation property gives $b = 0$ as required. □

Definition 3.2 (**Divisibility**)**.** Let $a, b \in R$. We say that a *divides* b (equivalently, that b is divisible by a) if there exists $c \in R$ such that $b = ac$. We write simply $a|b$.

Any statement about divisibility can be rephrased in terms of ideals as follows:

Lemma 3.3. *For $a, b \in R$ we have $a|b \iff b \in Ra \iff Rb \subseteq Ra$.*

Proof. If $a|b$ then there exists $c \in R$ such that $b = ca \in Ra$. Since Ra is an ideal, it follows that $rb \in Ra$ for all $r \in R$, giving $Rb \subseteq Ra$. Conversely, if $Rb \subseteq Ra$, then in particular, $b \in Rb$ lies in Ra, and hence there exists $c \in R$ such that $b = ca$, so $a|b$. □

Recall that an element $a \in R$ is a *unit* if there exists $b \in R$ satisfying $ab = 1 = ba$.

Lemma 3.4 (**Units don't change the ideal**)**.** *Let R be an integral domain and let $a, b \in R$. Then*

$$Ra = Rb \iff a|b \text{ and } b|a \iff a = ub \text{ for some unit } u \in R.$$

In particular, $R = Ru$ if and only if u is a unit in R.

Proof. If $Ra = Rb$, then we have both $Ra \subseteq Rb$ and $Rb \subseteq Ra$, hence $b|a$ and $a|b$. Thus there exist $u, v \in R$ such that $a = ub$ and $b = va$. Putting these equations together shows that $1a = a = ub = uva$. If $a = 0$, then $b = 0$ and there's nothing to prove. Otherwise, the cancellation law in the integral domain R gives $uv = 1$, so u is a unit in R. Conversely, suppose $a = ub$ for some unit $u \in R$. Then $a \in Rb$, so $Ra \subseteq Rb$. Since u is a unit, we may multiply $a = ub$ by u^{-1} to obtain $b = u^{-1}a$. This gives $b \in Ra$ and hence $Rb \subseteq Ra$. These two inclusions together give $Ra = Rb$ as required. The final statement of the lemma follows from the special case $a = 1$. □

Definition 3.5 (**Primes and irreducibles**)**.** Let R be an integral domain. Let $p \in R$ be nonzero and not a unit. Then we say:

(1) p is *prime* iff for all $a, b \in R$, we have $p|ab \implies p|a$ or $p|b$.
(2) p is *irreducible* iff for all $a, b \in R$, we have $p = ab \implies a$ or b is a unit.

We say that p is *reducible* if it's not irreducible, i.e., if there exists $a, b \in R$ such that $p = ab$ where neither a nor b is a unit.

Examples 3.6. (1) The prime elements in $\mathbb{Z}$ are $\{\ldots, -7, -5, -3, -2, 2, 3, 5, 7, \ldots\}$, i.e., ± 1 times the positive prime numbers. The irreducible elements are identical.
(2) Let $\Bbbk$ be a field. Every nonzero element in $\Bbbk$ is a unit, so $\Bbbk$ contains neither primes nor irreducibles.

Proposition 3.7. *Let R be an integral domain. Then every prime element is irreducible.*

Proof. Let $p \in R$ be prime, and suppose $p = ab$. Then either $p|a$ or $p|b$. Assume without loss of generality (we may swap the letters a and b if we want) that $p|a$, i.e., there exists $c \in R$ such that $a = pc$. Then $p \cdot 1 = p = ab = pcb$, and the cancellation property gives $cb = 1$, so b must be a unit. This shows that p is irreducible. □

The converse is not true in general, see Exercise Sheet 5.

3.2. **Euclidean domains and PIDs.** We start by formalising a notion that you met in Algebra 1A when studying the rings $\mathbb{Z}$ and $\Bbbk[x]$ where $\Bbbk$ is a field.

Definition 3.8 (**Euclidean domain**)**.** Let R be an integral domain. A *Euclidean valuation* on R is a map $\nu\colon R \setminus \{0\} \to \{0, 1, 2, \ldots\}$ such that:

(1) for $f, g \in R \setminus \{0\}$ we have $\nu(f) \leq \nu(fg)$; and
(2) for all $f, g \in R$ with $g \neq 0$, there exists $q, r \in R$ such that

$$f = qg + r$$

and either $r = 0$ or $r \neq 0$ and $\nu(r) < \nu(g)$.

We say that R is a *Euclidean domain* if it has a Euclidean valuation.

Examples 3.9.
(1) Let $\Bbbk$ be any field, and define $\nu\colon \Bbbk\backslash\{0\} \to \{0, 1, 2, \ldots\}$ by setting $\nu(a) = 1$. Then ν is a Euclidean valuation (check it!), so $\Bbbk$ is a Euclidean domain.
(2) Absolute value $\nu(n) = |n|$ provides a Euclidean valuation on the ring of integers, so $\mathbb{Z}$ is a Euclidean domain.
(3) For $\Bbbk$ a field, the degree of a polynomial $\nu(f(x)) = \deg f(x)$ provides a Euclidean valuation on $\Bbbk[x]$ (see Algebra 1A), so $\Bbbk[x]$ is a Euclidean domain.
(4) In Exercise Sheet 1 you saw that $\mathbb{Z}[i] = \{a + bi \in \mathbb{C} : a, b \in \mathbb{Z}\}$ are a subring of the field $\mathbb{C}$, so $\mathbb{Z}[i]$ is an integral domain. On Exercise Sheet 5, you're asked to show that $\mathbb{Z}[i]$ is a Euclidean domain.

We now introduce Principal Ideal Domains. Let R be an integral domain. Since R is necessarily a commutative ring, Example 1.29 shows that each $a \in R$ determines an ideal

$$Ra := \langle a \rangle = \{r \cdot a \mid r \in R\}.$$

called the *ideal generated by* a.

Definition 3.10 (**PID**)**.** An ideal I of R is a *principal ideal* if $I = Ra$ for some $a \in R$. An integral domain R is a *Principal Ideal Domain* (PID) if every ideal in R is principal.

Theorem 3.11 (**Euclidean domains are PIDs**)**.** *Let R be a Euclidean domain. Then R is a PID.*

Proof. Let R be a Euclidean domain with Euclidean valuation ν. Let I be an ideal in R. If $I = \{0\}$ then $I = R0$, so I is principal. Otherwise we have $I \neq \{0\}$. Define

$$\mathcal{S} = \{\nu(a) \in \mathbb{Z}_{\geq 0} \mid a \in I, a \neq 0\}.$$

Since I is nonzero, this is a nonempty subset of $\{0, 1, 2, \ldots\}$ and hence we may choose g to be an element of I that achieves the minimum value in $\mathcal{S}$, i.e., $g \neq 0$ and $\nu(f) \geq \nu(g)$ for all $f \in I$. Now let $f \in I$. Since R is a Euclidean domain there exist $q, r \in R$ such that $f = qg + r$ and $r = 0$ or $\nu(r) < \nu(g)$. If $r \neq 0$ then $r = f - qg \in I$ which contradicts minimality in our choice of g. Thus $r = 0$, so $f = qg \in Rg$. Hence $I \subseteq Rg$. On the other hand, since $g \in I$ we have $Rg \subseteq I$. Hence $I = Rg$ and so I is principal. □

Examples 3.12. Theorem 3.11 implies that the following rings are PID's:

(1) any field;
(2) the ring of integers $\mathbb{Z}$;
(3) the polynomial ring $\Bbbk[x]$ with coefficients in a field $\Bbbk$; and
(4) the ring of Gaussian integers $\mathbb{Z}[i]$.

Examples 3.13.
(1) Exercise Sheet 5 asks you to prove that the integral domain $R = \mathbb{Z}[x]$ is not a PID, so it can't be a Euclidean domain.
(2) It is harder to produce a PID that is not a Euclidean domain. One example is the subring $R = \{a(\frac{1}{2} + \frac{\sqrt{19}}{2}i) \mid a \in \mathbb{Z}\}$ of $\mathbb{C}$. We shan't prove this.

3.3. **Key properties of PIDs.** We now examine some key properties of PIDs that go some way to explaining why they're an important class of rings.

Definition 3.14. Let R be a PID. Two elements $a, b \in R$ are said to be *coprime* if every common factor is a unit; by this, we mean that if $d|a$ and $d|b$, then d is a unit.

Lemma 3.15. *Let R be a PID and let $a, b \in R$ be coprime. There exists $r, s \in R$ such that $1 = ra + sb$.*

Proof. Consider the ideal $Ra + Rb$. Since R is a PID, there exists $d \in R$ such that

$$Ra + Rb = Rd.$$

In particular, $a, b \in Rd$, so d divides both a and b. Since a and b are coprime, it follows that d is a unit. Lemma 3.4 gives $Rd = R$ and hence $Ra + Rb = R$. Since R is a ring with 1, there exists $r, s \in R$ such that $1 = ra + sb$ as required. □

Proposition 3.16. *Let R be a PID. Then every irreducible element in R is prime.*

Proof. Suppose that $p|ab$ and that p does not divide a. Let d be a common factor of both a and p. In particular, we have $p = cd$ for some $c \in R$. Since p is irreducible, either d is a unit, or c is a unit. If c is a unit, then combining $d = c^{-1}p$ with the fact that $d|a$ would imply that $p|a$ which is a contradiction. Therefore d is a unit, so a and p are coprime. Lemma 3.15 gives $r, s \in R$ such that $1 = ra + sp$. Then

$$b = 1 \cdot b = (ra + sp) \cdot b = rab + psb.$$

We know ab is divisible by p, so b is divisible by p as required. □

Theorem 3.17. *Let R be a PID. If p is irreducible then R/Rp is a field.*

Proof. The ring R is commutative with 1, hence so is the quotient ring R/Rp. Lemma 3.4 implies that $Rp \neq R$ because p is not a unit, so R/Rp is not the zero ring. It remains to show that every nonzero element of R/Rp is a unit.

For this, let $a + Rp \in R/Rp$ be nonzero, i.e., $a + Rp \neq 0 + Rp$, i.e., $a \notin Rp$, i.e., p does not divide a. Let d be a common factor of a and p. In particular, $p = cd$ for some $c \in R$. Since p is irreducible, either d is a unit, or c is a unit. In fact c cannot be a unit (otherwise the equation $d = c^{-1}p$ shows that $p|d$ and since $d|a$ it follows that $p|a$ which is a contradiction), so d must be a unit. Therefore a and p are coprime, and Lemma 3.15 gives $r, s \in R$ such that $1 = ra + sp$. Then

$$1 + Rp = (ra + sp) + Rp = ra + Rp = (r + Rp) \cdot (a + Rp).$$

This shows that $a + Rp$ has a multiplicative inverse as required. □

Remark 3.18. **It is impossible to overstate how important Theorem 3.17 is.**

3.4. **Unique factorisation domains.** Finally, we're in a position to introduce the special class of integral domains that we really care about.

Definition 3.19 (UFD). An integral domain R is called a *Unique Factorisation Domain* (UFD) if

(1) every nonzero nonunit element in R can be written as the product of finitely many irreducibles in R; and
(2) given two such decompositions, say $r_1 \cdots r_s = r'_1 \cdots r'_t$ we have that $s = t$ and, after renumbering if necessary, we have $Rr_i = Rr'_i$ for $1 \leq i \leq s$.

Proposition 3.20. *Let R be a UFD. Then $p \in R$ is irreducible if and only if it is prime.*

Proof. Every prime is irreducible by Proposition 3.7 since R is an integral domain. Conversely, let $p \in R$ be irreducible, and suppose $a, b \in R$ satisfy $p|ab$, i.e.,

$$ab = cp$$

for some $c \in R$. If a is a unit, then $b = a^{-1}cp$ and so $p|b$ which is what we want to prove; and similarly for b. Therefore, we may assume that neither a nor b is a unit. Also, if a is zero, then we have $p|0$ (simply because every element divides the zero element), that is, $p|a$ which again is what we want to prove; and similarly b is nonzero. Therefore we may assume that a and b are nonzero and nonunit. By applying part (1) from Definition 3.19, we may take the irreducible factorisations of both $a = p_1 \cdots p_k$ and $b = p'_1 \cdots p'_\ell$ to get

$$p_1 \cdots p_k p'_1 \cdots p'_\ell = ab = cp.$$

We can further factorise the element c to get a pair of irreducible decompositions, so part (2) from Definition 3.19 and Lemma 3.4 together give a unit $u \in R$ such that either $p_i = up$ for some $1 \leq i \leq k$ or $p'_j = up$ for some $1 \leq j \leq \ell$. Substitute the expression into the decomposition of a or b to see that either $p|a$ or $p|b$ as required. □

Theorem 3.21. *Let R be a PID. Then R is a UFD.*

Proof. We first show part (1) of Definition 3.19. Let $a \in R$ be a nonzero, nonunit element and suppose for a contradiction a cannot be written as a finite product of irreducibles. In particular, a itself is reducible, so there exists a decomposition

$$a = a_1 b_1$$

for some $a_1, b_1 \in R$ where both a_1 and b_1 are nonunits (and nonzero because a is nonzero). If both a_1 and b_1 can be expressed as products of irreducibles then a can as well which is absurd, so at least one of them cannot be written in this way. Without loss of generality, suppose that this is a_1. Notice that

$$Ra \subseteq Ra_1 \text{ (because } a_1|a) \text{ and } Ra \neq Ra_1 \text{ (because } b \text{ is not a unit), hence } Ra \subsetneqq Ra_1.$$

Applying the same argument to a_1 produces an element $a_2 \in R$ that cannot be expressed as a product of irreducibles such that $Ra_1 \subsetneqq Ra_2$. Repeat to obtain a strictly increasing chain of ideals in R:

$$Ra \subsetneqq Ra_1 \subsetneqq Ra_2 \subsetneqq Ra_3 \cdots$$

This completes the first step of the proof. As a second step, we show that the union

$$I = Ra_1 \cup Ra_2 \cup \cdots$$

is an ideal. Indeed, $0 \in Ra \subseteq I$, so I is nonempty. Let $b, c \in I$ and $r \in R$. There exists $i \geq 1$ such that $b, c \in Ra_i$, therefore $b - c, rb, br \in Ra_i \subseteq I$. Thus I is an ideal. For step three, since R is a principal ideal domain we have that $I = Rd$ for some $d \in R$. Then $d = 1 \cdot d \in I$ and thus $d \in Ra_i$ for some $i \geq 1$. But then

$$Ra_{i+1} \subseteq I = Rd \subseteq Ra_i \subsetneqq Ra_{i+1}$$

which is absurd. This contradiction proves Definition 3.19(1).

For part (2) of Definition 3.19, suppose

$$p_1 \cdots p_s = p'_1 \cdots p'_t \tag{3.1}$$

are two such decompositions where we may assume without loss of generality that $s \leq t$. Equation (3.1) shows that p_1 divides $p'_1 \cdots p'_t$. Proposition 3.16 shows that the irreducible element p_1 is prime, so $p_1 | p'_i$ for some $1 \leq i \leq t$. Thus $p'_i = ap_1$, and since p'_i is irreducible it follows that a must be a unit and hence $Rp_1 = Rp'_i$ by Lemma 3.4. Relabel p'_i as p'_1 and vice-versa, giving $Rp_1 = Rp'_1$, so there exists a unit $u_1 \in R$ such that $p'_1 = u_1 p_1$, giving

$$p_1 \cdots p_s = p'_1 \cdots p'_t = u_1 p_1 p'_2 \cdots p'_t.$$

The cancellation property in the integral domain R leaves

$$p_2 \cdots p_s = p'_1 \cdots p'_t = u_1 p'_2 \cdots p'_t.$$

Repeat for each element on the left hand side, giving $Rp_i = Rp'_i$ for all $1 \leq i \leq s$ and

$$1 = u_1 \cdots u_s p'_{s+1} \cdots p'_t.$$

But the p'_j are prime and hence nonunits, so we must have $s = t$. □

As an application, we obtain the following result which you saw in Algebra 1A:

Corollary 3.22 (**Fundamental Theorem of Arithmetic**). *Every natural number greater than 1 is of the form* $\Pi p_i^{n_i}$ *for distinct prime numbers* p_i *and positive integers* n_i*. The primes* p_i *and their exponents* n_i *are uniquely determined (up to order).*

Proof (non-examinable, just for interest). The ring $\mathbb{Z}$ is a Euclidean domain, so it's a PID by Theorem 3.11 and hence a UFD by Theorem 3.21. Therefore every integer, and in particular every positive integer, can be written as the product of finitely many irreducible (=prime) elements in $\mathbb{Z}$, each of which is prime by Proposition 3.20. Some of these primes could be negative (because -1 is a unit in $\mathbb{Z}$), but if we factor out all the minus signs then we get the desired decomposition as a product of primes. The only units in $\mathbb{Z}$ are ± 1, so having $\mathbb{Z}p = \mathbb{Z}q$ with $p, q > 0$ forces $p = q$, so the decomposition is unique up to the order in which we write the factors. □

Example 3.23. The ring $\mathbb{Z}[x]$ is not a PID (see Exercise Sheet 5), but we will shortly see that it is a UFD.

To summarise, we've shown that

$$\text{Euclidean domain} \implies \text{PID} \implies \text{UFD} \implies \text{integral domain.}$$

In particular, each ring listed in Examples 3.9 is a UFD.

End of Week 5.

3.5. **Polynomials over a UFD.** Let R be an integral domain. A common factor c of $a_1, \ldots, a_m \in R$ is called a *highest common factor* (hcf) if for any other common factor b of $a_1, \ldots, a_m$ we have $b|c$. Evidently if both b and c are hcfs of $a_1, \ldots, a_m$, then $b|c$ and $c|b$, so $b = uc$ for a unit u by Lemma 3.4.

If all common factors of $a_1, \ldots, a_m$ are units, we say $a_1, \ldots, a_m$ are coprime.

Lemma 3.24 (**Highest common factors in a UFD**). *If* R *is a UFD and* $a_1, \ldots, a_m \in R$ *are not all zero, then they have an hcf* c*.*

Proof. We induct on m, assuming $a_1, \ldots, a_m$ are not all zero. If $m = 1$ then we can take $c = a_1$. Otherwise, we may assume by reordering that $a_1, \ldots, a_{m-1}$ are not all zero and have an hcf $a \neq 0$. Observe that an hcf c of a and $b = a_m$ is an hcf of $a_1, \ldots, a_m$ since it is clearly a common factor, and any common factor d of $a_1, \ldots, a_m$ divides both a and b, hence c. If $b = 0$ or a is a unit, we may take $c = a$, and if b is a unit, we may take $c = b$.

Otherwise, since R is a UFD, we may write $a = p_1 \cdots p_k$ and $b = q_1 \cdots q_\ell$ as products of irreducibles. Now if some p_i divides b, it must divide some q_j (since irreducibles are prime) in which case, by reordering, we may assume $i = j = 1$ and $q_1 = u_1 p_1$ for a unit u_1. Repeating this process with a/p_1 and b/p_1, we eventually obtain that $a = ca'$ and $b = cb'$ where, for some $0 \leq j \leq \min\{k, \ell\}$ and some unit u, $a' = p_{j+1} \cdots p_k$ and $b' = u q_{j+1} \cdots q_\ell$ have only units as common factors, and $c = p_1 \cdots p_j$. Hence c is a common factor of a and b, and if $d = r_1 \cdots r_m$ is another, written as a product of irreducibles, then r_1

cannot divide both a' and b', so it must divide some p_i with $1 \leq i \leq j$. By reordering, we may assume $p_1 = v_1 r_1$ for a unit v_1, and repeating this process with d/r_1 and c/r_1, we eventually obtain that $d|c$. Hence c is an hcf of a and b. □

Definition 3.25 (**Primitive polynomial**). A nonzero polynomial $f \in R[x]$, with R a UFD, is *primitive* if its coefficients are coprime. More generally, we say f has *content* $c \in R$ if c is an hcf of the coefficients of f (thus if c is a unit, f is primitive).

Example 3.26. In $\mathbb{Z}[x]$, $f = 3x^3 + 6x - 3$ has content 3 and $g = x^3 + 2x - 1$ is primitive.

Lemma 3.27 (**Pulling out the content**). *Let R be a UFD. A nonzero polynomial $f \in R[x]$ has content c if and only if $f = c \cdot g$ with $g \in R[x]$ primitive. In particular c and g are uniquely determined by f up to multiplication by a unit.*

Proof. We may write $f = a \cdot g$ with $a \in R$ and $g \in R[x]$ if and only if a is a common factor of the coefficients of f. Then if g has content b, f has content ab. Thus $f = c \cdot g$ has content c if and only if g is primitive.

Now if $f = c \cdot g = d \cdot h$ with g, h primitive, then $d = uc$ for some unit u, so $c \cdot g = d \cdot h = (uc) \cdot h$ and hence $g = u \cdot h$ by cancellation. □

Since every UFD R is an integral domain, it has a field of fractions F (see section 2.6). For example, $\mathbb{Z}$ has field of fractions $\mathbb{Q}$. We now relate factorizations in $R[x]$ and $F[x]$.

Lemma 3.28. *Let R be a UFD with field of fractions F, and suppose $h \in R[x]$.*

(1) *If $h = fg$ with $f, g \in R[x]$ primitive, then h is primitive.*
(2) *If $h = f_1 f_2 \cdots f_k$ where $f_j \in R[x]$ has content c_j, then h has content $c_1 c_2 \cdots c_k$.*
(3) *If h is irreducible in $F[x]$ and primitive in $R[x]$, then h is irreducible in $R[x]$.*
(4) *If $h = g_1 g_2 \cdots g_k$ where $g_j \in F[x]$, then $h = c \cdot f_1 f_2 \cdots f_k$ where $c \in R$, each $f_j \in R[x]$ is primitive and $g_j = u_j \cdot f_j$ for some unit $u_j \in F$.*

Proof. (1) Let

$$f = \sum_{i=0}^{n} a_i x^i \quad \text{and} \quad g = \sum_{i=0}^{m} b_i x^i$$

be primitive and suppose $h = fg$ has content c. If p is any irreducible factor of c, then since f, g are primitive, there is a least i such that p does not divide the coefficient a_i of f and a least j such that p does not divide the coefficient b_j of g. The coefficient of x^{i+j} in $h = fg$ is

$$(a_0 b_{i+j} + \cdots + a_{i-1} b_{j+1}) + a_i b_j + (a_{i+1} b_{j-1} + \cdots + a_{i+j} b_0). \tag{3.2}$$

Minimality of i implies that p divides $a_0 b_{i+j} + \cdots + a_{i-1} b_{j+1}$, while minimality of j implies that p divides $a_{i+1} b_{j-1} + \cdots + a_{i+j} b_0$. But p divides the content c of fg, so it must divide the coefficient (3.2) and hence it must divide $a_i b_j$. But p is irreducible and hence prime by Proposition 3.20, so p must divide either a_i or b_j, which is a contradiction. Hence c has no irreducible factors, so it must be a unit, since R is a UFD.

(2) By Lemma 3.27, $f_j = c_j \cdot f_j'$ with f_j' primitive. Hence $h = f_1 \cdots f_k = (c_1 \cdots c_k) \cdot (f_1' \cdots f_k')$, and $f_1' \cdots f_k'$ is primitive by (1) and induction on k. Hence h has content $c_1 \cdots c_k$ by Lemma 3.27.

(3) If $h \in R[x]$ is irreducible in $F[x]$, then it is nonzero and a nonunit in $F[x]$, so it has positive degree, and is neither zero nor a unit in $R[x]$. Suppose now that $h = fg$ for $f, g \in R[x]$. Then either f or g must be a unit in $F[x]$, say f, i.e., f is a nonzero constant a in $F \cap R[x] = R$. Now $h = a \cdot g$ is primitive and so its content is a unit divisible by a. Thus a is a unit in R, i.e., f is a unit in $R[x]$. Hence h is irreducible in $R[x]$.

(4) If $h = g_1 g_2 \cdots g_k$, then each g_j has coefficients of the form a/v for $a, v \in R$ with $v \neq 0$. Clearing denominators, for some nonzero $v_j \in R \subseteq F$, $v_j g_j = h_j \in R[x]$. Thus

$$d \cdot h = h_1 h_2 \cdots h_k \tag{3.3}$$

in $R[x]$ with $d = v_1 v_2 \cdots v_k \in R$. Now suppose $h = c' \cdot f$ and $h_j = c_j \cdot f_j$ for each j, where $f, f_1, \ldots, f_k \in R[x]$ are primitive and $c', c_1, \ldots, c_k \in R$. Then $d \cdot h$ has content dc' and also, by (3.3) and (2), content $c_1 \cdots c_k$. Hence $c_1 \cdots c_k = dc$, where $c = c'u$ for some unit $u \in R$, and

$$d \cdot h = (c_1 \cdots c_k) \cdot (f_1 \cdots f_k) = dc \cdot (f_1 \cdots f_k).$$

Hence $h = c \cdot (f_1 \cdots f_k)$ by cancellation, and each f_j is primitive with $v_j g_j = c_j f_j$ and $v_j, c_j \in R$ nonzero. Thus v_j and c_j are units in F, as is $u_j = c_j / v_j$, and $g_j = u_j f_j$. □

Corollary 3.29 (Gauss' Lemma). *Let R be a UFD with field of fractions F, and let $h \in R[x]$. Then h is irreducible in $R[x]$ if and only if either it is an irreducible element of R, or it is primitive in $R[x]$ and irreducible in $F[x]$.*

Proof. ($\Rightarrow$) Lemma 3.27 gives $h = c \cdot g$ for $c \in R$ and $g \in R[x]$ primitive. Since h is irreducible in $R[x]$, either:

(1) g is a unit in $R[x]$, in which case $g \in R$, hence $h \in R$, and irreducibility of h in $R[x]$ forces irreducibility of h in R as required; or

(2) c is a unit, in which case g being primitive implies h is primitive. If $h \in R$, then h is a unit in R, hence a unit in $R[x]$, contradicting irreducibility. Thus h has positive degree and hence is not a unit in $F[x]$. Now if $h = g_1 g_2$ in $F[x]$ then by Lemma 3.28 (4), $h = d \cdot f_1 f_2$ with $d \in R$, $f_1, f_2 \in R[x]$ primitive, and $g_j = u_j f_j$ for units $u_1, u_2 \in F$. By irreducibility of h in $R[x]$, either f_1 or f_2 is a unit in $R[x]$, hence either g_1 or g_2 is a unit in $F[x]$. We conclude that h is irreducible in $F[x]$.

($\Leftarrow$) If $h \in R$ is irreducible then it's irreducible in $R[x]$ for degree reasons, while if $h \in R[x]$ is primitive in $R[x]$ and irreducible in $F[x]$, it is irreducible in $R[x]$ by Lemma 3.28 (3). □

Since F is a field, Examples 3.12 and Theorem 3.21 show that $F[x]$ is a UFD. Now suppose $h \in R[x]$ is a nonzero nonunit. If h is not in R, it is a nonzero nonunit in $F[x]$ hence has a factorization $h = g_1 \cdots g_k$ into irreducibles in $F[x]$. Thus by Lemma 3.28(4), $h = c \cdot f_1 \cdots f_k$, where $c \in R$ and each $f_j \in R[x]$ is primitive and $g_j = u_j \cdot f_j$ for some units $u_j \in F$. If $h \in R$, this also holds with $k = 0$. Irreducibility of g_j in $F[x]$ forces

irreducibility of f_j in $F[x]$, hence in $R[x]$ by Lemma 3.28(3). Now since R is a UFD, we may factorize $c = r_1 \cdots r_\ell$ into irreducibles in R, so that

$$h = r_1 \cdots r_\ell \cdot f_1 \cdots f_k \tag{3.4}$$

is a factorization into irreducibles in $R[x]$ by Gauss' Lemma (Corollary 3.29).

Theorem 3.30. *If R is a UFD, then the polynomial ring $R[x]$ is also a UFD.*

Proof. Definition 3.19(1), i.e., existence, follows from (3.4) above. For the uniqueness condition of Definition 3.19(2), suppose $h \in R[x]$ admits two such decompositions

$$r_1 \cdots r_\ell \cdot f_1 \cdots f_k = r'_1 \cdots r'_m \cdot f'_1 \cdots f'_n.$$

The content of f is unique up to multiplication by a unit, so $r_1 \cdots r_\ell = u \cdot r'_1 \cdots r'_m$ for some unit $u \in R$. Since R is a UFD, we have $\ell = m$ and (after permuting indices) $Rr_i = Rr'_i$ for $1 \le i \le \ell$. Similarly, the primitive part of h is unique up to multiplication by a unit, so there is a unit $u' \in R$ such that $f_1 \cdots f_k = u' \cdot f'_1 \cdots f'_n$, and each f_i, f'_j is irreducible in $F[x]$ by Gauss' Lemma (Corollary 3.29). Since $F[x]$ is a UFD, $k = n$ and (after permuting indices) $Ff_i = Ff'_i$ for $1 \le i \le k$. Now Lemma 3.4 gives a unit $u_i \in F$ such that $f_i = u_i f'_i \in R[x]$. Write $u_i = a_i/b_i$ and multiply $f_i = u_i f'_i$ by b_i, then use Lemma 3.27 and primitivity of f_i, f'_i to see that f_i is a unit times f'_i as required. □

End of Week 6.

---

4. Algebras and fields

In this chapter we study a class of rings with 1 that are simultaneously vector spaces.

4.1. Algebras.

Throughout this chapter we let $\Bbbk$ be a field.

Definition 4.1 (**$\Bbbk$-algebra**)**.** A $\Bbbk$-vector space V is called a $\Bbbk$-*algebra* if it's also a ring, where the scalar product and the ring multiplication are compatible in the following sense:

$$\lambda(u \cdot v) = (\lambda u) \cdot v = u \cdot (\lambda v) \quad \text{for all } u, v \in V, \lambda \in \Bbbk. \tag{4.1}$$

The *dimension* of a $\Bbbk$-algebra V is the dimension of V as a vector space over $\Bbbk$, and a nonempty subset W of V is a *subalgebra* if it is both a subring and a vector subspace.

Remarks 4.2. (1) For $v \in V$, the 'multiply on the left by v' map $T_v \colon V \to V$ given by $T_v(u) = v \cdot u$ is a $\Bbbk$-linear map; the same is true for 'multiply on the right'.

(2) Suppose that $(v_i)_{i \in I}$ is a basis for the $\Bbbk$-algebra V. To determine the multiplication on V, it suffices to know only the values of $v_i \cdot v_j$ for all $i, j \in I$, because

$$\left(\sum_{i \in I} \alpha_i v_i\right) \cdot \left(\sum_{j \in I} \beta_j v_j\right) = \sum_{i,j \in I} (\alpha_i \beta_j)(v_i \cdot v_j).$$

Examples 4.3. (1) Let $\Bbbk$ be a field. Then $\Bbbk = \Bbbk \cdot 1$ is a $\Bbbk$-algebra of dimension 1.

(2) For $n \geq 1$, the set $M_n(\Bbbk)$ of $n \times n$ matrices with coefficients in $\Bbbk$ is a $\Bbbk$-algebra of dimension n^2.
(3) The field $\mathbb{C} = \mathbb{R} + \mathbb{R}i$ is an $\mathbb{R}$-algebra that is a 2-dimensional vector space over $\mathbb{R}$.

Example 4.4 (**The quaternions**). Consider the vector space of dimension 4 over $\mathbb{R}$ with basis $1, i, j, k$, that is

$$\mathbb{H} = \mathbb{R} + \mathbb{R}i + \mathbb{R}j + \mathbb{R}k = \big\{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\big\},$$

where the $\mathbb{R}$-bilinear product is determined from

$$i^2 = j^2 = k^2 = -1, \quad ij = k, \ jk = i, \ ki = j, \ ji = -k, \ kj = -i, \ ik = -j.$$

Exercise Sheet 7 asks you to show that $\mathbb{H}$ is a noncommutative ring. Since the product was defined to be $\mathbb{R}$-bilinear, it follows that $\mathbb{H}$ is an $\mathbb{R}$-algebra of dimension 4; this is the *quaternionic algebra*, or simply, *the quaternions*. You'll see in Exercise Sheet 8 that $\mathbb{H}$ is a division ring; this is the first time that you've seen a division ring that is not a field!

4.2. **General polynomial rings.** To introduce a large class of $\Bbbk$-algebras, we study polynomials in several variables. For this, let $n \geq 1$, let $x_1, \ldots, x_n$ be variables and let R be a ring. A *polynomial* f in $x_1, \ldots, x_n$ with coefficients in R is a formal sum

$$f(x_1, \ldots, x_n) = \sum_{i_1, \ldots, i_n \geq 0} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n}, \tag{4.2}$$

with coefficients $a_{i_1,\ldots,i_n} \in R$ for all tuples $(i_1, \ldots, i_n) \in \mathbb{N}^n$, where only finitely many of the $a_{i_1,\ldots,i_n}$ are nonzero. To avoid having to write so many indices, let's write $a_I := a_{i_1,\ldots,i_n}$ and $x^I := x_1^{i_1} \cdots x_n^{i_n}$ for any n-tuple $I = (i_1, \ldots, i_n) \in \mathbb{N}^n$. Then every polynomial in $x_1, \ldots, x_n$ can be written in the form

$$f = \sum_{I \in \mathbb{N}^n} a_I x^I$$

where only finitely many of the elements $a_I \in R$ are nonzero and where $x^I := x_1^{i_1} \cdots x_n^{i_n}$.

Definition 4.5 (**General polynomial ring**). For $n \geq 1$, the *polynomial ring in n variables with coefficients in R* is the set $R[x_1, \ldots, x_n]$ of all polynomials in $x_1, \ldots, x_n$ with coefficients in R, where for $f = \sum_{I \in \mathbb{N}^n} a_I x^I$ and $g = \sum_{I \in \mathbb{N}^n} b_I x^I$ we define

$$f + g = \sum_{I \in \mathbb{N}^n} (a_I + b_I) x^I \quad \text{and} \quad f \cdot g = \sum_{I \in \mathbb{N}^n} \left(\sum_{J+K=I} a_J \cdot b_K \right) x^I.$$

Example 4.6. To illustrate this, set $n = 3$ and write $\mathbb{R}[x, y, z]$ for the polynomial ring in three variables. Then for $f = x^2y + 3xz$ and $g = 2x - 3xz$, we have

$$f + g = x^2y + 2x \quad \text{and} \quad f \cdot g = 2x^3y + 6x^2z - 3x^3yz - 9x^2z^2.$$

Remarks 4.7. (1) This ring is a generalisation of the ring $R[x]$ from Example 1.21, and the proof that it is a ring can be carried out directly just as in Example 1.15.

(2) If the coefficient ring R is a field $\Bbbk$, then the general polynomial ring $\Bbbk[x_1, \ldots, x_n]$ is a $\Bbbk$-algebra with basis as a vector space given by all monomials

$$x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n} : i_1, \ldots, i_n \in \mathbb{N};$$

this vector space is not finite dimensional! As in Remark 4.2, multiplication of polynomials is determined by the bilinearity and multiplication of monomials:

$$(x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}) \cdot (x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n}) = x_1^{i_1+j_1} x_2^{i_2+j_2} \cdots x_n^{i_n+j_n}.$$

Proposition 4.8. *The polynomial ring $R[x_1, \ldots, x_n]$ in n variables is isomorphic to the polynomial ring $S[x_n]$ in the variable x_n with coefficients in $S = R[x_1, \ldots, x_{n-1}]$.*

Proof. The idea is that for any $f = \sum_{i_1,\ldots,i_n \geq 0} a_{i_1,\ldots,i_n} x_1^{i_1} \cdots x_n^{i_n}$ in the ring $R[x_1, \ldots, x_n]$, gathering all terms involving $x_n^{i_n}$ for each power $i_n \geq 0$ gives an expression

$$f(x_1, \ldots, x_n) = \sum_{i_n \geq 0} \left(\sum_{i_1,\ldots,i_{n-1} \geq 0} a_{i_1,\ldots,i_n} x_1^{i_1} \cdots x_{n-1}^{i_{n-1}} \right) x_n^{i_n}, \tag{4.3}$$

which we may regard as an element of $S[x_n]$ if we view the elements in the parentheses as coefficients in S. See Exercise Sheet 7 for details. □

Corollary 4.9. *Let $\Bbbk$ be a field. Then $\Bbbk[x_1, \ldots, x_n]$ is a UFD.*

Proof. We know $\Bbbk$ is a PID, hence a UFD. Assume by induction that $S := \Bbbk[x_1, \ldots, x_{n-1}]$ is a UFD, then $S[x_n]$ is a UFD by Theorem 3.30 and we're done by Proposition 4.8. □

Remark 4.10. Note that $\Bbbk[x_1, \ldots, x_n]$ is not a PID for $n \geq 2$, see Exercise Sheet 7.

4.3. **Constructing field extensions.** We now construct new fields from old.

Definition 4.11 (**Subfield and field extension**)**.** A non-zero subring $\Bbbk \neq \{0\}$ of a field K is a *subfield* if for each nonzero element $a \in \Bbbk$, the multiplicative inverse of a in K lies in $\Bbbk$. We also refer to $\Bbbk \subseteq K$ as a *field extension.*

In this case, choose non-zero $a \in \Bbbk$ to write $1_K = a \cdot a^{-1}$ whence $1_K \in \Bbbk$ and then it is easy to see that $\Bbbk$ is a field in its own right with $1_\Bbbk = 1_K$. Conversely, if $\Bbbk$ is a non-zero subring of a field K that is a field (so there is a multiplicative identity in $\Bbbk$ and each non-zero $a \in \Bbbk$ has a multiplicative inverse in $\Bbbk$) then $\Bbbk$ is a subfield: $1_\Bbbk = 1_K$ and the inverses in $\Bbbk$ and K coincide. Moreover, K gets structure too:

Lemma 4.12. *Let $\Bbbk \subseteq K$ be a field extension. Then K is a $\Bbbk$-algebra.*

Proof. K is a field so that $(K, +)$ is already an abelian group. We now restrict the multiplication $K \times K \to K$ to obtain a scalar multiplication $\Bbbk \times K \to K$. We then have

$$\begin{aligned}
\lambda(\mu v) &= (\lambda\mu)v, && \text{as multiplication is associative} \\
1_\Bbbk \cdot v &= 1_K \cdot v = v && \text{as } 1_\Bbbk = 1_K \\
(\lambda + \mu)v &= \lambda v + \mu v && \text{as the distributive laws hold in } K, \\
\lambda(v + w) &= \lambda v + \lambda w && \text{as the distributive laws hold in } K
\end{aligned}$$

for $v \in K$ and $\lambda, \mu \in \Bbbk$, so K is a vector space over $\Bbbk$. In addition, multiplication in K is associative and commutative, so $(\lambda v) \cdot w = v \cdot (\lambda w) = \lambda(vw)$ for $v, w \in K$ and $\lambda \in \Bbbk$. $\square$

Given a field extension $\Bbbk \subseteq K$, we now construct intermediate fields $\Bbbk \subseteq \Bbbk[a] \subseteq K$.

Theorem 4.13 (**Constructing intermediate fields**). *Let $\Bbbk \subseteq K$ be a field extension, and let $a \in K$ be a root of some nonzero polynomial in $\Bbbk[x]$. The set*

$$\Bbbk[a] := \big\{ f(a) \in K \mid f \in \Bbbk[x] \big\}$$

is a field, with field extensions $\Bbbk \subseteq \Bbbk[a] \subseteq K$. In fact $(1, a, a^2, \ldots, a^{n-1})$ is a basis for $\Bbbk[a]$ over $\Bbbk$ where $n = \min\{\deg(p) \mid p \in \Bbbk[x] \text{ satisfies } p(a) = 0\}$.

Proof. Consider the evaluation homomorphism $\phi_a \colon \Bbbk[x] \to K$ given by $\phi_a(f) = f(a)$ from Example 2.6. Since $\Bbbk$ is a field, $\Bbbk[x]$ is a PID and hence $\mathrm{Ker}(\phi_a)$ is a principal ideal, that is, $\mathrm{Ker}(\phi_a) = \Bbbk[x]p$ for some $p \in \Bbbk[x]$.

We claim that p is irreducible. If $p = 0$, then $\mathrm{Ker}(\phi_a) = 0$ which is absurd because a is a root of some nonzero polynomial by assumption; and if p were a unit, then $\mathrm{Ker}(\phi_a) = \Bbbk[x]$ which is absurd because nonzero constant polynomials do not lie in $\mathrm{Ker}(\phi_a)$. Finally, supose there exists $f, g \in \Bbbk[x]$ such that $p = fg$ where neither f nor g is a unit, i.e., where f, g of degree smaller than that of p. But as $f(a)g(a) = p(a) = 0$, so (at least) one of $f(a)$ or $g(a)$ is 0, say $f(a)$. Then $f \in \mathrm{Ker}(\phi_a) = \Bbbk[x]p$ and hence $p|f$ which is absurd as f is a non-zero polynomial of smaller degree than p. This proves that p is irreducible after all. In particular, $n = \deg p$.

Now observe that $\Bbbk[a]$ is the image of the ring homomorphism ϕ_a so that it is a subring of K isomorphic to $\Bbbk[x]/\mathrm{Ker}\phi_a = \Bbbk[x]/\Bbbk[x]p$ by the First Isomorphism Theorem 2.14. However, by Theorem 3.17, $\Bbbk[x]/\Bbbk[x]p$ is a field so that $\Bbbk[a]$ is too.

Lemma 4.12 shows that $\Bbbk[a]$ is a $\Bbbk$-algebra, so it remains to show $(1, a, a^2, \ldots, a^{n-1})$ is a basis of $\Bbbk[a]$ over $\Bbbk$. To show spanning, let $f(a) \in \Bbbk[a]$. Since $\Bbbk[x]$ is a Euclidean domain, division of f by p gives $q, r \in \Bbbk[x]$ such that $f = qp + r$ where either $r = 0$ or $\deg(r) < \deg(p) = n$, say $r = b_0 + b_1 x + \cdots + b_{n-1}x^{n-1}$. In either case

$$\begin{aligned} f(a) &= q(a)p(a) + r(a) \\ &= r(a) \\ &= b_0 \cdot 1 + b_1 a + \cdots + b_{n-1}a^{n-1}. \end{aligned}$$

Thus $f(a)$ is a linear combination of $1, a, \ldots, a^{n-1}$. To show that $1, a, \ldots, a^{n-1}$ are linearly independent, suppose $c_0 \cdot 1 + c_1 a + \cdots + c_{n-1}a^{n-1} = 0$. Then $h := c_0 + c_1 x + \cdots + c_{n-1}x^{n-1}$ lies in $\mathrm{Ker}(\phi_a) = \Bbbk[x]p$, so $p|h$. Since $\deg(h) < \deg(p)$, this is possible only if $h = 0$, that is, only if $c_0 = c_1 = \cdots = c_{n-1} = 0$. $\square$

Examples 4.14.

(1) We have that $\mathbb{R} \subseteq \mathbb{C}$ and that $i \in \mathbb{C}$ is a root of the irreducible polynomial $x^2 + 1 \in \mathbb{R}[x]$. Here $\mathbb{R}[i] = \mathbb{R} + \mathbb{R}i = \mathbb{C}$ has basis $(1, i)$.

(2) We have that $\mathbb{Q} \subseteq \mathbb{R}$ and that $\sqrt[3]{2}$ is a root of the irreducible polynomial $x^3 - 2 \in \mathbb{Q}[x]$. Here $\mathbb{Q}[\sqrt[3]{2}] = \mathbb{Q} + \mathbb{Q}\sqrt[3]{2} + \mathbb{Q}(\sqrt[3]{2})^2$ has basis $(1, \sqrt[3]{2}, (\sqrt[3]{2})^2)$.

We now prove a kind of converse to Theorem 4.13. Suppose that we have only the field $\Bbbk$ and an irreducible polynomial $p \in \Bbbk[x]$. We now construct a field extension $\Bbbk \subseteq K$ and an element $a \in K$ such that a is a root of p.

Theorem 4.15 (Constructing field extensions containing roots). *Let $p \in \Bbbk[x]$ be irreducible in $\Bbbk[x]$. The field extension $\Bbbk \subseteq K := \Bbbk[x]/\Bbbk[x]p$ has dimension $n := \deg(p)$ as a $\Bbbk$-vector space, and the element $a := [x] \in K$ in this new field is a root of p.*

Proof. Since $\Bbbk$ is a field, $\Bbbk[x]$ is a PID, so Theorem 3.17 shows that irreducibility of p implies that $K = \Bbbk[x]/\Bbbk[x]p$ is a field. The multiplicative identity in K is $[1] \in K$, so if we identify $\Bbbk$ with the subfield $\Bbbk[1] \subseteq K$ then we have that $\Bbbk \subseteq K$ is a field extension.

Now let $a = [x] \in K$ and let $f \in \Bbbk[x]$. Write $f = \sum_i c_i x^i$. Then

$$f(a) = \sum_i c_i a^i = \sum_i c_i [x^i] = [\sum_i c_i x^i] = [f].$$

In particular, $p(a) = [p] = [0]$ so that a is a root of p in K and

$$\Bbbk[a] = \{f(a) \colon f \in \Bbbk[x]\} = \{[f] \colon f \in \Bbbk[x]\} = K.$$

Hence by Theorem 4.13, the dimension of K as a vector space over $\Bbbk$ is the minimum degree of a polynomial in $\Bbbk[x]$ that vanishes at a. Now p is such a polynomial and if h is another, then $h(a) = [h] = [0]$ so $p|h$ and hence $\deg h \geq \deg p$. Thus $\dim K = \deg p$. □

Corollary 4.16 (Construction of splitting fields). *Let $\Bbbk$ be a field and let $f \in \Bbbk[x]$ be nonconstant. Then there exists a field extension $\Bbbk \subseteq K$ and an element $a \in K$ such that $f(a) = 0$. Moreover, f can be written as product of polynomials of degree 1 in $K[x]$.*

Proof. See Exercise Sheet 7. □

Examples 4.17.

(1) The polynomial $p = x^2 + 1 \in \mathbb{R}[x]$ is irreducible in $\mathbb{R}[x]$, so Theorem 4.15 gives a root a in the field

$$\mathbb{R}[x]/\mathbb{R}[x](x^2 + 1) = \mathbb{R} + \mathbb{R}a,$$

where $a = [x]$. Now $a^2 + 1 = 0$ and thus $a^2 = -1$. This field is isomorphic to $\mathbb{C}$.

(2) Consider the polynomial $x^2 - 3 \in \mathbb{Q}[x]$. This is an irreducible polynomial in $\mathbb{Q}[x]$ and Theorem 4.15 gives a root a in the field

$$\mathbb{Q}[x]/\mathbb{Q}[x](x^2 - 3) = \mathbb{Q} + \mathbb{Q}a$$

where $a = [x]$. This field is isomorphic to the subfield $\mathbb{Q} + \mathbb{Q}\sqrt{3}$ of $\mathbb{R}$.

(3) Consider $p = x^2 + x + 1$ in $\mathbb{Z}_2[x]$. If the polynomial were not irreducible there would be a linear factor in $\mathbb{Z}_2[x]$. But as $p(0) = p(1) = 1$ this is not the case, so p is irreducible and has a root $a = [x]$ in the field $\mathbb{Z}_2[x]/\mathbb{Z}_2[x]p = \mathbb{Z}_2 + \mathbb{Z}_2 a$.

4.4. **Normed $\mathbb{R}$-algebras.** Recall from Algebra 2A that an *inner product* on a real vector space V is a positive definite symmetric bilinear form

$$\langle \cdot\,,\ \cdot \rangle \colon V \times V \to \mathbb{R}.$$

The corresponding *norm* is $\|\cdot\| \colon V \to \mathbb{R}$ given by $\|v\| = \sqrt{\langle v, v\rangle}$. Positive definiteness gives that $\|v\| = 0 \implies v = 0$.

Definition 4.18 (**Normed $\mathbb{R}$-algebra**)**.** Let V be an $\mathbb{R}$-algebra with 1 such that $V \neq \{0\}$. We say that V is a *normed* $\mathbb{R}$-algebra if it is equipped with an inner product such that the corresponding norm satisfies $\|u \cdot v\| = \|u\| \cdot \|v\|$ for all $u, v \in V$.

Remark 4.19. The $V \neq \{0\}$ assumption gives $1_V \neq 0$ and hence $\|1_V\| \neq 0$. We have $\|1_V\| = \|1_V \cdot 1_V\| = \|1_V\| \cdot \|1_V\|$. Since the norm takes values in the integral domain $\mathbb{R}$, the resulting equality $\|1_V\| \cdot (1 - \|1_V\|) = 0$ implies that $\|1_V\| = 1$.

Examples 4.20 ($\mathbb{R}$**,** $\mathbb{C}$ **and** $\mathbb{H}$ **are normed** $\mathbb{R}$**-algebras**)**.** Examples 4.3–4.4 shows that $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ are $\mathbb{R}$-algebras of dimension one, two and four respectively, and in each case a basis over $\mathbb{R}$ is given. With respect to these bases, the standard dot product on $\mathbb{R}^n$ gives a norm on each algebra.

(1) The norm of $a \in \mathbb{R}$ is absolute value $|a| = \sqrt{a^2}$, and since $|a \cdot b| = |a| \cdot |b|$ for all $a, b \in \mathbb{R}$ we have that $\mathbb{R}$ is a normed $\mathbb{R}$-algebra.

(2) The norm of $z = z_1 + z_2 i \in \mathbb{C}$ is $\|z\| = \sqrt{z\bar{z}} = \sqrt{z_1^2 + z_2^2}$ so $\|z \cdot w\|^2 = z\bar{z}w\bar{w} = \|z\|^2 \cdot \|w\|^2$, i.e., $\mathbb{C}$ is a normed $\mathbb{R}$-algebra and if $w = w_1 + w_2 i$ we have

$$(z_1^2 + z_2^2)(w_1^2 + w_2^2) = (z_1 w_1 - z_2 w_2)^2 + (z_1 w_2 + z_2 w_1)^2. \tag{4.4}$$

(3) The norm of $z = z_1 + z_2 i + z_3 j + z_4 k \in \mathbb{H}$ us $\sqrt{z_1^2 + z_2^2 + z_3^2 + z_4^2}$ and if $w = w_1 + w_2 i + w_3 j + w_4 k$ we have

$$\begin{aligned} zw \ = \ & (z_1 w_1 - z_2 w_2 - z_3 w_3 - z_4 w_4) + (z_1 w_2 + z_2 w_1 + z_3 w_4 - z_4 w_3) i \\ & + (z_1 w_3 - z_2 w_4 + z_3 w_1 + z_4 w_2) j + (z_1 w_4 + z_2 w_3 - z_3 w_2 + z_4 w_1) k \end{aligned}$$

and can show (see Exercise Sheet 8) that $\|z\|^2 \cdot \|w\|^2 = \|z \cdot w\|^2$. Hence $\mathbb{H}$ is a normed $\mathbb{R}$-algebra and

$$\begin{aligned} &(z_1^2 + z_2^2 + z_3^2 + z_4^2)(w_1^2 + w_2^2 + w_3^2 + w_4^2) \\ &\quad = (z_1 w_1 - z_2 w_2 - z_3 w_3 - z_4 w_4)^2 + (z_1 w_2 + z_2 w_1 + z_3 w_4 - z_4 w_3)^2 \\ &\qquad + (z_1 w_3 - z_2 w_4 + z_3 w_1 + z_4 w_2)^2 + (z_1 w_4 + z_2 w_3 - z_3 w_2 + z_4 w_1)^2. \end{aligned} \tag{4.5}$$

Theorem 4.21 (**Classification of normed $\mathbb{R}$-algebras**)**.** *There are exactly three normed $\mathbb{R}$-algebras up to isomorphism, namely, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$.*

Idea of proof. Let V be a normed $\mathbb{R}$-algebra.

- If $1, t$ are orthonormal in V, then $t^2 = -1$.
- If $1, i, j$ are orthonormal in V, then so are $1, i, j, ij$. Moreover $ji = -ij$.

- If $1, i, j, ij, e$ are orthonormal in V, then $(ij)e = -e(ij) = iej = -ije$ so $(ij)e = 0$ which is absurd.

Thus $\dim V \in \{1, 2, 4\}$ and we get $V \cong \mathbb{R}$, $\mathbb{C}$ or $\mathbb{H}$ accordingly. □

There is a beautiful application in Number Theory of normed algebras, obtained by using complex numbers and quaternions with integer coefficients.

Theorem 4.22 (Fermat's two square theorem and Lagrange's four square theorem)**.** *Let $n \in \mathbb{N}$. Then n is a sum of four integer squares, and n is a sum of two integer squares provided it has no prime factors congruent to 3 modulo 4.*

The idea is to use (4.4)–(4.5) to show that if we have two sums of two squares or of four squares then their product is also a sum of two squares or four squares respectively. Hence we are reduced to the case that n is prime, but the proof in this case is beyond the scope of the course.

Example 4.23. To give the idea, consider a simple example: $21 = 3 \cdot 7$. We have

$$3 = 1^2 + 1^2 + 1^2 + 0^2 \quad \text{and} \quad 7 = 2^2 + 1^2 + 1^2 + 1^2,$$

so

$$21 = 3 \cdot 7 = (1^2 + 1^2 + 1^2 + 0^2) \cdot (2^2 + 1^2 + 1^2 + 1^2) = 0^2 + 4^2 + 2^2 + 1^2.$$

Similarly $2 = 1^2 + 1^2$ and $5 = 2^2 + 1^2$ so $10 = 2 \cdot 5 = (1^2 + 1^2)(2^2 + 1^2) = (2-1)^2 + (1+2)^2$.

End of Week 7.

---

5. The structure of linear operators

Let V be an n-dimensional vector space over $\Bbbk$. Let $\alpha\colon V \to V$ be a linear operator and let A be the matrix representing α with respect to a given basis $(v_1, v_2, \dots, v_n)$ of V.

5.1. **Minimal polynomials.** Given a polynomial $f = \sum_{i=0}^{n} a_i t^i \in \Bbbk[t]$, we write

$$f(A) = a_0\mathbb{I}_n + a_1 A + a_2 A^2 + \cdots + a_n A^n$$

for the $n \times n$ matrix obtained by substituting A for t (and formally replacing $t^0 = 1$ by the $n \times n$ matrix identity $\mathbb{I}_n$). It is not hard to show that the map $\Bbbk[t] \to M_n(\Bbbk)$ defined by sending $f \mapsto f(A)$ is a ring homomorphism. Recall from Exercise 3.4 that the rings $\mathrm{End}\,(V)$ and $M_n(\Bbbk)$ are isomorphic as rings as well as vector spaces over $\Bbbk$ of dimension n^2, and by precomposing with this isomorphism we obtain a ring homomorphism

$$\Phi_\alpha\colon \Bbbk[t] \to \mathrm{End}(V),\ f \mapsto f(\alpha), \tag{5.1}$$

where the multiplication in $\mathrm{End}(V)$ is the composition of maps.

Lemma 5.1. *The kernel of the ring homomorphism Φ_α is not the zero ideal.*

Proof. The dimension of $\mathrm{End}(V)$ as a $\Bbbk$-vector space is n^2, so the list $\mathrm{id}, \alpha, \alpha^2, \dots, \alpha^{n^2}$ comprising n^2+1 linear operators, or equivalently, the list $(\mathbb{I}_n, A, A^2, \dots, A^{n^2})$ of matrices, is linearly dependent. If $a_0, \dots, a_{n^2} \in \Bbbk$ (not all zero) satisfy $a_0\mathbb{I}_n + \cdots + a_{n^2}A^{n^2} = 0$, then the polynomial $f = \sum_{i=0}^{n^2} a_i t^i$ satisfies $\Phi_\alpha(f) = 0$, so $f \in \mathrm{Ker}(\Phi_\alpha)$ is nonzero. □

Since $\Bbbk[t]$ is a PID, there exists a monic polynomial $m_\alpha \in \Bbbk[t]$ of degree at least one such that $\mathrm{Ker}(\Phi_\alpha) = \Bbbk[t]m_\alpha$. Recall from the proof of Theorem 3.11 that $m_\alpha \in \Bbbk[t]$ is the unique monic polynomial of smallest degree such that $m_\alpha(\alpha) = m_\alpha(A) = 0$.

Definition 5.2 (Minimal polynomial). The *minimal polynomial* of $\alpha\colon V \to V$ is the monic polynomial $m_\alpha \in \Bbbk[t]$ of lowest degree such that $m_\alpha(\alpha) = 0$. We also write m_A and refer to the minimal polynomial of an $n \times n$ matrix A representing α.

Examples 5.3.
(1) If $\alpha = \lambda\mathrm{id}$ then $p(\alpha) = 0$ where $p(t) = t - \lambda$, so $m_\alpha(t) = t - \lambda$.
(2) If $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, then $A^2 = \mathbb{I}_2$ and $p(A) = 0$ where $p(t) = t^2 - 1$. As A is not a diagonal matrix, we have that $q(A) \neq 0$ for any $q = t - \lambda$. Hence $m_A(t) = t^2 - 1$.

Lemma 5.4. *Let p be a polynomial such that $p(\alpha) = 0$. Then every eigenvalue of α is a root of p. In particular every eigenvalue of α is a root of m_α.*

Proof. Let $v \neq 0$ be an eigenvector for eigenvalue λ and suppose $p(t) = \sum_{i=0}^{k} a_i t^i$. Then $p(\alpha) = 0$ gives

$$0 = p(\alpha)\, v = (a_0\mathrm{id} + a_1\alpha + \cdots + a_k\alpha^k)v = (a_0 + a_1\lambda + \cdots + a_k\lambda^k)v = p(\lambda)v.$$

As $v \neq 0$ it follows that $p(\lambda) = 0$. □

Definition 5.5 (Characteristic polynomial and multiplicities of eigenvalues). The *characteristic polynomial* of $\alpha\colon V \to V$ is $\Delta_\alpha(t) = \det(\alpha - t\mathrm{id}) = \det(A - t\mathbb{I}_n)$, where A is a matrix representing α with respect to some basis. The *algebraic multiplicity*, $\mathrm{am}(\lambda)$, of an eigenvalue λ is the multiplicity of λ as a root of $\Delta_\alpha(t)$. The *geometric multiplicity* $\mathrm{gm}(\lambda)$ is the dimension of the eigenspace $E_\alpha(\lambda) = \mathrm{Ker}(\alpha - \lambda\,\mathrm{id}) = \mathrm{Ker}(A - \lambda\mathbb{I}_n)$.

Remarks 5.6. (1) This characteristic polynomial of a linear operator α does not depend on the choice of matrix A representing α, so it's well-defined.
(2) We have $\mathrm{am}(\lambda) \geq \mathrm{gm}(\lambda)$.

Theorem 5.7 (Cayley-Hamilton). *For any $A \in M_n(\Bbbk)$ we have $\Delta_A(A) = 0 \in M_n(\Bbbk)$. Equivalently, for any linear $\alpha\colon V \to V$ we have $\Delta_\alpha(\alpha) = 0 \in \mathrm{End}(V)$.*

Remark 5.8. One can't argue that $\det(A - A\mathbb{I}_n) = \det(0) = 0$ and thus $\Delta_A(A) = 0$ because $\Delta_\alpha(A)$ is a matrix whereas $\det(0)$ is a scalar. To illustrate this for $n = 2$:

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \text{has} \quad \Delta_A(t) = \det\begin{pmatrix} a-t & b \\ c & d-t \end{pmatrix} = t^2 - (a+d)t + (ad - bc),$$

so the Cayley–Hamilton Theorem is the generalisation to arbitrary n of the calculation

$$\begin{aligned} \Delta_A(A) &= A^2 - (a+d)A + (ad-bc)\cdot\mathbb{I}_2 \\ &= \begin{pmatrix} a^2+bc & ab+bd \\ ca+cd & bc+d^2 \end{pmatrix} - \begin{pmatrix} a^2+ad & ab+bd \\ ac+cd & ad+d^2 \end{pmatrix} + (ad-bc)\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}. \end{aligned}$$

If you don't think this is remarkable, check the case $n = 3$ for yourself!

Corollary 5.9. *The minimal polynomial m_α divides the characteristic polynomial Δ_α. In fact the roots of m_α are precisely the eigenvalues of α.*

Proof. The Cayley–Hamilton theorem gives that the characteristic polynomial Δ_α lies in the kernel of the ring homomorphism Φ_α from (5.1). Since $\mathrm{Ker}(\Phi_\alpha) = \Bbbk[t]m_\alpha$, we have that m_α divides Δ_α. Therefore every root of m_α is a root of Δ_α, and hence an eigenvalue of α. Conversely, every eigenvalue of α is a root of m_α by Lemma 5.4. □

Remark 5.10. When working over $\mathbb{C}$, Corollary 5.9 says that if $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of λ and $\Delta_\alpha(t) = (\lambda_1 - t)^{r_1} \cdots (\lambda_k - t)^{r_k}$, then

$$m_\alpha(t) = (t - \lambda_1)^{s_1} \cdots (t - \lambda_k)^{s_k}$$

with $1 \leq s_i \leq r_i$ for all $1 \leq i \leq k$.

Proof of Theorem 5.7. Suppose $\Delta_A(t) = \det(A - t\mathbb{I}_n) = a_0 + a_1 t + \cdots + a_n t^n$. We must show that $\Delta_A(A) = a_0\mathbb{I}_n + a_1 A + \cdots + a_n A^n$ is equal to the zero matrix. Recall the adjugate formula from [Algebra 1B]:

$$\mathrm{adj}(A - t\mathbb{I}_n)(A - t\mathbb{I}_n) = \det(A - t\mathbb{I}_n)\mathbb{I}_n = \Delta_A(t)\mathbb{I}_n. \tag{5.2}$$

Write $\mathrm{adj}(A - t\mathbb{I}_n) = B_0 + B_1 t + \cdots + B_{n-1}t^{n-1}$ for $B_i \in M_n(\Bbbk)$. Substite into (5.2) gives

$$(B_0 + B_1 t + \cdots + B_{n-1}t^{n-1})(A - t\mathbb{I}_n) = (a_0 + a_1 t + \cdots + a_n t^n)\mathbb{I}_n. \tag{5.3}$$

Comparing terms involving t^i for any $1 \leq i \leq n$, we have that

$$\big(B_iA - B_{i-1}\big)t^i = (B_it^i)A + (B_{i-1}t^{i-1})(-t\mathbb{I}_n) = a_i\mathbb{I}_nt^i \tag{5.4}$$

Notice that in gathering terms here, we used the fact that the monomial t^i commutes with A (after all, these equations involve elements in the ring $R[t]$ where $R = M_n(\Bbbk)$, so we have $At^i = t^iA$). If we now subsitute any matrix $T \in M_n(\Bbbk)$ into equation (5.3), the left hand side will become a polynomial in T in which the coefficient of T^i is given by equation (5.4) if and only if $AT^i = T^iA$. For any such matrix T satisfies

$$(B_0 + B_1T + \cdots + B_{n-1}T^{n-1})(A - T) = a_0\mathbb{I}_n + a_1T + \cdots + a_nT^n.$$

Since A satisfies $A \cdot A^i = A^i \cdot A$, we may substitute $T = A$ to obtain

$$\Delta_A(A) = a_0\mathbb{I}_n + a_1A + \cdots a_nA^n = (B_0 + B_1A + \cdots + B_{n-1}A^{n-1})(A - A) = 0$$

as required. □

5.2. **Invariant subspaces.** Let $\alpha\colon V \to V$ be a linear operator over a field $\Bbbk$.

Definition 5.11 (Invariant subspace). For a linear operator $\alpha : V \to V$, we say that a subspace W of V is *α-invariant* if $\alpha(W) \subseteq W$. If W is α-invariant, then the *restriction* of α to W, denoted $\alpha|_W \in \mathrm{End}(W)$, is the linear operator $\alpha|_W : W \to W : w \mapsto \alpha(w)$.

Examples 5.12. (1) The subspaces $\{0\}$ and V are always α-invariant.

(2) Let λ be an eigenvalue of α. If v is an eigenvector for λ, then the one dimensional subspace $\Bbbk v$ is α-invariant because $\alpha(av) = a\alpha(v) = a\lambda v \in \Bbbk v$.

(3) For any $\theta \in \mathbb{R}$ with $\theta \neq 2\pi k$ for $k \in \mathbb{Z}$, the linear operator $\alpha\colon \mathbb{R}^3 \to \mathbb{R}^3$ that rotates every vector by θ radians anticlockwise around the z-axis has $V_1 := \mathbb{R}e_1 \oplus \mathbb{R}e_2$ and $V_2 := \mathbb{R}e_3$ as α-invariant subspaces. The restriction $\alpha|_{V_1}\colon V_1 \to V_1$ is simply rotation by θ radians in the plane, while $\alpha|_{V_2}\colon V_2 \to V_2$ is the identity on the real line. Notice that the matrix for α in the basis e_1, e_2, e_3 is the 'block' matrix

$$A = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Notice that this matrix has two square non-zero 'blocks' (the top left 2×2 matrix and the bottom right 1×1 matrix). These two blocks are precisely the matrices for the linear maps $\alpha|_{V_1}$ and $\alpha|_{V_2}$ in the given bases on V_1 and V_2 respectively.

Definition 5.13 (Direct sum of linear maps and matrices). For $1 \leq i \leq k$, let V_i be a vector space and let $\alpha_i \in \mathrm{End}(V_i)$. The *direct sum* of $\alpha_1, \ldots, \alpha_k$ is the linear map

$$(\alpha_1 \oplus \cdots \oplus \alpha_k)\colon \bigoplus_{1\leq i\leq k} V_i \to \bigoplus_{1\leq i\leq k} V_i$$

defined as follows: each $v \in \bigoplus_{1\leq i\leq k} V_i$ can be written uniquely in the form $v = v_1 + \cdots + v_k$ for some $v_i \in V_i$, and we define

$$(\alpha_1 \oplus \cdots \oplus \alpha_k)(v_1 + \cdots + v_k) := \alpha_1(v_1) + \cdots + \alpha_k(v_k).$$

Remark 5.14. For $1 \leq i \leq k$, let $A_i \in M_{n_i}(\Bbbk)$ be the matrix for a linear map α_i with respect to some basis $\mathcal{B}_i$ of V_i. Then the matrix for the direct sum $\alpha_1 \oplus \cdots \oplus \alpha_k$ with respect to the concatenated basis $\mathcal{B}_1, \mathcal{B}_2, \ldots \cup \mathcal{B}_k$ of $\bigoplus_{1 \leq i \leq k} V_i$ is the *direct sum* (block matrix)

$$A_1 \oplus \cdots \oplus A_k := \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}$$

(zeros everywhere else in the matrix) of the matrices $A_1, \ldots, A_k$.

Lemma 5.15. *For $\alpha \in \operatorname{End}(V)$, suppose $V = V_1 \oplus V_2 \oplus \cdots \oplus V_k$ where $V_1, \ldots, V_k$ are α-invariant subspaces. For $1 \leq i \leq k$, write $\alpha_i := \alpha|_{V_i} \in \operatorname{End}(V_i)$. Then*

(1) *$\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k \in \bigoplus_{i=1}^{k} \operatorname{End}(V_i)$; and*
(2) *the minimal polynomial m_α is the least common multiple of $m_{\alpha_1}, \ldots, m_{\alpha_k}$.*

Proof. For (1), each $v \in V$ can be written uniquely as $v = v_1 + \cdots + v_k$ for $v_i \in V_i$, and

$$\alpha(v) = \alpha(v_1) + \cdots + \alpha(v_k) = \alpha_1(v_1) + \cdots + \alpha_k(v_k)$$

where $\alpha_i(v_i) \in V_i$ which proves (1). For (2), we claim first that any $f \in \Bbbk[t]$ satisfies

$$f(\alpha) = f(\alpha_1) \oplus f(\alpha_2) \oplus \cdots \oplus f(\alpha_k). \tag{5.5}$$

Indeed, for $i > 0$ and $v = v_1 + \cdots + v_k$ we have $\alpha^i(v_1 + \cdots + v_k) = \alpha_1^i(v_1) + \cdots + \alpha_k^i(v_k)$, so $\alpha^i = \alpha_1^i \oplus \cdots \oplus \alpha_k^i$. For any scalar $c \in \Bbbk$, it follows that $c\alpha^i = c\alpha_1^i \oplus \cdots \oplus c\alpha_k^i$. We add in $\operatorname{End}(V)$ using the formula from Exercise 3.4, so any polynomial $f = \sum_i c_i t^i$ satisfies (5.5) as claimed. Then m_α divides f if and only if $f(\alpha) = 0$ which holds if and only if $f(\alpha_i) = 0$ for all $1 \leq i \leq k$, which holds if and only if $m_{\alpha_i} | f$ for all $1 \leq i \leq k$. Equivalently m_α is the least common multiple of $m_{\alpha_1}, \ldots, m_{\alpha_k}$ as required. □

End of Week 8.

5.3. **Jordan blocks.** To begin our study of the structure of α, we first consider the special case where $\alpha \colon V \to V$ has only one eigenvalue λ, so Remark 5.10 tells us that

$$\Delta_\alpha(t) = (\lambda - t)^r \quad \text{and} \quad m_\alpha(t) = (t - \lambda)^s \tag{5.6}$$

where $1 \leq s \leq r$. From now on we assume that our field $\Bbbk$ contains λ to ensure that $\Delta_\alpha(t)$ splits into a product of linear factors as in (5.6). We can choose $\Bbbk$ to be $\mathbb{Q}[\lambda]$ (see Theorem 4.13), but for simplicity we choose to work with the field $\Bbbk = \mathbb{C}$. We also set $\alpha_\lambda = \alpha - \lambda \operatorname{id}$.

Definition 5.16 (Cyclic subspace generated by v). For any vector space V and any $v \in V$, the *cyclic subspace generated by v* is the subspace

$$\mathbb{C}[\alpha]v = \{p(\alpha)v \in V \mid p \in \mathbb{C}[t]\}.$$

Remark 5.17. Note that $\mathbb{C}[\alpha]v$ is an α-invariant subspace of V. Indeed, for $p_1, p_2 \in \mathbb{C}[t]$ and $\lambda_1, \lambda_2 \in \Bbbk$, we have $\lambda_1(p_1(\alpha)v)+\lambda_2 p_2(\alpha)v = (\lambda_1 p_1+\lambda_2 p_2)(\alpha)v$, so $\mathbb{C}[\alpha]v$ is a subspace of V. It is also α-invariant since $\alpha p(\alpha)v = u(\alpha)v$ where u is the polynomial $tp(t)$.

Example 5.18. If $v \in E_\alpha(\lambda)$, that is, if $\alpha(v) = \lambda v$, then $\mathbb{C}[\alpha]v = \mathbb{C}v$. Thus, for every eigenvector v of α we have that $\mathbb{C}v$ is the cyclic α-invariant subspace generated by v.

Proposition 5.19. *Let $\alpha \in \mathrm{End}(V)$ be such that $\Delta_\alpha(t) = (\lambda - t)^r$ and $m_\alpha(t) = (t-\lambda)^s$. For any nonzero vector $v \in V$, define $e := e(v) \in \mathbb{Z}_{>0}$ to be the smallest positive integer such that ${\alpha_\lambda}^e(v) = 0$ where $\alpha_\lambda := \alpha - \lambda\,\mathrm{id}$, and write*

$$v_1 = {\alpha_\lambda}^{e-1}(v),\ v_2 = {\alpha_\lambda}^{e-2}(v),\ \ldots,\ v_{e-1} = \alpha_\lambda(v),\ v_e = v.$$

Then

(1) *$(v_1, v_2, \ldots, v_e)$ is a basis for the $\mathbb{C}$-vector space $W := \mathbb{C}[\alpha]v$;*

(2) *in this basis, the matrix for the linear map $\beta := \alpha|_W \in \mathrm{End}(W)$ is the $e \times e$ matrix*

$$J(\lambda, e) = \begin{pmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \ddots & \ddots & \\ & & & \lambda & 1 \\ & & & & \lambda \end{pmatrix}; \text{and}$$

(3) *we have that $E_\beta(\lambda) = \mathbb{C}v_1$, that $m_\beta(t) = (t-\lambda)^e$ and that $\Delta_\beta(t) = (\lambda - t)^e$.*

Proof. Note first that since $m_\alpha(t) = (t-\lambda)^s$, we have that ${\alpha_\lambda}^s(v) = m_\alpha(\alpha)v = 0$ and therefore $1 \le e \le s$ is well-defined.

To see that $v_1, \ldots, v_e$ span W, let $w \in W$. By hypothesis $w = f(\alpha)v$ for some $f \in \mathbb{C}[t]$. Exercise Sheet 9 gives $a_0, \ldots, a_k \in \mathbb{C}$ such that $f(t) = a_0 + a_1(t-\lambda) + \cdots + a_k(t-\lambda)^k$ for some $k \ge 0$, and hence

$$w = f(\alpha)v = a_0 v + a_1 \alpha_\lambda(v) + a_2 {\alpha_\lambda}^2(v) + \cdots,$$

so W is spanned by $v_1, \ldots, v_e$ because ${\alpha_\lambda}^e(v) = 0$. The fact that $v_1, \ldots, v_e$ are linearly independent is immediate from Exercise 9.3, so statement (1) holds.

For (2), we compute that

$$\alpha(v_1) = \lambda\, v_1 + \alpha_\lambda(v_1) = \lambda\, v_1 + {\alpha_\lambda}^e(v) = \lambda\, v_1$$

and for $2 \le i \le e$ we have

$$\alpha(v_i) = \lambda\, v_i + \alpha_\lambda(v_i) = \lambda\, v_i + v_{i-1} = v_{i-1} + \lambda\, v_i.$$

Therefore we have expressed the image under α of each basis vector v_i in terms of the basis $(v_1, v_2, \ldots, v_e)$, and the coefficients in this expansion provide the entries in each column of the matrix for β; the resulting matrix is therefore $J(\lambda, e)$.

The statements from part (3) follow from Exercise 9.4. □

Definition 5.20 (Jordan block). We call $J(\lambda, e)$ a *Jordan block* of α.

Example 5.21. Consider the linear operator $\alpha : \mathbb{C}^2 \to \mathbb{C}^2,\ v \mapsto Av$ where

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

satisfies $\Delta_\alpha(t) = (1-t)^2$, and $m_\alpha(t) = (t-1)^2$. Following Proposition 5.19 we first compute an eigenvector v_1 for $\lambda = 1$, i.e., solve

$$(A - \mathbb{I})v_1 = 0, \text{ that is } \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & -1/2 \end{pmatrix} v_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

giving, say, $v_1 = (1, -1)^t$. Next solve

$$(A - \mathbb{I})v_2 = v_1, \text{ that is } \begin{pmatrix} 1/2 & 1/2 \\ -1/2 & -1/2 \end{pmatrix} v_2 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

giving, say, $v_2 = (0, 2)^t$. The matrix required to change basis so that A can be written in the form of Proposition 5.19 is the matrix whose columns are v_1, v_2, namely

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

Please check for yourself that

$$P^{-1}AP = \begin{pmatrix} 1 & 0 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 3/2 & 1/2 \\ -1/2 & 1/2 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} = J(1, 2).$$

Theorem 5.22 (**Jordan normal form - special case**). *Let $\alpha \in \mathrm{End}(V)$ be such that $\Delta_\alpha(t) = (\lambda - t)^r$ and $m_\alpha(t) = (t - \lambda)^s$. Then there exists a basis for V such that the matrix for α with respect to this basis is*

$$A := JNF(\alpha) := \begin{pmatrix} J(\lambda, e_1) & & & \\ & J(\lambda, e_2) & & \\ & & \ddots & \\ & & & J(\lambda, e_m) \end{pmatrix} = J(\lambda, e_1) \oplus \cdots \oplus J(\lambda, e_m),$$

where

(1) *$m = gm(\lambda)$ is the number of Jordan blocks;*
(2) *$s = \max\{e_1, \ldots, e_m\}$; and*
(3) *$r = e_1 + \cdots + e_m$.*

The proof uses the following Lemma.

Lemma. *Let $\alpha \in \mathrm{End}(V)$ with $m_\alpha(t) = (t-\lambda)^s$. Then there exist nonzero $v_1, \ldots, v_m \in V$ such that*

$$V = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_m \tag{5.7}$$

Proof of Theorem 5.22. Let $W_j := \mathbb{C}[\alpha]v_j$ as in the lemma, and $e_j := \dim W_j$. Each W_j is α-invariant, so if we let α_j be the restriction of α to W_j, then $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_m$ by Lemma 5.15, and we choose the basis on each subspace W_j by applying Proposition 5.19 which gives the required form for the matrix A and gives $m_{\alpha_j} = (t - \lambda)^{e_j}$ for $1 \le j \le m$. Moreover:

(1) Each $v \in V$ is $v = w_1 + \cdots + w_m \in V$ for $w_j \in W_j$ by (5.7), so if $v \in E_\alpha(\lambda)$, then

$$\alpha_1(w_1) + \cdots + \alpha_m(w_m) = \alpha(v) = \lambda(v) = \lambda w_1 + \cdots + \lambda w_m$$

and thus $\alpha_j(w_j) = \lambda w_j$ for $1 \le j \le m$. Then $E_\alpha(\lambda) = E_{\alpha_1}(\lambda) \oplus \cdots \oplus E_{\alpha_m}(\lambda)$. By Proposition 5.19, we have $\dim E_{\alpha_i}(\lambda) = 1$, so

$$m = \dim E_{\alpha_1}(\lambda) + \cdots + \dim E_{\alpha_m}(\lambda) = \dim E_\alpha(\lambda) = \mathrm{gm}\,(\lambda).$$

This proves (1).

(2) Lemma 5.15 shows that $m_\alpha(t)$ is the least common multiple of $m_{\alpha_1}(t), \ldots, m_{\alpha_m}(t)$. Proposition 5.19 shows that $m_{\alpha_i}(t) = (t-\lambda)^{e_i}$, so (2) follows immediately.

(3) This says simply that $\dim V = \dim W_1 + \cdots + \dim W_m$. □

Proof of the lemma (not examinable). We use induction on s. If $s = 1$, then $\alpha = \lambda\mathrm{id}$. Pick any basis $v_1, \ldots, v_r$ for V and apply Proposition 5.19 with $e = 1$ for each basis vector to see that

$$V = \mathbb{C}v_1 \oplus \cdots \oplus \mathbb{C}v_r = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_r.$$

This proves the case $s = 1$. Now suppose that $s \ge 2$ and that the claim holds for smaller values of s. Consider the α-invariant subspace

$$W = \alpha_\lambda(V) = \{\alpha_\lambda(v) \in V \mid v \in V\}.$$

Notice that ${\alpha_\lambda}^{s-1}(w) = 0$ for all $w \in W$ and the minimal polynomial of $\alpha|_W$ is $(t-\lambda)^{s-1}$. The inductive hypothesis gives $\alpha_\lambda(v_1), \ldots, \alpha_\lambda(v_\ell) \in W \setminus \{0\}$ with

$$W = \mathbb{C}[\alpha]\alpha_\lambda(v_1) \oplus \cdots \oplus \mathbb{C}[\alpha]\alpha_\lambda(v_\ell). \tag{5.8}$$

Let β_i be the restriction of α to $\mathbb{C}[\alpha]v_i$. Proposition 5.19 shows that $E_{\beta_i}(\lambda) = \mathbb{C}w_i$ where $w_i = {\alpha_\lambda}^{e_i-1}(v_i) \in \mathbb{C}[\alpha]\alpha_\lambda(v_i)$ for some $e_i \ge 2$. Since the sum from (5.8) is direct, it follows as in the proof of (1) above that $(w_1, \ldots, w_\ell)$ is a basis for $E_{\alpha|_W}(\lambda)$. Extend this to a basis $(w_1, \ldots, w_\ell, v_{\ell+1}, \ldots, v_{\ell+m})$ for $E_\alpha(\lambda) \subseteq \mathbb{C}[\alpha]v_1 + \cdots + \mathbb{C}[\alpha]v_\ell + \mathbb{C}v_{\ell+1} + \cdots + \mathbb{C}v_{\ell+m}$. We can now throw away that w_i's completely, because we claim that

$$V = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_\ell \oplus \mathbb{C}[\alpha]v_{\ell+1} \oplus \cdots \oplus \mathbb{C}[\alpha]v_{\ell+m}.$$

Since $v_{\ell+1}, \ldots, v_{\ell+m}$ are eigenvectors for λ, this is the same as saying that

$$V = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_\ell \oplus (\mathbb{C}v_{\ell+1} \oplus \cdots \oplus \mathbb{C}v_{\ell+m}). \tag{5.9}$$

The right hand side is by definition contained in the left. For the opposite inclusion, let $v \in V$. Then $\alpha_\lambda(v) \in W$, so by (5.8) there exist $p_1, \ldots, p_e \in \mathbb{C}[t]$ such that

$$\alpha_\lambda(v) = p_1(\alpha)\alpha_\lambda(v_1) + \cdots + p_e(\alpha)\alpha_\lambda(v_\ell).$$

Gather all terms on one side to obtain $\alpha_\lambda(v - (p_1(\alpha)v_1 + \cdots + p_e(\alpha)v_e)) = 0$, so

$$v - (p_1(\alpha)v_1 + \cdots + p_\ell(\alpha)v_\ell) \in E_\alpha(\lambda) \subseteq \mathbb{C}[\alpha]v_1 + \cdots + \mathbb{C}[\alpha]v_\ell + \mathbb{C}v_{\ell+1} + \cdots + \mathbb{C}v_{\ell+m}.$$

Now we know that the decomposition

$$v = (p_1(\alpha)v_1 + \cdots + p_\ell(\alpha)v_\ell) + \big(v - (p_1(\alpha)v_1 + \cdots + p_\ell(\alpha)v_\ell)\big)$$

presents v as the sum of an element of $\mathbb{C}[\alpha]v_1 + \cdots + \mathbb{C}[\alpha]v_\ell$ and an element of the space $\mathbb{C}[\alpha]v_1 + \cdots + \mathbb{C}[\alpha]v_\ell + \mathbb{C}v_{\ell+1} + \cdots + \mathbb{C}v_{\ell+m}$, so it lies in the right hand side of (5.9) as required. It remains to show that the sum from (5.9) is direct. Suppose

$$0 = p_1(\alpha)v_1 + \cdots + p_\ell(\alpha)v_\ell + a_{\ell+1}v_{\ell+1} + \cdots + a_{\ell+m}v_{\ell+m}.$$

Applying α_λ to both sides gives

$$0 = p_1(\alpha)\alpha_\lambda(v_1) + \cdots + p_\ell(\alpha)\alpha_\lambda(v_\ell).$$

Since W is a direct sum in equation (5.8), we have $\alpha_\lambda(p_i(\alpha)v_i) = 0$ for $1 \leq i \leq \ell$, so $p_i(\alpha)v_i$ is an eigenvector that lies in $\mathbb{C}[\alpha]v_i$, so it must be a multiple of w_i. Since $w_1, \ldots, w_\ell$ are linearly independent, it follows that $p_i(\alpha)v_i = 0$ for $1 \leq i \leq \ell$. Hence

$$0 = a_{\ell+1}v_{\ell+1} + \cdots + a_{\ell+m}v_{\ell+m}$$

and as $v_{\ell+1}, \ldots, v_{\ell+m}$ are linearly independent, it follows that $a_{\ell+1} = \ldots = a_{\ell+m} = 0$. This finishes the proof. □

Example 5.23. For a complex vector space V of dimension 4, suppose that $\alpha \in \mathrm{End}(V)$ has $m_\alpha(t) = (t-5)^2$ and $\Delta_\alpha(t) = (t-5)^4$. Since the degree of $m_\alpha(t)$ is 2, we must have at least one largest block $J(5,2)$, so the possible decompositions of the 4-dimensional space V are $J(5,2) \oplus J(5,2)$ and $J(5,2) \oplus J(5,1) \oplus J(5,1)$. If we know in addition that $\mathrm{gm}(5) = 3$ then we must have three blocks, so the second possibility applies.

5.4. **Primary Decomposition.** What if a linear map $\alpha\colon V \to V$ has more than one eigenvalue? Our goal is to choose a basis in which the matrix for α is a block matrix (see Remark 5.14), and Lemma 5.15 tells us that we can achieve this by writing V as the direct sum of α-invariant subspaces. But does such a decomposition exist?

Our goal now is produce one such decomposition of V. The key is to factor the minimal polynomial m_α in the ring $\Bbbk[t]$ as a product of irreducible factors. We begin with the case where the minimal polynomial has two coprime factors (see Definition 3.14). In this section, $\Bbbk$ is any field.

Proposition 5.24 (Primary decomposition in the case $k = 2$). *Let $\alpha\colon V \to V$ be a linear operator and suppose that the minimal polynomial satisfies $m_\alpha = q_1 q_2$, where q_1, q_2 are monic and coprime. For $1 \leq i \leq 2$, let $V_i = \mathrm{Ker}(q_i(\alpha))$. Then:*

(1) *the subspaces V_1, V_2 are α-invariant and satisfy $V = V_1 \oplus V_2$; and*
(2) *the maps $\alpha_i = \alpha|_{V_i}$ for $1 \leq i \leq 2$ satisfy $\alpha = \alpha_1 \oplus \alpha_2$ and $m_{\alpha_i} = q_i$.*

Proof. Define subspaces $W_1 = \mathrm{Im}(q_2(\alpha))$ and $W_2 = \mathrm{Im}(q_1(\alpha))$. Our first goal is to prove a modified version of parts (1) and (2) with W_i in place of V_i.

We first prove that W_i are α-invariant and $V = W_1 \oplus W_2$. Since $q_i(\alpha)$ commutes with α, Exercise Sheet 10 shows that $\mathrm{Im}(q_i(\alpha))$ is α-invariant, i.e., W_i is α-invariant. Since q_1, q_2 are coprime, Lemma 3.15 gives $f, g \in \Bbbk[t]$ such that $1 = fq_1 + gq_2$, so

$$\mathrm{id} = f(\alpha)q_1(\alpha) + g(\alpha)q_2(\alpha).$$

Any $v \in V$ satisfies

$$v = \mathrm{id}(v) = q_2(\alpha)\big(g(\alpha)(v)\big) + q_1(\alpha)\big(f(\alpha)(v)\big) \in W_1 + W_2,$$

so $V = W_1 + W_2$. To see that the sum is direct, suppose $v \in W_1 \cap W_2$, say $v = q_1(\alpha)(v_1) = q_2(\alpha)(v_2)$. Then using the equation above, we have that

$$\begin{aligned} v &= f(\alpha)q_1(\alpha)(v) + g(\alpha)q_2(\alpha)(v) \\ &= [f(\alpha)q_1(\alpha)q_2(\alpha)](v_2) + [g(\alpha)q_2(\alpha)q_1(\alpha)](v_1) \\ &= [f(\alpha)m_\alpha(\alpha)](v_2) + [g(\alpha)m_\alpha(\alpha)](v_1) \\ &= 0. \end{aligned}$$

Hence $W_1 \cap W_2 = \{0\}$ and $V = W_1 \oplus W_2$, so the modified version of (1) holds.

For the modified version of (2), the fact that $\alpha_i = \alpha|_{W_i}$ satisfy $\alpha = \alpha_1 \oplus \alpha_2$ follows from Lemma 5.15. For the statement about the minimal polynomial, fix $i = 1$ and note that

$$\begin{aligned} m_{\alpha_1} \text{ divides } f &\iff f(\alpha_1)(w) = 0 \text{ for all } w \in W_1 \\ &\iff f(\alpha)(w) = 0 \text{ for all } w \in W_1 && \text{as } \alpha(w) = \alpha_1(w) \text{ for } w \in W_1 \\ &\iff f(\alpha)\big(q_2(\alpha)(v)\big) = 0 \text{ for all } v \in V && \text{as } W_1 = \mathrm{Im}(q_2(\alpha)) \\ &\iff m_\alpha \text{ divides } fq_2 && \text{by definition of } m_\alpha \\ &\iff q_1 \text{ divides } f && \text{as } m_\alpha = q_1q_2 \\ &\iff q_1 \text{ is the minimal polynomial of } \alpha_1 \end{aligned}$$

as required. Similarly q_2 is the minimal polynomial of α_2.

We've now proved the result for W_i in place of V_i, so it remains to show that $W_i = V_i$ for $1 \le i \le 2$. Since each $v \in V$ satisfies $q_1(\alpha)q_2(\alpha)(v) = m_\alpha(\alpha)(v) = 0$, we have that $W_1 = \mathrm{Im}(q_2(\alpha)) \subseteq \mathrm{Ker}(q_1(\alpha)) = V_1$. The rank-nullity theorem from Linear Algebra gives

$$\dim \mathrm{Ker}(q_1(\alpha)) + \dim \mathrm{Im}(q_1(\alpha)) = \dim V = \dim W_1 + \dim W_2.$$

Subtract $\dim \mathrm{Im}(q_1(\alpha)) = \dim W_2$ to leave $\dim V_1 = \dim \mathrm{Ker}(q_1(\alpha)) = \dim W_1$, so in fact the inclusion $W_1 \subseteq V_1$ must be equailty. Showing $W_2 = V_2$ is similar. □

Theorem 5.25 (**Primary Decomposition**). *Let $\alpha\colon V \to V$ be a linear operator and write $m_\alpha = p_1^{n_1} \cdots p_k^{n_k}$, where $p_1, \ldots, p_k$ are the distinct monic irreducible factors of m_α in $\Bbbk[t]$. Let $q_i = p_i^{n_i}$ and let $V_i = \mathrm{Ker}(q_i(\alpha))$. Then:*

(1) *the subspaces $V_1, \ldots, V_k$ are α-invariant and $V = V_1 \oplus \cdots \oplus V_k$; and*
(2) *the maps $\alpha_i = \alpha|_{V_i}$ for $1 \le i \le k$ satisfy $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$ and $m_{\alpha_i} = q_i$.*

Proof. We use induction on k. For $k = 1$, we have $m_\alpha = p_1^{n_1} = q_1$. Then

$$V_1 = \mathrm{Ker}\,\big(q_1(\alpha)\big) = \mathrm{Ker}\,\big(m_\alpha(\alpha)\big) = V$$

because $m_\alpha(\alpha)$ is the zero map by Definition 5.2. This proves the case $k = 1$. For $k \ge 2$, suppose the result holds for any linear operator whose minimal polynomial has less than k

distinct irreducible factors. Suppose now that $m_\alpha = p_1^{n_1} \cdots p_k^{n_k}$. Define $r_1 = p_1^{n_1} \cdots p_{k-1}^{n_{k-1}}$ and $r_2 = p_k^{n_k}$, so $m_\alpha = r_1 r_2$. Note that r_1 and r_2 are coprime, Proposition 5.24 gives

$$V = \operatorname{Ker}\big(r_1(\alpha)\big) \oplus \operatorname{Ker}\big(r_2(\alpha)\big),$$

where $\beta_i := \alpha|_{\operatorname{Ker}(r_i(\alpha))}$ satisfies $\alpha = \beta_1 \oplus \beta_2$ and $m_{\beta_i} = r_i$ for $1 \le i \le 2$. In particular, β_1 is a linear operator on $\operatorname{Ker}(r_1(\alpha))$ whose minimal polynomial $r_1 = p_1^{n_1} \cdots p_{k-1}^{n_{k-1}}$ has $k-1$ irreducble factors, so by induction there exist β_1-invariant subspaces such that

$$\operatorname{Ker}\big(r_1(\alpha)\big) = V_1 \oplus \cdots \oplus V_{k-1}$$

and maps $\alpha_i = \beta_1|_{V_i}$ for $1 \le i \le k-1$ satisfy $\beta_1 = \alpha_1 \oplus \cdots \oplus \alpha_{k-1}$ and $m_{\alpha_i} = p_i^{n_i}$. Since $\beta_1 := \alpha|_{\operatorname{Ker}(r_1(\alpha))}$ and $\alpha_i = \beta_1|_{V_i}$, we have that $\alpha_i = \alpha|_{V_i}$ for $1 \le i \le k-1$. Defining $V_k := \operatorname{Ker}\big(r_2(\alpha)\big)$ gives $V = V_1 \oplus \cdots \oplus V_k$, and if we set $\alpha_k := \beta_2|_{V_k} = \alpha|_{V_k}$, then we have that $\alpha_i = \alpha|_{V_i}$ and $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$ for $1 \le i \le k$. It remains to note that $m_{\alpha_i} = p_i^{n_i}$ for all $1 \le i \le k$. □

End of Week 9.

Example 5.26. Consider rotation by θ radians about the z-axis from Examples 5.12(3), and let's work over the field $\mathbb{C}$. The characteristic polynomial of α is

$$\Delta_\alpha(A) = \det(A - t\mathbb{I}_3) = (e^{i\theta} - t)(e^{-i\theta} - t)(1 - t).$$

Since each root has multiplicity one, Remark 5.10 shows that

$$m_\alpha(t) = (t - e^{i\theta})(t - e^{-i\theta})(t - 1).$$

If we now work over $\mathbb{R}$, as we should since $V = \mathbb{R}^3$ is a vector space over $\mathbb{R}$, we obtain

$$m_\alpha(t) = (t^2 - 2\cos\theta t + 1)(t - 1) \tag{5.10}$$

as the factorisation of m_α into irreducibles $q_1 = (t^2 - 2\cos\theta t + 1)$ and $q_2 = t - 1$ in $\mathbb{R}[t]$ (which is a UFD). Now compute

$$\begin{aligned} q_1(\alpha) &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}^2 - 2\cos\theta \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2 - 2\cos(\theta) \end{pmatrix} \end{aligned}$$

and

$$q_2(\alpha) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos(\theta) - 1 & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) - 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Notice that

$$\operatorname{Ker}(q_1(\alpha)) = \left\{ \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \in \mathbb{R}^3 \mid x, y \in \mathbb{R} \right\} \quad \text{and} \quad \operatorname{Ker}(q_2(\alpha)) = \left\{ \begin{pmatrix} 0 \\ 0 \\ z \end{pmatrix} \in \mathbb{R}^3 \mid z \in \mathbb{R} \right\}$$

are the α-invariant subspaces V_1 and V_2 that we considered in Examples 5.12(3). Thus, even if we had not noticed that $V = V_1 \oplus V_2$ as in Examples 5.12(3), we could nevertheless have computed the factorisation (5.10) of the minimal polynomial m_α and obtained the following direct sum decomposition:

$$V = \operatorname{Ker}(m_{\alpha_1}(\alpha)) \oplus \operatorname{Ker}(m_{\alpha_2}(\alpha))$$

with $\alpha = \alpha|_{\operatorname{Ker}(m_{\alpha_1}(\alpha))} \oplus \alpha|_{\operatorname{Ker}(m_{\alpha_2}(\alpha))}$.

Corollary 5.27 (Diagonalisability). *A linear map $\alpha : V \to V$ is diagonalisable iff*

$$m_\alpha(t) = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_k)$$

for distinct $\lambda_1, \ldots, \lambda_k \in \Bbbk$.

Proof. Suppose first that $\alpha \in \operatorname{End}(V)$ is diagonalisable with distinct eigenvalues $\lambda_1, \ldots, \lambda_k$. Let $V_i = E_\alpha(\lambda_i)$ be the λ_i-eigenspace. Pick a basis $\mathcal{B}_i$ for V_i. Then the matrix for α with respect to the concatenated basis $\mathcal{B} = \mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_k$ of V is

$$A = \begin{pmatrix} A_1 & & & \\ & A_2 & & \\ & & \ddots & \\ & & & A_k \end{pmatrix}$$

where $A_i = \lambda_i \mathbb{I}_{n_i}$ for $n_i = \operatorname{gm}(\lambda_i)$. Now $m_{A_i} = t - \lambda_i$ and Lemma 5.15 gives

$$m_A(t) = m_{A_1}(t) m_{A_2}(t) \cdots m_{A_k}(t) = (t - \lambda_1)(t - \lambda_2) \cdots (t - \lambda_k).$$

For the converse, we apply Theorem 5.25 with $q_i := t - \lambda_i$ for $1 \le i \le k$ to obtain

$$V = \operatorname{Ker}(\alpha - \lambda_1 \mathrm{id}) \oplus \cdots \oplus \operatorname{Ker}(\alpha - \lambda_k \mathrm{id}) = E_\alpha(\lambda_1) \oplus \cdots \oplus E_\alpha(\lambda_k),$$

so V must therefore have a basis comprising eigenvectors of α, i.e., α is diagonalisable. □

5.5. **Jordan Decomposition.** We now tackle the general case, where $\alpha \in \operatorname{End}(V)$ need not have a single eigenvalue. Let's work over a field $\Bbbk$ that contains all of the eigenvalues of α, in which case we can decompose the minimal polynomial as

$$m_\alpha(t) = (t - \lambda_1)^{s_1} \cdot (t - \lambda_2)^{s_2} \cdots (t - \lambda_k)^{s_k}$$

where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of α (recall that the roots of m_α are exactly the eigenvalues of α); for example, we could use $\mathbb{C}$, but using the results of Section 4 one can often get away with a much smaller field.

In any event, the Primary Decomposition Theorem 5.25 implies that

$$V = \operatorname{Ker}(\alpha - \lambda_1 \mathrm{id})^{s_1} \oplus \operatorname{Ker}(\alpha - \lambda_2 \mathrm{id})^{s_2} \oplus \cdots \oplus \operatorname{Ker}(\alpha - \lambda_k \mathrm{id})^{s_k}$$

is a decomposition of V as a direct sum of α-invariant subspaces.

Definition 5.28 (Generalised eigenspace). Let $\alpha \colon V \to V$ be a linear map with eigenvalue λ. A nonzero vector $v \in V$ is a *generalised eigenvector* with respect to λ if $(\alpha - \lambda\,\mathrm{id})^s v = 0$ for some positive integer s. The *generalised λ-eigenspace* of α is

$$G_\alpha(\lambda) = \big\{ v \in V : (\alpha - \lambda \mathrm{id})^s v = 0 \text{ for some positive integer } s \big\}.$$

Remark 5.29. We have $E_\alpha(\lambda) \subseteq G_\alpha(\lambda)$.

Lemma 5.30. *Let s be the multiplicity of the eigenvalue λ as a root of m_α. Then*

$$G_\alpha(\lambda) = \mathrm{Ker}(\alpha - \lambda \mathrm{id})^t \quad \text{for all } t \geq s.$$

Proof. The right hand side is contained in the left by Definition 5.28. For the opposite inclusion, suppose $m_\alpha(t) = (t-\lambda_1)^{s_1}(t-\lambda_2)^{s_2}\cdots(t-\lambda_k)^{s_k}$. By the Primary Decomposition Theorem we have that

$$V = V_1 \oplus V_2 \oplus \cdots \oplus V_k,$$

where $V_i = \ker(\alpha - \lambda_i \mathrm{id})^{s_i}$, and the minimal polynomial of $\alpha_i = \alpha|_{V_i}$ is $(t-\lambda_i)^{s_i}$. Now suppose that $\lambda = \lambda_i$. The map α_j only has the eigenvalue λ_j, so for $j \neq i$ we have $\ker(\alpha_j - \lambda_i \mathrm{id}) = \{0\}$ and $\alpha_j - \lambda_i \mathrm{id}$ is a bijective linear operator on V_j. Now let

$$v = v_1 + v_2 + \cdots + v_k$$

be any element in $G_\alpha(\lambda)$ with $v_i \in V_i$. Suppose that $(\alpha - \lambda_i \mathrm{id})^t v = 0$. Then

$$0 = (\alpha - \lambda_i \mathrm{id})^t v = (\alpha_1 - \lambda_i \mathrm{id})^t v_1 + \cdots + (\alpha_k - \lambda_i \mathrm{id})^t v_k.$$

This happens if and only if $(\alpha_j - \lambda_i \mathrm{id})^t v_j = 0$ for all $j = 1, \ldots, k$. As $(\alpha_j - \lambda_i \mathrm{id})^t$ is bijective if $j \neq i$, we must have that $v_j = 0$ for $j \neq i$. Hence $v = v_i \in V_i = \ker(\alpha - \lambda_i)^{s_i}$. This shows that $G_\alpha(\lambda_i) \subseteq \ker(\alpha - \lambda_i \mathrm{id})^{s_i}$ and as $(\alpha - \lambda_i \mathrm{id})^{s_i} v = 0$ clearly implies that $(\alpha - \lambda_i \mathrm{id})^t v = 0$ for any $t \geq s_i$, it follows that $G_\alpha(\lambda_i) \subseteq \ker(\alpha - \lambda_i \mathrm{id})^t$ as required. □

Remark 5.31. This last lemma implies in particular that $G_\alpha(\lambda) = \ker(\alpha - \lambda \mathrm{id})^r$ where r is the algebraic multiplicity of λ. This is useful for calculating $G_\alpha(\lambda)$ as it is easier to determine $\Delta_\alpha(t)$ than $m_\alpha(t)$.

Theorem 5.32 (Jordan Decomposition). *Suppose that the characteristic and minimal polynomials are $\Delta_\alpha(t) = \prod_{1\leq i\leq k}(\lambda_i - t)^{r_i}$ and $m_\alpha(t) = \prod_{1\leq i\leq k}(t-\lambda_i)^{s_i}$ respectively. Then*

$$V = G_\alpha(\lambda_1) \oplus \cdots \oplus G_\alpha(\lambda_k),$$

and if $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$ is the corresponding decomposition of α, then $\Delta_{\alpha_i}(t) = (\lambda_i - t)^{r_i}$ and $m_{\alpha_i}(t) = (t-\lambda_i)^{s_i}$.

Proof. Almost everything follows directly from the Primary Decomposition Theorem 5.25 and Lemma 5.30. It remains to prove that $\Delta_{\alpha_i}(t) = (\lambda_i - t)^{r_i}$. To see this, Corollary 5.9 shows that the roots of m_{α_i} are exactly the eigenvalues of α_i, so $\Delta_{\alpha_i}(t) = (\lambda_i - t)^{t_i}$ for some positive integer t_i. We have that $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$ from Theorem 5.25, and hence

$A = A_1 \oplus \cdots \oplus A_k$ where $A_i \in M_{\ell_i}(\Bbbk)$ is any matrix for the map α_i. Therefore

$$\begin{aligned}(\lambda_1 - t)^{r_1} \cdots (\lambda_k - t)^{r_k} &= \Delta_\alpha(t) \\ &= \det(A - t\mathbb{I}_n) \\ &= \det\left(A_1 \oplus \cdots \oplus A_k - t(\mathbb{I}_{\ell_1} \oplus \cdots \oplus \mathbb{I}_{\ell_k})\right) \\ &= \det\left((A_1 - t\mathbb{I}_{\ell_1}) \oplus \cdots \oplus (A_k - t\mathbb{I}_{\ell_k})\right) \\ &= \det(A_1 - t\mathbb{I}_{\ell_1}) \cdot \det(A_2 - t\mathbb{I}_{\ell_2}) \cdots \det(A_k - t\mathbb{I}_{\ell_k}) \\ &= \Delta_{\alpha_1}(t) \cdots \Delta_{\alpha_k}(t) \\ &= (\lambda_1 - t)^{t_1} \cdots (\lambda_k - t)^{t_k}\end{aligned}$$

where the fact that the determinant of a direct sum equals the product of the determinants is Exercise 9.6 (it was also on the final exercise sheet of Algebra 2A!). Comparing exponents gives $t_i = r_i$ for $i = 1, \ldots, k$ as required. □

The next result achieves the main goal of Algebra 2B:

Corollary 5.33 (**Jordan normal form**). *For $\alpha \in \mathrm{End}(V)$, write the characteristic polynomial as $\Delta_\alpha(t) = (\lambda_1 - t)^{r_1} \cdots (\lambda_k - t)^{r_k}$. Then there exists a basis on V such that the matrix A for α expressed in this basis is*

$$JNF(\alpha) := JNF(\alpha_1) \oplus \cdots \oplus JNF(\alpha_k),$$

where for $1 \leq i \leq k$, the map α_i is the restriction of α to $G_\alpha(\lambda_i)$.

Proof. Theorem 5.32 gives the decompositions $V = G_\alpha(\lambda_1) \oplus G_\alpha(\lambda_2) \oplus \cdots \oplus G_\alpha(\lambda_k)$, $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$, and the characteristic and minimal polynomials of each α_i. The α_i each satisfy the hypotheses of Theorem 5.22 and we conclude that there is a basis $\mathcal{B}_i$ of $G_\alpha(\lambda_i)$ for which α_i has matrix $\mathrm{JNF}(\alpha_i)$. Concatenating these bases gives a basis $\mathcal{B} = \mathcal{B}_1, \ldots, \mathcal{B}_k$ for which α has matrix $\mathrm{JNF}(\alpha) := \mathrm{JNF}(\alpha_1) \oplus \cdots \oplus \mathrm{JNF}(\alpha_k)$. □

Remark 5.34. The matrix A in Theorem 5.33 is called a *Jordan Normal Form* for α. One can show that the Jordan blocks in $\mathrm{JNF}(\alpha)$ are unique up to the order in which we write the blocks.

Example 5.35. We'll conclude the unit by discussing in complete detail how to compute a basis for $\mathbb{C}^4$ that puts the matrix

$$A = \begin{pmatrix} 2 & -4 & 2 & 2 \\ -2 & 0 & 1 & 3 \\ -2 & -2 & 3 & 3 \\ -2 & -6 & 3 & 7 \end{pmatrix},$$

into Jordan Normal Form.

First, we expand the determinant to compute that

$$\Delta_\alpha(t) = (2 - t)^2(4 - t)^2.$$

Using trial and error, we notice that $(A-2\mathbb{I})(A-4\mathbb{I}) \neq 0_{4\times 4}$, but that

$$(A-2\mathbb{I})(A-4\mathbb{I})^2 = 0_{4\times 4},$$

so the minimal polynomial of α is $m_\alpha(t) = (t-2)(t-4\mathbb{I})^2$. By Primary decomposition, we may treat each eigenvalue in turn, namely:

(1) $\lambda = 2$. The exponent of $t-2$ in the minimal polynomial is 1, so Theorem 5.22 says that the maximal Jordan block for eigenvalue 2 is 1×1.
(2) $\lambda = 4$. The exponent of $t-4$ in the minimal polynomial is 2, so Theorem 5.22 says that the maximal Jordan block for eigenvalue 4 is 2×2.

This means we already known that the Jordan Normal Form of the matrix A is of the form

$$\mathrm{JNF}(\alpha) = \begin{pmatrix} 4 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & \lambda \end{pmatrix}$$

where λ equals either 2 or 4. Notice that by Remark 5.34, it doesn't matter which order you put the blocks in, but you are not allowed to split up the 2×2 block with 4's on the diagonal.

To compute $\mathrm{JNF}(\alpha)$, and indeed, to compute a matrix P satisfying $P^{-1}AP = \mathrm{JNF}(\alpha)$, we compute the eigenspaces of both eigenvalues. First, for eigenvalue 2, solve

$$(A-2\mathbb{I})v = 0,$$

that is

$$\begin{pmatrix} 0 & -4 & 2 & 2 \\ -2 & -2 & 1 & 3 \\ -2 & -2 & 1 & 3 \\ -2 & -6 & 3 & 5 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Perform row operations to obtain

$$\begin{pmatrix} 2 & -2 & -1 & -3 \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

The solution set $E_\alpha(2)$ therefore satisfies $2x-2y-z-3w = 0$ and $2y-z-w = 0$, i.e.,

$$E_\alpha(2) = \left\{ \begin{pmatrix} w \\ \frac{1}{2}(z+w) \\ z \\ w \end{pmatrix} \mid z, w \in \mathbb{C} \right\} = \left\{ z \begin{pmatrix} 0 \\ 1 \\ 2 \\ 0 \end{pmatrix} + w \begin{pmatrix} 2 \\ 1 \\ 0 \\ 2 \end{pmatrix} \mid x, y \in \mathbb{C} \right\},$$

so $E_\alpha(2)$ has basis $(0,1,2,0)^T$ and $(2,1,0,2)^T$. In particular, the geometric multiplicity of 2 is two, and therefore there are two Jordan blocks for the eigenvalue 2, that is, 2 is the final unknown entry in the Jordan Normal Form of the matrix.

Next, we repeat for the eigenvalue 4. We already know that the eigenspace must be of dimension one, because there is only one Jordan block for eigenvalue 4. We now solve

$$(A - 4\mathbb{I})v = 0,$$

that is

$$\begin{pmatrix} -2 & -4 & 2 & 2 \\ -2 & -4 & 1 & 3 \\ -2 & -2 & -1 & 3 \\ -2 & -6 & 3 & 3 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}.$$

Perform row operations (add each row to -1 times the top row) to obtain

$$\begin{pmatrix} 2 & 4 & -2 & -2 \\ 0 & 0 & -1 & 1 \\ 0 & 2 & -3 & 1 \\ 0 & -2 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

and then rearrange to get

$$\begin{pmatrix} 2 & 4 & -2 & -2 \\ 0 & 2 & -3 & 1 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The solution set $E_\alpha(4)$ therefore satisfies $2x + 4y - 2z - 2w = 0$, $2y - 3z + w = 0$ and $-z + w = 0$. Therefore

$$E_\alpha(4) = \left\{ \begin{pmatrix} 0 \\ w \\ w \\ w \end{pmatrix} \mid w \in \mathbb{C} \right\} = \left\{ w \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mid w \in \mathbb{C} \right\},$$

so $E_\alpha(2)$ has basis $v_1 = (0, 1, 1, 1)^T$. To compute the matrix P, we must find the second basis vector in the generalised eigenspace for eigenvalue 4 (as in Proposition 5.19), so we compute v_2 given by

$$(A - 4\mathbb{I})v_2 = v_1,$$

that is, we solve

$$\begin{pmatrix} -2 & -4 & 2 & 2 \\ -2 & -4 & 1 & 3 \\ -2 & -2 & -1 & 3 \\ -2 & -6 & 3 & 3 \end{pmatrix} \cdot \begin{pmatrix} x \\ y \\ z \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Perform the same operations as you did in computing the eigenspace gives similar equations (just the right hand sides are different), namely $2x+4y-2z-2w = 0$, $2y-3z+w = 1$ and $-z + w = 1$. A vector satisfying these equations is $v_2 = (1, 0, 0, 1)^T$.

Since we put the Jordan block for eigenvalue 4 first in our Jordan Normal form, we must collect v_1, v_2 as the first columns (in that order!) in our matrix P, and then we can

feed in the basis for the eigenspace $E_\alpha(2)$ as columns three and four, giving

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

If we let J denote the Jordan Normal Form matrix, then you can check that

$$P^{-1}AP = J$$

by checking more simply that $AP = PJ$.

End of Algebra 2B.
