LECTURE NOTES FOR MA20217: ALGEBRA 2B

ALASTAIR CRAW (2013/14)

Abstract. This course introduces abstract ring theory and provides a thorough structure theory of linear operators on finite dimensional vector spaces.

Contents

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 set of $n \times n$ matrices over a fixed field with the usual matrix addition and multiplication, the set of polynomials with coefficients in some fixed field with the usual polynomial addition and multiplication, and the integers. At the same time we want the definition to be somewhat restricted so that we can build a general theory that deals with all these examples at once.

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:

(a) (The associative law)

$$(a * b) * c = a * (b * c) \text{ for all } a, b, c \in G.$$

(b) (Existence of an identity) There exist an element $e \in G$ with the property that

$$e * a = a \text{ and } a * e = a \text{ for all } a \in G.$$

(c) (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 called 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, Prop 6.3):

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. $\square$

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:

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

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

(2) $(R, \cdot)$ satisfies the associative law, that is, we have

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

(3) R satisfies the distributive laws:

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

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

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

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. $\square$

Definition 1.9 (Rings with additional properties). Let $(R, +, \cdot)$ be a ring. Then:

(1) R a *ring with* 1 if there is an element $1 := 1_R \in R$ satisfying

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

(2) R is a *commutative ring* if

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

(3) R a *division ring* if it is a ring with 1 such that

$$\text{for all } a \in R \setminus \{0\}, \text{ there exists } b \in R \text{ such that } ab = 1 = ba.$$

(4) R is a *field* if it is a commutative division ring in which $0 \neq 1$.

Remark 1.10. If R is a ring with 1, then 1 is the unique multiplicative identity. The same argument as before works, i.e., if $\bar{1}$ was another multiplicative identity, then $\bar{1} = \bar{1} \cdot 1 = 1$.

Definition 1.11 (Unit). Let R be a ring with 1. An element $a \in R$ is called 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 1.12. (1) In a division ring, every nonzero element is a unit.
(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}.

(3) If 0 is a unit then Lemma 1.8(1) implies that $1 = 0 \cdot 0^{-1} = 0$. Therefore, for $a \in R$, we have $a = a \cdot 1 = a \cdot 0 = 0$, i.e., R is the zero ring $\{0\}$.

Definition 1.13 (Group of units). Let R be a ring with 1 and write $R^* := \{a \in R \mid a \text{ is a unit}\}$ for the set of all units of R. Then $(R^*, \cdot)$ is a group - the *group of units* of R.

Proof. See Exercise 1.3 for the fact that this is indeed a group. □

Examples 1.14. We have $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$ and $\mathbb{Z}^* = \{1, -1\}$.

1.3. **Examples of rings.** By definition, every field is a commutative ring and hence so are $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ with respect to the usual addition and multiplication. The ring of integers $\mathbb{Z}$ is a commutative ring with 1 that is not a field.

Example 1.15 (The ring of $n \times n$ matrices over R). Let R be a ring with 1. Exercise 1.1 shows that the set $M_n(R)$ of all $n \times n$ matrices over R is a ring with 1 with respect to matrix addition and multiplication. Ask yourself: when is this ring commutative?

Example 1.16 (The endomorphism ring of V). Let V be a finite dimensional vector space over a field $\Bbbk$. An *endomorphism on V* is a linear operator on V, that is, a linear map $\alpha\colon V \to V$. Let $\mathrm{End}(V)$ denote the set of all endomorphisms on V. Define addition and multiplication on $\mathrm{End}(V)$ as follows.

(+) For $\alpha, \beta \in \mathrm{End}(V)$ we let $[\alpha+\beta]\colon V \to V$ be the map that takes v to $\alpha(v)+\beta(v)$. This map is linear, because for $v, w \in V$ and $\lambda \in \Bbbk$ we have

$$\begin{aligned} [\alpha + \beta](\lambda v + w) &= \alpha(\lambda v + w) + \beta(\lambda v + w) && \text{by definition} \\ &= \lambda\alpha(v) + \alpha(w) + \lambda\beta(v) + \beta(w) && \text{as } \alpha, \beta \text{ are linear} \\ &= \lambda\big(\alpha(v) + \beta(v)\big) + \big(\alpha(w) + \beta(w)\big) \\ &= [\alpha + \beta](v) + [\alpha + \beta](w). \end{aligned}$$

This means that $[\alpha + \beta] \in \mathrm{End}(V)$.

($\cdot$) Define multiplication on $\mathrm{End}(V)$ to be composition of maps. Thus for $\alpha, \beta \in \mathrm{End}(V)$ we let $[\alpha \cdot \beta]\colon V \to V$ be the map that takes v to $(\alpha \circ \beta)(v) = \alpha(\beta(v))$. In Algebra 1B you saw that the composition of two linear maps is linear. This means that $[\alpha \cdot \beta] \in \mathrm{End}(V)$.

Exercise 1.2 shows that $\mathrm{End}(V)$ is a ring with 1 with respect to this addition and multiplication. This ring is typically not commutative.

Example 1.17 (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). Let $R[[x]]$ be the set of all formal power series over R. Define addition and multiplication on $R[[x]]$ as follows: given formal power series $\sum_{k=0}^{\infty} a_k x^k, \sum_{k=0}^{\infty} b_k x^k \in R[[x]]$, define

(+) the sum to be the formal power series

$$\sum_{k=0}^{\infty} a_k x^k + \sum_{k=0}^{\infty} b_k x^k := \sum_{k=0}^{\infty} (a_k + b_k) x^k;$$

$(\cdot)$ the product to be the formal power series

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

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.

Notice that if R is a ring with 1, then the power series $1 = 1 + 0x + 0x^2 + 0x^3 + \cdots$ provides a multiplicative identity for $R[[x]]$, and if R is commutative then so is $R[[x]]$.

Remarks 1.18. (1) The formal power series $\sum_{k=0}^{\infty} a_k x^k$ depends only on the sequence (a_k), i.e., the variable x really is superfluous. Indeed, power series $\sum_{k=0}^{\infty} a_k x^k$ and $\sum_{k=0}^{\infty} b_k x^k$ are the same if and only if $(a_k) = (b_k)$.

(2) We're doing algebra rather than analysis, so we're not interested in questions of convergence. Indeed, R is any ring, so it doesn't have a metric in general.

End of Week 1.

1.4. **When do equivalence classes form a ring?** 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 [Proposition 3.5]).

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 1.19. *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 [Lecture 10, "*The algebra of* $\mathbb{Z}_n$"]:

Example 1.20 (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

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

The crucial point for us is that $\mathbb{Z}_n$ is more than a set: in Algebra 1A [Proposition 4.13][1], addition and multiplication were 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

[1] Don't worry, we'll prove this again shortly!

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

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

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 1.21 (**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 1.22. The key point is 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 1.20 is a congruence. More generally, we have the following:

Theorem 1.23 (**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 so is $R/\sim$ (the multiplicative identity is $[1]$); 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, 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 $([a]+[b])\cdot[c] = [a]\cdot[c]+[b]\cdot[a]$ 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

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

so $R/\sim$ is commutative. □

1.5. **Subrings and ideals.** We now introduce subrings and ideals of a ring which leads to a simple method for constructing congruence relations on a ring R.

Definition 1.24 (Subring). A nonempty subset S of a ring R is called 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 1.25. *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. See Exercise 2.2. □

Examples 1.26. (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 $\mathbb{C}$.

(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 1.25. 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 $\mathbb{Z}[i] := \{a + bi \in \mathbb{C} \mid a, b \in \mathbb{Z}\}$ are a subring of the field $\mathbb{C}$, see Exercise 2.1.

Example 1.27 (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 $R[x] \subset R[[x]]$ 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 $a_n \neq 0$ (the degree of the zero polynomial is defined to be $-\infty$).

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 1.25.

The concept of a subring isn't as important as you might guess. After all, Lemma 1.25 says that every subring is a ring in its own right. However, if we strengthen slightly the notion of a subring we obtain the following fantastically useful notion:

Definition 1.28 (Ideal). A nonempty subset I of a ring R is an *ideal* if and only if

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

Remark 1.29. Notice that every ideal I in R is a subring of R. In particular, Lemma 1.25 implies that every ideal contains 0_R.

Example 1.30. Let R be a commutative ring and let $a \in R$. We claim that the set

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

is an ideal of R. Indeed, $0 = 0 \cdot a \in I$, so $I \neq \emptyset$. Also, I is closed under subtraction and multiplication by elements of R because $r \cdot a - s \cdot a = (r - s) \cdot a$ and $s \cdot (r \cdot a) = (rs) \cdot a$.

The following result illustrates one reason why we like ideals so much!

Proposition 1.31. *Let S be a subring in R, and define $\sim$ on R by setting*

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

Then

(1) *the relation $\sim$ is an equivalence relation in which the equivalence classes are the cosets of S in R, i.e., we have $[a] = a + S$ for all $a \in R$; and*
(2) *$\sim$ is a congruence iff S is an ideal.*

Proof. We first show that $\sim$ is an equivalence relation. Let $a, b, c \in R$. Then $a - a = 0 \in S$ means $a \sim a$, so $\sim$ is reflexive. If $a \sim b$ then $a - b \in S$ and hence $b - a = -(a - b) \in S$ by Lemma 1.25. This gives $b \sim a$, so $\sim$ is symmetric. Finally if $a \sim b$ and $b \sim c$ then $a - b, b - c \in S$. As S is closed under addition, it follows that $(a - b) + (b - c) = a - c \in S$ and hence $a \sim c$. This shows that $\sim$ is transitive, so $\sim$ is an equivalence relation.

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

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

as claimed. This proves part (1).

To prove the final statement, suppose first that S is an ideal. Let $a, b, a', b' \in R$ and suppose that $a \sim a'$ and $b \sim b'$. Then $a - a', b - b' \in S$. Since S is an ideal, we have

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

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

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

by the second defining property of an ideal, so $ab \sim a'b'$ as required. Conversely, let S be a subring and suppose that $\sim$ is a congruence relation. Let $a \in S$ and $r \in R$. Then $a \sim 0$, and since $\sim$ is a congruence, we have $r \cdot a \sim r \cdot 0 = 0$ and $a \cdot r \sim 0 \cdot r = 0$. This gives $r \cdot a, a \cdot r \in S$, so the subring S is an ideal. □

Remark 1.32. Proposition 1.31 says that ideals determine congruence relations. Exercise 3.1 establishes the converse statement, i.e., that every congruence relation $\sim$ on a ring R arises from an ideal I in R as described by Proposition 1.31.

Definition 1.33 (Quotient ring). Let I be an ideal in a ring R and let $\sim$ be the corresponding congruence. The *quotient ring* R/I is the ring $R/\sim$ constructed in Theorem 1.23. Explicitly, the ring

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

is the set of cosets of I in R (the congruence classes for $\sim$), and we define addition and multiplication on R/I by

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

Remark 1.34. Remember that these addition and multiplication formulas simply mean that we add and multiply the representatives as if we're adding and multiplying in R, and then we take the coset (=congruence class) of the resulting element of R.

Example 1.35. In Example 1.20, 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 1.36 (The quotient ring $R[x]/I$ for the ideal $I = R[x]x^2$). The ideal $R[x]x^2$ in the ring $R[x]$ determines the congruence relation $\sim$ on $R[x]$, where for $f, g \in R[x]$

$$f \sim g \iff f-g \in R[x]x^2 \iff x^2 | f-g.$$

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

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

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

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

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 $4nm = (2n)(2m) = \phi(n)\phi(m)$.

Example 2.3 (**The quotient map**)**.** Let I be an ideal in a ring R. The *quotient map* asociated to I is the map $\phi\colon R \to R/I$ defined by setting

$$\phi(a) = a + I.$$

This is a ring homomorphism, because

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

and

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

This is the most important example of a ring homomorphism; we'll soon see why!

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

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

Proof. For part (1), we have $\phi(0_R) + 0_S = \phi(0_R) = \phi(0_R + 0_R) = \phi(0_R) + \phi(0_R)$. Now add $-\phi(0_R)$ to both sides to get $\phi(0_R) = 0_S$. For part (2), notice that

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

Since S is an abelian group under addition, we also have $\phi(-a) + \phi(a) = 0_S$, so $\phi(-a)$ is the additive inverse of $\phi(a)$, i.e., $\phi(-a) = -\phi(a)$. For (3), let $a, b \in R$ and compute

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

as required. $\square$

Example 2.5 (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 formal power series $f = \sum_{k=0}^{\infty} a_k x^k \in S[[x]]$, we don't know in general whether or not the element

$$f(r) = \sum_{k=0}^{\infty} a_k r^k$$

lies in R. However, if $f \in S[x]$, that is, if only finitely many of the coefficients a_k are nonzero, then $f(z) \in R$, and hence 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. This is a ring homomorphism, because for $f = \sum_{k=0}^{\infty} a_k x^k$ and $g = \sum_{k=0}^{\infty} b_k x^k$, we have

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

where the third equals sign uses commutativity of addition and distributivity in R. Also

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

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

Definition 2.6 (Ring isomorphism). If a ring homomorphism ϕ is bijective as a map of sets, then we say that ϕ is a *ring isomorphism*. If there exists a ring isomorphism from R to S then we say that R is *isomorphic* to S and write $R \cong S$.

Lemma 2.7. *Let $\phi\colon R \to S$ and $\psi\colon S \to T$ be ring homomorphisms. Then $\psi \circ \phi\colon R \to T$ is a ring homomorphism. Furthermore if ϕ is an isomorphism then so is ϕ^{-1}.*

Proof. See Exercise 3.1 $\square$

Remarks 2.8. (1) If R is isomorphic to S then there is no structural difference between the two rings, i.e., the ring S can be thought of as a copy of R.

(2) Exercise 3.1 shows that 'is isomorphic to' is an equivalence relation, so we're allowed to say that R and S are isomorphic without having to worry about whether we say R first or S first.

Example 2.9 (**Square matrices and Endomorphisms**). Let V be an n-dimensional vector space over a field $\Bbbk$. We claim that the ring $M_n(\Bbbk)$ of $n \times n$ matrices over $\Bbbk$ is isomorphic to the ring $\mathrm{End}(V)$ of linear operators on V. To write down the map between these rings, we recall some results from Algebra 1B. Choose a basis $(v_1, \ldots, v_n)$ of V and consider the invertible linear map

$$\alpha\colon \Bbbk^n \to V \ : \ \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \mapsto a_1v_1 + \cdots + a_nv_n.$$

This map is the bridge between $n \times n$ matrices with entries in $\Bbbk$ and linear maps $V \to V$. Indeed, on one hand, left multiplication by a square matrix $A \in M_n(\Bbbk)$ defines a linear map $A\colon \Bbbk^n \to \Bbbk^n$. On the other hand, the composition

$$a_1v_1 + \cdots + a_nv_n \xrightarrow{\alpha^{-1}} \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \xrightarrow{\text{left mult by } A} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \xrightarrow{\alpha} b_1v_1 + \cdots + b_nv_n,$$

defines the linear map $T_A\colon V \to V$ given by $T_A(v) = \alpha A \alpha^{-1}$. Our claim is that the map

$$\phi\colon M_n(\Bbbk) \longrightarrow \mathrm{End}\,(V) \ : \ A \mapsto T_A$$

is a ring isomorphism. To prove the claim, notice that

$$\phi(A + B) = \alpha(A + B)\alpha^{-1} = \alpha A\alpha^{-1} + \alpha B\alpha^{-1} = T_A + T_B = \phi(A) + \phi(B)$$

and

$$\phi(AB) = \alpha AB\alpha^{-1} = (\alpha A\alpha^{-1})(\alpha B\alpha^{-1}) = T_A \circ T_B = \phi(A)\phi(B),$$

so ϕ is a ring homomorphism. Finally, it's bijective as a map of sets with inverse given by the matrix $\phi^{-1}(f)$ corresponding to the map $\alpha^{-1}f\alpha\colon \Bbbk^n \to \Bbbk^n$. Explicitly, $\phi^{-1}(f)$ is the $n \times n$ matrix whose ith column is $(\alpha^{-1}f\alpha)(e_i)$, where e_i denotes the basis vector of $\Bbbk^n$ with 1 in the ith entry and 0 elsewhere. This shows that ϕ is an isomorphism.

2.2. **The fundamental isomorphism theorem.** We now work towards what is probably the most important results in ring theory.

Definition 2.10 (**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 2.11 (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(1) gives $\phi(0) = 0 = \phi(a)$, and injectivity of ϕ forces $a = 0$, hence $\mathrm{Ker}(\phi) = \{0\}$ as required. $\square$

Lemma 2.12 (Properties of the image). *The image* $\mathrm{Im}(\phi)$ *is a subring of S, and if R is a ring with 1 then so is* $\mathrm{Im}(\phi)$. *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(2), 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. If R is a ring with 1, then the element $\phi(1) \in \mathrm{Im}(\phi)$ satisfies

$$\phi(a) \cdot \phi(1) = \phi(a \cdot 1) = \phi(a) = \phi(1 \cdot a) = \phi(1) \cdot \phi(a)$$

for all $\phi(a) \in \mathrm{Im}(\phi)$, so $\phi(1)$ is a multiplicative identity in $\mathrm{Im}(\phi)$, i.e., the subring $\mathrm{Im}(\phi)$ is a ring with 1. Finally, the fact that ϕ is surjective if and only if $\mathrm{Im}(\phi) = S$ is immediate from the definitions. $\square$

Theorem 2.13 (The fundamental isomorphism theorem). *Let $\phi\colon R \to S$ be a ring homomorphism. Then there is a ring isomorphism*

$$\big(R/\,\mathrm{Ker}(\phi)\big) \cong \mathrm{Im}(\phi).$$

Proof. Consider the map $\overline{\phi}\colon R/\,\mathrm{Ker}(\phi) \to \mathrm{Im}(\phi)$ defined by setting[2]

$$\overline{\phi}\big([a]\big) = \phi(a).$$

To see that this map is well-defined, notice that

$$[a] = [b] \iff a - b \in \mathrm{Ker}(\phi) \iff 0 = \phi(a-b) = \phi(a) - \phi(b) \iff \phi(a) = \phi(b) \tag{2.1}$$

[2]Here we use the equivalence class notation $[a]$ for elements in $R/\,\mathrm{Ker}(\phi)$, but one may equally use coset notation $a + \mathrm{Ker}(\phi)$.

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

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

and

$$\overline{\phi}([a]\cdot[b]) = \overline{\phi}([ab]) = \phi(ab) = \phi(a)\cdot\phi(b) = \overline{\phi}([a])\cdot\overline{\phi}([b]).$$

Notice that $[a] \in \operatorname{Ker}(\overline{\phi})$ iff there exists $a' \in R$ satisfying $[a'] = [a]$ with $\phi(a') = 0$ iff there exists $a' \in \operatorname{Ker}(\phi)$ with $[a] = [a']$ iff $[a] = [0] \in R/\operatorname{Ker}(\phi)$. Thus $\overline{\phi}$ is injective by Lemma 2.11. Also, $\overline{\phi}$ is surjective by definition of $\operatorname{Im}(\phi)$. This finishes the proof. $\square$

Remark 2.14. **It is impossible to overstate how important Theorem 2.13 is**. It 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. I'll draw the relevant diagram in the lecture!!

End of Week 3.

2.3. **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 2.15 (**Characteristic of a ring with 1**)**.** Let R be a ring with 1. The *characteristic* of R, denoted $\operatorname{char}(R)$, is a non-negative integer defined as follows; if there is a positive integer m such that $m1_R = 0_R$, then $\operatorname{char}(R)$ is the smallest such positive integer; otherwise, there is no such positive integer and we say that $\operatorname{char}(R) = 0$.

Examples 2.16.
(1) The zero ring $R = \{0\}$ is actually a ring with 1 (!!), and it's the only ring for which $\operatorname{char}(R) = 1$.
(2) For any positive integer n, we have that $\operatorname{char}(\mathbb{Z}_n) = n$.
(3) The field $\mathbb{C}$ has characteristic zero, and hence so do $\mathbb{Z}, \mathbb{Q}, \mathbb{R}$.

Lemma 2.17. *Let R be a ring of characteristic $n > 0$. Then $n \cdot a = 0$ for all $a \in R$.*

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

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

as required. $\square$

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

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

Lemma 2.18. *Let R be a ring with 1. Then either:*

(1) *char$(R) = 0$, in which case $\mathbb{Z}1_R$ is isomorphic to $\mathbb{Z}$; or*
(2) *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 $\phi(nm) = nm1_R = n1_R \cdot m1_R = \phi(n) \cdot \phi(m)$. Moreover, the image of ϕ is clearly $\mathbb{Z}1_R$.

Suppose first that $\mathrm{char}(R) = 0$. Then $\phi(n) = n \cdot 1_R$ equals 0_R if and only if $n = 0$. Therefore $\mathrm{Ker}(\phi) = \{0\}$, and ϕ is injective by Lemma 2.11. 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. □

2.4. **The Chinese remainder theorem.** In this section we revisit the fabulously named 'Chinese remainder theorem' that you met in Algebra 1A [Propositions 4.18, 4.19, 4.20]. We first introduce and study two new ideals that we can associate to a pair of ideals.

Definition 2.19 (Sum and product 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 product of I and J is the subset

$$IJ := \left\{ \sum_{i=1}^{k} a_i b_i \in R \mid k \in \mathbb{N}, a_i \in I, b_i \in J \text{ for all } 1 \leq i \leq k \right\}$$

of all ab with $a \in I, b \in J$.

Lemma 2.20. *The sets $I \cap J, I + J$ and IJ are ideals of R, and*

$$IJ \subseteq I \cap J \subseteq I + J.$$

Moreover, if R is a commutative ring with 1 satisfying $I + J = R$, then we have $IJ = I \cap J$

Proof. We first show that each of the given subsets of R is an ideal.

- For IJ, we have $0 = 0 \cdot 0 \in IJ$, so $IJ \neq \emptyset$. Consider $\sum_{i=1}^{k} a_i b_i, \sum_{i=1}^{\ell} c_i d_i \in IJ$, for elements $a_i \in I, b_i \in J$ for $1 \leq i \leq k$, and for $c_i \in I, d_i \in J$ for $1 \leq j \leq \ell$. Consider also $r \in R$. Since I and J are ideals, we have that $ra_i \in I$ and $b_i r \in J$ for $1 \leq i \leq k$. It follows that

$$\sum_{i=1}^{k} a_i b_i - \sum_{i=1}^{\ell} c_i d_i = a_1 b_1 + \cdots + a_k b_k + (-c_1)d_1 + \cdots + (-c_\ell)d_\ell \in IJ,$$

$$r\sum_{i=1}^{k} a_i b_i = \sum_{i=1}^{k}(ra_i)b_i \in IJ, \quad \text{and} \quad \left(\sum_{i=1}^{k} a_i b_i\right)\cdot r = \sum_{i=1}^{k} a_i(b_i r) \in IJ.$$

This shows that IJ is also an ideal.

- For $I \cap J$, we have $0 \in I \cap J$, so $I \cap J \neq \emptyset$. Let $a, b \in I \cap J$ and let $r \in R$. As I, J are ideals of R, it follows that $a-b, ra, ar$ lie in both I and J, so $a-b, ra, ar \in I \cap J$. This shows $I \cap J$ is an ideal of R.
- For $I+J$, we have $0 = 0+0 \in I+J$, so $I+J \neq \emptyset$. Let $a_1+b_1, a_2+b_2 \in I+J$ for elements $a_1, a_2 \in I$, $b_1, b_2 \in J$. Consider also $r \in R$. Since I, J are ideals, we have that $a_1 - a_2, ra_1, a_1 r \in I$ and $b_1 - b_2, rb_1, b_1 r \in J$, we have that

$$(a_1 + b_1) - (a_2 + b_2) = (a_1 - a_2) + (b_1 - b_2) \in I + J,$$

and that $r(a_1 + b_1) = ra_1 + rb_1 \in I + J$ and $(a_1 + b_1)r = a_1 r + b_1 r \in I + J$. This shows that $I + J$ is an ideal of R.

For the inclusions, notice that each element $a \in I \cap J$ can be written as $a = a + 0 \in I + J$, so $I \cap J \subseteq I + J$. Let $a_i \in I$ and $b_i \in J$ for $1 \leq i \leq k$ and consider $\sum_{i=1}^{k} a_i b_i \in IJ$. Since both I and J are ideals we have that $a_i b_i \in I$ and $a_i b_i \in J$, so $a_1 b_1, \ldots, a_n b_n \in I \cap J$. Since $I \cap J$ is an ideal, we have that $\sum_{i=1}^{k} a_i b_i \in I \cap J$, so $IJ \subseteq I \cap J$.

For the final statement, we already know that $IJ \subseteq I \cap J$, so it remains to show the opposite inclusion. Let $t \in I \cap J$. Notice first that $I + J = R$ iff $1 = x + y$ for $x \in I$ and $y \in J$. Then we can write

$$t = t \cdot 1 = t(x + y) = tx + ty = xt + ty \in IJ,$$

because commutativity of R gives $tx = xt$. This shows $I \cap J \subseteq IJ$ as required. □

Remark 2.21. A common mistake is to believe that the product of ideals IJ consists only of products of the form ab for $a \in I, b \in J$; it consists of *finite sums* of such elements. The point is that the set $\{ab \in R \mid a \in I, b \in J\}$ is not closed under addition and therefore it cannot be an ideal. Note that IJ is the smallest ideal that contains this set.

Definition 2.22 (Direct product). The *direct product* of rings R and S is the set

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

where addition and multiplication are given by

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

Remark 2.23. All the algebraic laws hold in $R \times S$ since they hold for both R and S; clearly $(0_R, 0_S)$ is the zero element, while $(-a, -b)$ is the additive inverse of (a, b). If both R and S have a 1, then $(1_R, 1_S)$ makes $R \times S$ into a ring with 1, in which case $(a, b) \in R \times S$ is unit if and only if a is a unit in R and b is a unit in S, i.e., $(R \times S)^* = R^* \times S^*$.

Theorem 2.24 (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*

$$\overline{\phi}\colon R/IJ \longrightarrow R/I \times R/J.$$

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

$$\begin{aligned}
\phi(a+b) &= (a+b+I, a+b+J) \\
&= \big((a+I)+(b+I), (a+J)+(b+J)\big) && \text{by Definition 1.33} \\
&= (a+I, a+J) + (b+I, b+J) && \text{by Definition 2.22} \\
&= \phi(a) + \phi(b)
\end{aligned}$$

and

$$\begin{aligned}
\phi(a\cdot b) &= (a\cdot b+I, a\cdot b+J) \\
&= \big((a+I)\cdot(b+I), (a+J)\cdot(b+J)\big) && \text{by Definition 1.33} \\
&= (a+I, a+J) \cdot (b+I, b+J) && \text{by Definition 2.22} \\
&= \phi(a) \cdot \phi(b).
\end{aligned}$$

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 $\operatorname{Ker}(\phi) = I \cap J$. Since $I + J = R$, the final statement of Lemma 2.20 gives $I \cap J = IJ$, hence $\operatorname{Ker}(\phi) = IJ$. Apply the Fundamental Isomorphism Theorem 2.13 to ϕ to see that

$$\overline{\phi}\colon R/IJ \longrightarrow \operatorname{Im}(\phi)$$

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. □

Example 2.25. Let $m, n \in \mathbb{Z}$ be coprime natural numbers. This means that there exists $\lambda, \mu \in \mathbb{Z}$ such that $1 = \lambda m + \mu n$, that is, we have $\mathbb{Z} = \mathbb{Z}m + \mathbb{Z}n$. Apply Lemma 2.20 to the ideals $I = \mathbb{Z}m$ and $J = \mathbb{Z}m$ to see that $IJ = I \cap J = \mathbb{Z}mn$, in which case Theorem 2.24 gives an isomorphism $\overline{\phi}\colon \mathbb{Z}_{mn} \longrightarrow \mathbb{Z}_m \times \mathbb{Z}_n$ which recovers the Chinese Remainder Theorem from Algebra 1A [Proposition 4.18].

3. Factorisation in integral domains

Throughout this section we let R be a commutative ring with 1 such that $0 \neq 1$. We introduce several special classes of such rings and study factorisation properties.

3.1. **Integral domains and Euclidean domains.** We now restrict attention to a class of *commutative* rings that have a very strong cancellation property.

Definition 3.1 (**Integral domain**). Let R be a commutative ring with 1 such that $0 \neq 1$. We say that R is an *integral domain* if for $a, b \in R$,

$$ab = 0 \implies (a = 0 \text{ or } b = 0).$$

Examples 3.2.
(1) Every field $\Bbbk$ is an integral domain. Indeed, if $a, b \in \Bbbk$ satisfy $ab = 0$ and if $a \neq 0$, then $b = 1 \cdot b = a^{-1}ab = a^{-1} \cdot 0 = 0$.
(2) The ring of integers $\mathbb{Z}$ is an integral domain that is not a field.
(3) Every subring of an integral domain is an integral domain.
(4) Let R be an integral domain. By inspecting the formula for multiplication in the ring of formal power series $R[[x]]$, we see that $R[[x]]$ is an integral domain. Part (3) above then implies that the polynomial ring $R[x]$ is an integral domain.

Example 3.3. 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.

End of Week 4.

Lemma 3.4 (**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. □

Proposition 3.5. *The characteristic of an integral domain is either 0 or a prime number.*

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., $n = r \cdot s$ for some $1 < r, s < n$. Then $0 = n \cdot 1_R = rs \cdot 1_R = (r \cdot 1_R) \cdot (s \cdot 1_R)$, but since R is an integral domain it follows that either $r \cdot 1_R = 0$ or $s \cdot 1_R = 0$. Either case is impossible in a ring of characteristic n because $r, s < n$. Thus, the characteristic must be zero or prime after all. □

We concluded this section by formalising another notion that you met in Algebra 1A when studying the rings $\mathbb{Z}$ and $\Bbbk[x]$ where $\Bbbk$ is a field.

Definition 3.6 (Euclidean domain). Let R be an integral domain. A *Euclidean valuation* on R is a map $\nu\colon R \smallsetminus \{0\} \to \{0, 1, 2, \dots\}$ such that:

(1) for $f, g \in R \smallsetminus \{0\}$ 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 3.7. (1) Let $\Bbbk$ be any field, and define $\nu\colon \Bbbk\backslash\{0\} \to \{0, 1, 2, \dots\}$ 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 [Lecture 14]), so $\Bbbk[x]$ is a Euclidean domain.
(4) Recall from Example 1.26 that the Gaussian integers $\mathbb{Z}[i] = \{a+bi \in \mathbb{C} : a, b \in \mathbb{Z}\}$ are a subring of the field $\mathbb{C}$, so $\mathbb{Z}[i]$ is an integral domain. Exercise 5.1 establishes that the map $\nu\colon \mathbb{Z}[i]\smallsetminus\{0\} \to \{0, 1, 2, \dots\}$ given by $\nu(a+bi) = a^2+b^2$ (the absolute value when viewed as a complex number) is a Euclidean valuation, so $\mathbb{Z}[i]$ is a Euclidean domain.

3.2. **Principal ideal domains.** Let R be an integral domain. Since R is necessarily a commutative ring, Example 1.30 shows that each $a \in R$ determines an ideal
$$Ra := \{r \cdot a \mid r \in R\}.$$

Definition 3.8 (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.

Lemma 3.9. *Let R be a nonzero commutative ring with 1. Then R is a field if and only if the only ideals of R are $\{0\}$ and R. In particular, every field is a PID.*

Proof. First let R be a field. For a nonzero ideal I in R, choose $a \in I \setminus \{0\}$. Then any $b \in R$ can be written as $b = (ba^{-1})a \in I$, so $R \subseteq I$ and hence $R = I$ as required. Conversely, let R be a nonzero commutative ring with 1, and suppose $\{0\}$ and R are the only ideals. For $a \in R\setminus\{0\}$, the ideal Ra contains $a = 1a$, so $Ra \neq \{0\}$. Our assumption gives $Ra = R$. In particular $1 = ba$ for some $b \in R$, so a has a multiplicative inverse. This shows that R is a field. The final statement follows from the observation that both $\{0\} = R0$ and $R = R1$ are principal ideals. $\square$

Theorem 3.10 (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 $I \subseteq Rd$. On the other hand, since $g \in I$ we have $Rg \subseteq I$. Hence $I = Rg$ and so I is principal. □

Examples 3.11. Theorem 3.10 implies that the following rings are PID's:

(1) any field (which we proved directly in Lemma 3.9 above);
(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 $\mathbb{Z}[i]$.

Example 3.12. Exercise 5.3 shows that the integral domain $R = \mathbb{Z}[x]$ is not a PID, so it can't be a Euclidean domain.

Example 3.13. It is harder to produce a PID that is not a Euclidean domain. One example is the subring $R = \{a + b(1 + \sqrt{-19})/2 \mid a, b \in \mathbb{Z}\}$ of $\mathbb{C}$. We shan't prove this.

3.3. **Irreducible elements in an integral domain.** Let R be an integral domain.

Definition 3.14 (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 3.15. *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 3.16 (Units don't change the ideal). *Let R be an integral domain and let $a, b \in R$. Then*

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

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 $1a = a = ub = uva$. Since R is a domain the cancellation law 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 3.17 (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* if $p|ab \implies p|a$ or $p|b$ for $a, b \in R$.
(2) p is *irreducible* if $p = ab \implies a$ or b is a unit.

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

Examples 3.18. (1) The prime elements in $\mathbb{Z}$ are $\{\dots, -7, -5, -3, -2, 2, 3, 5, 7, \dots\}$, 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 3.19. *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. □

Remark 3.20. The converse is not true in general, see Exercise 5.5. However, we have:

Proposition 3.21. *Let R be a principal ideal domain. Every irreducible $p \in R$ is prime.*

Proof. Suppose that $p|ab$ and that p does not divide a. We want to show that $p|b$. Since R is a PID, there exists an element $d \in R$ such that

$$Ra + Rp = Rd.$$

In particular, $a, p \in Rd$. Write $p = cd$ for some $c \in R$. Irreducibility of p implies that either c or d is a unit. However, if c were a unit then $a \in Rd = Rp$ by Lemma 3.16, contradicting the fact that p does not divide a. Thus d is a unit, so $Rd = R$ and hence

$$Ra + Rp = R.$$

Since R is a ring with 1, there exists $r, s \in R$ such that $1 = ra + sp$, so

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

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

Corollary 3.22. *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 3.16 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.

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. Since p is irreducible and R is a PID, we proceed precisely as in the previous proof: consider the ideal $Ra+Rp$ and (quoting verbatim from above) we eventually deduce that there exists $r, s \in R$ such that $1 = ra + sp$. Now consider the corresponding cosets:

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

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

3.4. **Unique factorisation domains.** Recall the Fundamental Theorem of Arithmetic from Algebra 1A [Theorem 4.11]:

Theorem 3.23 (**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 each n_i is a positive integer. The primes p_i and their exponents n_i are uniquely determined (up to order).*

Definition 3.24 (**UFD**)**.** An integral domain R is 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 $1 \leq i \leq s$.

Example 3.25. The fundamental theorem of arithmetic implies that $\mathbb{Z}$ is a UFD. This is almost obvious, but we should take care with minus signs. To this end, every nonzero nonunit in $\mathbb{Z}$ is of the form $\pm m$ where m is a natural number greater than 1, so $\pm m = \pm\Pi p_i^{n_i}$ by Theorem 3.23. If this integer is negative then we pull out a single copy of p_1 to help us deal with the minus sign, i.e.,

$$\pm m = -(p_1^{n_1} p_2^{n_2} \cdots p_k^{n_k}) = (-p_1)(p_1)^{n_1-1} p_2^{n_2} \cdots p_k^{n_k}. \tag{3.1}$$

Each prime p_i is irreducible by Proposition 3.19, and irreducibility of p_1 forces irreducibilty of $-p_1$, so (3.1) is the decomposition as in Definition 3.24(1). The fact that the primes p_i and their exponents n_i are uniquely determined (up to order) gives Definition 3.24(2).

End of Week 5.

Rather then relying on Theorem 3.23 to deduce that $\mathbb{Z}$ is a UFD, we provide the following much more general result from which we can recover the fact that $\mathbb{Z}$ is a UFD.

Theorem 3.26. *Let R be a PID. Then R is a UFD.*

Proof. We first establish that part (1) of Definition 3.24 holds. Let $a \in R$ be a nonzero, non-unital 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

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

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

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

which is absurd. This contradiction proves Definition 3.24(1). For part (2), suppose

$$p_1 \cdots p_s = p_1' \cdots p_t' \tag{3.2}$$

are two such decompositions where we may assume without loss of generality that $s \leq t$. Equation (3.2) shows that p_1 divides $p_1' \cdots p_t'$. We know p_1 is prime by Proposition 3.21, 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 3.16. Relabel p_i' as p_1' and vice-versa. We now have $Rp_1 = Rp_1'$, 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

$$1 = u_1 \cdots u_s p_{s+1}' \cdots r_t'.$$

But the p_j' are prime and hence nonunits, so we must have $s = t$. $\square$

Remark 3.27. 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 3.7 is a UFD.

3.5. **General polynomial rings.** We now introduce a beautiful class of integral domains that are UFD's but not PIDs.

Definition 3.28 (**General polynomial ring**). For $n \geq 1$, 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

$$f(x_1, \ldots, x_n) = \sum_{i_1, \ldots, i_n \geq 0} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n}, \tag{3.3}$$

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. The *polynomial ring* $R[x_1, \ldots, x_n]$ is the set of all such polynomials, where addition and multiplication of polynomials f, g are defined as follows:

- the sum $f + g$ is defined by gathering terms and adding coefficients, i.e.,

$$\sum_{i_1, \ldots, i_n \geq 0} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n} + \sum_{i_1, \ldots, i_n \geq 0} b_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_n^{i_n} = \sum_{i_1, \ldots, i_n \geq 0} (a_{i_1, \ldots, i_n} + b_{i_1, \ldots, i_n}) x_1^{i_1} \cdots x_n^{i_n};$$

- the product $f \cdot g$ is defined as usual by distributivity (you write down the formula!) together with multiplication of monomials given by

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

These operations generalise the operations familiar in the case $n = 1$.

Example 3.29. 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

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

Proposition 3.30. *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

$$f(x_1, \ldots, x_n) = \sum_{i_n \geq 0} \left(\sum_{i_1, \ldots, i_{n-1} \geq 0} a_{i_1, \ldots, i_n} x_1^{i_1} \cdots x_{n-1}^{i_{n-1}} \right) x_n^{i_n}, \tag{3.4}$$

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 6.4 for details. $\square$

Remark 3.31. For any field $\Bbbk$, the ring $\Bbbk[x_1]$ is a Euclidean domain and hence a PID. However, for any $n \geq 2$, the ring $\Bbbk[x_1, \ldots, x_n]$ is not a PID, see Exercise 6.5.

3.6. **Field of fractions and Gauss' lemma.** Let R be an integral domain.

Theorem 3.32 (Polynomial rings are UFD's). *If R is a UFD then $R[x]$ is a UFD.*

Examples 3.33.
(1) $\mathbb{Z}$ is a UFD, hence so is $\mathbb{Z}[x]$ (yet it's not a PID by Exercise 5.3).
(2) Let $\Bbbk$ be a field, so $\Bbbk$ is a UFD. Exercise 6.4 shows that $\Bbbk[x_1, \ldots, x_n] \cong S[x_n]$ for $S = \Bbbk[x_1, \ldots, x_{n-1}]$, so induction and Theorem 3.32 implies $\Bbbk[x_1, \ldots, x_n]$ is a UFD (Exercise 6.5 shows that $\Bbbk[x_1, \ldots, x_n]$ is not a PID for $n \geq 2$).

We need two ingredients to prove Theorem 3.32. First, Exercise 6.3 shows that the set

$$\mathrm{Frac}(R) = \left\{ \frac{a}{b} \mid a, b \in R \text{ with } b \neq 0 \right\}$$

together with the relation $\frac{a}{b} \sim \frac{c}{d} \iff ad = bc$ is such that the set of equivalence classes $F(R) := \mathrm{Frac}(R)/\sim$ admits addition and multiplication given by

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

With these binary operations, the set $F(R)$ becomes field.

Definition 3.34 (Field of fractions of an integral domain). The *field of fractions* of an integral domain R is the field $F(R) := \mathrm{Frac}(R)/\sim$.

Remark 3.35. The map $R \to F(R)$ given by $a \mapsto \frac{a}{1}$ is an injective ring homomorphism, so R is a subring of the field $F(R)$.

Example 3.36. The field of fractions of the ring $\mathbb{Z}$ is the field $\mathbb{Q}$ (!), and we know $\mathbb{Z} \subset \mathbb{Q}$.

The second ingredient we need for the proof of Theorem 3.32 is:

Definition 3.37 (Primitive polynomial). Let R be a UFD. A nonconstant polynomial $f = \sum_{i=0}^{n} a_i x^i \in R[x]$ is *primitive* if the only common divisors of all the coefficients of f are units in R.

Remark 3.38. In light of unique factorisation, it's equivalent to say that f is primitive if and only if no irreducible $p \in R$ divides all coefficients of f.

Example 3.39. $x^3 + 2x - 1 \in \mathbb{Z}[x]$ is primitive, whereas $3x^3 + 6x - 3 \in \mathbb{Z}[x]$ is not.

Lemma 3.40 (Pulling out the content). *Let R be a UFD. For every nonconstant $f \in R[x]$, there exists $c \in R$ (unique upto multiplication by a unit) and a primitive polynomial $g \in R[x]$ (unique upto multiplication by a unit of R) such that $f = c \cdot g$.*

Proof. Write $f = \sum_{i=0}^{n} a_i x^i \in R[x]$. Since R is a UFD we may decompose each $a_i \in R$ as a product of irreducibles in R. Let p be irreducible in R. If the decomposition of each a_i involves an irreducible q_i with $Rq_i = Rp$, write $q_i = u_i p$ for some unit $u_i \in R$ by Lemma 3.16, and replace each occurrence of q_i in the decomposition of a_i by $u_i p$. Now factor out the highest possible power of p that is common to all a_i, i.e., let n be such that each a_i is divisible by p^n and is not divisible by p^{n+1}. Repeat for the next irreducible in the decomposition, and so on. If we let $c \in R$ denote the product of all such irreducibles, then $f = c \cdot g$ for some $g \in R[x]$ which is primitive by construction.

For uniqueness, suppose $f = d \cdot h$ with $c \in R$ and $h \in R[x]$ primitive. Each irreducible factor of c divides $f = d \cdot h$, and since h is primitive, the factor divided d. Symmetrically, each irreducible factor of d divides c. Cancelling all such factors in the expression $c \cdot g = d \cdot h$ removes all irreducibles factors of c and d, leaving only units in their place, i.e., $u \cdot g = v \cdot h$ for units $u, v \in R$. Then $h = (uv^{-1})g$, so g is unique up to multiplication by a unit. Moreover, if k is the product of all irreducible factors that we just cancelled, then $c = uk$ and $d = vk$, so $c = u(v^{-1}d) = (uv^{-1})d$, so c is unique up to multiplication by a unit. $\square$

Lemma 3.41 (**Gauss' Lemma**)**.** *Let R be a UFD. The product of finitely many primitive polynomials in $R[x]$ is primitive.*

Proof. It suffices to prove the result for two polynomials and apply induction. To this end, let $f = \sum_{i=0}^{n} a_i x^i$ and $g = \sum_{j=0}^{m} b_i x^i$ be primitive in $R[x]$ and let p be irreducible in R. Our goal is to find a coefficient of fg that is not divisible by p. Since f and g are primitive, we know p doesn't divide each a_i, nor does it divide each b_j. Let k be minimal such that a_k is not divisible by p, and similarly, let ℓ be minimal such that b_j is not divisible by p. The coefficient of $x^{k+\ell}$ in the product fg is

$$(a_0 b_{k+\ell} + \cdots + a_{k-1} b_{\ell+1}) + a_k b_\ell + (a_{k+1} b_{\ell-1} + \cdots + a_{k+l} b_0). \tag{3.5}$$

Minimality of k implies that p divides $a_0 b_{k+\ell} + \cdots + a_{k-1} b_{\ell+1}$, while minimality of ℓ implies that p divides $a_{k+1} b_{\ell-1} + \cdots + a_{k+l} b_0$. However, p does not divide a_k or b_ℓ, so by unique factorisation in $R[x]$, it doesn't divide the product $a_k b_\ell$ and in particular, it doesn't divide the coefficient (3.5) of $x^{k+\ell}$ in fg. Thus (3.5) is the required coefficient. $\square$

Proof of Theorem 3.32. We first establish the decomposition into irreducibles in $R[x]$ as in Definition 3.24(1). Let $f \in R[x]$ be a nonzero, non-unit. If $\deg(f) = 0$ then $f \in R$, and since R is a UFD we obtain a decomposition of f as a product of irreducible elements of R, each of which must be irreducible in $R[x]$ for degree reasons. Otherwise $\deg(f) \geq 1$. Write F for the field of fractions of the integral domain R, and regard f as an element of $F[x]$. Since F is a field, $F[x]$ is a UFD by Remark 3.27, so we can write $f = p_1 p_2 \cdots p_s$ for irreducible elements $p_1, \ldots, p_s \in F[x]$. The coefficients of each p_i lie in F, so every such coefficient is of the form a/b for some $a, b \in R$, and clearing denomenators gives

$$r \cdot f = q_1 q_2 \cdots q_s \tag{3.6}$$

for some $r \in R$ and $q_1, \ldots, q_s \in R[x]$. Notice that each $q_i = u_i p_i$ for some nonzero $u_i \in R$. Every nonzero element in R is a unit in F, so since p_i is irreducible in $F[x]$ it follows that q_i is irreducible when regarded as an element of $F[x]$. Now apply Lemma 3.40 to draw the content out of each polynomial in equation (3.6), giving

$$r(c\overline{f}) = (c_1 \overline{q_1}) \cdots (c_s \overline{q_s}) = (c_1 \cdots c_s) \overline{q_1} \cdots \overline{q_s} \tag{3.7}$$

where $c, c_1, \ldots, c_s \in R$ are the contents of $f, q_1, \ldots, q_s \in R[x]$ respectively. The product of primitive polynomials is primitive by Gauss' Lemma 3.41, so in fact this equation provides two apparently different ways to draw the content out of a polynomial. The uniqueness statement from Lemma 3.40 shows that these two expressions for the content must be related by a unit, i.e., there exists a unit $u \in R$ such that $rcu = c_1 \cdots c_s$. Substitute into (3.7) to get $rc\overline{f} = rcu \prod_{i=1}^{s} \overline{q_i}$ and cancel r by Lemma 3.4 to get that

$$f = c\overline{f} = cu\overline{q_1} \cdots \overline{q_s}.$$

Now, $cu \in R$ admits a decomposition into irreducibles in R (since R is a UFD) and hence irreducibles in $R[x]$ (for degree reasons). Moreover, each $\overline{q_i}$ is irreducible in $R[x]$, because each is both primitive in $R[x]$ and irreducible in $F[x]$. This gives our desired decomposition, so Definition 3.24(1) holds.

To show uniqueness as in Definition 3.24(2), consider a decomposition of $f \in R[x]$ as a product of irreducibles as above. If $\deg(f) = 0$, then the decomposition is unique because we used the UFD property of the ring R to produce the decomposition in that case. Otherwise, $\deg(f) \geq 1$. Every irreducible in $R[x]$ is also irreducible when regarded as an element of $F[x]$, so our decomposition of f in $R[x]$ may be regarded as a decomposition into a product of irreducibles in $F[x]$. Since F is a field, the ring $F[x]$ is a UFD, so the polynomials appearing in the decomposition are unique up to multiplication by units in $F[x]$, that is, by nonzero elements of R. To ensure that these nonzero elements of R don't ruin uniqueness in $R[x]$, notice that a given irreducible factor in our decomposition is either noconstant, in which case it's primitive and hence (by Lemma 3.40) it's unique up to multiplication by a unit, or it's constant, in which case notice that the product of all such irreducibles equals the content of f and this product is therefore unique up to multiplication by a unit in R by Lemma 3.40. □

End of Week 6.

---

4. Associative algebras with 1 over a field

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

4.1. **Algebras.** We'll first give the most general definition of an algebra over a field, even though we're primarily interested in a smaller class of algebras.

Definition 4.1 ($\Bbbk$**-algebra**). Let $\Bbbk$ be a field, and let V be a $\Bbbk$-vector space V that has a bilinear product, that is, a map $\cdot\colon V \times V \to V$ that is bilinear over $\Bbbk$:

$$\begin{aligned}(\lambda u_1 + u_2) \cdot v &= \lambda(u_1 \cdot v) + u_2 \cdot v \\ u \cdot (\lambda v_1 + v_2) &= \lambda(u \cdot v_1) + u \cdot v_2.\end{aligned}$$

We say that V is a $\Bbbk$-*algebra* if the product is associative[3]. If in addition the product has a multiplicative identity then V is a $\Bbbk$-*algebra with 1*.

Lemma 4.2 (**$\Bbbk$-algebras are rings**). *A nonempty set V is a $\Bbbk$-algebra if and only if V is a ring that admits a map $\Bbbk \times V \to V$ which makes V into a vector space, such that*

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

Proof. ($\Longrightarrow$) Let V be a $\Bbbk$-algebra. Since V is a vector space, it is already an abelian group under addition. The product is an associative binary operation by definition, and the formulae from Definition 4.1 in the special case $\lambda = 1$ show that it satisfies the distributive laws, so V is a ring. Formula (4.1) holds by substituting $u_2 = v_2 = 0$ into the formulae from Definition 4.1. ($\Longleftarrow$) For the opposite direction one need only check that the multiplication operation in the ring V is bilinear over $\Bbbk$, but this follows from the distributivity laws and the equations (4.1), e.g.,

$$(\lambda u_1 + u_2) \cdot v = (\lambda u_1) \cdot v + u_2 \cdot v = \lambda(u_1 \cdot v) + u_2 \cdot v.$$

The other distributivity law is similar. □

Definition 4.3 (**Subalgebra**). A *subalgebra* of a $\Bbbk$-algebra V is a nonempty subset $W \subseteq V$ that is both a subring and a vector subspace of V.

Remarks 4.4.
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
$$\left(\sum_{i \in I} \alpha_i v_i\right) \cdot \left(\sum_{j \in I} \beta_j v_j\right) = \sum_{i \in I,\, j \in J} (\alpha_i \beta_j)(v_i \cdot v_j).$$

Examples 4.5.
1. Let $\Bbbk$ be a field. Then $\Bbbk = \Bbbk \cdot 1$ is a $\Bbbk$-algebra of dimension 1.
2. The field $\mathbb{C} = \mathbb{R} + \mathbb{R}i$ is an $\mathbb{R}$-algebra that is a 2-dimensional vector space over $\mathbb{R}$.
3. [**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 bilinear product is determined from
$$i^2 = j^2 = k^2 = -1, \;\; ij = k, \;\; jk = i, \;\; ki = j, \;\; ji = -k, \;\; kj = -i, \;\; ik = -j.$$
Exercise 7.1 shows that $\mathbb{H}$ is a (noncommutative!) ring with 1; this is the *quaternionic algebra*, or simply, *the quaternions*. Both $\mathbb{R}$ and $\mathbb{C}$ are subalgebras of $\mathbb{H}$.

[3]In defining a $\Bbbk$-algebra, some people drop the requirement that the multiplication is associative, because many such examples arise naturally (e.g., Lie algebras $\mathfrak{g}$, the Octonion algebra $\mathbb{O}$). However, our $\Bbbk$-algebras will always be associative because, as Lemma 4.2 shows, this extra assumption enables us to think ring-theoretic thoughts.

(3) Let $\Bbbk$ be a field. For $n \geq 1$, the general polynomial ring $\Bbbk[x_1, \ldots, x_n]$ is a $\Bbbk$-algebra with basis as a vector space equal to the set of all monomials

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

this vector space is not finite dimensional! (As in Remark 4.4, multiplication of polynomials is determined by the bilinearity of the product and multiplication of monomials, namely $(x_1^{i_1} x_2^{i_2} \cdots x_n^{i_n}) \cdot (x_1^{j_1} x_2^{j_2} \cdots x_n^{j_n}) = x_1^{i_1+j_1} x_2^{i_2+j_2} \cdots x_n^{i_n+j_n}$.)

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

Definition 4.6 (**Subfield and field extension**). A subring $\Bbbk$ of a field K is a *subfield* if for each $a \in \Bbbk \setminus \{0\}$, the multiplicative inverse of a in the field K lies in $\Bbbk$. We also refer to $\Bbbk \subseteq K$ as a *field extension.*

Lemma 4.7. *Let $\Bbbk \subseteq K$ be a field extension. Then K is a $\Bbbk$-algebra.*

Proof. Exercise 8.1 implies that $1_\Bbbk = 1_K$. By restricting the multiplication $K \times K \to K$, we obtain a map $\Bbbk \times K \to K$ given by $(\lambda, v) \mapsto \lambda v$. Since K is a field, $(K, +)$ is an abelian group, and hence

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

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 $\lambda \in \Bbbk$. Therefore K is a $\Bbbk$-algebra. □

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

Theorem 4.8 (**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*

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

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 $n = \min\{\deg(p) \mid p \in \Bbbk[x]$ satisfies $p(a) = 0\}$.

Proof. Consider the evaluation homomorphism $\phi_a \colon \Bbbk[x] \to K$ given by $\phi_a(f) = f(a)$ (see Example 2.5). Since $\Bbbk$ is a field, $\Bbbk[x]$ is a PID and hence $\mathrm{Ker}(\phi_a)$ is a principal ideal, that is, $\mathrm{Ker}(\phi_a) \cong \Bbbk[x]p$ for some $p \in \Bbbk[x]$. The fundamental isomorphism theorem gives

$$\Bbbk[x]/\Bbbk[x]p \cong \mathrm{Im}(\phi_a) = \left\{f(a) \in K \mid f \in \Bbbk[x]\right\} = \Bbbk[a]. \tag{4.2}$$

Notice that the polynomial p is a nonzero nonunit element, because

- $p \neq 0$, otherwise $\mathrm{Ker}(\phi_a) = \{0\}$, so the only element of $\Bbbk[x]$ having a as a root is the zero polynomial which is absurd; and
- p is not a unit, otherwise $\mathrm{Ker}(\phi_a) \cong \Bbbk[x]p = \Bbbk[x]$, so $0 = \phi(1) = 1$ which is absurd.

Examples 3.2 show that the field K is an integral domain, and that every subring of an integral domain is an integral domain, so $\Bbbk[a] := \mathrm{Im}(\phi_a)$ is an integral domain. It follows from the isomorphism (4.2) that $\Bbbk[x]/\Bbbk[x]p$ is an integral domain. The key step is to apply Exercise 6.2 to deduce that (p is irreducible and) $\Bbbk[x]/\Bbbk[x]p$ is a field (!). Since $\Bbbk \subseteq K$ is a field extension, we have $1_K = 1_\Bbbk \in \Bbbk$ and hence $1_K = 1_\Bbbk \in \Bbbk[a]$, so both inclusions of fields $\Bbbk \subseteq \Bbbk[a] \subseteq K$ are actually field extensions by Exercise 8.3.

Lemma 4.7 shows $\Bbbk[x]/\Bbbk[x]p$ is a $\Bbbk$-algebra, so it remains to show $(1, a, a^2, \dots, a^{n-1})$ 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 $f = qg + r$ 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 $1, a, \dots, a^{n-1}$. To show that $1, a, \dots, a^{n-1}$ 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 4.9. (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 $x^3 - 2 \in \mathbb{R}[x]$. 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 4.8. 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 4.10 (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 Corollary 3.22 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 identitfy $\Bbbk$ with the subfield $\Bbbk[1] \subseteq K$ then we have that $\Bbbk \subseteq K$ is a field extension.

We will show that $[1], [x], \dots, [x]^{n-1}$ is a basis for the $\Bbbk$-vector space $K = \Bbbk[x]/\Bbbk[x]p$. To show spanning, let $[f] \in \Bbbk[x]/\Bbbk[x]p$. Since $\Bbbk[x]$ is a Euclidean domain, there exists $q, r \in \Bbbk[x]$ such that $f = qp + r$, where r is either zero or a nonzero polynomial of degree less than $\deg(p) = n$. If we write $r = b_0 + b_1x + \cdots + b_{n-1}x^{n-1}$, then

$$\begin{aligned}
[f] &= [q][p] + [r] \\
&= [r] && \text{as } [p] = [0] \\
&= [b_0 + b_1x + \cdots + b_{n-1}x^{n-1}] \\
&= b_0[1] + b_1[x] + \cdots + b_{n-1}[x]^{n-1},
\end{aligned}$$

so $[1], [x], \ldots, [x]^{n-1}$ span K over $\Bbbk$. To show linear independence, if

$$[0] = c_0[1] + c_1[x] + \cdots + c_{n-1}[x]^{n-1} = [c_0 + c_1x + \cdots + c_{n-1}x^{n-1}],$$

then $h := c_0 + c_1x + \cdots + c_{n-1}x^{n-1}$ lies in $\Bbbk[x]p$. In particular, p divides f, but since $\deg(f) = n - 1 < n = \deg(p)$, we must have $f = 0$ and hence $c_0 = c_1 = \ldots = c_{n-1} = 0$, so $[1], [x], \ldots, [x]^{n-1}$ are linearly independent over $\Bbbk$.

Finally, to see that $a = [x]$ is a root of p, write $p = \sum_i \alpha_i x^i$, so

$$p(a) = \sum_i \alpha_i a^i = \sum_i \alpha_i [x]^i = \Big[\sum_i \alpha_i x^i\Big] = [p] = [0]$$

as required. □

Corollary 4.11. *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. This is Exercise 7.4. □

End of Week 7.

Examples 4.12. (1) The polynomial $p = x^2 + 1 \in \mathbb{R}[x]$ is irreducible in $\mathbb{R}[x]$, so Theorem 4.10 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 $x^2 - 2 \in \mathbb{Q}[x]$. This is an irreducible polynomial in $\mathbb{Q}[x]$ and Theorem 4.10 gives a root a in the field

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

where $a = [x]$. This field is isomorphic to the subfield $\mathbb{Q} + \mathbb{Q}\sqrt{2}$ 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

$$\mathbb{Z}_2[x]/\mathbb{Z}_2 f = \mathbb{Z}_2 + \mathbb{Z}_2 a.$$

Notice that this new field has $2^2 = 4$ elements (compare Exercise 3.4).

4.3. **Normed $\mathbb{R}$-algebras.** Recall from [Algebra 2A, Section 2.1] 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 4.13 (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 $u, v \in V$.

Remarks 4.14.

(1) 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 $\|1_V\| = 1$.

(2) Recall from Remarks 4.4(2) that the structure of an $\mathbb{R}$-algebra V is determined by the dimension of V over $\Bbbk$ and the product of elements in some chosen basis.

Examples 4.15 ($\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ are normed $\mathbb{R}$-algebras). Examples 4.5 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. That is:

(1) on $\mathbb{R}$ the norm 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) on $\mathbb{C}$ the norm is $\|a + bi\| = \sqrt{a^2 + b^2}$, so for $a + bi, c + di \in \mathbb{C}$ we have

$$\begin{aligned} \|(a+bi)\cdot(c+di)\| &= \sqrt{(ac-bd)^2 + (bc+ad)^2} \\ &= \sqrt{(ac)^2 + (bc)^2 + (ad)^2 + (bd)^2} \\ &= \sqrt{(a^2+b^2)}\sqrt{(c^2+d^2)} = \|a+bi\| \cdot \|c+di\|, \end{aligned}$$

so $\mathbb{C}$ is a normed $\mathbb{R}$-algebra.

(3) on $\mathbb{H}$ the norm is $\|a + bi + cj + dk\| = \sqrt{a^2 + b^2 + c^2 + d^2}$. Exercise 8.2 shows that $\|u \cdot v\| = \|u\| \cdot \|v\|$ for all $u, v \in \mathbb{H}$, so $\mathbb{H}$ is a normed $\mathbb{R}$-algebra.

Lemma 4.16. *Let V be a normed $\mathbb{R}$-algebra.*

(1) *If $(1, t) \in V$ are orthonormal, then $t^2 = -1$.*

(2) *If $(1, i, j) \in V$ are orthonormal, then so are $(1, i, j, ij)$. Moreover $ji = -ij$.*

Proof. (Nonexaminable) For (1), we have $\|t^2\| = \|t\|^2 = 1$, so

$$\|t^2 + (-1)\| = \|(t-1)(t+1)\| = \|t-1\| \cdot \|t+1\| = \sqrt{2}\sqrt{2} = 1 + 1 = \|t^2\| + \|-1\|.$$

According to the triangle inequality we should only get equality here if t^2 is a positive multiple of -1 and, as $\|t^2\| = 1$, this can only happen if $t^2 = (-1)$. For (2), we have that

$\frac{i+j}{\sqrt{2}}$ is orthogonal to 1 and of length 1. By part (1), it follows that

$$-1 = \Big(\frac{i+j}{\sqrt{2}}\Big)^2 = \frac{i^2+j^2+ij+ji}{2} = \frac{(-1)+(-1)+ij+ji}{2} = -1 + \frac{ij+ji}{2}.$$

Hence $ji = -ij$. Notice that $\|ij\| = \|i\| \cdot \|j\| = 1$, so

$$\|ij + (-i)\|^2 = \|i(j-1)\|^2 = \|i\|^2 \cdot \|j-1\|^2 = 1 \cdot 2 = 1 + 1 = \|ij\|^2 + \|-i\|^2.$$

The Pythogoras theorem implies that ij is orthogonal to i. Similarly, write $\|ij+(-j)\|^2 = \|ij\|^2 + \|-j\|^2$ to see that ij is orthogonal to j. Finally

$$\|ij - 1\|^2 = \|ij + i^2\|^2 = \|i(j+i)\|^2 = \|i\|^2 \cdot \|j+i\|^2 = 1 \cdot 2 = 2 = \|ij\|^2 + \|-1\|^2$$

gives that ij is orthogonal to 1 as well. □

Theorem 4.17 (Classification of normed $\mathbb{R}$-algebras). *There are exactly three normed $\mathbb{R}$-algebras up to isomorphism, namely, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{H}$ (see Examples 4.15).*

Proof. Let V be a normed $\mathbb{R}$-algebra. We check case-by-case according to the dimension of V as a vector space over $\mathbb{R}$.

If $\dim V = 1$, then $V = \mathbb{R}1_V$. Since $1_V \cdot 1_V = 1_V$, we have that V is isomorphic as an $\mathbb{R}$-algebra (that is, as a ring and as an $\mathbb{R}$-vector space) to $\mathbb{R}$. If $\dim V = 2$, we may choose an orthonormal basis $(1, i)$ and Lemma 4.16(1) shows that $i^2 = -1$. Thus, V is isomorphic as an $\mathbb{R}$-algebra to $\mathbb{C}$. If $\dim V \geq 3$, then Lemma 4.16(2) shows that if $(1, i, j) \in V$ are orthonormal, then so are $(1, i, j, ij)$ and hence $\dim V \geq 4$.

If $\dim(V) = 4$, we may choose an orthonormal basis $(1, i, j, ij)$ of V. The linear map $\phi\colon V \to \mathbb{H}$ sending $1, i, j, ij$ to $1, i, j, k$ respectively preserves the product and hence shows that V is isomorphic to $\mathbb{H}$ as an $\mathbb{R}$-algebra. Indeed, we have $i^2 = j^2 = (ij)^2 = -1$ on V by Lemma 4.16(1) and $i^2 = j^2 = k^2 = -1$ on $\mathbb{H}$ by definition. As for the other products in V, Lemma 4.16(2) shows that $ji = -ij$ (and similarly, for any pair among i, j, ij) while in $\mathbb{H}$ we have $ji = -ij = -k$ by definition (and similarly, for any pair among i, j, k). Thus, the product of any two basis elements, and hence the structure of the algebra, is uniquely determined.

If $\dim(V) > 4$, we derive a contradiction, i.e., no such V exists. For this, take an orthonormal set of vectors $1, i, j$ and apply Lemma 4.16 to get a subspace $\mathbb{R}+\mathbb{R}i+\mathbb{R}j+\mathbb{R}ij$ of V. Now pick $e \in V$ with $\|e\| = 1$ that is orthogonal to $1, i, j, ij$. Lemma 4.16(2) gives

$$(ij)e = -e(ij) = iej = -ije$$

and thus we get $ije = 0$ but $\|ije\| = \|i\| \cdot \|j\| \cdot \|e\| = 1$ so this is absurd. □

4.4. **Application to number theory.** Exercise 7.2 studies the link between geometry in $\mathbb{R}^3$ and $\mathbb{H}$, where the inner product and cross product in $\mathbb{R}^3$ can be interpreted via $\mathbb{H}$. Now we investigate a beautiful application in Number Theory. Consider the subring

$$\mathbb{Z} + \mathbb{Z}i + \mathbb{Z}j + \mathbb{Z}k := \big\{ z_1 + z_2 i + z_3 j + z_4 k \in \mathbb{H} \mid z_1, z_2, z_3, z_4 \in \mathbb{Z} \big\}$$

of the quaternions. For $z = z_1 + z_2 i + z_3 j + z_4 k$ and $w = w_1 + w_2 i + w_3 j + w_4 k$, we have

$$\begin{aligned} zw &= (z_1w_1 - z_2w_2 - z_3w_3 - z_4w_4) + (z_1w_2 + z_2w_1 + z_3w_4 - z_4w_3)i \\ &\quad +(z_1w_3 - z_2w_4 + z_3w_1 + z_4w_2)j + (z_1w_4 + z_2w_3 - z_3w_2 + z_4w_1)k. \end{aligned}$$

Exercise 8.2 gives that $\|z\|^2\|w\|^2 = \|z \cdot w\|^2$, so

$$\begin{aligned} &(z_1^2 + z_2^2 + z_3^2 + z_4^2)(w_1^2 + w_2^2 + w_3^2 + w_4^2) = \\ &\quad (z_1w_1 - z_2w_2 - z_3w_3 - z_4w_4)^2 + (z_1w_2 + z_2w_1 + z_3w_4 - z_4w_3)^2 \\ &\quad +(z_1w_3 - z_2w_4 + z_3w_1 + z_4w_2)^2 + (z_1w_4 + z_2w_3 - z_3w_2 + z_4w_1)^2. \end{aligned} \tag{4.3}$$

It follows that if we have two sums of four squares, then their product is also a sum of four squares that we can find explicitly using this formula. We are now going to prove that every natural number can be written as sum of four integer squares.

Theorem 4.18 (**Lagrange's four square theorem**). *Every natural number can be written as a sum of four integer squares.*

Proof. We break the proof down into a number of steps.

STEP 1: (IT SUFFICES TO CONSIDER ODD PRIMES) Notice first that $1 = 1^2 + 0^2 + 0^2 + 0^2$ and that $2 = 1^2 + 1^2 + 0^2 + 0^2$. Since the set consisting of sum of four squares is closed under multiplication and since $\mathbb{Z}$ is a UFD, it suffices to show that every odd prime p can be written as a sum of four squares.

STEP 2: (AN EQUATION INVOLVING z_i'S) We claim that we can define an integer m to be the smallest positive integer in the range $0 < m < p$ such that

$$pm = z_1^2 + z_2^2 + z_3^2 + z_4^2. \tag{4.4}$$

To justify the claim, we must exhibit $z_1, \ldots, z_4$ and m such that the equation holds. For this, we show that for any odd (positive) prime p, there exists $x, y, m \in \mathbb{Z}$ such that

$$pm = x^2 + y^2 + 1^2 + 0^2 \text{ where } 0 < m < p.$$

For this we calculate modulo p. If $[x]^2 = [y]^2$ for some $0 \leq y < x \leq (p-1)/2$, then $p|(x^2 - y^2) = (x-y)(x+y)$, so $p|(x-y)$ or $p|(x+y)$ because p is prime. This is absurd since $1 \leq x - y, x + y \leq p - 1$, so $[0]^2, [1]^2, \ldots, [\frac{p-1}{2}]^2$ are distinct. Thus, we get two lists

$$[1 + x^2], \quad 0 \leq x \leq (p-1)/2 \quad \text{ and } [-y^2], \quad 0 \leq y \leq (p-1)/2$$

each of which has $(p+1)/2$ distinct values. There are $p+1 > p$ values in total, so the two lists must have a value in common, say $[1 + x^2] = [-y^2]$. Then $[1 + x^2 + y^2] = [0]$. Hence $pm = 1 + x^2 + y^2$ for some integer m. Now $pm = 1 + x^2 + y^2 \leq 1 + (\frac{p-1}{2})^2 + (\frac{p-1}{2})^2 < 1 + 2(p/2)^2 < p^2$, so $m < p$ as required.

STEP 3: (SET UP THE CONTRADICTION) The aim now is to show that $m = 1$. We argue by contradiction and suppose that $m > 1$.

STEP 4: (m IS ODD). Otherwise an even number of z_1, z_2, z_3, z_4 are odd. By rearranging the order of terms if needed we can assume that both z_1, z_2 are even/odd and both z_3, z_4 are even/odd. Hence $z_1 + z_2, z_1 - z_2, z_3 + z_4, z_3 - z_4$ are all even. It follows that

$$\frac{pm}{2} = \frac{2(z_1^2 + z_2^2 + z_3^2 + z_4^2)}{4} = \Big(\frac{z_1 - z_2}{2}\Big)^2 + \Big(\frac{z_1 + z_2}{2}\Big)^2 + \Big(\frac{z_3 - z_4}{2}\Big)^2 + \Big(\frac{z_3 + z_4}{2}\Big)^2$$

which contradicts the minimality of m. Hence m is odd.

STEP 5: (WE DO NOT HAVE $[z_1] = [z_2] = [z_3] = [z_4] = [0] \in \mathbb{Z}_m$.) Otherwise m would divide all of $z_1, \dots, z_4$, so the right hand side of (4.4) would be divisible by m^2. But then $m|p$ and as $m < p$, we would have $m = 1$ contracting our assumption that $m > 1$.

STEP 6: (FIND $0 < r < m$ SATISFYING EQUATION IN w_i'S.) For each $i \in \{1, 2, 3, 4\}$ pick w_i such that $-(m-1)/2 \le w_i \le (m-1)/2$ and $[w_i] = [z_i]$ (needs m odd!). We have $[w_1^2 + w_2^2 + w_3^2 + w_4^2] = [z_1^2 + z_2^2 + z_3^2 + z_4^2] = [0] \in \mathbb{Z}_m$, so there exists r such that

$$mr = w_1^2 + w_2^2 + w_3^2 + w_4^2. \tag{4.5}$$

Since $|w_i| \le (m-1)/2$, this expression is bounded above by $4(\frac{m-1}{2})^2 = (m-1)(m-1)$, so $r < m$. Since $[w_i] = [z_i]$ for $1 \le i \le 4$, Step 5 implies that we do not have $[w_1] = [w_2] = [w_3] = [w_4] = [0] \in \mathbb{Z}_m$, so the right hand side of (4.5) is non-zero. Thus $0 < r < m$.

STEP 7: (PUTTING BOTH EQUATIONS TOGETHER.) Multiply (4.4) and (4.5) and use our understanding of multiplying quaternions from (4.3) to obtain

$$\begin{aligned}
prm^2 &= (z_1^2 + z_2^2 + z_3^2 + z_4^2)(w_1^2 + (-w_2)^2 + (-w_3)^2 + (-w_4)^2) \\
&= (z_1w_1 + z_2w_2 + z_3w_3 + z_4w_4)^2 + (-z_1w_2 + z_2w_1 - z_3w_4 + z_4w_3)^2 \\
&\quad +(-z_1w_3 + z_2w_4 + z_3w_1 - z_4w_2)^2 + (-z_1w_4 - z_2w_3 + z_3w_2 + z_4w_1)^2.
\end{aligned}$$

Since $[w_i] = [z_i] \in \mathbb{Z}_m$ for $1 \le i \le 4$, we calculate in $\mathbb{Z}_m$ that

$$\begin{aligned}
[z_1w_1 + z_2w_2 + z_3w_3 + z_4w_4] &= [z_1^2 + z_2^2 + z_3^2 + z_4^2] = [pm] = [0] \\
[-z_1w_2 + z_2w_1 - z_3w_4 + z_4w_3] &= [-z_1z_2 + z_2z_1 - z_3z_4 + z_4z_3] = [0] \\
[-z_1w_3 + z_2w_4 + z_3w_1 - z_4w_2] &= [-z_1z_3 + z_2z_4 + z_3z_1 - z_4z_2] = [0] \\
[-z_1w_4 - z_2w_3 + z_3w_2 + z_4w_1] &= [-z_1z_4 - z_2z_3 + z_3z_2 + z_4z_1] = [0].
\end{aligned}$$

Thus, all of these integers are divisible by m, so dividing by m^2 in the above gives

$$\begin{aligned}
pr &= \Big(\frac{z_1w_1 + z_2w_2 + z_3w_3 + z_4w_4}{m}\Big)^2 + \Big(\frac{-z_1w_2 + z_2w_1 - z_3w_4 + z_4w_3}{m}\Big)^2 \\
&\quad \Big(\frac{-z_1w_3 + z_2w_4 + z_3w_1 - z_4w_2}{m}\Big)^2 + \Big(\frac{-z_1w_4 - z_2w_3 + z_3w_2 + z_4w_1}{m}\Big)^2.
\end{aligned}$$

As $r < m$, we get a contradiction about our minimality assumption on m. It follows that the smallest m given in (4.4) must be 1 and thus p is a sum of integer squares. □

End of Week 8.

---

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 $(v_1, v_2, \ldots, v_n)$ 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 Example 2.9 that the rings $\mathrm{End}\,(V)$ and $M_n(\Bbbk)$ are isomorphic 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 $\mathrm{id}, \alpha, \alpha^2, \ldots, \alpha^{n^2}$ comprising n^2+1 linear operators, or equivalently, the list $(\mathbb{I}_n, A, A^2, \ldots, A^{n^2})$ of matrices, is linearly dependent. If $a_0, \ldots, a_{n^2} \in \Bbbk$ (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. $\square$

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 3.10 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 $m_A(t) = t^2 - 1$.

Definition 5.4 (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 5.5.

(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 $\mathrm{am}(\lambda) \geq \mathrm{gm}(\lambda)$.

Lemma 5.6. *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 $p(\lambda) = 0$. $\square$

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 $\Delta_\alpha(\alpha) = 0 \in M_n(\Bbbk)$.*

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

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

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

Proof of Theorem 5.7. Suppose $\Delta_A(t) = \det(A - t\mathbb{I}_n) = a_0 + a_1t + \cdots + a_nt^n$. We must show that $\Delta_A(A) = a_0\mathbb{I}_n + a_1A + \cdots + a_nA^n$ is equal to the zero matrix. Recall the adjugate formula from [Algebra 1B]:

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

Write $\operatorname{adj}(A - t\mathbb{I}_n) = B_0 + B_1t + \cdots + B_{n-1}t^{n-1}$ for $B_i \in M_n(\Bbbk)$. Substite into (5.2) gives

$$(B_0 + B_1t + \cdots + B_{n-1}t^{n-1})(A - t\mathbb{I}_n) = (a_0 + a_1t + \cdots + a_nt^n)\mathbb{I}_n. \tag{5.3}$$

Comparing terms involving t^i for any $1 \leq i \leq q$, 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. $\square$

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 $\operatorname{Ker}(\Phi_\alpha) = \Bbbk[t]m_\alpha$, 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 5.6. $\square$

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 $1 \leq i \leq k$.

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 $\alpha|_W \in \operatorname{End}(W)$, 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 $\alpha|_1\colon V_1 \to V_1$ is simply rotation by θ radians in the plane, while $\alpha|_2\colon V_2 \to V_2$ 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 $\alpha|_1$ and $\alpha|_2$ in the given bases on V_1 and V_2 respectively.

In Examples 5.12(3) it is convenient to think of the map α as the sum of $\alpha|_1$ and $\alpha|_2$, and think of the matrix A as the sum of the corresponding block matrices as follows.

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 $\alpha_i \in \operatorname{End}(V_i)$. 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

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

The *direct sum of matrices* $A_1, \ldots, A_k$, where $A_i \in M_{n_i}(\Bbbk)$ for $1 \leq i \leq k$, is the matrix

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

Remark 5.14. To see the link between these notions, let A_i be the matrix for α_i with respect to some basis pick a basis $\mathcal{V}_i$ of V_i. Then the matrix for the direct sum $\alpha_1\oplus\cdots\oplus\alpha_k$ with respect to the basis $\mathcal{V}_1 \cup \mathcal{V}_2 \cup \cdots \cup \mathcal{V}_k$ of $\bigoplus_{1\leq i\leq k} V_i$ is the matrix $A_1 \oplus \cdots \oplus A_k$.

Lemma 5.15. *For $\alpha \in \mathrm{End}(V)$ and 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 $\alpha_i := \alpha|_{V_i} \in \mathrm{End}(V_i)$. Then*

(1) *$\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k \in \bigoplus_{i=1}^k \mathrm{End}(V_i)$; 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

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

which proves (1). It follows that any $f \in \Bbbk[t]$ satisfies $f(\alpha) = f(\alpha_1)\oplus f(\alpha_2)\oplus\cdots\oplus f(\alpha_k)$. In particular, 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. □

5.3. **Primary Decomposition.** The rotation map α from Examples 5.12(3) was simple in the sense that we could easily compute α-invariant subspaces V_1 and V_2 such that $V = V_1 \oplus V_2$ and $\alpha = \alpha|_{V_1} \oplus \alpha|_{V_2}$. This is good, because for any basis on V_1 and V_2, the matrix for α in the corresponding basis of V is a block matrix and so has many zeroes.

More generally, given $\alpha\colon V \to V$, how do we find α-invariant subspaces $V_1, \ldots, V_k$ of V such that $V = V_1 \oplus\cdots\oplus V_k$ and $\alpha = \alpha_1 \oplus\cdots\oplus\alpha_k$ where $\alpha_i := \alpha|_{V_i}$? The key is to obtain the α-invariant subspaces V_i using the factorisation of the minimal polynomial m_α.

Example 5.16. Consider rotation by θ radians about the z-axis from Examples 5.12(3). If for a moment we work over $\mathbb{C}$, we compute that 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 over $\mathbb{C}$ we have

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

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

$$m_\alpha(t) = (t^2 - 2\cos\theta t + 1)(t - 1)$$

as the factorisation of m_α into irreducibles in $\mathbb{R}[t]$ (which is a UFD). In fact, we have that

$$m_{\alpha_1} = t^2 - 2\cos\theta t + 1 \quad \text{and} \quad m_{\alpha_2} = t - 1$$

are the minimal polynomials of $\alpha_1 = \alpha|_{V_1}$ and $\alpha_2 = \alpha|_2$ respectively. We now construct the α-invariant subspaces V_1 and V_2 in V *purely* from these factors of m_α. First compute

$$\begin{aligned} m_{\alpha_1}(\alpha) &= \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix}^2 - 2\cos\theta \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} . + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 2-2\cos(\theta) \end{pmatrix} \end{aligned}$$

and

$$m_{\alpha_2}(\alpha) = \begin{pmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{pmatrix} - \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos(\theta)-1 & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta)-1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

Notice that

$$\operatorname{Ker}(m_{\alpha_1}(\alpha)) = \left\{ \begin{pmatrix} x \\ y \\ 0 \end{pmatrix} \in \mathbb{R}^3 \mid x, y \in \mathbb{R} \right\} \quad \text{and} \quad \operatorname{Ker}(m_{\alpha_2}(\alpha)) = \left\{ \begin{pmatrix} 0 \\ 0 \\ z \end{pmatrix} \in \mathbb{R}^3 \mid z \in \mathbb{R} \right\}$$

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

Our next result shows that the phenomenon we noticed above holds in general.

Theorem 5.17 (**Primary Decomposition**). *Let $\alpha\colon V \to V$ be a linear operator and write factorise $m_\alpha = p_1^{n_1} \cdots p_k^{n_k}$, where the $n_i \in \mathbb{N}$ are chosen so that the irreducible monic factors p_i satisfy $\Bbbk[t]p_i \neq \Bbbk[t]p_j$ for $i \neq j$. Let $q_i = p_i^{n_i}$ and let $V_i = \operatorname{Ker}(q_i(\alpha))$. 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 $1 \leq i \leq k$ satisfy $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$ and $m_{\alpha_i} = q_i$.*

Corollary 5.18 (**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 of Corollary 5.18. Let $\alpha\colon V \to V$ be diagonalisable with a basis of eigenvectors $(v_1, \ldots, v_n)$ and corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$. This means $V = \Bbbk v_1 \oplus \cdots \oplus \Bbbk v_n$ is a decomposition into α-invariant subspaces, so Lemma 5.15(1) shows that $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_n$ for $\alpha_i := \alpha|_{V_i} \in \operatorname{End}(V_i)$. The map $\alpha_i\colon \Bbbk v_i \to \Bbbk v_i$ is simply multiplication by λ_i and hence

$m_{\alpha_i}(t) = t - \lambda_i$. Lemma 5.15(2) shows that the minimal polynomial $m_\alpha(t)$ is the least common multiple of the polynomials $\{t - \lambda_i \mid 1 \leq i \leq n\}$, that is, the product over those $t - \lambda_i$ that are distinct. If we reorder the eigenvalues so that the distinct eigenvalues are $\lambda_1, \ldots, \lambda_k$ for $k \leq n$, then $m_\alpha(t) = (t - \lambda_1)(t - \lambda_2)\cdots(t - \lambda_k)$.

For the converse, apply Theorem 5.17 with $q_i := t - \lambda_i$ for $1 \leq i \leq k$ to obtain

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

as required. □

End of Week 9.

Proof of Theorem 5.17. We use induction on k. For $k = 1$, we have $m_\alpha = p_1^{n_1} = q_1$. Then

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

because $m_\alpha(\alpha)$ is the zero map by Definition 5.2. This proves the case $k = 1$. For $k \geq 2$, suppose the result holds for any linear operator whose minimal polynomial decomposes as a product of fewer than k factors of the form $p_i^{n_i}$. Suppose now that $m_\alpha = p_1^{n_1} \cdots p_k^{n_k}$. Define $q_1 = p_1^{n_1} \cdots p_{k-1}^{n_{k-1}}$ and $q_2 = p_k^{n_k}$, so $m_\alpha = q_1 q_2$. Since $\Bbbk[t]$ is a PID, there exists nonzero $g \in \Bbbk[t]$ such that $\Bbbk[t]q_1 + \Bbbk[t]q_2 = \Bbbk[t]g$. It follows that g divides both q_1 and q_2. If g is not a unit then it has a (monic) irreducible factor, say $p \in \Bbbk[t]$, that divides both q_1 and q_2. But the factorisations of q_1 and q_2 are unique, so $\Bbbk[t]p = \Bbbk[t]p_k$ and $\Bbbk[t]p = \Bbbk[t]p_i$ for some $1 \leq i < k$. But then $\Bbbk[t]p_i = \Bbbk[t]p_k$ which is absurd because $i \neq k$. Thus $g \in \Bbbk[t]$ is a unit, in which case Lemma 3.16 implies that $\Bbbk[t] = \Bbbk[t]q_1 + \Bbbk[t]q_2$. Proposition 5.19 to follow shows that

$$V = \operatorname{Ker}\big(q_1(\alpha)\big) \oplus \operatorname{Ker}\big(q_2(\alpha)\big),$$

where $\alpha_i := \alpha|_{\operatorname{Ker}(q_i(\alpha))}$ satisfies $\alpha = \alpha_1 \oplus \alpha_2$ and $m_{\alpha_i} = q_i$ for $1 \leq i \leq 2$. In particular, α_1 is a linear operator on $\operatorname{Ker}(q_1(\alpha))$ whose minimal polynomial decomposes as a product $q_1 = p_1^{n_1} \cdots p_{k-1}^{n_{k-1}}$, so the result follows by applying the inductive hypothesis to α_1. □

Proposition 5.19. *Let $\alpha\colon V \to V$ be a linear operator whose minimal polynomial m_α satisfies $m_\alpha = q_1 q_2$ where q_1, q_2 are monic and satisfying $\Bbbk[t] = \Bbbk[t]q_1 + \Bbbk[t]q_2$. Then*

1. *the α-invariant subspaces $V_1 = \operatorname{Im}(q_2(\alpha))$ and $V_2 = \operatorname{Im}(q_1(\alpha))$ satisfy $V = V_1 \oplus V_2$;*
2. *the maps $\alpha_i = \alpha|_{V_i}$ for $1 \leq i \leq 2$ satisfy $\alpha = \alpha_1 \oplus \alpha_2$ and $m_{\alpha_i} = q_i$; and*
3. *we have $V_1 \cong \operatorname{Ker}(q_1(\alpha))$ and $V_2 \cong \operatorname{Ker}(q_2(\alpha))$.*

Proof. For (1), let $v = q_i(u) \in \operatorname{Im}(q_i(\alpha))$. Since $q_i(\alpha)$ commutes with α, we have

$$\alpha(v) = \alpha(q_i(u)) = q_i(\alpha(u)) \in \operatorname{Im}(q_i(\alpha)),$$

so $\operatorname{Im}(q_i(\alpha))$ is α-invariant for $1 \leq i \leq 2$. There exists $f, g \in \Bbbk[t]$ such that $1 = fq_1 + gq_2$, so $\mathrm{id} = f(\alpha)q_1(\alpha) + g(\alpha)q_2(\alpha)$. It follows that for any $v \in V$, we have

$$v = \mathrm{id}(v) = [g(\alpha)q_2(\alpha)](v) + [f(\alpha)q_1(\alpha)](v) \in \operatorname{Im}(q_2(\alpha)) + \operatorname{Im}(q_1(\alpha)) = V_1 + V_2.$$

This shows that $V = V_1 + V_2$. To see that the sum is direct, suppose $v \in V_1 \cap V_2$, say $v = q_2(\alpha)(v_2) = q_1(\alpha)(v_1)$. Then

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

Hence $V_1 \cap V_2 = \{0\}$ and $V = V_1 \oplus V_2$. For (2), the first statement follows from Lemma 5.15. For the second, we must show that q_i generates the kernel of the map $\Phi_{\alpha_i} \colon \Bbbk[t] \to \mathrm{End}(V_i)$, that is, each $f \in \Bbbk[t]$ satisfying $\Phi_{\alpha_1}(f) = 0$ is divisible by q_i. Now

$$\begin{aligned} f \in \mathrm{Ker}(\Phi_{\alpha_1}) &\iff f(\alpha_1)(v_1) = 0 \text{ for all } v_1 \in V_1 && \text{by definition of } \Phi_{\alpha_1} \\ &\iff f(\alpha)(v_1) = 0 \text{ for all } v_1 \in V_1 && \text{as } \alpha(v_1) = \alpha_1(v_1) \text{ for } v_1 \in V_1 \\ &\iff [f(\alpha)q_2(\alpha)](v) = 0 \text{ for all } v \in V && \text{as } V_1 = \mathrm{Im}(q_2(\alpha)) \\ &\iff m_\alpha \text{ divides } fq_2 && \text{as } m_\alpha \text{ generates } \mathrm{Ker}(\Phi_\alpha) \\ &\iff q_1 \text{ divides } f && \text{as } m_\alpha = q_1q_2 \\ &\iff q_1 \text{ is the minimal polynomial of } \alpha_1 \end{aligned}$$

as required. Similarly q_2 is the minimal polynomial of α_2. For (3), each $v \in V$ satisfies $q_1(\alpha)q_2(\alpha)(v) = m_\alpha(v) = 0$, we have that $V_1 = \mathrm{Im}(q_2(\alpha)) \subseteq \mathrm{Ker}(q_1(\alpha))$. The result will follow from [Algebra 1B] when we show these spaces have the same dimension. The rank-nullity theorem from [Algebra 1B] gives that

$$\dim \mathrm{Ker}(q_1(\alpha)) + \dim \mathrm{Im}(q_1(\alpha)) = \dim V = \dim V_1 + \dim V_2.$$

Now subtract $\dim \mathrm{Im}(q_1(\alpha)) = \dim V_2$ from each side to leave $\dim \mathrm{Ker}(q_1) = \dim V_1$ as required. Showing $V_2 = \mathrm{Ker}(q_2(\alpha))$ is similar. □

5.4. **The Jordan Decomposition over $\mathbb{C}$.** From now on we restrict to the case $\Bbbk = \mathbb{C}$. All polynomials in $\mathbb{C}[t]$ factor as a product of polynomials of degree 1. Now suppose that the linear operator $\alpha \colon V \to V$ has minimal polynomial

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

where $\lambda_1, \ldots, \lambda_k$ are the distinct eigenvalues of α (recall the roots of m_α are exactly the eigenvalues of α). The Primary Decomposition Theorem 5.17 implies that

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

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

Definition 5.20 (**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 V is

$$\begin{aligned} G_\alpha(\lambda) &= \left\{v \in V : (\alpha - \lambda\,\mathrm{id})^s v = 0 \text{ for some positive integer } s\right\} \cup \{0\} \\ &= \left\{v \in V : (\alpha - \lambda\,\mathrm{id})^s v = 0 \text{ for some } s \geq 0\right\} \end{aligned}$$

Remarks 5.21. (1) We have $E_\alpha(\lambda) \subseteq G_\alpha(\lambda)$.

(2) Since V has finite dimension, the chain of ideals

$$E_\alpha(\lambda) = \mathrm{Ker}(\alpha - \lambda \mathrm{id}) \subseteq \mathrm{Ker}(\alpha - \lambda \mathrm{id})^2 \subseteq \mathrm{Ker}(\alpha - \lambda \mathrm{id})^3 \subseteq \cdots$$

must stabilise at some point. The next Lemma tells us when.

Lemma 5.22. *Let s be the multiplicity of the eigenvalue λ as a root of m_α. Then*

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

Proof. The right hand side is contained in the left by Definition 5.20. 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 $\lambda = \lambda_i$. For $j \neq i$ we have that α_j only has the eigenvalue λ_j. Hence $\ker(\alpha_j - \lambda_i \mathrm{id}) = \{0\}$ 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. $\square$

Remark 5.23. 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 often easier to determine $\Delta_\alpha(t)$ than $m_\alpha(t)$.

Theorem 5.24 (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 5.17 and Lemma 5.22. 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 5.17, 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

$$\begin{aligned}
(\lambda_1 - t)^{r_1} \cdots (\lambda_k - t)^{r_k} &= \Delta_\alpha(t) \\
&= \det(A - t\mathbb{I}_n) \\
&= \det\big(A_1 \oplus \cdots \oplus A_k - t(\mathbb{I}_{\ell_1} \oplus \cdots \oplus \mathbb{I}_{\ell_k})\big) \\
&= \det\big((A_1 - t\mathbb{I}_{\ell_1}) \oplus \cdots \oplus (A_k - t\mathbb{I}_{\ell_k})\big) \\
&= \det(A_1 - t\mathbb{I}_{\ell_1}) \cdot \det(A_2 - t\mathbb{I}_{\ell_2}) \cdots \det(A_k - t\mathbb{I}_{\ell_k}) \quad \text{by Ex 10.3} \\
&= \Delta_{\alpha_1}(t) \cdots \Delta_{\alpha_k}(t) \\
&= (\lambda_1 - t)^{t_1} \cdots (\lambda_k - t)^{t_k}
\end{aligned}$$

Comparing exponents gives $t_i = r_i$ for $i = 1, \ldots, k$ as required. □

5.5. **Jordan normal form over $\mathbb{C}$.** Our study of the structure of α is now reduced to understanding each α_i, so we need only consider the special case $\alpha\colon V \to V$ such that

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

where $1 \leq s \leq r$. We work over $\Bbbk = \mathbb{C}$.

Definition 5.25 (**Cyclic subspace generated by** v)**.** For $v \in V$, the *cyclic α-invariant subspace generated by* v is the subspace

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

Remark 5.26. Note that $\mathbb{C}[\alpha]v$ is an α-invariant subspace of V. Indeed, for $p, q \in \mathbb{C}[t]$ and $\lambda \in \Bbbk$, we have $\lambda(p(\alpha)v) + q(\alpha)v = (\lambda p + q)(\alpha)v$, 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 5.27. 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 $\Bbbk v$ is the cyclic α-invariant subspace generated by v.

Proposition 5.28. *Let $\alpha\colon V \to V$ be any linear map such that $\Delta_\alpha(t) = (\lambda - t)^r$ and $m_\alpha(t) = (t - \lambda)^s$. For $v \in V \setminus \{0\}$, consider the $\mathbb{C}$-vector space $W := \mathbb{C}[\alpha]v$. Define e to be the smallest positive integer such that $(\alpha - \lambda id)^e v = 0$, and define*

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

The matrix for $\beta := \alpha|_W \in \mathrm{End}(W)$ in the basis $(v_1, v_2, \ldots, v_e)$ is the $e \times e$ matrix

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

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

Proof. As $m_\alpha(t) = (t-\lambda)^s$, we have that $(\alpha - \lambda\mathrm{id})^s v = m_\alpha(\alpha)v = 0$, so $1 \le e \le s$ is well-defined. To see that $v_1, \dots, v_e$ span W, let $u \in W$. By hypothesis $u = f(\alpha)v$ for some $f \in \mathbb{C}[t]$. Exercise 10.2 gives $a_0, \dots, a_e \in \mathbb{C}$ such that $f(t) = a_0 + a_1(t-\lambda) + \cdots + a_k(t-\lambda)^k$ for some $k \ge 0$, and hence

$$u = f(\alpha)v = a_0 v + a_1(\alpha - \lambda\mathrm{id})v + a_2(\alpha - \lambda\mathrm{id})^2 v + \cdots ,$$

so W is spanned by $v_1, \dots, v_e$ because $(\alpha - \lambda\mathrm{id})^e v = 0$. Exercise 9.4(b) shows that $v_1, \dots, v_e$ are linearly independent, so we have a basis.

Notice that

$$\alpha(v_1) = \lambda\, v_1 + (\alpha - \lambda\mathrm{id})v_1 = \lambda\, v_1 + (\alpha - \lambda\,\mathrm{id})^e v = \lambda\, v_1$$

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

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

the matrix for α with respect to the basis $v_1, \dots, v_e$ is therefore $J(\lambda, e)$. All other statements follow from Exercise 10.1. □

Definition 5.29 (Jordan block). We call $J(\lambda, e)$ a *Jordan block* of α.

Examples 5.30. (1) $J(\lambda, 1) = (\lambda)$ and $J(\lambda, 2) = \begin{pmatrix} \lambda & 1 \\ 0 & \lambda \end{pmatrix}$.

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

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

The characteristic polynomial is $(3/2-t)(1/2-t)+1/4 = 1-2t+t^2 = (1-t)^2$. As the matrix A is not the unit matrix the minimal polynomial is $(t-1)^2 = \Delta_\alpha(t)$. The situation is thus like Proposition 5.28 with $e = 2$. Following the recipe there, we seek a vector v such that $(A - I)v \ne 0$, say $v = (0\ 2)^T$. If we let $v_1 = (A-I)v = (1\ -1)^T$ and $v_2 = v$, the matrix for α in basis (v_1, v_2) is $J(1, 2)$.

The following is the key result that we've been aiming towards throughout Algebra 2B:

Theorem 5.31 (Jordan normal form). *Let* $\alpha : V \to V$ *be any linear map 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*

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

where

(1) $k = gm(\lambda)$ *is the number of Jordan blocks;*

(2) $s = \max\{e_1, \ldots, e_k\}$; *and*
(3) $r = e_1 + \cdots + e_k$.

Proof. By Proposition 5.28, showing A is a direct sum of Jordan blocks is equivalent to showing that there exist non-zero $v_1, \ldots, v_k \in V$ such that

$$V = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_k, \tag{5.5}$$

with $\dim \mathbb{C}[\alpha]v_i = e_i$. Suppose that we have already established this. Then:

(1) Let α_i be the restriction of α to $\mathbb{C}[\alpha]v$ so that $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_k$. By (5.5), every element of V can be written $v = v_1 + \cdots + v_k \in V$. If $v \in E_\alpha(\lambda)$, then

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

and thus $\alpha_i(v_i) = \lambda v_i$ for $1 \leq i \leq k$. It follows that $E_\alpha(\lambda) = E_{\alpha_1}(\lambda) \oplus \cdots \oplus E_{\alpha_k}(\lambda)$. By Proposition 5.28, we have $\dim E_{\alpha_i}(\lambda) = 1$, so

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

This proves (1).

(2) Lemma 5.15 shows that $m_\alpha(t)$ is the least common multiple of $m_{\alpha_1}(t), \ldots, m_{\alpha_k}(t)$. Proposition 5.28 shows that $m_{\alpha_i}(t) = (t - \lambda)^{e_i}$, so (2) follows immediately.

(3) Finally, (3) says nothing more than $\dim V = \dim \mathbb{C}[\alpha]v_1 + \cdots + \dim \mathbb{C}[\alpha]v_k$.

It remains to show that (5.5) holds. We establish this by induction on s.

If $s = 1$, then $\alpha = \lambda\mathrm{id}$. Pick any basis $v_1, \ldots, v_r$ for V and apply Proposition 5.28 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 $s \geq 2$ and that the claim holds for smaller values of s. Now consider the α-invariant subspace

$$W = (\alpha - \lambda\mathrm{id})V = \{(\alpha - \lambda\mathrm{id})(v) \in V \mid v \in V\}.$$

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

$$W = \mathbb{C}[\alpha](\alpha - \lambda\mathrm{id})v_1 \oplus \cdots \oplus \mathbb{C}[\alpha](\alpha - \lambda\mathrm{id})v_\ell. \tag{5.6}$$

Let β_i be the restriction of α to $\mathbb{C}[\alpha]v_i$. Proposition 5.28 shows that $E_{\beta_i}(\lambda)$ has dimension 1 and that it has a basis vector of the form $w_i = (\alpha - \lambda\mathrm{id})^{e_i - 1}v_i$ for some $e_i \geq 2$. Notice that $w_i \in \mathbb{C}[\alpha](\alpha - \lambda\mathrm{id})v_i$. Since the sum from (5.6) is direct, it follows 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 $E_\alpha(\lambda) \subseteq V$. 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

$$V = \mathbb{C}[\alpha]v_1 \oplus \cdots \oplus \mathbb{C}[\alpha]v_\ell \oplus (\mathbb{C}v_{\ell+1} \oplus \cdots \oplus \mathbb{C}v_{\ell+m}). \tag{5.7}$$

To see that the left hand side is contained in the right, let $v \in V$. Then $(\alpha - \lambda\mathrm{id})v \in W$, so by (5.6) there exist $p_1, \ldots, p_e \in \mathbb{C}[t]$ such that

$$(\alpha - \lambda\mathrm{id})v = p_1(\alpha)(\alpha - \lambda\mathrm{id})v_1 + \cdots + p_e(\alpha)(\alpha - \lambda\mathrm{id})v_\ell.$$

Gather all terms on one side to obtain $(\alpha - \lambda\mathrm{id})(v - (p_1(\alpha)v_1 + \cdots + p_e(\alpha)v_e)) = 0$, 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 (5.7) as required. It remains to show that the sum from (5.7) 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 $\alpha - \lambda\mathrm{id}$ to both sides gives

$$0 = p_1(\alpha)(\alpha - \lambda\mathrm{id})v_1 + \cdots + p_\ell(\alpha)(\alpha - \lambda\mathrm{id})v_\ell.$$

Since W is a direct sum in equation (5.6), we have $(\alpha - \lambda\mathrm{id})p_i(\alpha)v_i = 0$ for $1 \le i \le \ell$, 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 $1 \le i \le \ell$. 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. ☐

Remarks 5.32.
(1) The matrix A in Theorem 5.31 is called a *Jordan Normal Form* for α, sometimes denoted $\mathrm{JNF}(\alpha)$. One can show that the Jordan blocks in $\mathrm{JNF}(\alpha)$ are unique up to order.
(2) This generalises as follows. If $\alpha\colon V \to V$ has $\Delta_\alpha(t) = (\lambda_1 - t)^{r_1} \cdots (\lambda_m - t)^{r_m}$ and $V = G_\alpha(\lambda_1) \oplus G_\alpha(\lambda_2) \oplus \cdots \oplus G_\alpha(\lambda_m)$ with the corresponding decomposition $\alpha = \alpha_1 \oplus \cdots \oplus \alpha_m$,, then $\mathrm{JNF}(\alpha) = \mathrm{JNF}(\alpha_1) \oplus \cdots \oplus \mathrm{JNF}(\alpha_m)$.

Example 5.33. Suppose that $\alpha : V \to V$ is a linear map with $m_\alpha(t) = (t-5)^2$ 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.

End of Algebra 2B.
