MA20217 Algebra2B, February-May 2024

G.K. Sankaran

May 10, 2024

What this course is about

This course introduces abstract algebra, starting with group theory and then moving on to rings.

Here is a non-exhaustive list of conventions for these notes:

- $\mathbb{N}$ denotes the set of non-negative integers, i.e. including 0.
- Unless otherwise stated, group multiplication will be denoted by juxtaposition, so gh means g multiplied by h (on the right). If ϕ and ψ are maps, $\phi\psi$ will mean the composition ψ followed by ϕ. Sometimes both these conventions will apply at once, i.e. the group multiplication will actually be composition.
- The integers mod n are denoted $\mathbb{Z}/n$ (whether we think of them as an additive group or as a ring). The notation $\mathbb{Z}_n$ will not be used.

I Basic theory of groups

In this section we introduce the basic definitions and axioms of groups, and see some examples and consequences.

Groups and subgroups

I.1 One way to look at a group – not the only way, and not always the best way – is to say that it is a set where you are allowed to multiply elements together. There is more to it than that, but that's a starting point.

The corresponding basic informal idea of a ring is that in a ring you are allowed to multiply and you are allowed to add. Again, there's more to it than that.

A field is a special kind of ring: multiplication is commutative and you are allowed to divide by anything that isn't actually zero. Examples of fields include

$\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ and $\mathbb{Z}/p$ (also called $\mathbb{F}_p$) for p a prime. In a few places we shall mention fields before we have defined them precisely: these examples will suffice.

Groups are more basic than rings (it is convenient to be able to use the word "group" when defining a ring) so we begin by looking at groups first.

I.2 Groups are there to capture in a mathematical way the idea of symmetry. So we are going to use this to guess what the axioms of a group should be.

I.3 The general idea of a group is this (a more formal definition will follow). We have a set G and a way of "multiplying" together two elements a and b of G to get an element $ab \in G$. (If we imagine that a and b are symmetries of something, this says "do b and then do a".) We want abc to make sense, without having to write brackets, but quite possibly $ab \neq ba$: again, that is what we expect from thinking about symmetry. There is also a symmetry that says "leave it alone", and so we want a neutral element e with the property that $ae = ea = a$ for every $a \in G$. And finally, if we have done the symmetry a and then we wish we hadn't, we can undo it with another symmetry (or just possibly the same one).

Definition I.4 A *group* is a set G equipped with a multiplication such that $ab \in G$ if a, $b \in G$ and such that

(a) multiplication is associative: $(ab)c = a(bc)$ for all a, b, $c \in G$;

(b) there is an identity element $e \in G$ such that $ae = ea = a$ for all $a \in G$;

(c) for any $a \in G$ there exists an element $a^{-1} \in G$ such that $aa^{-1} = a^{-1}a = e$.

I.5 There is nothing wrong with Definition I.4 but it uses the word "multiplication" without explaining it. Also, we want to think about symmetries but you don't talk about multiplying symmetries. Without forgetting or abandoning Definition I.4, let us look at it in a slightly more precise way.

Definition I.6 A *binary operation* on a set S is a function

$$*\colon S \times S \to S.$$

Binary operations that we are interested in tend to be called addition, denoted by $+$, or multiplication, denoted by either $\cdot$ or nothing at all. In this context, by ancient tradition one writes $a * b$ rather than $*(a, b)$ (Polish notation): thus $7 + 6$ rather than $+(7, 6)$ and $a \cdot b$ or ab rather than $\cdot(a, b)$.

Definition I.7 A *group* is a pair $(G, *)$, where G is a set, $*$ is a binary operation on G and the following axioms hold:

(a) **(associative)** $(a * b) * c = a * (b * c)$ for all a, b, $c \in G$.

(b) **(identity)** There is an element $e \in G$ with the property that $e * a = a = a * e$ for all $a \in G$.

(c) **(inverse)** For each $a \in G$ there exists $b \in G$ such that $a * b = b * a = e$.

I.8 Often, in fact usually, we know what the group operation $*$ is, and then we don't need to mention it and we just refer to the group as G rather as $(G, *)$.

Notice that Definition I.7 says exactly the same thing as Definition I.4, but without saying "multiplication".

I.9 In a group, both the identity element and inverses are unique.

(i) if $e, e' \in G$ are two elements satisfying the identity property from I.4, then $e * e' = e'$ because e is an identity and $e * e' = e$ because e' is. So $e = e'$.

(ii) Given $a \in G$, if $b, b' \in G$ are both elements satisfying the inverse property for a in I.4, then

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

This unique element b from (ii) is called the *inverse* of a. It is often denoted by a^{-1}, especially if $*$ has been denoted by $\cdot$ or by nothing, in which case the identity element e is often denoted by 1, or by 1_G if we need to distinguish it from the integer 1.

I.10 A neat way to say this is that a group has three operations: there is a binary operation (usually called *the* group operation), $*$, which takes two elements of G and gives you a third one; there is a unary operation, $^{-1}$, which takes one element of G and gives you a second one; and there is a nullary operation, e, which takes no element of G and gives you a first one. Then there are the group axioms, which tells you how these operations interact.

Definition I.11 A group $(G, *)$ is *abelian* (or *commutative*) 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 $-a$.

Example I.12 We already know lots of examples of groups, and of structures that are not groups.

(i) $(\mathbb{Z}, +)$, the ordinary integers with addition, is a group (in fact an abelian group).

(ii) $(\mathbb{Z}, \cdot)$, the ordinary integers with multiplication, is *not* a group because inverses are missing. Most obviously, 0^{-1} is missing, but so is 2^{-1}: you can undo multiplication by 2 by multiplying by $\frac{1}{2}$, but $\frac{1}{2} \notin \mathbb{Z}$.

(iii) $(\mathbb{C}, +)$ is a group: so are $(\mathbb{R}, +)$ and $(\mathbb{Q}, +)$.

(iv) $(\mathbb{C}^*, \cdot)$, the non-zero complex numbers with multiplication, is a group. So are $(\mathbb{R}^*, \cdot)$ and $(\mathbb{Q}^*, \cdot)$, but not $(\mathbb{C}, \cdot)$ etc., because of 0 not having an inverse.

(v) $(\mathbb{Z}/n, +)$, the integers modulo n with addition (modulo n) is a group for any $n > 0$. This is sometimes called the cyclic group C_n.

(vi) $(\mathbb{Z}/n, \cdot)$ is not a group, even if n is prime, but if p is prime and $(\mathbb{Z}/p)^*$ denotes the non-zero elements of $\mathbb{Z}/p$ then $((\mathbb{Z}/p)^*, \cdot)$ is a group. If n is not prime we can still get a group but we have to take $(\mathbb{Z}/n)^*$ to be the set of elements that are coprime to n – we should be a bit careful about what this means.

(vii) $(\mathrm{GL}(2, \mathbb{C}), \cdot)$, the set of 2×2 matrices with complex entries and non-zero determinant, and matrix multiplication, is a group. It is the first non-abelian group we have seen.

(viii) $(M_{2\times 2}(\mathbb{C}), +)$, the set of 2×2 matrices with complex entries, and matrix addition, is a group.

(ix) D_{2n}, the group of symmetries of a $2n$-gon, is a group with $2n$ elements (the group operation is composition: ab means do symmetry b followed by symmetry a). It is non-abelian.

(x) S_n, the symmetric (permutation) group on n symbols (i.e. the symmetries of a set of size n, is a group. (We could allow n to be an infinite cardinal here but we won't.) It is non-abelian.

(xi) A_n, the even permutations of n symbols, form a group, also non-abelian.

(xii) The set of continuous bijections from the circle to itself is a group, called the group of homeomorphisms of S^1. This is a very big group: too big for use really.

(xiii) If X is any set then the set of bijective maps $f\colon X \to X$ forms a group called the symmetry group or automorphism group or permutation group $\mathrm{Sym}(X)$ of X. This is also uncomfortably big unless X is finite, in which case it is just S_n again.

Definition I.13 A subset H of a group G is called a *subgroup* of G if and only if it is closed under all the group operations. That means that $1 \in H$ (so, in particular, $H \neq \varnothing$); that $a^{-1} \in H$ if $a \in H$; and that $ab \in H$ if $a, b \in H$.

In other words, H is a subgroup if it is a subset of G that is also a group (with the same binary operation).

In this case, we write $H < G$. If $H < G$ and $H \neq G$ we say that H is a *proper subgroup* of G.

Lemma I.14 Suppose that H is non-empty subset of G. Then H is a subgroup of G if and only if

$$\text{for all } a, b \in H, \text{ we have } ab^{-1} \in H.$$

Proof: One implication is immediate from Definition I.13: if H is a subgroup and $a, b \in H$ then $b^{-1} \in H$ and hence $ab^{-1} \in H$, by I.4.

Conversely, suppose we have this condition. First choose $a \in H$ (so we do need to know that $H \neq \varnothing$) and take $b = a$. That tells us that $1 \in H$, so now we can start again and take $a = 1$. That tells us that if $b \in H$ then $b^{-1} \in H$. Finally, if we want to show that $ab \in H$ we take $c = b^{-1}$ and then the criterion tells us that $H \ni ac^{-1} = ab$.

Example I.15 We also know many examples of subgroups, and of things that are not subgroups.

(i) $A_n < S_n$: the product and inverse of even permutations are even.

(ii) $2\mathbb{Z} < \mathbb{Z}$: even integers are integers; there are some; and the difference of even integers is even (we are using Lemma I.14 here).

(iii) $\mathbb{R}_{>0}$, the positive real numbers with multiplication, form a subgroup of $\mathbb{R}^*$.

(iv) S^1, the set of complex numbers of modulus 1, is a subgroup of $\mathbb{C}^*$. (This group goes by many different names.)

(v) $\mathrm{SL}(2,\mathbb{C}) < \mathrm{GL}(2,\mathbb{C})$: it is by definition the set of 2×2 matrices of determinant 1.

(vi) $\mathbb{Z}/n$ is *not* a subgroup of $\mathbb{Z}$: this is important. It is not even a subset of $\mathbb{Z}$, in fact, but more than that: there isn't even a subgroup of $\mathbb{Z}$ that "looks like" $\mathbb{Z}/n$ (except in the trivial case $n = 1$). We shall make this more precise later: see Definition I.32.

(vii) $\mathbb{Z}/2$ a subgroup of $\mathbb{Z}/6$?

(viii) Indisputably, $6\mathbb{Z} < 2\mathbb{Z}$: numbers divisible by 6 are even.

(ix) There is a subgroup R of D_6 that consists of 1 and a reflection in a (single, fixed) line of symmetry.

Definition I.16 A subset H of a group G is called a *normal subgroup* of G if it is a subgroup and

$$g^{-1}hg \in H \text{ whenever } h \in H \text{ and } g \in G.$$

In this case, we write $H \lhd G$.

Example I.17 Again we know many examples.

(i) If G is an abelian group and $H < G$ then $H \lhd G$.

(ii) $\mathrm{SL}(2,\mathbb{C}) \lhd \mathrm{GL}(2,\mathbb{C})$, because if $\det A = 1$ (so that $A \in \mathrm{SL}(2,\mathbb{C})$) and $B \in \mathrm{GL}(2,\mathbb{C})$, then $\det(B^{-1}AB) = \det(B^{-1}) \det A \det B = \det(B^{-1}) \det B = 1$, and therefore $B^{-1}AB \in \mathrm{SL}(2,\mathbb{C})$.

(iii) The subgroup of D_6 consisting of rotations of the triangle in the plane (three of them, one trivial) is a normal subgroup of D_6. However, the subgroup R of D_6 from Example I.15(ix) is not a normal subgroup.

I.18 It is not immediately obvious why Definition I.16 is important, but it is. In fact, non-normal subgroups are very common. It is not normal for a subgroup to be normal.

Maps between groups

I.19 We do not only want to talk about one group at a time. We want to be able to compare them, and that means having maps that go from one group to another. These maps could just be maps of sets, of course, but they you lose all the group information. We want to keep that.

Definition I.20 If G and H are groups then a map $\varphi\colon G \to H$ is called a *group homomorphism* if

(a) $\varphi(1_G) = 1_H$;

(b) if $g \in G$ then $\varphi(g^{-1}) = (\varphi(g))^{-1}$;

(c) if g_1, $g_2 \in G$ then $\varphi(g_1 g_2) = \varphi(g_1)\varphi(g_2)$.

I.21 What Definition I.20 says is that φ is a group homomorphism if it respects all the group operations. In effect, what this means is that the result of doing a calculation in G and sending the answer to H by using φ is the same as the result of sending all the components to H by φ and doing the corresponding computation in H.

I.22 A very similar definition is the definition of linear map for vector spaces. The only difference is that in groups you can multiply and take inverses, and a group homomorphism respects those operations; in a vector space you are allowed to add and to multiply by scalars, and a linear map respects those operations. A linear map is just a vector space homomorphism.

Proposition I.23 Let G and H be groups and $\varphi\colon G \to H$ a map. Then φ is a group homomorphism if and only if $\varphi(ab) = \varphi(a)\varphi(b)$ for all a, $b \in G$, i.e. if and only if Definition I.20(iii) holds.

Proof: Suppose we know Definition I.20(iii). Let us first prove that Definition I.20(i) holds. We have

$$\varphi(a) = \varphi(1_G a) = \varphi(1_G)\varphi(a)$$

and if we multiply on the right by $\varphi(a)^{-1}$ we get

$$1_H = \varphi(a)\varphi(a)^{-1} = \varphi(1_G)\varphi(a)\varphi(a)^{-1} = \varphi(1_G)$$

as required. Now we can prove Definition I.20(ii) because

$$1_H = \varphi(1_G) = \varphi(aa^{-1}) = \varphi(a)\varphi(a^{-1})$$

and now multiplying on the left by $\varphi(a)^{-1}$ gives us what we want.

I.24 Proposition I.23 is very convenient and sometimes it is made the definition. If we know we are talking about groups, we often just say "homomorphism".

Definition I.25 If $\varphi\colon G \to H$ is a group homomorphism then

(a) the *image* of φ is $\{\varphi(g) \mid g \in G\}$, a subset of H;

(b) the *kernel* of φ is $\{g \in G \mid \varphi(g) = 1_H\}$, a subset of G.

Lemma I.26 If $\varphi\colon G \to H$ is a group homomorphism then the image $\operatorname{Im}\varphi = \varphi(G)$ is a subgroup of H.

Proof: Exercise.

Lemma I.27 If $\varphi\colon G \to H$ is a group homomorphism then the kernel $\operatorname{Ker}\varphi$ is a normal subgroup of G.

Proof: Let us write $K = \operatorname{Ker}\varphi$ and suppose that $g \in G$ and k, k_1, $k_2 \in K$.

First, according to Lemma I.14, we can check that K is a subgroup by checking that $k_1k_2^{-1} \in K$. But for this it is enough to compute

$$\varphi(k_1k_2^{-1}) = \varphi(k_1)\varphi(k_2^{-1}) = \varphi(k_1)\varphi(k_2)^{-1} = 1_H.$$

Then we want to check that K is a normal subgroup. But $\varphi(g^{-1}) = \varphi(g)^{-1}$ and $\varphi(k) = 1_H$, so

$$\varphi(g^{-1}kg) = \varphi(g^{-1})\varphi(k)\varphi(g) = \varphi(g)^{-1} \cdot 1_H \cdot \varphi(g) = \varphi(g)^{-1}\varphi(g) = 1_H$$

so $g^{-1}kg \in K$ if $k \in K$, as required.

Lemma I.28 A group homomorphism $\varphi\colon G \to H$ is injective if and only if $\operatorname{Ker}\varphi = \{1_G\}$.

Proof: Exercise.

Definition I.29 Let G and G' be groups. A homomorphism $\varphi\colon G \to G'$ is called an *isomorphism* if there is a homomorphism $\psi\colon G' \to G$ such that $\psi\varphi\colon G \to G$ is the identity on G and $\varphi\psi\colon G' \to G'$ is the identity on G'.

I.30 Notice that this definition is symmetric: $\psi\colon G' \to G$ is also an isomorphism.

The important part of the definition is the requirement that ψ is a homomorphism. If we drop that and just allow ψ to be any map, then the rest of the definition just says that φ is bijective and ψ is its two-sided inverse map. So if we are given φ we have no choice about what map ψ is: either φ is not bijective, in which case it cannot be an isomorphism because ψ doesn't exist at all, or it is, in which case ψ is its inverse. But the definition also insists that ψ should be a homomorphism.

In fact this is no extra requirement in this case. But beware: that is true about groups, but this definition generalises, and in some of the other cases we cannot be sure that the inverse map preserves all the structure that we want to preserve.

Proposition I.31 Let $\varphi\colon G \to G'$ be a bijective group homomorphism. Then φ is an isomorphism.

Proof: The only choice for ψ is the inverse map to φ: if $\varphi(a) = b$, then $\psi(b) = a$. This is defined for every b because φ is surjective, and a is unique because φ is injective. It satisfies $\varphi\psi = \mathrm{id}_{G'}$ and $\psi\varphi = \mathrm{id}_G$.

The task, then, is to show that ψ, thus defined, is a group homomorphism. According to Proposition I.23, it is enough to show that if b_1, $b_2 \in G'$ then $\psi(b_1 b_2) = \psi(b_1)\psi(b_2)$.

But $\varphi\psi(b_1 b_2) = b_1 b_2$, and $\varphi(\psi(b_1)\psi(b_2)) = \varphi\psi(b_1)\varphi\psi(b_2) = b_1 b_2$, where the first equality holds because φ is a homomorphism. So $\psi(b_1 b_2)$ and $\psi(b_1)\psi(b_2)$ are mapped to the same element by φ, so they must be equal because φ is injective.

Definition I.32 Two groups G and H are said to be *isomorphic* if there exists an isomorphism $\varphi\colon G \to H$: we write $G \cong H$ in this case.

I.33 Isomorphism – that is, the relation of being isomorphic – is an equivalence relation among groups. Sometimes one thinks of isomorphic groups as being "the same", but sometimes that is the wrong thing to do. The difficulty is that the definition of "isomorphic" only promises that an isomorphism exists: it doesn't supply it, the isomorphism may not be unique (in particular, an isomorphism from G to G does not have to be the identity) and even if you know φ you may not know ψ.

Much of cryptography rests on this last possibility. Alice knows both halves of an isomorphism between isomorphic groups G and H but she publishes only G, H and φ. She tells Bob to encode his message as an element of G, and (perhaps after some computations) apply φ and send her the result. Eve knows that to find the message all she needs to do is apply ψ, but she doesn't know ψ: that's Alice's secret, and it is computationally hard to work out ψ from φ even though in principle φ does determine ψ.

Order, generators and cyclic groups

Definition I.34 The *order of a group* is its cardinality: that is, the order of G is $|G|$.

I.35 We will usually only use this if G is finite: otherwise we shall say that G is infinite or that $|G| = \infty$. We could also distinguish different cardinalities among infinite groups. Thus we shall say that $|\mathbb{Z}| = \infty$ and $|\mathbb{R}| = \infty$, simply meaning that neither is finite, although in fact $|\mathbb{Z}| \neq |\mathbb{R}|$.

Definition I.36 The *order of an element* $g \in G$ is the smallest natural number n such that $g^n = 1_G$, or ∞ is there is no such number. We write $o(g)$ for the order of g.

I.37 It is important to note that Definition I.34 and Definition I.36 are entirely different: notice the different notation for order of an element and order of a group. In particular, the order of a group is ascertained by counting: it is a set-theoretic property. On the other hand, to determine the order of an element one must know how to multiply, so it it is a group-theoretic property. The use of the same word is not unintentional, though: they are related, as we shall see in Proposition I.46.

Lemma I.38 Let G be a group and $g \in G$. Then $o(g) = 1$ if and only if $g = 1_G$.

This is a trivial observation, but a very useful one.

Definition I.39 Suppose that G is a group and $S \subset G$ (note that this means a subset, not a subgroup). We say that S *generates* G if no proper subgroup of G contains S.

I.40 There are many ways to rephrase this. One is to say that G is the smallest subgroup of G that contains S. This is all right, but then we need to ask what is meant by "the smallest". Another way is via the following proposition, which as well as being practically useful also gives a different way of thinking about what Definition I.39 means.

Proposition I.41 Suppose that G is a group and $S \subset G$. Then S generates G if and only if every $g \in G$ can be written in the form

$$g = s_1 \dots s_k$$

for some $k \in \mathbb{N}$, where $s_i \in S$ or $s_i^{-1} \in S$ for every i.

Proof: Consider the set $\langle S \rangle = \{s_1 \dots s_k \mid k \in \mathbb{N},\ s_i \in S \text{ or } s_i^{-1} \in S\}$. This clearly includes S: we claim that it is a group. We use Lemma I.14, noticing that $\langle S \rangle$ is not empty, even if S is empty, because if we take $k = 0$ we get the empty product, which is 1 by definition. (If $S = \varnothing$ there is nothing to check anyway.) So we just need to check that if g, $h \in \langle S \rangle$ then $gh^{-1} \in \langle S \rangle$. But this is immediate: if $g = s_1 \dots s_k$ and $h = r_1 \dots r_l$ with s_i or s_i^{-1} in S and also r_j or r_j^{-1} in S for each i and j, then

$$gh^{-1} = s_1 \dots s_k r_l^{-1} \dots r_1^{-1}$$

which obviously satisfies the same condition.

So the statement that every $g \in G$ can be written in this form is the same as saying that $\langle S \rangle = G$. So we want to show that no proper subgroup of G contains S if and only if $G = \langle S \rangle$.

If $G \neq \langle S \rangle$ then $\langle S \rangle$ is a proper subgroup of G containing S, so S does not generate G.

Conversely, if S does not generate G, then there is a proper subgroup H of G containing S, but then $\langle S \rangle < H$ because H is closed under the group operations, and so $\langle S \rangle \neq G$.

Definition I.42 The subgroup $\langle S \rangle$ in the proof of Proposition I.41 is called the *subgroup generated by* S. We shall use this notation, and if $S = \{g\}$ consists of a single element we shall usually write $\langle g \rangle$ instead of $\langle \{g\} \rangle$. Notice that $\langle g \rangle = \{g^i \mid i \in \mathbb{Z}\}$.

Example I.43 We already know many examples of this.

(i) If G is a group, $\varnothing$ generates the trivial subgroup 1. So does $S = \{1_G\}$.

(ii) If $G = \mathbb{Z}$ (with addition, of course) then $\{1\}$, which is not the identity element, generates $\mathbb{Z}$.

(iii) If $G = \mathbb{Z}/p$ with p a prime then any non-identity element of G generates G.

(iv) If $G = \mathbb{Z}/6$ then $\{1\}$ generates G but $\{2\}$ does not.

(v) If $G = D_6$ or S_3 (in fact $D_6 \cong S_3$) then there is no element $g \in G$ such that $\{g\}$ generates G.

Definition I.44 We say that a group G is *cyclic* if $G = \langle g \rangle$ for some $g \in G$.

I.45 Cyclic groups are abelian. We know all the cyclic groups up to isomorphism: they are $\mathbb{Z}/n$ for $n > 0$, and $\mathbb{Z}$.

Proposition I.46 If $g \in G$ then $|\langle g \rangle| = o(g)$: that is, the order of an element is the same as the order of the (cyclic) group that it generates.

Proof: Suppose that $o(g) = r \in \mathbb{N}$ (if the order is infinite there is nothing to prove) and consider the set $R := \{g^i \mid 0 < i \leq r\}$. Recall that $\langle g \rangle = \{g^i \mid i \in \mathbb{Z}\}$. Obviously $R \subseteq \langle g \rangle$. We claim that $R = \langle g \rangle$ and that $|R| = r$.

For the first, it is enough to note that if $i \in \mathbb{Z}$ and $i \equiv j \bmod r$ then $g^i = g^j$, because we have $i = j + kr$ for some $k \in \mathbb{Z}$ so $g^i = g^{j+kr} = g^j g^{kr} = g^j g^{rk} = g^j 1^k = g^j$.

For the second we know by Definition I.36 that $g^r = 1$ and $g^s \neq 1$ if $1 < s < r$ (if $r = 1$ there is nothing to prove). If $1 \leq s < s' \leq r$ and $g^s = g^{s'}$ then $1 = g^{s'}(g^s)^{-1} = g^{s'-s}$ which is impossible as $1 < s' - s < r$ so the r elements $g, g^2, \ldots, g^r = 1$ are all different so $|R| = r$.

II Structure of groups

In this section we shall prove two major theorems about the structure of groups: Lagrange's theorem, which is about the orders of finite groups, and the (first) isomorphism theorem, which describes group homomorphisms.

Cosets and Lagrange's theorem

Definition II.1 Suppose that G is a group and H is a subgroup of G. A *left coset* of H in G is a subset of G of the form

$$gH := \{gh \mid h \in H\}$$

where $g \in G$ is fixed. A *right coset* Hg is defined similarly.

II.2 In general, cosets are not subgroups: for one thing, if $g \notin H$ then $1_G \notin g^{-1}H$. They are analogous to affine linear subspaces, e.g. planes in $\mathbb{R}^3$ that don't go through the origin. Those aren't vector subspaces but we are still allowed to think about them.

Notice that gH can be equal to $g'H$ even if $g \neq g'$. For example, if $g' = 1$ and $g \in H$ then both of these are equal to H.

Lemma II.3 If $H < G$ and $g \in G$ then $|gH| = |H| = |Hg|$.

Proof: The map $\mu_g \colon H \to gH$ given by $\mu_g(h) = gh$ is bijective, because it has the two-sided inverse map $\mu_{g^{-1}}$.

Theorem II.4 If $H < G$ then every element of G belongs to exactly one left coset of H. In particular if g_1H and g_2H are left cosets, then either $g_1H = g_2H$ or $g_1H \cap g_2H = \varnothing$.

Proof: If $g \in G$ then $g = g1 \in gH$, so every $g \in G$ is contained in at least one coset. Suppose that $g \in g'H$: we claim that $g'H = gH$, so that gH is the only coset that contains g. To check this, first we write $g = g'h$ (since $g \in g'H$ this is true for some $h \in H$) and then we consider an arbitrary element $g'h' \in g'H$. We have $g' = gh^{-1}$, so $g'h' = gh^{-1}h'$ and that is an element of gH because H is a group so $h^{-1}h' \in H$. So $g'H \subseteq gH$. Similarly, if $gh' \in gH$ then $gh' = g'hh' \in g'H$ so $gH \subseteq g'H$.

II.5 Another way to state Theorem II.4 is to say that the cosets *partition* G: that is, $G = \coprod_g gH$ where g runs over a set of representatives, one from each left coset of H.

A third way to state this is to say that the relation $g_1 \sim g_2 \iff g_1H = g_2H$ is an equivalence relation.

Corollary II.6 If $g_1, g_2 \in G$ and $H < G$ then $g_1H = g_2H$ if and only if $g_1 \in g_2H$.

Definition II.7 The number (possibly infinite) of different cosets of H in G is called the *index* of H in G, written $|G : H|$. That is

$$|G : H| = |\{gH \mid g \in G\}|.$$

It is easy to check that if we look at right cosets instead of left cosets the number of cosets is the same, even though the cosets themselves are not ($gH \neq Hg$ in general, unless G is abelian).

Theorem II.8 [Lagrange's Theorem] If G is a finite group and H is a subgroup of G, then $|H|$ divides $|G|$. In fact, $|G| = |H| \cdot |G : H|$, so the index of H also divides $|G|$.

Proof: By Theorem II.4, H has $r := |G : H|$ different left cosets, $g_1H, \ldots, g_rH$, and $G = g_1H \coprod \ldots \coprod g_rH$. By Lemma II.3 we then have

$$|G| = \sum_{i=1}^{r} |g_iH| = \sum_{i=1}^{r} |H| = |G : H| \cdot |H|.$$

Corollary II.9 If G is a finite group and $g \in G$ then $o(g)$ divides $|G|$.

Proof: By Proposition I.46, $o(g) = |\langle g \rangle|$ and the latter divides $|G|$ by Lagrange's theorem.

II.10 Lagrange's theorem is extremely important in both algebra and number theory. It is one of the few ways of showing that one number divides another without actually knowing what numbers they are.

The converse to Lagrange's theorem is false: it is possible for r to divide $|G|$ without G having a subgroup of order exactly r. In particular the converse to Corollary II.9 is false. Both converse statements are true, however, if $r = p$ is prime: this is a consequence of the Sylow theorems, one of which is a partial converse to Lagrange's theorem. The Sylow theorems will not be proved or even stated in this course, but they are not extremely hard.

Quotient groups

II.11 We have seen that in general the left cosets and right cosets of a subgroup $H < G$ are not the same thing, except if G is an abelian group. There is another important case where this happens, which we have already met: if H is a normal subgroup of G (Definition I.16).

Lemma II.12 Suppose G is a group and $H < G$. Then H is a normal subgroup if and only if every left coset is also a right coset.

Proof: If the left coset gH is also a right coset then it must be the right coset Hg, because $g \in gH$ and $g \in Hg$ and g belongs to exactly one coset of each handedness by Theorem II.4. So if that is true for every $g \in G$ then $gH = Hg$ for all $g \in G$. Now if $h \in H$ we have $hg \in Hg = gH$ so $hg = gh'$ for some $h' \in H$, so $g^{-1}hg \in H$ so $H \lhd G$ according to Definition I.16.

Conversely, if $H \lhd G$ and gH is a left coset, then I claim that $gH = Hg$. But if $hg \in Hg$ then $hg = gg^{-1}hg = gh' \in H$ for some $h' \in H$, by Definition I.16, so $Hg \subseteq gH$. The other inclusion follows by symmetry, or by replacing g by g^{-1} (or immediately if G is finite).

II.13 The importance of Lemma II.12 is that it allows us to put a group structure on the set of cosets of H in G, but only if H is a normal subgroup.

We first make a definition that is valid even if H is not normal, but gives us only a set, not a group.

Definition II.14 Let G be a group and H a subgroup. The *coset set* of G with respect to H is the set G/H given by

$$G/H := \{gH \mid g \in G\}.$$

II.15 Of course we can also define the coset set $H\backslash G$ of right cosets, but we do not need it: notice, however, that it will be the same as the coset set defined above if H is a normal subgroup. Notice also that $|G/H| = |G : H|$. In fact some people prefer to use $G : H$ for the set defined in Definition II.14 and reserve G/H for the group that we are about to produce.

Theorem II.16 Suppose that H is a normal subgroup of G. Then we may define a multiplication on G/H by

$$(g_1H)(g_2H) = (g_1g_2)H.$$

Moreover, with this multiplication G/H is a group, in which $e = H$ and $(gH)^{-1} = g^{-1}H$.

Proof: The main thing that has to be proved here is that the definition of multiplication makes sense. We need to check that if $g_1'H = g_1H$ and $g_2'H = g_2H$ then $g_1'g_2'H = g_1g_2H$.

By Corollary II.6, $g_1'H = g_1H$ if and only if $g_1' \in g_1H$, so $g_1' = g_1h_1$, so we assume this and also that $g_2' = g_2h_2$, for h_1, $h_2 \in H$. Now we need to show only that $g_1'g_2' \in g_1g_2H$, again by Corollary II.6. But $g_1'g_2' = g_1h_1g_2h_2$ and $h_1g_2 \in Hg_2 = g_2H$ (since H is normal in G). So $h_1g_2 = g_2h_1'$ for some $h_1' \in H$, so $g_1'g_2' = g_1g_2h_1'h_2 \in g_1g_2H$.

The rest is very similar. Notice that we do not have to check associativity: that is part of the definition. H is the identity because $H = 1H$ and $(1H)(gH) = (1g)H = gH$, and $g^{-1}H$ is the inverse of gH because $(g^{-1}H)(gH) = (g^{-1}g)H = 1H = H$. We do, though, have to check that $g^{-1}H$ is well defined: that is, that if $g_1H = g_2H$ then $g_1^{-1}H = g_2^{-1}H$. But $g_1H = g_2H$ if and only if $g_2 \in g_1H$, so $g_2 = g_1h$ for some $h \in H$: that happens if and only if $g_2^{-1} = h^{-1}g_1^{-1}$ for some h (that is, some h^{-1}) in H: but that says $g_2^{-1} \in Hg_1^{-1} = g_1^{-1}H$ as H is normal. So $g_2^{-1}H = g_1^{-1}H$ by Corollary II.6.

Definition II.17 The group G/H, with the operations (multiplication and inverse) defined in Theorem II.16, is called the *quotient group* of G by H.

II.18 This is one of the most important concepts in all of mathematics. There is plenty to think about.

One important thing to understand is that G/H is not a subgroup of G. (It might happen to be isomorphic to a subgroup of G, although it also might not, but that is a different thing.)

What we are doing in this construction is pretending that everything in H is the identity. For example, we might do that because we are not interested in H, or because H is acting trivially (see Example III.7(ix)). If you pretend that all elements of H are trivial, then you also have to think that any two elements of gH are the same, and that is exactly what this construction tells you to do. That also explains where the definition of normal subgroup comes from. If you think, however wrongly, that h_1 and h_2 are the identity, then you also have to think that $h_1 h_2^{-1}$ is the identity, but you also have to think that gh_1g^{-1} is the identity, whether $g \in H$ or not. We shall make this idea a bit more precise shortly.

II.19 If G is a group and H is a (proper, nontrivial) normal subgroup then one thinks of G/H and H as both being smaller that G: if G is finite, this is literally true. So we could try to understand finite groups, at least, by looking for normal subgroups inside them and trying to proceed by induction, understanding the smaller "pieces" G/H and H. There is still a question about how the pieces fit together. This stops working if G simply does not have any normal subgroups, apart from 1 and G. Such a group is called *simple*: they are the building blocks of all finite groups.

One of the major triumphs of 20th-century algebra was the classification, around 1980, of all the finite simple groups. A big first step was the Thompson-Feit theorem from the early 1960s, which says that the only simple groups of odd order are the cyclic groups $C_p \cong \mathbb{Z}/p$ with p an odd prime. Of course $\mathbb{Z}/2$ is also simple. Apart from those, the alternating group A_n is simple if $n \geq 5$, and there are families of finite simple groups called "of Lie type": these are (more or less) groups of matrices over a finite field, such as $\mathrm{PSL}(n, \mathbb{F}_p)$. And then there are twenty-six others, called the sporadic groups: most of them occur as subquotients, i.e. quotients of subgroups, of the biggest of the 26 sporadic groups, the Monster. The order of the Monster is

$$2^{46} \cdot 3^{20} \cdot 5^9 \cdot 7^6 \cdot 11^2 \cdot 13^3 \cdot 17 \cdot 19 \cdot 23 \cdot 29 \cdot 31 \cdot 41 \cdot 47 \cdot 59 \cdot 71.$$

That's about the mass of the observable universe in grams. (Dark matter not included. Other restrictions may apply.)

Example II.20 We already know a few examples of quotient groups.

(i) $n\mathbb{Z} \lhd \mathbb{Z}$ since the additive group $\mathbb{Z}$ is abelian, and $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/n$. This is the basic example. When you do addition modulo n, you are treating n as 0 (which is the identity) and when you talk about "6 mod 11" what you really mean is "any number that is 6 mod 11" or, what amounts to the same thing in that it carries the same information, "the class of numbers that are 6 mod 11". That's the coset $6 + 11\mathbb{Z}$ (additive notation). Notice that $\mathbb{Z}/n\mathbb{Z}$ is definitely not isomorphic to any subgroup of $\mathbb{Z}$, because every element in $\mathbb{Z}/n\mathbb{Z}$ has finite order but no element of $\mathbb{Z}$ apart from 0 has finite order.

(ii) There are two trivial cases: G/G is the trivial group $\{1\}$ and $G/\langle 1_G \rangle$ is G (strictly speaking "is naturally isomorphic to G" but even algebraists won't worry about that).

(iii) The order 3 subgroup of rotations in D_6 is normal, and the quotient is (isomorphic to) $\mathbb{Z}/2$, telling you which side of the triangle is facing you.

(iv) More generally, $A_n \lhd S_n$ with quotient $\mathbb{Z}/2$.

(v) If G is a group and $H < G$ is of index 2, then $H \lhd G$ and $G/H \cong \mathbb{Z}/2$.

(vi) The group $\mathrm{Aff}(\mathbb{R}^2)$ of affine linear transformations of a plane, $\mathbf{x} \mapsto M\mathbf{x}+\mathbf{a}$ where $M \in \mathrm{GL}(2,\mathbb{R})$ and $\mathbf{a} \in \mathbb{R}^2$, has a normal subgroup consisting of the translations. The quotient is isomorphic to $\mathrm{GL}(2,\mathbb{R})$, but the same thing cannot be done in the other direction because $\mathrm{GL}(2,\mathbb{R})$ is not a normal subgroup of $\mathrm{Aff}(\mathbb{R}^2)$.

Lemma II.21 If G is a group and H is a normal subgroup of G, then the map $\pi\colon G \to G/H$ given by $\pi(g) = gH$ is a surjective group homomorphism whose kernel is H. In particular, every normal subgroup is the kernel of a group homomorphism.

Proof: There is actually nothing to prove. π is surjective by the definition of G/H as a set, Definition II.14, and it is a group homomorphism by Theorem II.16: the definition of multiplication there exactly matches what is needed for π to be a group homomorphism according to Proposition I.23. The statement about the kernel is simply the statement that $gH = H$ if and only if $g \in H$, which is a special case of Corollary II.6.

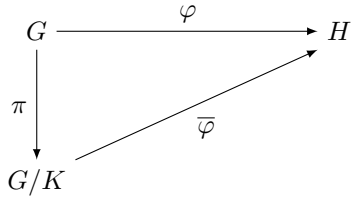
II.22 Notice that Lemma II.21 and Lemma I.27 between them say that kernels and normal subgroups are exactly the same thing.

In vector spaces, homomorphisms are linear maps and kernels are null spaces. Any linear subspace can be the null space of something, and null spaces are linear subspaces. The difference in the case of groups is that only some subgroups, the normal subgroups, are eligible to be kernels.

Definition II.23 Let G be a group and H a subgroup. If $H \lhd G$ then the map $\pi\colon G \to G/H$ defined in Lemma II.21 is called the *quotient map*. The map $\iota\colon H \to G$ defined by $\iota(h) = h$ is called the *inclusion map*.

It is obvious that ι is an injective group homomorphism.

Theorem II.24 Let $\varphi\colon G \to H$ be a group homomorphism and let K be a normal subgroup in G satisfying $K \subseteq \operatorname{Ker}\varphi$. Then there exists a unique group homomorphism $\overline{\varphi}\colon G/K \to H$ such that the diagram



commutes, i.e., $\overline{\varphi} \circ \pi = \varphi$.

Proof: The map $\overline{\varphi}\colon G/K \to H$ is defined by setting $\overline{\varphi}\big(gK\big) = \varphi(g)$. To see that this map is well-defined, independent of any choices, notice that

$$\begin{aligned} g_1K = g_2K &\iff g_1^{-1}g_2 \in K \\ &\Rightarrow\ g_1^{-1}g_2 \in \operatorname{Ker}(\varphi) \\ &\iff 1_H = \varphi(g_1^{-1}g_2) = \varphi(g_1)^{-1}\varphi(g_2) \\ &\iff \varphi(g_1) = \varphi(g_2). \end{aligned}$$

In particular, $g_1K = g_2K \Rightarrow \varphi(g_1) = \varphi(g_2)$, so $\overline{\varphi}$ does not depend on the choice of representative g in the coset gK.

To see that $\overline{\varphi}$ is a group homomorphism, notice that

$$\overline{\varphi}((g_1K)(g_2K)) = \overline{\varphi}(g_1g_2K) = \varphi(g_1g_2) = \varphi(g_1)\varphi(g_2) = \overline{\varphi}(g_1K)\overline{\varphi}(g_2K).$$

The group homomorphism $\overline{\varphi}$ satisfies $\overline{\varphi} \circ \pi = \varphi$, because for all $g \in G$ we have

$$\big(\overline{\varphi} \circ \pi\big)(g) = \overline{\varphi}(gK) = \varphi(g)$$

as required.

Finally, $\overline{\varphi}$ is the unique map (in particular, the unique homomorphism) satisfying the conditions. For if $\theta\colon G/K \to H$ satisfies $\theta \circ \pi = \varphi$ then $(\theta \circ \pi)(g) = \varphi(g)$ for all $g \in G$, and since $\pi(g) = gK$, this implies that $\theta(gK) = \varphi(g)$ for all $g \in G$: that is, $\theta = \overline{\varphi}$.

Theorem II.25 [First Isomorphism Theorem] Let $\varphi\colon G \to H$ be a group homomorphism with kernel $\operatorname{Ker}\varphi = K$. Then there is a group isomorphism

$$\overline{\varphi}\colon G/K \longrightarrow \operatorname{Im}\varphi.$$

Proof: Applying the universal property from Theorem II.24 to the normal subgroup K gives a homomorphism $\overline{\varphi}\colon G/K \to H$ given by $\overline{\varphi}(gK) = \varphi(g)$. From this we get a surjective homomorphism (with the same name!)

$$\overline{\varphi}\colon G/K \longrightarrow \operatorname{Im}\varphi$$

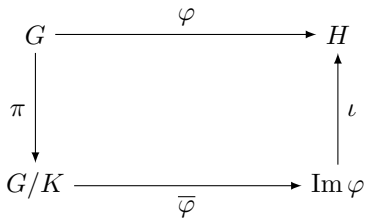
simply by changing the target of the morphism, making it be the image of φ rather than H.

To see that $\overline{\varphi}$ is injective, suppose $\overline{\varphi}(g_1K) = \overline{\varphi}(g_2K)$. Then $\varphi(g_1) = \varphi(g_2)$, so $\varphi(g_1^{-1}g_2) = \varphi(g_1)^{-1}\varphi(g_2) = 1_H$, giving $g_1^{-1}g_2 \in \operatorname{Ker}\varphi = K$ and hence $g_1K = g_2K$ as required.

Therefore the map $\overline{\varphi}\colon G/K \to \operatorname{Im}\varphi$ is a bijective group homomorphism, so it is an isomorphism by Proposition I.31.

Corollary II.26 Every homomorphism can be written as the composition of a surjective group homomorphism, then an isomorphism, and finally an injective group homomorphism.

Proof: We factorise $\varphi\colon G \to H$ as shown below:



where π is the quotient map, $\overline{\varphi}$ is the map from the First Isomorphism Theorem, and ι is the inclusion map. These maps have all the required properties.

III Group actions

We shall make groups work as symmetries of sets, and relate this idea to what has gone before.

Basics of group actions

III.1 The idea is that, given a group G, we want to realise it as the group of symmetries of something. Or, failing that, we want to get a relation between it and a group of symmetries of something: that is, a homomorphism $G \to \operatorname{Sym} X$ for some X. Actions are very far from unique and they can tell us a great deal about X as well as G, especially if X has some extra structure and G preserves it. For example, if we can write G as a group of matrices then we can make G act on a vector space so that the elements of G become linear maps.

Definition III.2 Suppose G is a group and X is a set. An *action* of G on X is a map $a\colon G \times X \to X$, in which we normally denote $a(g, x)$ by $g(x)$ or even gx, such that if g_1, $g_2 \in G$ and $x \in X$ then $g_1(g_2(x)) = (g_1g_2)(x)$ and $1_G(x) = x$. In this case we say that G *acts on* X.

Lemma III.3 If G acts on X and $g \in G$ then the map $a_g\colon X \to X$ given by $x \mapsto g(x)$ is a bijection.

Proof: The inverse map is given by the map $a_{g^{-1}}\colon x \mapsto g^{-1}(x)$.

Proposition III.4 If G acts on X then the map $\alpha\colon G \to \operatorname{Sym} X$ given by $g \mapsto a_g$ is a group homomorphism. Conversely, any group homomorphism $\alpha\colon G \to \operatorname{Sym} X$ induces an action of G on X by setting $g(x) = \alpha(g)(x)$.

Proof: For the first part, we need to check that $\alpha(g_1g_2) = \alpha(g_1)\alpha(g_2)$. But

$$\alpha(g_1g_2)(x) = a_{g_1g_2}(x) = (g_1g_2)(x) = g_1(g_2(x)) = a_{g_1}(a_{g_2}(x) = \alpha(g_1)(\alpha(g_2)(x))$$

as required. The second part leads to the same computation.

III.5 These are, technically, left actions. Left actions are pleasanter than right actions because then composition of maps goes the way you expect if you write maps on the left ($f(x)$ rather than xf).

Although a_g is a bijection, α is in general not: it need be neither injective nor surjective.

If G acts on X and H is a subgroup of G then H also acts on X, simply by restricting a to $H \times X$.

Definition III.6 If α is injective, so $a_g = \mathrm{id}_X$ only if $g = 1_G$, we say that G acts *faithfully* on X.

Example III.7 We have already seen many examples of group actions.

(i) $\mathbb{Z}$ acts on $\mathbb{R}$, by translations.

(ii) D_6 acts on the vertices of a triangle.

(iii) D_6 acts on the faces of a (cardboard) triangle.

(iv) The cube group (the group of symmetries of a cube) acts on the cube, but also on the octohedron.

(v) $\mathrm{GL}(2,\mathbb{R})$ acts on $\mathbb{R}^2$: more generally, any subgroup of $\mathrm{GL}(n,\mathbb{F})$ acts on $\mathbb{F}^n$.

(vi) $\mathrm{SL}(2,\mathbb{R})$ acts on the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im}\, z > 0\}$ by Möbius transformations: $\begin{pmatrix} a & b \\ c & d \end{pmatrix}(z) = \frac{az+b}{cz+d}$. (This action, and especially its restriction to $\mathrm{SL}(2,\mathbb{Z})$, is extremely important in number theory.)

(vii) Any group G acts on itself (i.e. we take $X = G$) by (say) left multiplication, $a(g,h) = gh$, but notice that $a_g \colon G \to G$ is not a group homomorphism except when $g = 1$.

(viii) Any group G acts on itself by conjugation, $a(g,h) = ghg^{-1}$: this time a_g is a homomorphism.

(ix) Any group G acts on any set X by the trivial action, $g(x) = x$ for all $g \in G$ and all $x \in X$.

Definition III.8 Suppose that G acts on X and $x \in X$. The *orbit* of x, denoted $\mathrm{orb}(x)$, is the set

$$\mathrm{orb}(x) = \{gx \mid g \in G\} \subseteq X.$$

We will sometimes refer to $\mathrm{orb}(x)$ as the G-orbit of x and write $\mathrm{orb}_G(x)$. A set that is $\mathrm{orb}_G(x)$ for some $x \in X$ is called a G-orbit.

III.9 The orbit of x is the set of places that G can take x to.

Obviously $|\mathrm{orb}(x)| \leq |G|$ but different orbits can be of different sizes. Consider, for example, the action of $\mathbb{C}^*$ on $\mathbb{C}$ given by multiplication: $a(w,z) = wz$. The orbit of $z \in \mathbb{C}$ is $\{wz \mid w \in \mathbb{C}^*\}$ and that is $\mathbb{C}^*$ if $z \neq 0$ but $\{0\}$ if $z = 0$.

Another example is conjugacy classes: these are the orbits of the action in Example III.7(viii). If G is not abelian, there is some pair of elements h and g such that $ghg^{-1} \neq h$, and then if G acts on itself by conjugation the orbit $\mathrm{orb}(h)$ contains at least those two elements. But $\mathrm{orb}(1) = \{1\}$ for this action, since $g1g^{-1} = 1$.

On the other hand, in the action in Example III.7(vii) the orbits are all the same size. In fact there is only one orbit, because any $g \in G$ is in $\mathrm{orb}(1) = \{g1 \mid g \in G\} = G$.

If we restrict the action in Example III.7(vii) to a subgroup $H < G$ then the orbit of g is $\{hg \mid h \in H\}$ which is the (right) coset Hg. In this case the orbits are all the same size by Lemma II.3.

Theorem III.10 [Orbit partition theorem] If G acts on X, then every $x \in X$ belongs to exactly one G-orbit.

Proof: We shall prove that G-orbits are equivalence classes for an equivalence relation. We define the relation by $x \sim y \iff \exists g \in G$ such that $g(y) = x$. This is reflexive (take $g = 1_G$), symmetric (because $y = g^{-1}(x)$) and transitive because if $x \sim y$ and $y \sim z$ then $x = g(y)$ and $y = g'(z)$ for some g, $g' \in G$ and then $x = (gg')(z)$ by Definition III.2.

III.11 We can now see Theorem II.4 and II.5 as a special case: in view of III.9 and Example III.7(vii) we can describe cosets as orbits.

Definition III.12 We say that the action of G on X is *transitive* (or G acts transitively on X) if $\mathrm{orb}_G(x) = X$ for some $x \in X$.

III.13 In this case there is only one orbit so $\mathrm{orb}_G(x) = X$ for any $x \in X$.

We saw in III.9 that the action of any G on itself by multiplication, Example III.7(vii), is transitive. Also any action on a 1-point set must be transitive.

It is easy to make many examples of transitive group actions. Take any group action of G on X, and pick a G-orbit X_0. Then G acts transitively on X_0. Sometimes this is interesting: for instance, $\mathrm{GL}(n, \mathbb{F})$ acts on $\mathbb{F}^n$ (Example III.7(v)) and it acts transitively on $\mathbb{F}^n \setminus \{0\}$, which is the orbit of any non-zero vector.

Definition III.14 Let G be a group acting on a set X and suppose $x \in X$. The *stabiliser* $\mathrm{Stab}_G(x)$ of x in G is

$$\mathrm{Stab}_G(x) := \{g \in G \mid g(x) = x\}.$$

Lemma III.15 The stabiliser $\mathrm{Stab}_G(x)$ is a subgroup of G.

Proof: Notice that $1 \in \mathrm{Stab}_G(x)$, so $\mathrm{Stab}_G(x) \neq \varnothing$. So according to Lemma I.14, we need to show that $g_1 g_2^{-1} \in \mathrm{Stab}_G(x)$ if g_1, $g_2 \in \mathrm{Stab}_G(x)$. If $g_i \in \mathrm{Stab}_G(x)$ then $g_i(x) = x$ and therefore also $g_i^{-1}(x) = g_i^{-1} g_i(x) = x$, so in this case we have $(g_1 g_2^{-1})(x) = g_1(g_2^{-1}(x)) = g_1(x) = x$ as required.

III.16 A very useful way to find subgroups of a group G is to make G act on something and look at the stabiliser, often choosing the element x carefully so as to get an interesting stabiliser. Unlike kernels, which will only find normal subgroups, this sees all the subgroups of G.

In particular, the stabiliser is not usually a normal subgroup, so you cannot say "choose $x \in X$ and take the quotient of G by $\mathrm{Stab}_G(x)$".

Example III.17 We again know many examples.

(i) $\mathbb{Z}$ acting on $\mathbb{R}$ has trivial stabilisers.

(ii) D_6 acting on the vertices of a triangle has stabilisers of order 2.

(iii) D_6 acting on the faces of a (cardboard) triangle has stabilisers of order 3.

(iv) In the action of $\mathrm{GL}(2,\mathbb{R})$ on $\mathbb{R}^2$, the stabiliser of the origin is $\mathrm{GL}(2,\mathbb{R})$ and the stabiliser of $\begin{pmatrix}1\\0\end{pmatrix}$ is $\left\{\begin{pmatrix}1 & b\\0 & d\end{pmatrix} \mid d \neq 0,\ b \in \mathbb{R}\right\}$.

(v) In the action of $\mathrm{SL}(2,\mathbb{Z})$ on the upper half-plane $\mathbb{H} = \{z \in \mathbb{C} \mid \mathrm{Im}\, z > 0\}$, the stabiliser of most $z \in \mathbb{H}$ is $\pm I$, but the stabiliser of $i \in \mathbb{H}$ is a group of order 4 generated by $\begin{pmatrix}0 & 1\\-1 & 0\end{pmatrix}$ and the stabiliser of $\omega = e^{2\pi i/3}$ is of order 6, generated by $\begin{pmatrix}0 & -1\\1 & 1\end{pmatrix}$. You wouldn't believe how much trouble, or the reverse, is caused by these two groups.

Proposition III.18 Suppose $H < G$. Then there exists a set X, and element $x \in X$ and an action of G on X such that $H = \mathrm{Stab}_G(x)$.

Proof: Take $X = \{gH \mid g \in G\}$ to be the coset set of H in G (Definition II.14). Then G acts on X by left multiplication, i.e. $a(g, g'H) = gg'H$. It is simple to check that this is indeed a group action, and the stabiliser of $H \in X$ (that is, the coset $1_G H$) is H.

Theorem III.19 [Orbit-Stabiliser Theorem] Suppose G is a group acting on a set X and $x \in X$. Then the size of the orbit of x is the index of its stabiliser:

$$|\,\mathrm{orb}_G(x)| = |G : \mathrm{Stab}_G(x)|.$$

Proof: For brevity, write $S = \mathrm{Stab}_G(x)$. We need to find a bijection between $\mathrm{orb}_G(x)$ and the coset set G/S of S in G (this is only a set, in general, not a group: see II.15 and Definition II.14). An element of the coset set is gS for some $g \in G$ and we define $\omega\colon G/S \to \mathrm{orb}_G(x)$ by $\omega(gS) = g(x)$. We need to check that this is well-defined, i.e. that if $g'S = gS$ then $g'(x) = g(x)$, but $g'S = gS$ if and only if $g' \in gS$, i.e. if and only if $g' = gh$ for some $h \in S$, and then $g'(x) = gh(x) = g(h(x)) = g(x)$ since h stabilises x.

We also need to check that ω is bijective. We do this by writing down its inverse. If $y \in \mathrm{orb}_G(x)$ then $y = g(x)$ for some $g \in G$ and we consider the

map $\gamma\colon \mathrm{orb}_G(x) \to G/S$ given by $\gamma(y) = gS$. Again we must check that γ is well-defined. If $y = g(x) = g'(x)$ then $g^{-1}g'(x) = x$ so $g^{-1}g' \in S$ so $g' = gg^{-1}g' \in gS$, so $g'S = gS$ as required. Then $\gamma(\omega(gS)) = \gamma(g(x)) = gS$ and $\omega(\gamma(y)) = \omega(gS) = g(x) = y$, so the maps are inverse so they must both by bijections.

III.20 Note that we have not assumed that either G or X is finite.

The orbit-stabiliser theorem is very useful for counting, which is difficult in general. It is often much easier to compute the index of a group than to compute the size of an orbit, or vice versa.

Representations

III.21 Suppose that G is a group: in principle any group, but we shall mainly want this construction for finite groups. Then G acts on itself by left multiplication (Example III.7(vii)) and this gives a homomorphism $\alpha\colon G \to \mathrm{Sym}\, G$ by Proposition III.4.

Recall (Example I.12(xii)) that the set of bijections from a set X to itself forms a group under composition. Such a group is called a permutation group.

Theorem III.22 [Cayley's Theorem] Every group is isomorphic to a subgroup of a permutation group.

Proof: Consider the left action of G on itself, so $a_g = \alpha(g)\colon h \to gh$ for $g, h \in G$. We claim that α is injective: then $G \cong \mathrm{Im}\,\alpha < \mathrm{Sym}\, G$ by Theorem II.25.

According to Lemma I.28 it is sufficient to check that $\mathrm{Ker}\,\alpha = 1$. But if $g \in \mathrm{Ker}\,\alpha$ then $a_g = \mathrm{id}$: that is, $gh = h$ for all $h \in G$. In particular if we take $h = 1_G$ we get $g = 1_G$, as required.

III.23 Despite its short proof, Cayley's theorem is important. It is the formal statement that captures the idea that groups are all about symmetry. It exhibits every possible group as a group of symmetries of something; namely, itself. Not the full group of symmetries: $\mathrm{Sym}\, G$ is typically much bigger than G.

Corollary III.24 Every finite group is isomorphic to a subgroup of $\mathrm{GL}(|G|, \mathbb{F})$, where $\mathbb{F}$ is any field. In particular, any finite group is isomorphic to a group of matrices.

Proof: In view of Cayley's Theorem it is enough to prove that S_n is isomorphic to a subgroup of $\mathrm{GL}(n, \mathbb{F})$. (Note that this is not the special case of this corollary for $G = S_n$: that says that S_n is isomorphic to a subgroup of $\mathrm{GL}(n!, \mathbb{F})$.) We take the usual basis $e_, \ldots, e_n$ for $\mathbb{F}^n$ and we define $M\colon S_n \to \mathrm{GL}(n, \mathbb{F})$ by

$$M(\sigma)e_i = e_{\sigma(i)}$$

and extending by linearity to other elements of $\mathbb{F}^n$. This is a group homomorphism because

$$M(\sigma\sigma')(i) = e_{(\sigma\sigma')(i)} = e_{\sigma(\sigma'(i))} = M(\sigma)(e_{\sigma'(i)}) = M(\sigma)(M(\sigma')(i))$$

for every i, and since linear maps are determined by their effect on a basis, that proves that $M(\sigma\sigma') = M(\sigma)M(\sigma')$. It is injective because if $M(\sigma) = \mathrm{id}$ then $e_i = M(\sigma)(e_i) = e_{\sigma(i)}$ for all i, so σ is the trivial permutation.

III.25 This is completely false for infinite groups. They could, for example, simply be bigger than the set of all matrices over $\mathbb{F}$, but in fact it can fail even for countable groups.

III.26 A homomorphism $\varphi\colon G \to \mathrm{GL}(n, \mathbb{F})$ for some n and some field $\mathbb{F}$ is called a representation of G over the field $\mathbb{F}$. A representation is called faithful if its kernel is trivial, i.e. its image is isomorphic to G. Corollary III.24 says that every finite group has a faithful representation. Studying representations of groups is one of the main ways to get information about them.

IV Rings

We introduce rings, which have both an addition and a multiplication, but not multiplicative inverses.

Basic theory of rings

IV.1 The basic idea of a ring is that you are allowed to add and subtract and multiply, but not necessarily divide: addition is commutative, but multiplication might not be. Then there have to be rules about how multiplication and addition interact. So, looking at I.10, we expect two binary operations, one unary operation (subtraction, or rather minus), and one or two nullary operations (0 and, optionally, 1).

IV.2 The basic example of some things that you can add and multiply by not necessarily divide is the integers. You can't expect to be able to divide an integer by 2, not if you want the answer to be an integer. So $\mathbb{Z}$ should be one basic example of a ring. In that case multiplication is commutative. Another good example is the ring of polynomials with (say) real coefficients in one variable, $\mathbb{R}[t]$. You can add two of these and you can multiply them, but you can't divide: $\frac{1}{t}$ isn't a polynomial. In fact, these two turn out to have much more in common than just being rings: they are very similar rings. In other words, a polynomial may be nothing like an integer, but polynomials collectively behave remarkably like integers collectively.

If you want an example of a noncommutative ring, consider the set of (say) 2×2 (say) real matrices, $M_2(\mathbb{R})$. Here you have addition and multiplication, and you can take inverses as long as the determinant isn't zero; but sometimes it is zero.

Definition IV.3 A *ring* is a triple $(R, +, \cdot)$, where R is a set with two binary operations called addition (denoted $+$) and multiplication (denoted $\cdot$, or just nothing), such that the following axioms hold:

(i) **(addition)** $(R, +)$ is an abelian group.

(ii) **(associative)** The binary operation $\cdot$ is associative.

(iii) **(distributive)** Multiplication is *distributive over addition*, from the left or the right: that is

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

(iv) **(identity)** There is a *multiplicative identity*, an element $1 \in R$ such that $1 \cdot a = a \cdot 1 = a$ for all $a \in R$.

We write 0 for the (unique) additive identity, and $-a$ for the (unique) additive inverse of $a \in R$.

IV.4 The last axiom **(identity)** in IV.3 is in a sense optional. Many writers omit it, and allow rings without 1. The disadvantage is that when you want to talk about a ring with a 1 (also called a unital ring), you have to say so; and that happens a lot, so you would rather keep the simpler phrasing for the commonest case. Here we have made the opposite decision, and the disadvantage is that we now have no name for a ring without a 1, i.e. something that satisfies IV.3(i)–(iii) but not necessarily IV.3(iv).

IV.5 We often omit $\cdot$ and write ab instead of $a \cdot b$. Sometimes we will need to wite 0_R and 1_R for the identities in R, to distinguish them from identities in other rings or groups.

As with groups, we will very often know what the operations are and just talk about the ring R, rather than calling it $(R, +, \cdot)$.

For simplicity we often avoid brackets when there is no ambiguity. Here the same conventions hold as for real numbers: $\cdot$ 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 (see V.30 below) $2a$ for $a + a$ and so on.

Lemma IV.6 In any ring $(R, +, \cdot)$, we have

(i) $a \cdot 0 = 0$ and $0 = 0 \cdot a$ for all $a \in R$; and

(ii) $a \cdot (-b) = -(a \cdot b)$ and $-(a \cdot b) = (-a) \cdot b$ for all $a, b \in R$.

Definition IV.7 Let R be a ring. 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 $a \cdot b = b \cdot a = 1$.

Lemma IV.8 Let R be a ring. Then the multiplicative identity is unique, and if $a \in R$ is a unit then the multiplicative inverse of a is unique.

Proof: The same argument as in I.9 still works.

IV.9 Notice that if R is a ring then $(R, \cdot)$ is *never* a group, except in the trivial case where $R = \{0\}$ and $0 = 1$. This is because 0 cannot possibly have a multiplicative inverse, because of IV.6(i).

The set of units in R is denoted by R^* or sometimes $R^\times$. It is easy to see that R^* does form a group under multiplication (but it behaves very badly under addition: the sum of two units does not have to be a unit). In fact, there is nothing to check because associativity and identity are already in the ring axioms.

Example IV.10 Some examples of rings and their units are:

(i) $R = \mathbb{Z}$; then $\mathbb{Z}^* = \{\pm 1\}$.

(ii) $R = \mathbb{C}$; then $\mathbb{C}^* = \{z \in \mathbb{C} \mid z \neq 0\}$.

(iii) $R = M_2(\mathbb{R})$; then $R^* = \{A \in M_2(\mathbb{R}) \mid \det A \neq 0\}$.

Notice how different these three are. In (i) there are very few units, but more than just 1. In (ii) everything is a unit, except for 0 of course. In (iii) most elements are units, but there are still exceptions.

Types of rings

Definition IV.11 A ring R is a *commutative ring* if $a \cdot b = b \cdot a$ for all a, $b \in R$.

One does not say "abelian ring"! In contexts where only commutative rings are important, one sometimes drops the word "commutative", and explicitly says "noncommutative ring" where necessary, but that will not be done here.

Definition IV.12 A ring R is an *integral domain* if it is a commutative ring in which $0 \neq 1$, such that if a, $b \in R$ and $ab = 0$, then $a = 0$ or $b = 0$.

Sometimes an integral domain is just called a domain.

Definition IV.13 A ring R is a *division ring* if $0 \neq 1$, and every non-zero element is a unit.

That is, a^{-1} exists as long as $a \neq 0$. Note that a division ring is not required to be commutative.

Definition IV.14 R is a *field* if it is a commutative division ring.

IV.15 Informally, in a field you can add and multiply and divide by anything that isn't actually zero, and mutiplication is commutative.

Fields are often called K (German *Körper*). Every field K is an integral domain: if a, $b \in K$ and $ab = 0$, then if $a \neq 0$ we have $b = 1 \cdot b = a^{-1}ab = a^{-1} \cdot 0 = 0$.

Example IV.16 We start with a few familiar examples.

(i) Every field is a commutative ring. In particular $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are commutative rings.

(ii) Division rings need not be commutative, so division rings need not be fields. An example of this is the quaternions.

(iii) The ring $\mathbb{Z}$ is an integral domain (that's why they are called "integral"), but it is not a division ring, so it is not a field.

(iv) The set $\mathbb{R}[t]$ of polynomials in one variable with real coefficients is also an integral domain, but not a field.

(v) The commutative ring $\mathbb{Z}/4$ consists of the integers with arithmetic mod 4. It is a commutative ring, but it is not an integral domain because $2 \times 2 = 4 = 0$ but $2 \neq 0$.

(vi) On the other hand, $\mathbb{Z}/5$ is a field; in fact $\mathbb{Z}/p\mathbb{Z}$ is a field, called $\mathbb{F}_p$, whenever p is prime.

There is also a field with 4 elements, called $\mathbb{F}_4$, but it is not the same as $\mathbb{Z}/4\mathbb{Z}$. They are both commutative rings and they both have four elements, but they are otherwise different.

Definition IV.17 If K is a field, a *K-algebra* is a ring that is also a K-vector space, and such that $(\lambda a)(\mu b) = (\lambda\mu)(ab)$ for any λ, $\mu \in K$ and a, $b \in A$.

Example IV.18 For any ring R, let $M_n(R)$ denote the set of all $n \times n$ matrices with coefficients in the ring R. Then $M_n(R)$ is a ring with respect to usual addition and multiplication of square matrices. However, even if R is a commutative ring, M_n is not commutative, unless $n = 1$ (when it is just R in a very light disguise).

We saw this ring in IV.10(iii) in the case $R = \mathbb{R}$. Note that if $R = K$ is a field then $M_n(K)$ is a K-algebra.

Example IV.19 Let R be a ring and let t be a variable. Let $d \geq 0$ be a non-negative integer. A *polynomial f over R of degree d* is a formal expression

$$f = \sum_{k=0}^{d} a_k t^k = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \cdots + a_d t^d,$$

with $a_k \in R$ for $0 \leq k \leq d$ and $a_d \neq 0$. The a_k are called the *coefficients*.

We let $R[t]_d$ denote the set of all polynomials of degree d, and we set

$$R[t] = \{0\} \cup \bigcup_{d=0}^{\infty} R[t]_d.$$

Thus $1 + t$, $39 - 62t + 3t^{19}$ and 94 are all elements of $\mathbb{Z}[t]$, and so is 0, but $\frac{1}{t}$ is not and neither is e^t.

The *degree* of a polynomial is the highest k such that $a_k \neq 0$, i.e. the highest power of t that actually occurs: a polynomial of degree 0 is called a non-zero constant (it is just a non-zero element of R). The degree of the polynomial 0 is not defined – sometimes one can save some writing by declaring that it is 0, or that it is -1 or even $-\infty$, but only temporarily. 0 is also called a constant, so the *constants* are just the elements of R.

It is easy to check that $R[t]$ is a ring. If $R = K$ is a field, than $K[t]$ is a K-algebra.

Definition IV.20 A nonempty subset S of a ring R is called a *subring* if and only if, for any a, $b \in S$ we have both $a - b \in S$ and $ab \in S$.

Lemma IV.21 Let S be a subset of a ring $(R, +, \cdot)$. Then S is a subring of R if and only if $(S, +, \cdot)$ is satisfies IV.3(i)–(iii) of a ring.

In other words, a subring is a subset that happens to be a ring, but not necessarily with a 1. This is an inconvenience of our convention that "ring" means "unital ring", but it is what we want.

Example IV.22 We have seen several examples of subrings already.

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

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

(iii) The even integers $2\mathbb{Z}$ are a subring of $\mathbb{Z}$. By Lemma IV.21 the even integers therefore form a ring but not necessarily with a 1: and indeed 1 is odd. (But 0 is even, a fact that seems to be poorly understood.) So a subring does not have to be a ring in the full sense of satisfying IV.3(iv) as well.

(iv) If R is any ring and t is a variable, then R is a subring of $R[t]$. In this case, the subring really is a ring.

(v) The Gaussian integers $\mathbb{Z}[i] = \{a + bi \in \mathbb{C} \mid a, b \in \mathbb{Z}\}$ form a subring of the field $\mathbb{C}$: also $1 = 1 + 0i \in \mathbb{Z}[i]$, so $\mathbb{Z}[i]$ is a ring.

The notation $\mathbb{Z}[i]$ is not incompatible with the notation for polynomial rings: unlike t, i is not a variable, but in any case,

$$\sum_{k=0}^{d} a_k i^k = \left(\sum_{l=0}^{\lceil d/4 \rceil} (a_{4l} - a_{4l+2}) \right) + i \left(\sum_{l=0}^{\lceil d/4 \rceil} (a_{4l+1} - a_{4l+3}) \right) \in \mathbb{Z}[i].$$

Lemma IV.23 If a subring S of an integral domain R contains the element $1 \in R$, then S is an integral domain.

Maps between rings

Definition IV.24 Let R, and S be rings. A map $\varphi\colon R \to S$ is said to be a *ring homomorphism* if and only if for all a, $b \in R$, we have

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

Example IV.25 Here are a few examples and unexamples.

(i) The map $\varphi\colon \mathbb{Z} \to \mathbb{Z}/2\mathbb{Z}$ defined by

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

is a ring homomorphism. Indeed, if we compare the rules for adding and multiplying even and odd integers with the addition and multiplication tables for $\mathbb{Z}/2\mathbb{Z}$, we see that computing in $\mathbb{Z}$ and then applying φ is the same as applying φ and then computing in $\mathbb{Z}/2\mathbb{Z}$.

(ii) More generally, the map $\varphi\colon \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ that takes an integer a to its residue class $[a]$ mod n is a ring homomorphism, for any $n \in \mathbb{N}$.

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

(iv) Let R be a commutative ring and choose $r \in R$. a polynomial $f \in R[t]$ with coefficients in R, then $f(r) \in R$, so we obtain a map

$$\mathrm{ev}_r\colon R[t] \longrightarrow R$$

by taking $ev_r(f) = f(r)$. In other words, we *evaluate* each polynomial at $r \in R$, i.e., substitute $r \in R$ into each polynomial. We shall see shortly in Proposition IV.29 that ev_r is a ring homomorphism, for any $r \in R$.

(v) The map $\varphi\colon \mathbb{R} \to M_{2\times 2}(\mathbb{R})$ given by $\varphi(x) = \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix}$ is a ring homomorphism.

Lemma IV.26 The composition of two ring homomorphisms is a ring homomorphism.

Lemma IV.27 If $\varphi\colon R \to S$ is a ring homomorphism then

(i) for a, $b \in R$, we have $\varphi(b-a) = \varphi(b) - \varphi(a)$;

(ii) $\varphi(0_R) = 0_S$;

(iii) for $a \in R$, we have $\varphi(-a) = -\varphi(a)$.

Proof: For part (i), we have

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

and we add $-\varphi(a)$ to both sides.

For (ii), substitute $b=a$ in (i) to obtain

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

For part (iii), substitute $b=0$ into (i) and use (ii) to obtain

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

as required.

IV.28 Notice that in Lemma IV.27 we have not claimed that $\varphi(1_R)=1_S$, and indeed that does not have to happen. For an example, see IV.25(vi). In that case we have $\varphi(1)=\begin{pmatrix}1&0\\0&0\end{pmatrix}$, which is not the identity matrix.

Proposition IV.29 Let R be a commutative ring and $r\in R$. Then the evaluation map $\mathrm{ev}_r\colon R[t]\to R$ is a ring homomorphism.

Proof: Given any two polynomials $f=\sum_{k=0}^{d_1}a_kt^k$ and $g=\sum_{k=0}^{d_2}b_kt^k$, we have

$$\begin{aligned}
\mathrm{ev}_r(f+g) &= \mathrm{ev}_r\left(\sum_{k=0}^{\max\{d_1,d_2\}}(a_k+b_k)t^k\right)\\
&= \sum_{k=0}^{\max\{d_1,d_2\}}(a_k+b_k)r^k\\
&= \sum_{k=0}^{d_1}a_kr^k+\sum_{k=0}^{d_2}b_kr^k\\
&= \mathrm{ev}_r(f)+\mathrm{ev}_r(g),
\end{aligned}$$

using commutativity of addition and distributivity in R to get from the second line to the third. Similarly

$$\begin{aligned}
\mathrm{ev}_r(fg) &= \mathrm{ev}_r\left(\sum_{k=0}^{d_1+d_2}\left(\sum_{i+j=k}a_ib_j\right)t^k\right)\\
&= \sum_{k=0}^{d_1+d_2}\left(\sum_{i+j=k}a_ib_j\right)r^k\\
&= \sum_{i=0}^{d_1+d_2}a_ir^i\cdot\sum_{j=0}^{n}b_jr^j\\
&= \mathrm{ev}_r\left(\sum_{i=0}^{d_1}a_it^i\right)\cdot\mathrm{ev}_r\left(\sum_{j=0}^{d_2}b_jt^j\right)\\
&= \mathrm{ev}_r(f)\cdot\mathrm{ev}_r(g),
\end{aligned}$$

where the first line comes from the definition of multiplication in $R[t]$ and getting from the second line to the third uses everything: the distributive laws, commutativity of addition and associativity of both addition and multiplication in the ring R.

V Modules, ideals and quotient rings

In this section we introduce modules, ideals and quotients, all of which are analogous to structures that are already familiar.

Modules

V.1 A module is to a ring what a vector space is to a field: that is, it is an abelian group on which the ring acts linearly. For simplicity, we shall limit ourselves to the case of commutative rings.

Definition V.2 Let R be a commutative ring. An *R-module* is an abelian group M (written additively) together with a map $R \times M \to M$, called *scalar multiplication* and written $(r, m) \mapsto rm$, such that or all r, $s \in R$ and m, $n \in M$ we have

(a) $r(m + n) = rm + rn$;

(b) $r(sm) = (rs)m$;

(c) $(r + s)m = rm + sm$;

(d) $1m = m$.

Some texts omit the last condition. Leaving it out allows the possibility that there is a non-zero subgroup $N < M$ for which multiplication by $r \in R$ is simply 0: that is, $rn = 0$ for every $n \in N$ and every $r \in R$. In practice this makes very little difference.

V.3 Informally (but completely accurately) Definition V.2 means that a module is a structure in which we are able to take linear combinations with coefficients taken from R: if r, $s \in R$ and m, $n \in M$ then $rm + sn \in M$. Addition in M is the case $r = s = 1_R$ and scalar multiplication is the case $s = 0$.

Example V.4 We know a few examples.

(i) If $R = K$ is a field, then a K-module is the same thing as a K-vector space.

(ii) Any commutative ring R is an R-module, by taking scalar multiplication to be ring multiplication. This is the same as thinking of K as a 1-dimensional K-vector space. More generally, R^n is an R-module by

multiplication in each factor: that is, $r(r_1, \ldots, r_n) = (rr_1, \ldots, rr_n)$: if $R = K$ is a field, this is the same as thinking of K^n as an n-dimensional vector space. Such a module is called a *free R-module*.

(iii) If S is a subring of R (and $1_R \in S$) then R is an S-module, again by taking scalar multiplication to be ring multiplication in R. In particular, $R[t]$ is an R-module.

(iv) $\mathbb{Z}/N$ is a $\mathbb{Z}$-module, where the multiplication is given by multiplication in $\mathbb{Z}$. That is, if $r \in \mathbb{Z}$ and $[m]$ is the class of m mod N, then we define $r[m] = [rm]$: since $[r(m + kN)] = [m]$ this is well defined, and the module axioms follow at once from the ring axioms in $\mathbb{Z}$.

(v) Abelian groups and $\mathbb{Z}$-modules are the same thing.

V.5 Notice that Example V.4(iv) shows that even for $R = \mathbb{Z}$ there is not a simple classification theorem for R-modules as there is for vector spaces: there is no such thing as the dimension of an R-module.

Definition V.6 If N is a subgroup of M that is closed under scalar multiplication (that is, $rn \in N$ if $r \in R$ and $n \in N$) then N is called a *submodule* of M.

Definition V.7 An *ideal* of R is an R-submodule of R.

V.8 A more direct way to give Definition V.7 is to say that an ideal is a subgroup of $(R, +)$ such that $ra \in I$ and $ar \in I$ for any $a \in I$ and $r \in R$. Another way to say the same this is to say that I is a subring of R with the extra property that $ar \in I$ if just one of a and r belongs to I (the definition of subring only guarantees this if they both do).

V.9 To check that a nonempty set $I \subset R$ is an ideal, it is enough to check that

(i) $a - b \in I$ if $a, b \in I$, and

(ii) $ra \in I$ if $a \in I$ and $r \in R$.

or more simply still, that linear combinations stay in I: that is, $ra + sb \in I$ if $r, s \in R$ and $a, b \in I$. This is a short cut similar to Lemma I.14.

V.10 Ideals play for rings the role that normal subgroups play for groups. Indeed we shall write $I \lhd R$ to mean that I is an ideal of R (one also says "ideal in R").

V.11 If R is non-commutative there are left and right modules and left and right ideals, and one wants to work mostly with two-sided ideals, i.e. left ideals that are also right ideals.

Example V.12 There is one motivating example of an ideal, and we do know some others.

(i) Suppose $N \in \mathbb{Z}$ and $N \geq 0$. Then $N\mathbb{Z}$, the set of multiples of N, is an ideal. This is the example everybody knows, and the historically first one.

(ii) More generally, R be a commutative ring and let $a \in R$. The set $aR = \{ar \mid r \in R\}$, also sometimes denoted $\langle a \rangle$ if the ring R is clear from the context, is an ideal in R. As we are assuming that R is a commutative ring, we may also write Ra if that is convenient.

(iii) Let $R = K[t]$ for some field K (say $K = \mathbb{C}$). Then the set of polynomials with $a_0 = 0$ forms an ideal. It is in fact the ideal $\langle t \rangle = tR$. More generally, for any d, the set of polynomials with $a_0 = \ldots = a_{d-1} = 0$ is the ideal $\langle t^d \rangle$.

(iv) $\{0\}$ is an ideal of R, usually called simply 0: it is equal to $0R$.

(v) R is an ideal of R: it is equal to $1R$.

(vi) Consider the ring $\mathbb{C}[x, y]$ of polynomials in two variables. Again the set of polynomials whose constant term is zero forms an ideal, but this is not of the form pR for any polynomial $p \in R$.

(vii) Again with $R = \mathbb{C}[x, y]$, we may regard any $p \in R$ as a function $p \colon \mathbb{C}^2 \to \mathbb{C}$. If we take X to be the union of the x- and y-axes in $\mathbb{C}^2$ and consider

$$I(X) = \{p \in R \mid p(a, b) = 0 \text{ if } (a, b) \in X\}$$

then $I(X)$ is an ideal of R.

(viii) In (vii), if we instead take $X = \{(0, 0)\}$ then $I(X)$ is the set of polynomial functions on $\mathbb{C}^2$ that vanish at the origin, which is the same ideal as in (vi).

Definition V.13 If $a \in R$ then $\langle a \rangle$ (also written aR or Ra) is called the *ideal generated by* a, and every ideal of this form is called a *principal ideal*.

Definition V.14 Let I be an ideal in a commutative ring R. The *quotient ring* R/I is the set

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

of cosets of I in R. Addition and multiplication in the ring R/I are defined by

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

That is, we add and multiply cosets by choosing representatives of them and doing the addition and multiplication in R, and then taking the coset of the result.

Example V.15 We already know several examples of quotient rings.

(i) If $n \in \mathbb{Z}$ then $n\mathbb{Z}$ is an ideal in $\mathbb{Z}$ (Example V.12(i)) and $\mathbb{Z}/n\mathbb{Z} = \mathbb{Z}/n$.

(ii) If $R = \mathbb{C}[t]$ then the quotient by the ideal $\langle t^2 \rangle$ from Example V.12(iii) is described as follows. Any polynomial f can be written in the form $f = t^2h+at+b$ for unique a, $b \in \mathbb{C}$ (and $h \in \mathbb{C}[t]$), so $f+\langle t^2 \rangle = (at+b)+\langle t^2 \rangle$ for some unique a, $b \in \mathbb{C}$. Therefore $\mathbb{C}[t]/\langle t^2 \rangle = \{[at+b] \mid a, b \in R\}$, where $[f]$ denotes $f + \langle t^2 \rangle$, and addition and multiplication are given by

$$[at + b] + [ct + d] = [(a + c)t + (b + d)]$$

and

$$[at + b] \cdot [ct + d] = [act^2 + (ad + bc)t + bd] = [(ad + bc)t + bd].$$

respectively. In other words we work with polynomials and then discard all the terms involving t^2, because that is zero in $R[t]/\langle t^2 \rangle$.

V.16 V.15(ii) provides a respectable way to say the thing that you said when learning calculus: "ε^2 is small so we'll ignore it". You were actually working in $R[\varepsilon]/\langle \varepsilon^2 \rangle$.

Kernel and image

Definition V.17 Let $\varphi\colon R \to S$ be a ring homomorphism. The *kernel* of φ is the subset $\operatorname{Ker}\varphi$ of R given by

$$\operatorname{Ker}\varphi = \{a \in R \mid \varphi(a) = 0\}.$$

This is very similar to the definition of null space in vector spaces, or kernel in groups. Remember, though, that the kernel consists of all those elements that are mapped to 0, not 1. This is because the only operation that makes a ring into a group is addition: we are missing multiplicative inverses.

Definition V.18 Again let $\varphi\colon R \to S$ be a ring homomorphism. The *image* of φ is the subset $\operatorname{Im}\varphi$ of S given by

$$\operatorname{Im}\varphi = \{\varphi(a) \in S \mid a \in R\}.$$

Lemma V.19 Let $\varphi\colon R \to S$ be a ring homomorphism. Then $\operatorname{Ker}\varphi$ is an ideal of R. Moreover, φ is injective if and only if $\operatorname{Ker}\varphi = \{0\}$.

Proof: A ring homomorphism is in particular an additive group homomorphism, so $\operatorname{Ker}\varphi$ is an additive subgroup by Lemma I.27 and the statement about injectivity follows from Lemma I.28. Also, if $r \in R$ and $a \in \operatorname{Ker}\varphi$ then $\varphi(ra) = \varphi(r)\varphi(a) = \varphi(r) \cdot 0$ and similarly for ar, so $\operatorname{Ker}\varphi$ is an ideal.

Lemma V.20 Let $\varphi\colon R \to S$ be a ring homomorphism. Then $\operatorname{Im}\varphi$ is a subring of S. Moreover, φ is surjective if and only $\operatorname{Im}\varphi = S$.

This is trivial, but notice that $\varphi(1_R) \neq 1_S$ in general.

Definition V.21 Let I be an ideal in a ring R. The *quotient map* $\pi\colon R \to R/I$ is defined by setting $\pi(a) = a + I$ (exactly as in Definition II.23).

Lemma V.22 Let R be a ring, let I be an ideal in R and let S be a subring of R with $1 \in S$. Then

(i) $\pi\colon R \to R/I$ is a surjective ring homomorphism, so $\operatorname{Im}\pi = R/I$, and $\operatorname{Ker}\pi = I$;

(ii) $\iota\colon S \to R$ is an injective ring homomorphism, so $\operatorname{Ker}\iota = \{0\}$, and $\operatorname{Im}\iota = S$.

Proof: π is a ring homomorphism, because

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

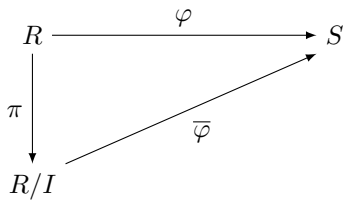
and

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

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

For ι there is nothing to prove.

Theorem V.23 Let $\varphi\colon R \to S$ be a ring homomorphism and let I be an ideal in R satisfying $I \subseteq \operatorname{Ker}\varphi$. Then there exists a unique ring homomorphism $\overline{\varphi}\colon R/I \to S$ such that the diagram



commutes, i.e., $\overline{\varphi} \circ \pi = \varphi$.

Proof: This follows from II.24 applied to the additive groups, except that we need to check that $\bar{\varphi}$ is a ring homomorphism, i.e. that $\bar{\varphi}$ preserves multiplication. But

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

Definition V.24 Let R, S be rings. A homomorphism $\varphi\colon R \to S$ is called an *isomorphism* if there is a ring homomorphism $\psi\colon S \to R$ such that $\psi(\varphi(r)) = r$ for all $r \in R$ and $\varphi(\psi(s)) = s$ for all $s \in S$. Given an isomorphism $\varphi\colon R \to S$, we say that R is *isomorphic* to S and write $R \cong S$.

V.25 As with groups, it is in fact the case that an bijective ring homomorphism is an isomorphism. We are fully entitled to use this fact, but it is not the definition.

Theorem V.26 Let $\varphi\colon R \to S$ be a ring homomorphism. Then there is a ring isomorphism

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

Proof: Exactly as for Theorem II.25.

Corollary V.27 Every ring homomorphism can be written as the composition of a surjective ring homomorphism, then an isomorphism, and finally an injective ring homomorphism.

Proof: Exactly as for Theorem II.26.

Corollary V.28 Let $\varphi\colon K \to R$ be a ring homomorphism where K is a field. Then φ is either the zero map or an isomorphism from K to a subring of R.

Proof: The kernel of φ is an ideal in K, and a field can only have two ideals: $\{0\}$ (when φ is an isomorphism onto its image) and K (when $\varphi = 0$). For if I is a nonzero ideal in K then we have $0 \neq b \in I$, so $1 = b^{-1}b \in I$, so for any $a \in K$ we have $1a \in I$, so $I = K$.

Characteristic

Definition V.29 The *characteristic* of a ring R, denoted $\operatorname{char} R$, is the order $o(1_R)$ of 1_R in the abelian group $(R, +)$, if that is finite. If $o(1_R) = \infty$ we say that $\operatorname{char} R = 0$.

V.30 Given a ring R, an element $a \in R$ and $n \in \mathbb{Z}$ we define an element $na \in R$ by

$$na = \underbrace{a + \cdots + a}_{n} \text{ if } n \geq 0, \text{ and } (-n)a = -(na).$$

In particular, zero copies of an element $a \in R$ is the zero element $0_R \in R$, that is $0a = 0_R$, where 0 is the zero element in $\mathbb{Z}$.

This is just notation. It is tempting to using the phrase "multiply a by n" for it, and everybody does, but what it really means is "add together n copies of a". We are not doing any ring multiplication, only addition.

Slightly more formally, we could say that we are defining a map $\mathbb{Z} \times R \to R$ and calling this multiplication, even though we shouldn't do because we have already used the word "multiplication" as the name of a map $\cdot\colon R \times R \to R$.

Notice that $0_R \cdot a = 0_R$ is a fact that we proved in Lemma IV.6 but $0a = 0_R$ is just a natural notation when 0 is the zero integer.

Example V.31 It is usually obvious what the characteristic is.

(i) The zero ring $R = \{0\}$ has $\operatorname{char} R = 1$, because $1_R = 0_R = 0$. No other ring has $\operatorname{char} R = 1$.

(ii) For any $n \in \mathbb{N}$ we have $\operatorname{char} \mathbb{Z}/n\mathbb{Z} = n$.

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

Lemma V.32 Let R be a ring of positive characteristic $n > 0$. Then $na = 0$ for all $a \in R$.

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

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

as required.

Definition V.33 Let R be a ring. Then

$$\mathbb{Z}1_R = \{n1_R \mid n \in \mathbb{Z}\} = \{\ldots, -2 \cdot 1_R, -1_R, 0_R, 1_R, 2 \cdot 1_R, \ldots\}.$$

is (obviously) a subring of R (with 1), called the *prime subring* of R.

Lemma V.34 Let R be a ring. Then either:

(i) $\operatorname{char} R = 0$, in which case $\mathbb{Z}1_R$ is isomorphic to $\mathbb{Z}$; or

(ii) $\operatorname{char} R = n > 0$, in which case $\mathbb{Z}1_R$ is isomorphic to $\mathbb{Z}/n\mathbb{Z}$.

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

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

and the distributive law gives

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

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

Suppose first that $\operatorname{char} R = 0$. Then $\varepsilon(n) = n1_R$, which equals 0_R if and only if $n = 0$, so $\operatorname{Ker} \varepsilon = \{0\}$. Applying V.26 to ε gives $\mathbb{Z} \cong \mathbb{Z}1_R$ which proves (i).

Otherwise, $\operatorname{char}(R) = n > 0$. Then $\varepsilon(m) = m1_R$, which equals 0_R if and only if $n|m$, so $\operatorname{Ker} \varepsilon = n\mathbb{Z}$. Applying V.26 to ε gives $\mathbb{Z}_n \cong \mathbb{Z}1_R$, which proves (ii).

Proposition V.35 The characteristic of an integral domain is either 0 or a prime.

Proof: Let R be an integral domain. Notice first that $\operatorname{char} R \neq 1$, because $R \neq \{0\}$. So if $n = \operatorname{char} R$ is neither 0 nor a prime, it must be composite: that is, we can write $n = rs$ for some $1 < r \leq s < n$. Then

$$0_R = n1_R = (rs)1_R = (r1_R) \cdot (s1_R),$$

but since R is an integral domain it follows that either $r1_R = 0$ or $s1_R = 0$. But then we have found k with $0 < k < n$ such that $k1_R = 0_R$, which is impossible because n was supposed to be the least such positive integer.

The Chinese remainder theorem

Definition V.36 Let I and J be ideals of R.

(i) The *sum* $I + J$ of I and J is the subset

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

(ii) The *product* IJ of I and J is the subset

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

(iii) The *intersection* $I \cap J$ of I and J is the subset

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

Lemma V.37 If I and J are ideals of R then $I + J$, IJ and $I \cap J$ are all ideals of R. Moreover, $IJ \subseteq I \cap J \subseteq I + J$.

V.38 IJ is the smallest ideal containing the set of products ab for $a \in I$ and $b \in J$. The set of such products fails to be an ideal itself because it is not closed under addition.

Example V.39 Take $R = \mathbb{Z}$ and consider m, $n \in \mathbb{Z}$. If we take $I = m\mathbb{Z} = \langle m \rangle$ and $J = n\mathbb{Z} = \langle n \rangle$, then $IJ = \langle mn \rangle$, $I+J = \langle \mathrm{hcf}(m, n) \rangle$ and $I \cap J = \langle \mathrm{lcm}(m, n) \rangle$.

Notice that $mn \geq \mathrm{lcm}(m, n) \geq \mathrm{hcf}(m, n)$: compare this with V.37.

Definition V.40 Let R and S be rings. The *direct product* of R and S is the ring

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

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

Theorem V.41 (Chinese Remainder Theorem) Let I, J be ideals in a ring R satisfying $I + J = R$. Then there is a ring isomorphism

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

Proof: Consider the map $\varphi\colon R \to R/I \times R/J$ defined by setting $\varphi(a) = (a + I, a + J)$. This is a ring homomorphism because

$$\begin{aligned} \varphi(a + b) &= (a + b + I, a + b + J) \\ &= \big((a + I) + (b + I), (a + J) + (b + J)\big) && \text{by V.14} \\ &= (a + I, a + J) + (b + I, b + J) && \text{by V.40} \\ &= \varphi(a) + \varphi(b) \end{aligned}$$

and

$$\begin{aligned}\varphi(a\cdot b) &= (a\cdot b+I, a\cdot b+J)\\ &= \big((a+I)\cdot(b+I),(a+J)\cdot(b+J)\big) && \text{by V.14}\\ &= (a+I,a+J)\cdot(b+I,b+J) && \text{by V.40}\\ &= \varphi(a)\cdot\varphi(b).\end{aligned}$$

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

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

so $\operatorname{Ker}\varphi = I\cap J$. The first isomorphism theorem, V.26, applied to φ gives an isomorphism $\overline{\varphi}\colon R/(I\cap J)\to\operatorname{Im}\varphi$.

So it remains to show that φ is surjective, i.e. that $\operatorname{Im}\varphi = R/I\times R/J$, so we choose an arbitrary $(a+I,b+J)\in R/I\times R/J$ and we need to show that this is in the image of φ. Since $R=I+J$, there exist $x\in I$ and $y\in J$ such that $1=x+y$: we set $r=ay+bx\in R$. Then

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

as required.

Corollary V.42 Let m, $n\in\mathbb{N}$ be coprime natural numbers: that is, there exist λ, $\mu\in\mathbb{Z}$ such that $\lambda m+\mu n=1$. Then $\mathbb{Z}/mn\mathbb{Z}\cong\mathbb{Z}/m\mathbb{Z}\times\mathbb{Z}/n\mathbb{Z}$.

Proof: In this case we have $\mathbb{Z}=m\mathbb{Z}+n\mathbb{Z}$ (because the right-hand side is an ideal that contains 1). Now V.41 gives the isomorphism: this is the Chinese Remainder Theorem for the integers.

Prime ideals and maximal ideals

Definition V.43 Suppose that R is a commutative ring. A proper ideal I in R is called a *prime ideal* if, whenever $ab\in I$, then $a\in I$ or $b\in I$.

V.44 This is very similar to the condition in Definition IV.12. It is also the condition satified by prime numbers in $\mathbb{Z}$: an integer $p>1$ is prime if and only if $ab\in p\mathbb{Z}$ implies $a\in p\mathbb{Z}$ or $b\in p\mathbb{Z}$. Thus $p\mathbb{Z}$ is a prime ideal of $\mathbb{Z}$. Note that in Definition V.43 we do not allow $I=R$ (that is, we insist that I should be a proper ideal) but we do allow $I=0$.

Theorem V.45 Let R be a commutative ring and I an ideal of R. Then I is a prime ideal if and only if R/I is an integral domain.

Proof: If I is a prime ideal and $(a+I)(b+I) = 0_{R/I}$ then $ab + I = I$ so $ab \in I$ so $a \in I$ or $b \in I$, meaning that $a + I$ or $b + I$ is $0_{R/I}$. Therefore R/I is an integral domain.

Conversely, if R/I is an integral domain and $ab \in I$, then $(a+I)(b+I) = 0_{R/I}$ so $a + I = 0_{R/I}$ or $b + I = 0_{R/I}$, i.e. $a \in I$ or $b \in I$.

Definition V.46 Let R be a commutative ring. A proper ideal I is called a *maximal ideal* if every ideal $J \supseteq I$ is either I or R.

Theorem V.47 Let R be a commutative ring and I an ideal of R. Then I is a maximal ideal if and only if R/I is a field.

Proof: Suppose that I is maximal: we shall show that $a + I$ has an inverse unless $a \in I$. Consider the ideal $J = \langle a, I \rangle$: that is, the smallest ideal of R that contains both a and all of I. Note that $J = \{ra + b \mid r \in R,\ b \in I\}$. Since I is maximal, J must be equal to either R or I. If $J = I$ then $a \in I$. If $J = R$ then $1_R \in J$ so $1_R = ra + b$ so $1 + I = ra + I = (r + I)(a + I)$ so $r + I$ is the inverse of $a + I$.

Conversely, suppose that R/I is a field and that $J \supseteq I$ is an ideal containing I. If $J \neq I$ choose $a \in J \setminus I$ and consider $a + I \in R/I$. Since $a + I \neq 0_{R/I}$ it has an inverse $r + I$, say; but then $1_R \in (r + I)(a + I) = ra + I$ and $ra + I \subseteq J$ so $1_R \in J$ so $J = R$.

Corollary V.48 If I is a maximal ideal then I is a prime ideal.

Field of fractions

V.49 The basic integral domain is $\mathbb{Z}$ and we can think of $\mathbb{Z}$ as determining $\mathbb{Q}$. We want to do something similar starting with an arbitrary integral domain.

Consider the set $T = R \times R \setminus \{0\}$: a typical member of this is (a, b) with $b \neq 0$, but it is preparing to be $\frac{a}{b}$ really. With this in mind we define two binary operations $T \times T \to T$ given by

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

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

Notice that T with these operations is *not* a ring: for instance, if $b \in R$ is not a unit then $(1, b)$ does not have an additive inverse. This is because we are treating fractions as numerator/denominator pairs, without cancelling: we still think that $\frac{1}{2}$ and $\frac{2}{4}$ are different.

Lemma V.50 Define a relation $\sim$ on T by setting $(a,b) \sim (c,d) \iff ad = bc$. Then for all a, a', b, b', c, c', d, $d' \in R$ with b, b', d, $d' \neq 0$, we have

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

Proof: Notice that

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

(The second line follows from the first by using the conditions $(a,b) \sim (a',b')$, i.e. $ab' = a'b$, and the same for c and d.) But this says that $(ad+bc, bd) \sim (a'd'+b'c', b'd')$ and those are $(a,b)+(c,d)$ and $(a',b')+(c',d')$, so we have proved the part about addition.

Similarly for multiplication we notice that $ab'cd' = a'bcd' = a'bc'd$, and that says that $(ac,bd) \sim (a'c', b'd')$, i.e., that $(a,b)\cdot(c,d) = (a',b')\cdot(c',d')$ as required.

Definition V.51 We define $\mathcal{Q}(R) = T/\sim$, and we give the name a/b (or $\frac{a}{b}$) to the equivalence class $[(a,b)]$.

a/b is not quite the same thing as ab^{-1}, even if b^{-1} happens to exist, but it almost is.

Theorem V.52 Let R be an integral domain. The operations $+$ and $\cdot$ on T induce operations (also called $+$ and $\cdot$) on $\mathcal{Q}(R) = T/\sim$. These rules make $\mathcal{Q}(R)$ into a field, called the *field of fractions* of R, and the map $R \to \mathcal{Q}(R)$ defined by $a \mapsto \frac{a}{1}$ is an injective ring homomorphism.

Proof: It is easy to check that the operations $+$ and $\cdot$ are well defined. However, we are not taking the quotient of a ring by an ideal (T is not even a ring) so we have to check the ring axioms in $\mathcal{Q}(T)$ by hand.

Our convention that the classes are called $\frac{a}{b}$ does mean that addition and multiplication, as we have just defined them, are expressed in the usual way as addition and multiplication of fractions:

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

To check that $(\mathcal{Q}(R), +)$ is an abelian group, take $\frac{a}{b}$, $\frac{c}{d}$, $\frac{e}{f} \in \mathcal{Q}(R)$: then

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

so addition is associative. Addition is commutative in $\mathcal{Q}(R)$ because multiplication in the integral domain R is commutative (and addition is commutative) and hence

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

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

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

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

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

Next multiplication is associative because multiplication in R is associative, so

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

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

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

The other distributive law is similar, and the multiplicative identity is $\frac{1}{1}$. This proves that $\mathcal{Q}(R)$ with the given operations is a ring.

To check that $\mathcal{Q}(R)$ is a field, we need to check that it is commutative and not $\{0\}$ and nontrivial and that every nonzero element is a unit. Commutativity is easy: $\frac{a}{b}\frac{c}{d}=\frac{ac}{bd}=\frac{ca}{db}=\frac{c}{d}\frac{a}{c}$. It is not $\{0\}$ because $\frac{0}{1}\neq\frac{1}{1}$ (otherwise $0=1$ in R which is excluded).

It remains to show that every nonzero element $\frac{a}{b}$ has a multiplicative inverse. But

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

so $\mathcal{Q}(F)$ is a field.

Finally, $a\mapsto\frac{a}{1}$ is a homomorphism because $\frac{a}{1}+\frac{a'}{1}=\frac{a1+1a'}{1\cdot 1}=\frac{a+a'}{1}$ and $\frac{a}{1}\frac{a'}{1}=\frac{aa'}{1}$, and it is injective because $a\in R$ is in its kernel if and only if $\frac{a}{1}=\frac{0}{1}$, which immediately gives $a=0$.

Example V.53 Apart from $\mathcal{Q}(\mathbb{Z})=\mathbb{Q}$, the most familiar example of this is $\mathcal{Q}(K[t])$ for K a field, which is the field $K(t)$ of rational functions.

VI Factorisation in integral domains

Unique factorisation of integers into primes is a very useful feature of $\mathbb{Z}$. We formulate an analogous statement for integral domains, and establish that it holds in some cases (but not all).

Primes and irreducibles

VI.1 Integral domains are very common and are significantly better behaved than general rings. For this whole section, R denotes an integral domain: in particular, $0_R \neq 1_R$.

Example VI.2 We've already seen many examples of integral domains:

(i) any field $\mathbb{F}$ is an integral domain (see IV.15);

(ii) the ring $\mathbb{Z}$ and the ring of Gaussian integers $\mathbb{Z}[i]$ are both integral domains by Lemma IV.23 because they are subrings of $\mathbb{C}$ and contain 1.

(iii) the ring $R[t]$ associated to an integral domain R is an integral domain.

Lemma VI.3 Let R be a commutative ring such that $0 \neq 1$. Then R is an integral domain if and only if it satisfies the *cancellation property*:

$$\text{if } a,\, b,\, c \in R \text{ and } a \neq 0, \text{ then } ab = ac \Rightarrow b = c.$$

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 $b = c$.

In the other direction, let R be a commutative ring such that $0 \neq 1$, and assume the cancellation property. Suppose $ab = 0$ and $a \neq 0$. Then $ab = 0 = a \cdot 0_R$, and since $a \neq 0$ the cancellation property gives $b = 0$, which shows that R is an integral domain.

Definition VI.4 Let $a,\, b \in R$. We say that a *divides* b (or that b is *divisible by* a), and write $a|b$, if there exists $c \in R$ such that $b = ac$.

The $|$ in $a|b$ is a verb, and $a|b$ is a sentence which may be true or false, but is not, ever, the name of an element of R.

Lemma VI.5 Suppose $a,\, b \in R$. The following are equivalent:

(i) $a|b$

(ii) $b \in aR$

(iii) $bR \subseteq aR$.

Proof: (i)⇒(ii): f $a|b$ then there exists $c \in R$ such that $b = ca = ac \in aR$. (ii)⇒(iii) aR is an ideal so if $b \in aR$ then $br \in aR$ for all $r \in R$, so $bR \subseteq aR$. (iii)⇒(i) If $bR \subseteq aR$ then $b = b1_R \in aR$ so $b = ac$ for some $c \in R$, so $a|b$.

VI.6 Lemma VI.5 says that "dividing" means "generating a bigger ideal": for example, 6 divides 12 and every multiple of 12 is also a multiple of 6

Lemma VI.7 Let R be an integral domain and let a, $b \in R$. Then $aR = bR$ if and only if $a = ub$ for some unit $u \in R^*$. In particular, $uR = R$ if and only if $u \in R^*$.

Proof: If $aR = bR$ then $aR \subseteq bR$ and $bR \subseteq aR$, so by Lemma VI.5 $b|a$ and $a|b$. Thus there exist u, $v \in R$ such that $a = ub$ and $b = va$, so $a = uva$. If $a = 0$, then $b = 0$ and there is nothing to prove. Otherwise, the cancellation law gives $uv = 1$, so u is a unit in R.

Conversely, if $a = ub$ for some unit $u \in R^*$ then $a \in bR$, so $aR \subseteq bR$. Since u is a unit, we may multiply $a = ub$ by u^{-1} to obtain $b = u^{-1}a = au^{-1}$: this gives $b \in aR$ and hence $bR \subseteq aR$. Hence $aR = bR$ as required. The final statement is just the case $a = 1$.

Definition VI.8 Let R be an integral domain. Let $p \in R$ be nonzero and not a unit. Then we say

(i) p is a *prime* if $p|ab \Rightarrow p|a$ or $p|b$.

(ii) p is *irreducible* if $p = ab \Rightarrow a \in R^*$ or $b \in R^*$.

We say that p is *reducible* if it is not irreducible, i.e., if there exist a, $b \in R$ such that $p = ab$ where neither a nor b is a unit: this is what is usually called "composite" in the case $R = \mathbb{Z}$.

Example VI.9 (i) The prime elements in $\mathbb{Z}$ are

$$\{\dots, -11, -7, -5, -3, -2, 2, 3, 5, 7, 11, \dots\},$$

i.e., ± 1 (a unit!) times the (positive) prime numbers. The irreducible elements are the same ones.

(ii) If $R = \mathbb{F}$ is a field then every nonzero element is a unit, so $\mathbb{F}$ contains neither primes nor irreducibles.

(iii) $t^2 + 1 \in \mathbb{R}[t]$ is irreducible (and, in fact, prime), but $t^2 + 1 \in \mathbb{C}[t]$ is reducible because $t^2 + 1 = (t + i)(t - i)$.

(iv) If R is an integral domain and $p \in R$ is nonzero then p is a prime if and only if pR is a prime ideal. (But $0R = \{0\}$ is a prime ideal, even though we have defined 0 not to be a prime: this just turns out to be convenient most of the time.)

Proposition VI.10 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 $p|b$, so without loss of generality $p|a$, say $a = pc$. Then $p = p \cdot 1_R = ab = pcb$, and the cancellation property gives $cb = 1$, so b is a unit. This shows that p is irreducible.

The converse is not true in general.

Euclidean domains and PIDs

Definition VI.11 Let R be an integral domain. A *Euclidean valuation* on R is a map $\nu\colon R \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ such that:

(i) if $f, g \in R \setminus \{0\}$ then $\nu(f) \leq \nu(fg)$; and

(ii) 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 $\nu(r) < \nu(g)$.

A *valuation* is a function satisfying (i) but not necessarily (ii).

Definition VI.12 We say that R is a *Euclidean domain* if it has a Euclidean valuation.

Example VI.13 (i) Any field is trivially a Euclidean domain because we may take $\nu(a) = 1$ for all $a \in \mathbb{F}$.

(ii) The absolute value $\nu(n) = |n|$ provides a Euclidean valuation on $\mathbb{Z}$, so $\mathbb{Z}$ is a Euclidean domain.

(iii) For $\mathbb{F}$ a field, the degree of a polynomial $\nu(f(t)) = \deg f(t)$ provides a Euclidean valuation on $\mathbb{F}[t]$, so $\mathbb{F}[t]$ is a Euclidean domain. This is the reason for the choice of the letters f, g, q and r in Definition VI.11.

(iv) $\mathbb{Z}[i]$ is a Euclidean domain.

Definition VI.14 Recall from Definition V.13 that an ideal I of a commutative ring R is a *principal ideal* if $I = aR$ (also denoted $\langle a \rangle$ if R is understood) for some $a \in R$. An integral domain R is called a *principal ideal domain* or *PID* if and only if every ideal in R is principal.

Lemma VI.15 Let R be a nonzero commutative ring. Then R is a field if and only if the only ideals of R are $\{0_R\} = \langle 0_R \rangle$ and $R = \langle 1_R \rangle$.

Proof: If R is a field and I is a nonzero ideal then choose any $0 \neq u \in I$. Since R is a field, u is a unit so $uR = R$ by VI.7. But $uR \subseteq I \subseteq R$, so $I = R$.

Conversely, if R is not a field then let $a \neq 0$ be any nonzero non-unit. Then $aR \neq R$ by VI.7 and $aR \neq \{0\}$ either.

Theorem VI.16 Let R be a Euclidean domain. Then R is a PID.

Proof: Denote the Euclidean valuation on R by ν and suppose I is a nonzero ideal in R.

Consider the image $\nu(I \setminus \{0\}$, i.e. $\{\nu(a) \in \mathbb{Z}_{\geq 0} \mid a \in I, a \neq 0\}$. This is a nonempty subset of $\mathbb{Z}_{\geq 0}$, so it has a least element σ.

Choose $g \in I$ such that $\nu(g) = \sigma$: in other words, we choose $0 \neq g \in I$ with $\nu(g)$ as small as possible, so $\nu(f) \geq \nu(g)$ for all $0 \neq f \in I$. If we take any $f \in I$, then since R is a Euclidean domain there exist $q, r \in R$ such that $f = qg + r$,

and $r = 0$ or $\nu(r) < \nu(g) = \sigma$. But if $r \neq 0$ then $r = f - qg \in I$, so $\nu(r) \geq \sigma$. This is a contradiction, so we must have $r = 0$, but then $f = qg \in gR$. Since f was arbitrary, that means $I \subseteq gR$; but $g \in I$ so we also have $gR \subseteq I$. Hence $I = gR$ and so I is a principal ideal.

Example VI.17 Theorem VI.16 implies that the following rings are PIDs:

(i) any field (this also follows from VI.15);

(ii) the ring of integers $\mathbb{Z}$;

(iii) the polynomial ring $\mathbb{F}[t]$ with coefficients in a field $\mathbb{F}$;

(iv) the ring of Gaussian integers $\mathbb{Z}[i]$.

Example VI.18 The integral domain $R = \mathbb{Z}[t]$ is not a PID, so it it is also not a Euclidean domain. It is slightly harder to produce a PID that is not a Euclidean domain. One example is $\mathbb{Z}[\theta]$ for $\theta = \frac{1}{2} + i\frac{\sqrt{19}}{2}$. We shall not prove this, but for reasons of space, not difficulty.

Properties of PIDs

Definition VI.19 Let R be a PID. Two elements $a, b \in R$ are said to be *coprime* if every common factor is a unit; by this, we mean that if $d|a$ and $d|b$, then d is a unit.

Lemma VI.20 Let R be a PID and let $a, b \in R$ be coprime. There exist $r, s \in R$ such that $1 = ra + sb$.

Proof: Consider the ideal $aR+bR$. Since R is a PID, there exists $d \in R$ such that $aR + bR = dR$: then $aR \subset dR$ so $d|a$ by VI.5 and similarly $d|b$. By hypothesis, then d is a unit, so $dR = R$ by VI.7: in particular, $1 \in dR = aR + bR$, which is to say that there exist $r, s \in R$ such that $1 = ra + sb$.

Proposition VI.21 Let R be a PID. Then every irreducible element in R is prime.

Proof: Suppose that p is irreducible and that $p|ab$, and that p does not divide a. We need to show that $p|b$.

We claim first that a and p are coprime. To see this, let d be a common factor of a and p: say $p = cd$ and $a = ed$. Since p is irreducible, we know that either c or d is a unit. But if c is a unit, then $a = ed = ec^{-1}cd = ec^{-1}p$ so $p|a$ which is contrary to the assumptions. Therefore d is a unit, so a and p are coprime.

Now by VI.20 there exist $r, s \in R$ such that $1 = ra + sp$. But now $b = 1 \cdot b = (ra+sp)\cdot b = rab+psb$, and we assumed that ab is divisible by p: so b is divisible by p, as required.

Theorem VI.22 Let R be a PID. If p is irreducible then R/pR is a field.

Proof: The ring R/pR is commutative because R is commutative, and it is nonzero because $pR \neq R$ by VI.7. It remains to show that every nonzero element of R/pR is a unit.

Choose $a \in R$ and consider the ideal $pR + aR = dR$ for some d (since R is a PID). This tells us that $pR \subseteq dR$, that is $d|p$, so $p = de$ for some e, and also that $d|a$. Since p is irreducible, either d or e must be a unit.

If e is a unit, then $a \in dR = deR = pR$ so $p|a$ and in that case $a + pR \in R/pR$ is zero.

If e is not a unit, then d is a unit; but $d \in pR + aR$ so $d = rp + sa$ for some r and s. So then $1 = d^{-1}rp + d^{-1}sa$, which means that $d^{-1}s + pR$ is an inverse of $a + pR$ in R/pR.

Unique factorisation domains

Definition VI.23 An integral domain R is called a *unique factorisation domain* or *UFD* if

(i) every nonzero nonunit element in R can be written as the product of finitely many irreducibles in R; and

(ii) 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 $r_i R = r'_i R$ for $1 \leq i \leq s$.

VI.24 By VI.7, the condition VI.23(ii) means that the factorisation into irreducibles is unique up to reordering and multiplying the factors by units.

Example VI.25 It is not usually easy to tell whether a given ring is a UFD. Some well-known examples are

(i) $\mathbb{Z}$ is a UFD;

(ii) $\mathbb{R}[t]$ is a UFD;

(iii) $\mathbb{Z}[t]$ is a UFD;

(iv) $\mathbb{Z}[i]$ is a UFD;

(v) $\mathbb{Z}[\sqrt{-5}]$ is not a UFD – in fact $6 = 2{\cdot}3 = (1+\sqrt{-5}){\cdot}(1-\sqrt{-5})$ is an example of two different factorisations of the same element into irreducibles.

Proposition VI.26 Let R be a UFD. Then $p \in R$ is irreducible if and only if it is prime.

Proof: Every prime is irreducible by Proposition VI.10, since R is an integral domain.

Conversely, let $p \in R$ be irreducible, and suppose that $p|ab$: say $ab = cp$ for some $c \in R$. We want to show that $p|a$ or $p|b$. We may as well assume that

neither a nor b is a unit, since if a is a unit then $b = a^{-1}cp$ and so $p|b$. Also $p|0$ so we may assume that a and b are nonzero.

Using Definition VI.23(i), we may take irreducible factorisations of a as $a = p_1 \cdots p_k$, of b as $b = q_1 \cdots q_l$ and of c as $c = r_1 \cdots r_m$. (Note that these irreducibles are not necessarily distinct!) Now we have two factorisations of ab:

$$p_1 \cdots p_k \cdot q_1 \cdots q_l = ab = cp = r_1 \cdots r_m \cdot p.$$

According to Definition VI.23(ii) and Lemma VI.7, there must be a a unit $u \in R$ such that either $p_i = up$ for some $1 \leq i \leq k$ or $q_j = up$ for some $1 \leq j \leq l$. In the first case $p|a$, and in the second case $p|b$.

Theorem VI.27 Let R be a PID. Then R is a UFD.

Proof: We need to check Definition VI.23(i) and Definition VI.23(ii). Both parts are nontrivial. We begin with VI.23(i), which says that factorisations exist.

Let $a \in R$ be a nonzero, nonunit element. Let us say, temporarily, that an element of R is *factorisable* if it can be written as the product of finitely many irreducibles, and suppose that a is unfactorisable. Then it is certainly reducible, so we can write $a = a_0 = a_1 b_1$, with a_1 and b_1 non-units: moreover, at least one of a_1 and b_1 must be unfactorisable as well. Without loss of generality we may assume that a_1 is unfactorisable, and we continue in this way: each a_{j+1} is chosen to be an unfactorisable factor of a_j, and $a_j = a_{j+1}b_{j+1}$ with b_{j+1} not a unit.

So we have $\ldots a_n|a_{n-1}|a_{n-2}|\ldots|a_1|a = a_0$, or, using Lemma VI.5

$$a_0R \subseteq a_1R \subseteq \ldots \subseteq a_{n-2}R \subseteq a_{n-1}R \subseteq a_nR \subseteq \ldots.$$

Next, we take $I = \bigcup_{j=0}^{\infty} a_jR$. We claim that I is an ideal in R: it obviously contains a_0 so it is not empty. Suppose that c, $d \in I$ and $r \in R$: it is enough to show that $rc \in I$ and $c - d \in I$, by V.9. But since $c \in I$ we know that $c \in a_nR$ for some n, and similarly $d \in a_mR$ for some m. Without loss of generality we assume $m \geq n$. Then $a_nR \subseteq a_mR$, so $c \in a_mR$ as well as $d \in a_mR$. But a_mR is an ideal, so $rc \in a_mR \subseteq I$ and $c - d \in a_mR \subseteq I$, so I is an ideal.

Because I is an ideal and R is a PID, we have $I = eR$ for some $e \in R$. But then $e \in I$, so $e \in a_iR$ for some i, and then

$$a_{i+1}R \subseteq I = eR \subseteq a_iR \subseteq a_{i+1}R$$

so $a_{i+1}R = a_iR$. According to Lemma VI.7 that implies that $a_i = ua_{i+1}$ for some unit u, but we also know that $a_i = b_{i+1}a_{i+1}$. By the cancellation property Lemma VI.3, that implies $b_{i+1} = u$; but b_{i+1} was chosen to be not a unit.

This is a contradiction, so no unfactorisable elements exist, and we have checked Definition VI.23(i).

Now we check Definition VI.23(ii). Suppose that there do exist elements of R violating this condition, so that we can find equations

$$vp_1 \cdots p_s = q_1 \cdots q_t$$

where the p_i and the q_j are irreducibles, v is a unit and the left-hand side is not just a rearrangement of the right-hand side, and we assume without loss of generality that $0 < s \leq t$. Among all such equations, we choose a shortest one: one for which t is as small as possible.

We have $p_s | q_1 \cdots q_t$. By Proposition VI.26, p_s is prime, so it divides one of the q_j: without loss of generality we may assume that $p_s | q_t$. However, q_t is irreducible, so $q_t = up_s$ for some unit. So now we have

$$u^{-1} v p_1 \cdots p_s = q_1 \cdots q_{t-1} \cdot p_s.$$

Cancelling the p_s gives a shorter violation of Definition VI.23(ii): one with only $t-1$ irreducibles q_j. This is a contradiction, so no such violations exist.

VI.28 It is somewhat easier to prove that a Euclidean domain is a UFD, because we can use the valuation for inductions, but the argument is essentially the same and it is worth taking the extra care to prove this stronger result.

Corollary VI.29 If $a \in \mathbb{Z}$ and $a > 1$ then a can be written as $a = \prod p_i^{n_i}$, where the p_i are distinct prime numbers and the n_i are positive integers. The primes p_i and their exponents n_i are uniquely determined (up to order).

This follows from Theorem VI.27 and the facts that $\mathbb{Z}$ is a Euclidean domain, hence a PID, and that the units in $\mathbb{Z}$ are ± 1 so every nonzero ideal has a unique positive generator.

Example VI.30 Not every UFD is a PID: in fact $\mathbb{Z}[t]$ is a UFD but not a PID.

VI.31 So far we have seen that shown that

(i) Fields are Euclidean domains — Example VI.13(i)

(ii) Euclidean domains are PIDs — Theorem VI.16

(iii) PIDs are UFDs — Theorem VI.27

(iv) UFDs are integral domains — Definition VI.23

and examples to show that the reverse implications do not hold.

UFDs and polynomial rings

VI.32 We want to investigate polynomial rings where the coefficients come from a UFD. For the rest of this section, we assume that R is a UFD.

Definition VI.33 An ideal in a commutative ring A is said to be *finitely generated* if it is of the form $a_1 A + \cdots + a_k A$ (also written $\langle a_1, \ldots, a_k \rangle$) for some finite set $\{a_1, \ldots, a_k\} \subset A$.

VI.34 Many important rings are *noetherian* (named after Emmy Noether): this means that all their ideals are finitely generated. We shan't need to impose this condition because the ideals we are about to consider are finitely generated anyway.

Lemma VI.35 Let R be a UFD and suppose that $I = \langle a_1, \ldots, a_k \rangle$ is a finitely generated ideal. Then there is a unique smallest principal ideal C containing I: that is, C is principal, $C \supseteq I$, and if $C' \supseteq I$ is another principal ideal containing I, then $C' \supseteq C$.

Proof: We proceed by induction on k. If $k = 1$ there is nothing to prove. Otherwise, suppose that $C_1 = \langle c_1 \rangle$ is the least principal ideal containing $I_1 = \langle a_1, \ldots, a_{k-1} \rangle$, which exists by the induction hypothesis. (Note that $I_1 \subseteq I$.) Since R is a UFD, there are finitely many irreducibles $p_1, \ldots, p_m$ such that $p_j | c_1$. For each p_j, we let $n_j \in \mathbb{Z}_{\geq 0}$ be the largest integer such that $p_j^{n_j} | c_1$ and $p_j^{n_j} | a_k$, and we take $c = p_1^{n_1} \cdots p_m^{n_m}$. I claim that $C = \langle c \rangle$ has the required properties.

Firstly, C is a principal ideal. Second, again because R is a UFD, $c | c_1$ so $C \supseteq C_1$, so $C \supseteq I_1$. Third, $c | a_k$, so $C \supseteq I$. Finally, suppose $C' = \langle c' \rangle \supseteq I$. Then $C' \supseteq I_1$ because $I \supseteq I_1$, so $C' \supseteq C_1$ by definition of C_1; so $c' | c_1$. Also, $C' \supseteq \langle a_k \rangle$, so $c' | a_k$. Now if p^n is a power of an irreducible and $p^n | c'$ then $p^n | c_1$ and $p^n | a_k$, so $p^n | p_j^{n_j}$ for some j: in other words, all the irreducible factors of c' divide c to at least the same power. So, by unique factorisation, $c' | c$: that is, $C' \supseteq C$.

VI.36 The element c in Lemma VI.35 is the product of all the powers of irreducibles that divide all of the a_i, so it should be thought of as the hcf of the a_i: we can't say it that way because we do not have a notion of "highest" until Lemma VI.35 gives us one.

Definition VI.37 A nonconstant polynomial $g = \sum_{i=0}^{n} a_i t^i \in R[t]$ is *primitive* if the only common divisors of all the coefficients of g are units in R.

VI.38 In light of unique factorisation in R, it is equivalent to say that g is primitive if and only if no irreducible element of R divides all coefficients of g.

If f is an arbitrary polynomial, then according to Lemma VI.35 the smallest principal ideal C_f containing all its coefficients is generated by an element $c \in R$, unique up to a unit factor, which is called the *content* of f, and then $f = cg$ with g primitive.

Another way to express Definition VI.37 is to say that g is primitive if the content of g is a unit (so the ideal C_g is R).

Example VI.39 $t^3 + 2t - 1 \in \mathbb{Z}[t]$ is primitive and so is $t^2 + 6t - 3$, whereas $3t^3 + 6t - 3 \in \mathbb{Z}[t]$ is not.

Proposition VI.40 Let R be a UFD. The product of finitely many primitive polynomials in $R[t]$ is primitive.

Proof: It suffices to prove the result for two primitive polynomials $f = \sum_{i=0}^{n} a_i t^i$ and $g = \sum_{i=0}^{m} b_i t^i$. Suppose that $fg = \sum_{l=0}^{m+n} d_l t^l$ is not primitive: then the

content $c \in R$ of fg is not a unit, so it has an irreducible factor $p|c$. Thus $p|d_l$ for any l. Since f and g are primitive, p does not divide all of the coefficients of either f or g. Let a_{i_0} and b_{j_0} be the first coefficients of f and g respectively that are not divisible by p, so $p|a_i$ if $i < i_0$ and $p|b_j$ if $j < j_0$.

Now we take $l_0 = i_0 + j_0$ and consider the coefficient of t^{l_0} in the product fg, namely

$$d_{l_0} = (a_0 b_{l_0} + \cdots + a_{i_0-1} b_{j_0+1}) + a_{i_0} b_{j_0} + (a_{i_0+1} b_{j_0-1} + \cdots + a_{l_0} b_0).$$

Rearranging this, we get

$$\begin{aligned} a_{i_0} b_{j_0} &= -d_{l_0} + (a_0 b_{l_0} + \cdots + a_{i_0-1} b_{j_0+1}) + (a_{i_0+1} b_{j_0-1} + \cdots + a_{l_0} b_0) \\ &= -d_{i_0+j_0} + \sum_{i<i_0} a_i b_{l_0-i} + \sum_{j<j_0} a_{l_0-j} b_j, \end{aligned}$$

and p divides all the terms on the right-hand side. So $p|a_{i_0} b_{j_0}$; but p is irreducible and hence prime by Proposition VI.26, so p must divide either a_{i_0} or b_{j_0}, which is a contradiction.

Corollary VI.41 Let R be a UFD with field of fractions $F = \mathcal{Q}(R)$, and let $f \in R[t]$. Then f is irreducible if and only if either

(i) f is an irreducible element of R, or

(ii) f is primitive in $R[t]$ and irreducible in $F[t]$.

Proof: ($\Rightarrow$) From VI.38 we have $f = c \cdot g$ for $c \in R$ and $g \in R[t]$ primitive. Since f is irreducible, either c or g must be a unit in $R[t]$.

(i) If g is a unit in $R[t]$, then $g \in R$ and hence $f \in R$. Then since f is irreducible in $R[t]$ it is also irreducibile in R.

(ii) If c is a unit then $f = cg$ is primitive because g is primitive. Moreover, $f \notin R$, since otherwise it would be primitive and thus a unit in R, so not irreducible. So f is primitive in $R[t]$ and of positive degree: such an f is irreducible in $F[t]$ (Exercise!).

($\Leftarrow$) If $f \in R$ is irreducible then it is irreducible in $R[t]$, so it remains to prove the result when $f \in R[t]$ is primitive in $R[t]$ and irreducible in $F[t]$. Since f is irreducible in $F[t]$, it is nonzero and a nonunit in $F[t]$, so it has positive degree and is therefore nonzero and a nonunit in $R[t]$ also.

Suppose now that there exist $g, h \in R[t]$ such that $f = g \cdot h$. We know f is irreducible in $F[t]$, so without loss of generality g is a unit in $F[t]$: that is, $g = a \in F^*$ because the units of $F[t]$ are the units of F. Since $F \cap R[t] = R$ that tells us that $a \in R$. From $f = ah$ we see that a divides the content of f, but f is primitive in $R[t]$, so its content is a unit. Hence $a = g$ is a unit in R, and thus a unit in $R[t]$, and we have shown that f is irreducible in $R[t]$.

Theorem VI.42 If R is a UFD, then the polynomial ring $R[t]$ is also a UFD.

Proof: For Definition VI.23(i), let $f \in R[t]$ be a nonzero nonunit. If $f \in R$, then it is a nonzero nonunit in R, and since R is a UFD, f can be written as the product of finitely many irreducibles in R which are necessarily irreducibles in $R[t]$.

Otherwise, f has positive degree. We may regard f as an element of $F[t]$ where $F = \mathcal{Q}(R)$. Since F is a field, $F[t]$ is a UFD by Example VI.17 and Theorem VI.27. Accordingly, we write $f = q_1 q_2 \cdots q_k$ for irreducible elements $q_1, \ldots, q_k \in F[t]$. We clear denominators: take d_i to be the product of all the denominators of all the coefficients of q_i and put $f_i = d_i q_i \in R[t]$. Then $df = f_1 f_2 \cdots f_k$ for some $d = d_1 \cdots d_k \in R$.

Then we use VI.38 to write $f_i = c_i g_i$ with $c_i \in R$ (the content of f_i) and $g_i \in R[t]$ primitive, and similarly $f = cg$, so that

$$dcg = df = f_1 \cdots f_k = (c_1 \cdots c_k) \cdot g_1 \cdots g_k.$$

Notice that in $F[t]$ we have $g_i = \frac{d_i}{c_i} q_i$ which is irreducible (in $F[t]$) as q_i is irreducible and $\frac{d_i}{c_i} \neq 0$ is a unit. By Corollary VI.41, therefore, g_i is irreducible in $R[t]$, and the product $g_1 \cdots g_k$ is primitive by Proposition VI.40, so the uniqueness part of Lemma VI.35 shows that there exists a unit $u \in R$ such that $dcu = c_1 \cdots c_k$.

So $df = dcug_1 \cdots g_k$, so $f = ucg_1 \cdots g_k$ and since $c \in R$ which is a UFD we can factorise $cu = r_1 \cdots r_l$ with $r_i \in R$ irreducible or a unit, and thus

$$f = r_1 \cdots r_l \cdot g_1 g_2 \cdots g_k,$$

with each factor irreducible or a unit in $R[t]$ as required for Definition VI.23(i).

To show Definition VI.23(ii), suppose $f \in R[t]$ admits two such decompositions

$$r_1 \cdots r_l \cdot g_1 \cdots g_k = r'_1 \cdots r'_m \cdot g'_1 \cdots g'_n.$$

These two polynomials have the same content (up to a unit factor), so $r_1 \cdots r_l = u \cdot r'_1 \cdots r'_m$ for some unit $u \in R$. Since R is a UFD, we have $l = m$ and (after permuting indices) $r_i R = r'_i R$ for $1 \leq i \leq l$. Similarly, the primitive part of f is unique up to multiplication by a unit, so there exists a unit $u' \in R$ such that $g_1 \cdots g_k = u' \cdot g'_1 \cdots g'_n$.

By Corollary VI.41, each g_i, $g'_j \in F[t]$ is irreducible: but $F[t]$ is a UFD because F is a field, so $k = n$ and (after permuting indices) $g_i F[t] = g'_i F[t]$ for $1 \leq i \leq k$. Now Lemma VI.7 gives a unit $u_i \in F$ such that $g_i = u_i g'_i \in R[t]$. But g_i and g'_i are primitive, so comparing contents shows that u_i is a unit in R.

Corollary VI.43 Let K be a field. Then $K[x_1, \ldots, x_n]$ is a UFD.

Proof: K is a PID, so it is a UFD. The result follows immediately by induction using Theorem VI.42.

Example VI.44 $K[x_1, \ldots, x_n]$ is not a PID for $n \geq 2$.

Irreducible polynomials

VI.45 Depending on the field K, we may be able to tell whether certain polynomials in $K[x]$ are irreducible or not. If K is a finite field, then deciding whether $f \in K[x]$ is irreducible is a finite search (we simply check, for each polynomial g of degree less than $\deg f$, whether $g|f$ or not). If K or $\deg f$ is very large this may be impractical, but it is possible in many useful cases.

VI.46 If $f \in K[x]$ is an irreducible polynomial of degree d, then the ideal $\langle f \rangle$ generated by f is a maximal ideal so $K[x]/\langle f \rangle$ is a field, according to Theorem V.47. Suppose that $K = \mathbb{F}_p$. It is fairly easy to see that $\mathbb{F}_p[x]/\langle f \rangle$ has p^d elements. Two other facts are less obvious and will merely be stated here. First, the isomorphism class of $\mathbb{F}[x]/\langle f \rangle$ depends only on d (and of course p, which is the characteristic): that is, all irreducible polynomials of degree d give the same quotient field by this construction. Second, for each p and d, irreducible polynomials of degree d do exist. Moreover, any finite field can be constructed in this way. It follows that for each $q = p^d$ there is up to isomorphism a unique finite field $\mathbb{F}_q$ with q elements, of charactristic p, and that there are no other finite fields. Notice that if $d > 1$ the rings $\mathbb{Z}/q$ and $\mathbb{F}_q$ are very different: the former is not even an integral domain.

VI.47 We should expect actually factorising polynomials over $\mathbb{Q}$ (or over $\mathbb{Z}$, which is the same thing by Corollary VI.41) to be quite difficult, but we might hope to be able to test individual polynomials for irreducibility.

Proposition VI.48 Suppose that $f \in \mathbb{Z}[x]$ and $p \in \mathbb{Z}$ is a prime. There map $\mathrm{red}_p \colon \mathbb{Z}[x] \to \mathbb{F}_p[x]$ given by $\mathrm{red}_p(\sum_{i=0}^d a_i x^i) = \sum_{i=0}^d \overline{a_i} x^i$, where $\overline{a_i} \in \mathbb{F}_p = \mathbb{Z}/p$ is the reduction mod p of $a_i \in \mathbb{Z}$, is a ring homomorphism and if $\mathrm{red}_p(f)$ is irreducible in $\mathbb{F}_p[x]$ and of the same degree as f, then f is irreducible in $\mathbb{Z}[x]$.

Proof: red_p is simply the homomorphism whose kernel is $\langle p \rangle \lhd \mathbb{Z}[x]$. Suppose that $f = gh$ is reducible in $\mathbb{Z}[x]$. Then $\mathrm{red}_p(f) = \mathrm{red}_p(g)\,\mathrm{red}_p(h)$ so since $\mathrm{red}_p(f)$ is irreducible one of the factors, say $\mathrm{red}_p(g)$, must be constant. However, g itself is not constant so its leading coefficient (and indeed all its coefficients apart from the constant) must be divisible by p. But then the leading coefficient of f is divisible by p as it is the product of the leading coefficients of g and h, so $\deg \mathrm{red}_p(f) < \deg f$.

Theorem VI.49 Eisenstein's criterion Suppose that $f = \sum_{i=0}^d a_i x^i \in \mathbb{Z}[x]$ is of degree d and for some prime $p \in \mathbb{Z}$ we have $p|a_i$ for $0 \leq i < d$ but p does not divide a_d and p^2 does not divide a_0. Then f is irreducible in $\mathbb{Z}[x]$ (and therefore irreducible in $\mathbb{Q}[x]$).

Proof: Suppose that f is reducible in $\mathbb{Z}[x]$, so $f = gh$. Then, as in Proposition VI.48, $\mathrm{red}_p f = (\mathrm{red}_p g)(\mathrm{red}_p h)$, but $\mathrm{red}_p f = \overline{a_d} x^d$ by the hypotheses. Since $\mathbb{F}_p[x]$ is a UFD it follows that $\mathrm{red}_p g = bx^{\deg g}$ and $\mathrm{red}_p h = cx^{\deg h}$ for some $b, c \in \mathbb{F}_p$. In particular the constant terms of $\mathrm{red}_p g$ and $\mathrm{red}_p h$ are both zero, so the constant terms of g and h are both divisible by p. But then the constant term a_0 of f is divisible by p^2.

VI.50 Eisenstein's criterion can often be applied directly but a useful trick is to apply it not to $f(x)$ but to $f(x+n)$ for some $n \in \mathbb{Z}$. The map that sends $f(x)$ to $f(x+n)$ is an isomorphism from $\mathbb{Z}[x]$ to $\mathbb{Z}[x]$ so $f(x)$ is irreducible if and only if $f(x+n)$ is irreducible.

Thus, for example, $x^5 - 76x^4 - 1002x^3 - 4630x^2 - 9437x + 194$ is irreducible because replacing x with $x-3$ gives $x^5 - 91x^4 + 7x^3 + 14x^2 - 98x + 7147$ which satisfies the conditions of Eisenstein's criterion with $p = 7$.