4. Rings

4.1. **Basic properties.**

Definition 4.1. A *ring* is a set R with two binary operations $+, \cdot$ and distinguished elements $0, 1 \in R$ such that: (1) $(R, +)$ is an abelian group with neutral element 0; (2) $(R, \cdot)$ is a monoid with neutral element 1; (3) the *distributive laws*

$$a(b + c) = ab + ac, \quad (a + b)c = ac + bc$$

hold for all $a, b, c \in R$.

Here, we have already implicitly assumed the usual convention that multiplication binds more strongly than addition, so $ab + ac$ is really short-hand for $(ab) + (ac)$.

We write $-a$ for the additive inverse of an $a \in R$, and we then use the familiar and convenient subtraction notation for what is really the addition of an additive inverse: we write $a - b := a + (-b)$.

Exercise 4.1. Show that $-(ab) = (-a)b = a(-b)$, so in particular $-a = (-1)a = a(-1)$. Moreover, subtraction in a ring also obeys distributive laws: $a(b - c) = ab - ac$, $(a - b)c = ac - bc$

Example 4.1. Familiar examples of rings are $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ with the usual operations. Another example is given by $\mathbb{Z}_k$; you showed in Exercise 1.11 that $+, \cdot$ on $\mathbb{Z}_k$ have the required properties. In the definition, we do not insist that $0 \neq 1$, so a very trivial example of a ring is $R = \{0\}$; for aesthetic reasons, we usually prefer the notation $R = 0$.

Exercise 4.2. Show that (much to our relief) $R = 0$ is the only ring with $0 = 1$.

Example 4.2. For any $k \in \mathbb{Z}$, we can form the ring

$$\mathbb{Z}[\sqrt{k}] := \{a + b\sqrt{k} : a, b \in \mathbb{Z}\}.$$

We add and multiply elements of this ring as real numbers (if $k \geq 0$) or as complex numbers (if $k < 0$; in this case, we interpret $\sqrt{k} = i\sqrt{-k}$).

The ring $\mathbb{Z}[\sqrt{-1}]$ is called the ring of *Gaussian integers.* Its members are the complex numbers $a + ib$, $a, b \in \mathbb{Z}$.

Example 4.3. All examples so far are *commutative* rings in the sense that $ab = ba$ (addition is, of course, always commutative in a ring, so it is not necessary to draw special attention to this). For a non-commutative example, let R be any ring, and consider $M_n(R)$, the $n \times n$ matrices with entries in R. This is a ring with the operations that suggest themselves: entrywise addition and matrix multiplication ("row times column").

Exercise 4.3. Check this please. What are the additive and multiplicative identities of $M_n(R)$? Then show that $M_n(R)$ is not commutative if $n \geq 2$ and $R \neq 0$ (even if the original ring R was).

We are often interested only in special types of rings, with extra properties. I already mentioned *commutative* rings. In any ring R, we have that $0a = (0+0)a = 0a + 0a$, so by subtracting $0a$ from both sides we see that $0a = 0$. Similarly, $a0 = 0$. So far, so good; however, in a general ring, there could be $a, b \neq 0$ with $ab = 0$. This will not happen if a (or b) is invertible in the multiplicative monoid $(R, \cdot, 1)$: in that case, $ab = 0$ implies that $b = a^{-1}ab = a^{-1}0 = 0$.

We call the invertible elements of (the multiplicative monoid of) R *units,* and we denote the collection of units by $U(R)$. Then $(U(R), \cdot)$ is a group, as we saw in Chapter 2 (but show it again perhaps). Also, we just showed that $0 \notin U(R)$ if $0 \neq 1$. These considerations motivate the following new definitions:

Definition 4.2. A *domain* is a ring $R \neq 0$ with the property that $ab = 0$ implies that $a = 0$ or $b = 0$.

We call R a *division ring* or *skewfield* if $R^\times = R \setminus 0$ is a subgroup of $(R, \cdot)$. A *field* is a commutative division ring.

In other words, R is a division ring if $1 \neq 0$ and $U(R) = R^\times$. We established above that every division ring is a domain, but the converse need not hold. Let's just go through our list of examples one more time: $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ are (very familiar) fields. $\mathbb{Z}$ is clearly not a field: for example $2 \in \mathbb{Z}$ does not have a multiplicative inverse. However, $\mathbb{Z}$ is a domain.

Next, let's take another look at $R = \mathbb{Z}_k$. If k is composite, say $k = jm$ with $2 \leq j, m < k$, then $\mathbb{Z}_k$ is not a domain because $jm \equiv 0 \mod k$, so $jm = 0$ in $\mathbb{Z}_k$. On the other hand, if $k = p$ is a prime, then $\mathbb{Z}_p$ is a field. This follows from Proposition 1.9, which says (in our new terminology) that every $a \in \mathbb{Z}_p$, $a \neq 0$, has a multiplicative inverse.

Exercise 4.4. Classify the rings $\mathbb{Z}[\sqrt{k}]$ and $M_2(\mathbb{R})$ in the same way.

Exercise 4.5. Find $U(R)$ for $R = \mathbb{Z}$ and $R = \mathbb{Z}[\sqrt{-1}]$. Then show that $m \in U(\mathbb{Z}_k)$ precisely if $(m, k) = 1$.

Proposition 4.3. *Let $R \neq 0$ be a ring. Then R is a domain if and only if R has the* cancellation property: *$ab = ac$, $a \neq 0$ implies that $b = c$. In this case, $ba = ca$, $a \neq 0$, also implies that $b = c$.*

Exercise 4.6. Prove Proposition 4.3.

Exercise 4.7. Show that every finite domain is a division ring. (In fact, it is a field, but this is harder to prove.)

If $n \in \mathbb{Z}$ and $a \in R$ is an element of a ring, we can define $na \in R$ in a natural way: if $n \geq 1$, then $na := a + a + \ldots + a$ is the n-fold sum of a with itself. We also set $0a := 0$ and $na = -(|n|a)$ for $n < 0$. (This is an exact analog of the exponential notation a^n in groups that we introduced earlier, it just looks different typographically because we are now using additive notation for the group operation.)

Similarly, we write a^n for the product of n factors of a, and we set $a^0 := 1$. Here, we assume that $n \geq 0$; if $n < 0$, then it would be natural to define $a^n := (a^{-1})^{|n|}$, but this we can only do if a is a unit.

Proposition 4.4 (The binomial formula)**.** *Let R be a commutative ring. Then, for $a, b \in R$, $n \geq 0$,*

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}.$$

This is proved in the same way as for numbers (by a combinatorial argument or by induction).

Exercise 4.8. Show that the binomial formula (for $n = 2$, say) can fail in a non-commutative ring.

Example 4.4. We still haven't seen an example of a non-commutative division ring. A very interesting example is provided by the *quaternions.* We will introduce these as a subring of $M_2(\mathbb{C})$. In general, given a ring R, a *subring* S is defined as a subset $S \subseteq R$ with $1 \in S$, and whenever $a, b \in S$, then also $a - b, ab \in S$; equivalently, S is a subring precisely if S is an additive subgroup and a multiplicative submonoid of R.

Exercise 4.9. A *submonoid* N of a monoid M is defined as a subset $N \subseteq M$ with: (a) $1 \in N$; (b) if $a, b \in N$, then also $ab \in N$. This is not (exactly) the same as asking that N satisfies (b) and is a monoid itself with the multiplication inherited from M. Please explain.

We denote the quaternion ring by $\mathbb{H}$, in honor of *William Rowan Hamilton* (1805–1865), the discoverer of the quaternions. Hamilton observed that the field $\mathbb{C}$ can be viewed as the vector space $\mathbb{R}^2$, endowed with a multiplication that is compatible with the vector space structure.

In modern terminology, such a structure (a ring A that is also a vector space over a field F, and $c(xy) = (cx)y = x(cy)$ for all $c \in F$, $x, y \in A$) is called an *algebra.* The multiplicative structure of an algebra is determined as soon as we know what the products $e_j e_k$ of basis vectors are equal to because then general products $\sum c_j e_j \sum d_j e_j$ can be evaluated by multiplying out in the expected way.

So $\mathbb{R}^2$ has at most one algebra structure with $e_1e_1 = e_1$, $e_1e_2 = e_2e_1 = e_2$, $e_2e_2 = -e_1$, and if you change names to $e_1 \to 1$, $e_2 \to i$, you see that there indeed is one, and it is isomorphic to $\mathbb{C}$, as algebras. (It is also not hard to show that $\mathbb{C}$ is the only field that is a two-dimensional $\mathbb{R}$-algebra, up to isomorphism, so this closes the case $n = 2$.)

Hamilton now tried to find other $\mathbb{R}$-algebras that are fields (though of course this terminology didn't exist at the time). In more concrete terms, this assignment reads: fix $n \geq 3$, and then try to come up with values of the products e_je_k, $1 \leq j, k \leq n$, where e_j are the standard basis vectors of $\mathbb{R}^n$, that make $\mathbb{R}^n$ a field. Some ten years later, all Hamilton had to show for his efforts was what seems to be a very partial success: for $n = 4$, there is a division ring (but not a field), namely $\mathbb{H}$.

It turns out that there are reasons for this failure. More precisely, one can show that: (1) $\mathbb{R}^n$ can not be made a field in this way for any $n > 2$; (2) if we are satisfied with division rings rather than fields, then the quaternions $\mathbb{H}$ are an example, and here $n = 4$, but no other value $n > 2$ works.

It is in fact very easy to see, definitely with modern tools, that there are no fields that are finite-dimensional $\mathbb{R}$-algebras other than $\mathbb{R}$ itself and $\mathbb{C}$ and that $n = 3$ is completely hopeless, even if one is satisfied with division rings rather than fields. We'll discuss this in the next chapter. Hamilton's quest looks rather quixotic from a modern point of view.

We don't follow the historical development here; we now introduce $\mathbb{H}$ as the subring of $M_2(\mathbb{C})$ with these elements:

$$\begin{pmatrix} w & z \\ -\overline{z} & \overline{w} \end{pmatrix}, \qquad w, z \in \mathbb{C}$$

The bar denotes complex conjugation: $\overline{x + iy} = x - iy$, $x, y \in \mathbb{R}$

Exercise 4.10. Show that $\mathbb{H} \subseteq M_2(\mathbb{C})$ is a subring. Then show that $\mathbb{H}$ is not commutative.

To show that $\mathbb{H}$ is a division ring, we must show that every $x \in \mathbb{H}$, $x \neq 0$, has an inverse. The condition that $x \neq 0$ means that w, z are not both zero. This means that $\det x = |w|^2 + |z|^2 \neq 0$, so x is definitely an invertible matrix, but of course that isn't quite good enough because it only says that x had an inverse in $M_2(\mathbb{C})$. We must really make sure that this inverse is in $\mathbb{H}$, but that's very easy too because we can just explicitly compute

$$\begin{pmatrix} w & z \\ -\overline{z} & \overline{w} \end{pmatrix}^{-1} = \frac{1}{|w|^2 + |z|^2} \begin{pmatrix} \overline{w} & -z \\ \overline{z} & w \end{pmatrix} = \begin{pmatrix} u & v \\ -\overline{v} & \overline{u} \end{pmatrix},$$

with $u = \overline{w}/(|w|^2 + |z|^2)$, $v = -z/(|w|^2 + |z|^2)$, so we do have that $x^{-1} \in \mathbb{H}$, as required.

Exercise 4.11. The discussion above suggests that $\mathbb{H}$ is a four-dimensional algebra over $\mathbb{R}$. Please make this explicit.

Much of the basic material on groups just carries over to rings (or other algebraic structures) in a very straightforward way. We already defined *subrings.* If R, R' are rings, then a map $\varphi : R \to R'$ is called a *homomorphism* if $\varphi(a + b) = \varphi(a) + \varphi(b)$, $\varphi(ab) = \varphi(a)\varphi(b)$, $\varphi(1) = 1'$. Equivalently, we ask that φ is a homomorphism for the additive groups and the multiplicative monoids. An *isomorphism* is a bijective homomorphism; as before, we write $R \cong R'$ to express the fact that R, R' are isomorphic rings.

As an application of these notions, let us find the smallest subring P of a given ring R. We call P the *prime ring.* We must in fact also show that such an object (the smallest subring) exists.

Clearly, we must have $0, 1 \in P$, and just to obtain an additive subgroup, we must then put $n1$ into P for all $n \in \mathbb{Z}$. This, however, will do: $P = \{n1 : n \in \mathbb{Z}\}$ is a subring of R. Indeed, P is closed under addition and additive inverses by construction, and $(m1)(n1) = (mn)1 \in P$, so P is also closed under multiplication.

Exercise 4.12. The identity $(m1)(n1) = (mn)1$ that I just used looks ridiculously obvious, but perhaps there is slightly more to it than meets the eye. Can you please explain how I really obtained it?

By construction, the prime ring P has the property that if $Q \subseteq R$ is any subring, then $P \subseteq Q$.

The elements $n1$, $n \in \mathbb{Z}$, of P are either all distinct, or there exist $m, n \in \mathbb{Z}$, $m \neq n$ such that $m1 = n1$. In the second case, we can then also find an $n \geq 1$ with $n1 = 0$ (why?). The smallest such n is called the *characteristic* of R. If there is no $n \neq 0$ with $n1 = 0$, then we say that R has characteristic 0.

Exercise 4.13. Show that $P \cong \mathbb{Z}$ if $\mathrm{char}(R) = 0$ and $P \cong \mathbb{Z}_n$ if $\mathrm{char}(R) = n \geq 1$. *Suggestion:* Proceed as in our (first) discussion of cyclic groups.

Exercise 4.14. Show that if R is a domain, then the characteristic can only be zero or a prime.

Exercise 4.15. Show that it is not possible to define a multiplication on the abelian group $A = \mathbb{Q}/\mathbb{Z}$ that makes A a ring.

Exercise 4.16. Let R be a commutative ring of prime characteristic $p \geq 2$. Show that $(a + b)^p = a^p + b^p$ for arbitrary $a, b \in R$.

Let's move on to the next items on our list (basic group theory material, adapted to rings). We define a *congruence* on R as an equivalence relation $\equiv$ that is compatible with the ring structure in the sense that if $a \equiv a'$, $b \equiv b'$, then $a+b \equiv a'+b'$, $ab \equiv a'b'$. A congruence on a ring must in particular be a congruence (in the group theory sense of the word) of its additive group. This already shows us that congruences come from normal subgroups of $(R,+)$, but not necessarily all of these since a congruence on a ring has additional properties. We also observe that in fact all subgroups of $(R,+)$ are normal because this is an abelian group.

Let's work this out in more detail. Suppose $\equiv$ is a congruence on R. Let $I \subseteq R$ be the corresponding subgroup of $(R,+)$. More precisely, we know from Theorem 2.22(a) that $I = \overline{0}$, the equivalence class of $0 \in R$. We also know that $a \equiv b$ precisely if $a-b \in I$. Now if $r \in R$, $a \in I$, then, since $r \equiv r$, $a \equiv 0$, we must have that $ra \equiv r0 = 0$ and $ar \equiv 0$, or, equivalently, $ra, ar \in I$. This motivates:

Definition 4.5. A non-empty subset $I \subseteq R$ is called a (two-sided) *ideal* if whenever $a, b \in I$, $r \in R$, then $a-b, ar, ra \in I$.

A *one-sided left ideal* would be defined by the condition that $a-b, ra \in I$ in the above situation, and of course right ideals can be defined similarly. For commutative rings the distinction disappears, and this is the case we will be most interested in. In the sequel, ideal will mean two-sided ideal.

We now have the following analog of Theorem 2.22 for rings.

Theorem 4.6. *(a) Let $\equiv$ be a congruence on a ring R. Then $I = \overline{0}$ is an ideal, and $a \equiv b$ precisely if $a-b \in I$.*
(b) Let $I \subseteq R$ be an ideal. Then $a \equiv b$ precisely if $a-b \in I$ defines a congruence on R, and $I = \overline{0}$, the equivalence class of 0 with respect to this congruence.

Proof. We just proved part (a). Suppose now, conversely, that we are given an ideal I. Then I is in particular a (normal) subgroup of $(R,+)$, so we know already that $a \equiv b$ defined by the condition that $a-b \in I$ is an equivalence relation and a congruence of the additive group; also, the final claim, that $I = \overline{0}$, is clear from this, by taking $b = 0$. It remains to show that if $a \equiv a'$, $b \equiv b'$, then also $ab \equiv a'b'$. By assumption, $a' = a+x$, $b' = b+y$, with $x, y \in I$, so $a'b' = ab + x(b+y) + ay = ab + z$ with $z \in I$, as required. □

As in the case of groups, given an ideal I, we can form the quotient ring $\overline{R} = R/I$. Its elements are the cosets $\overline{a} = a + I$, $a \in R$, and

the ring operations are performed on the representatives a. With these operations, R/I is indeed a ring: the algebraic laws from R just carry over automatically to R/I because the operations are performed on representatives. For example,

$$\overline{a}(\overline{b}+\overline{c}) = \overline{a}(\overline{b+c}) = \overline{a(b+c)} = \overline{ab+ac} = \overline{ab}+\overline{ac} = \overline{a}\,\overline{b}+\overline{a}\,\overline{c}.$$

The natural map $q : R \to R/I$, $q(a) = a + I$, is a surjective homomorphism.

Exercise 4.17. Show this please.

Exercise 4.18. Define addition and multiplication on subsets of a ring R as $A + B = \{a + b : a \in A, b \in B\}$, $AB = \{ab : a \in A, b \in B\}$, as expected. Show that if we multiply two cosets $(a + I)(b + I)$ as sets in this way, we are not guaranteed to obtain $ab + I$, the product of these factors taken in the quotient ring R/I. Also, show that the distributive laws fail for subsets. (So this interpretation is best abandoned here.)

Theorem 4.7. *Let $\varphi : R \to R'$ be a homomorphism. Then $I = \ker(\varphi) := \{a \in R : \varphi(a) = 0'\}$ is an ideal. Moreover, φ factors through R/I:*

$$\begin{array}{ccc} R & \xrightarrow{\varphi} & R' \\ {\scriptstyle q}\downarrow & \nearrow_{\overline{\varphi}} & \\ R/I & & \end{array}.$$

The unique induced map $\overline{\varphi}$ is an injective homomorphism, and $\varphi(R) \cong R/I$.

Proof. We already know that $I = \ker(\varphi)$ is a subgroup of $(R, +)$, so to establish that I is an ideal, we need only show that $rk, kr \in I$ for $k \in I$, $r \in R$. This is clear because $\varphi(rk) = \varphi(r)\varphi(k) = \varphi(r)0 = 0$, and similarly for kr.

The rest of the argument is exactly the same as in the group case. We are forced to set $\overline{\varphi}(a + I) = \varphi(a)$; this gives a well defined map because φ produces the same output on every $a' \in a + I$. It's easy to see that $\overline{\varphi}$ is a homomorphism: for example, we have that

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

Finally, $\overline{\varphi}$ is injective by construction: the collection of those elements of R that get sent to the same image by φ becomes one point of R/I. □

Exercise 4.19. Give a somewhat more explicit version of this last part of the argument. Also, show that a homomorphism φ is injective precisely if $\ker(\varphi) = 0$.

Let R be a ring, and fix a subset $S \subseteq R$. We denote by $I = (S)$ the ideal generated by S; this is defined as the smallest ideal I with $I \supseteq S$. First of all, we need to make sure that there always is such an object. As in the group case, this follows from the representation $I = \bigcap J$, where the intersection is taken over all ideals $J \supseteq S$.

Exercise 4.20. Show that an intersection of ideals is an ideal itself. Then conclude that $\bigcap J$ is an ideal with the required properties.

We can again ask ourselves if it is also possible to build $I = (S)$ from the bottom up, by gingerly putting elements into it, but only those that we are sure must be in this ideal. Since (S) is supposed to be an ideal, we must have $asb \in I$ for $s \in S$ and arbitrary $a, b \in R$. An ideal is an additive subgroup, so we must also put sums (and differences) of such terms into (S). Fortunately, the process stabilizes here:

Proposition 4.8. *Let $S \subseteq R$. Then the ideal generated by S can be described as follows:*

$$(S) = \left\{ \sum_{j=1}^{n} a_j s_j b_j : n \geq 0, s_j \in S, a_j, b_j \in R \right\} \tag{4.1}$$

If R is commutative, then this simplifies to $(S) = \{\sum a_j s_j\}$; in particular, $(x) = \{ax : a \in R\}$.

Exercise 4.21. Prove Proposition 4.8. *Suggestion:* Model your argument on the proof of Proposition 2.9.

Note that in the non-commutative case, we may have to repeat generators in (4.1). For example, $axb + cxd$ need not equal a single term of the form exf; in the commutative case, such repetitions become unnecessary because the expression simplifies to ex, $e = ab + cd$.

Theorem 4.9. *Let $R \neq 0$ be a commutative ring. Then R is a field precisely if R has no ideals $I \neq 0, R$.*

Proof. Suppose that R is a field. If I is an ideal and $x \in I$, $x \neq 0$, then for any $a \in R$, we have that $a = (ax^{-1})x \in I$. So $I = R$ if $I \neq 0$.

Conversely, suppose that R has this property, and let $x \in R$ be any non-zero element. Then $(x) = R$ by assumption, so in particular, $1 \in (x)$, and now Proposition 4.8 shows that there exists $a \in R$ with $ax = 1$. This says that x is invertible. □

Exercise 4.22. Formulate and prove an analog of Theorem 4.9 for non-commutative rings ("$R \neq 0$ is a division ring if and only if ...").

Exercise 4.23. Show that $R = M_2(\mathbb{R})$ has no ideals $\neq 0, R$ (or you can do it for $M_n(\mathbb{R})$ if feeling more ambitious). Now $M_2(\mathbb{R})$ is certainly not a division ring (why not?). Are there any contradictions to what you showed in the previous Exercise?

What are the ideals of $R = \mathbb{Z}$? We already know that the subgroups of $(\mathbb{Z}, +)$ are $k\mathbb{Z} = \{kn : n \in \mathbb{Z}\}$. It is now clear that all of these are ideals also because if $a \in k\mathbb{Z}$, say $a = kn$, and $x \in \mathbb{Z}$, then $ax = k(nx) \in k\mathbb{Z}$, as required. All these ideals are generated (already as subgroups) by a single element k.

Definition 4.10. A *principal ideal* is an ideal that is generated by a single element. A ring R is called a *principal ideal domain* (PID) if R is a commutative domain and every ideal $I \subseteq R$ is a principal ideal.

Exercise 4.24. Show that $I = 0, R$ are principal ideals for any ring R.

What we have just shown about $\mathbb{Z}$ can now be summarized as follows:

Theorem 4.11. *$\mathbb{Z}$ is a PID. Its ideals are $(k) = \{kn : n \in \mathbb{Z}\}$, $k = 0, 1, 2, \ldots$.*

What are the congruences that correspond to these ideals, via Theorem 4.6? Fix $k \geq 0$ and consider $I = (k)$. By part (b) of the Theorem, $a \equiv b$ if and only if $a - b \in (k)$, which happens if and only $a - b$ is a multiple of k. In other words, we recover congruence modulo k, as introduced in Section 1.2. We also obtain the satisfying conclusion that these are in fact the only (ring) congruences on $\mathbb{Z}$.

Exercise 4.25. Show that every field is a PID.

Here's the ring version of the first isomorphism theorem:

Theorem 4.12. *Let I be an ideal in a ring R. Then the ideals (subrings) $J \supseteq I$ of R are in one-to-one correspondence to the ideals (subrings) of $\overline{R} = R/I$, via $J \mapsto \overline{J} = J/I = \{x + I : x \in J\}$. Moreover, if J is such an ideal, then $R/J \cong \overline{R}/\overline{J}$. An isomorphism φ can be obtained from the diagram*

$$\begin{array}{ccc} R & \longrightarrow & R/I = \overline{R}\,; \\ \downarrow & \overset{\varphi}{\swarrow} & \downarrow \\ R/J & \leftarrow - - & \overline{R}/\overline{J} \end{array}$$

here, we use the natural quotient maps along the solid arrows.

Exercise 4.26. Prove this. Follow the same strategy as in our proof of Theorem 3.4.

As an application, let's take one more look at the rings $\mathbb{Z}_k$. We know that we can interpret $\mathbb{Z}_k \cong \mathbb{Z}/(k)$. Theorem 4.12 now says that the ideals of $\mathbb{Z}/(k)$ exactly correspond to the ideals $I \subseteq \mathbb{Z}$ with $I \supseteq (k)$. Since $\mathbb{Z}$ is a PID, we also know that any such I is of the form $I = (n)$, for some $n \geq 0$. Now $(m) = \{mx : x \in \mathbb{Z}\}$, so $(n) \supseteq (k)$ precisely if $n|k$. So the ideals of $\mathbb{Z}/(k)$ correspond to the divisors of k. In particular, $R = \mathbb{Z}/(k)$ has no ideals $\neq 0, R$ if and only if k is a prime. We recover what we found earlier by less abstract arguments: $\mathbb{Z}_k \cong \mathbb{Z}/(k)$ is a field precisely if k is a prime.

4.2. **The field of fractions.** Recall that we can construct (the field) $\mathbb{Q}$ from (the ring) $\mathbb{Z}$ by forming fractions $q = a/b$, $a, b \in \mathbb{Z}$, $b \neq 0$. Fractions of this type need not be reduced, so here we must be prepared to identify a/b with a'/b' when $ab' = a'b$. The algebraic operations on these fractions are then defined in the expected way, and the whole procedure delivers a field. This construction works the same way in an abstract setting.

For the remainder of this chapter, we will be interested in commutative rings almost exclusively, so it will be convenient to adopt the following

> *Convention:* From now on, ring (domain) will mean commutative ring (domain).

Let R be a domain. I want to build what we will call the *field of fractions* $F(R)$; so $F_0 = \{a/b : a, b \in R, b \neq 0\}$ looks like a good starting point. Of course, division is not defined in a ring, so a/b doesn't make sense as "a divided by b." I really view a/b as a convenient notation for the pair (a, b). Still taking the transition from $\mathbb{Z}$ to $\mathbb{Q}$ as our guide, we will want to identify two formal fractions a/b, a'/b' if $ab' = a'b$. In anticipation of this, we introduce the relation $\sim$ by declaring $a/b \sim a'/b'$ if $ab' = a'b$.

Exercise 4.27. Show that $\sim$ is an equivalence relation on F_0.

By the Exercise, we can form the quotient space $F(R) = F_0/\sim$. Its elements are equivalence classes of formal fractions a/b. We can now make $F(R)$ a field by introducing the operations in the expected way: we put

$$a/b + c/d := (ad + bc)/bd, \qquad (a/b)(c/d) := (ac)/(bd); \tag{4.2}$$

as usual, we do not distinguish between equivalence classes and representatives in the notation. Before we proceed, we must check that $+, \cdot$

are indeed well defined: we must make sure that the right-hand sides of (4.2) are independent of the choice of representatives. Let me do this for the product. So suppose that $a/b \sim a'/b'$, $c/d \sim c'/d'$. This means that $ab' = a'b$, $cd' = c'd$. It follows that $acb'd' = a'c'bd$, but this says that $(ac)/(bd) \sim (a'c')/(b'd')$, as required.

Exercise 4.28. Establish the analogous property for $+$.

Now a tedious but entirely straightforward direct verification shows that $(F(R), +, \cdot)$ is a ring, with neutral elements $0 = 0/1$, $1 = 1/1$.

Theorem 4.13. *$F(R)$ is a field. Moreover, $\iota : R \to F(R)$, $\iota(a) = a/1$, defines an injective homomorphism.*

The last statement is usually expressed by saying that R is *embedded* in $F(R)$; we really mean by this that $F(R)$ contains a ring that is isomorphic to R as a subring. Indeed, ι, thought of as a map $\iota : R \to \iota(R) \subseteq F(R)$, is an isomorphism, so $F(R)$ contains an isomorphic copy $\iota(R)$ of R as a subring, as claimed.

Proof. We already saw that $F(R)$ is a ring. To show that $F(R)$ is a field, let $a/b \in F(R)$, $a/b \neq 0$. This last condition really means that $a/b \not\sim 0/1$, or, equivalently, $a \neq 0$. Thus $b/a \in F(R)$, and clearly $(a/b)(b/a) = ab/ab \sim 1/1 = 1$, so a/b is invertible.

Exercise 4.29. Verify that $\iota(a) = a/1$ is an injective homomorphism.

□

Exercise 4.30. Let R be a field. Show that then $\iota(R) = F(R)$; so, in particular, $F(R) \cong R$.

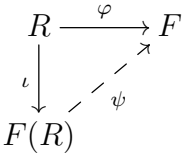
Exercise 4.31. Show that $a/b = (a/1)(b/1)^{-1}$. Or, since $a/1 \in F(R)$ is the element that corresponds to $a \in R$ under the embedding ι, we could write somewhat imprecisely but more intuitively $a/b = ab^{-1}$. So, as expected, the formal fraction a/b can be thought of as a multiplied by the inverse of b.

$F(R)$ is not just an arbitrary field that contains (an isomorphic copy of) R. Rather, the construction looks minimal in the following sense: suppose I want a field that contains R (really an isomorphic copy of R, but I'm going to ignore this distinction now). This field then definitely has to contain a multiplicative inverse b^{-1} of each $b \in R \setminus 0$, but since I can multiply in a field, I then obtain all products ab^{-1}, $a, b \in R$, $b \neq 0$. As you showed in the previous Exercise, these (formal, since b^{-1} need not be in R) products can be naturally identified with the elements of

$F(R)$. So in this sense, I will not be able to embed R into a smaller field.

These considerations were somewhat informal and vague. The asserted minimality of $F(R)$ finds its rigorous expression in a mapping property.

Theorem 4.14. *Let R be a domain and F a field. Then every injective homomorphism $\varphi : R \to F$ factors through $F(R)$: there is a unique homomorphism ψ such that the following diagram commutes:*



Proof. Just to make the diagram commute, we must set $\psi(a/1) = \varphi(a)$ for $a \in R$. By Exercise 4.31, if we want ψ to be a homomorphism, we are then forced to define

$$\psi(a/b) = \varphi(a)\varphi(b)^{-1}. \tag{4.3}$$

Exercise 4.32. Confirm that this definition makes sense. Address the following points: (a) Show that $\varphi(b)$ is invertible if $b \neq 0$; (b) verify that (4.3) defines $\psi(x)$ for $x \in F(R)$ unambiguously (what exactly do you need to show here?).

Now that you have made sure that (4.3) does give us a map ψ, it remains to check that this ψ works. Clearly, the diagram commutes, so we must check that ψ is a homomorphism. I'll only verify that ψ is multiplicative:

$$\begin{aligned}\psi((a/b)(c/d)) = \psi((ac)/(bd)) &= \varphi(ac)\varphi(bd)^{-1} \\ &= \varphi(a)\varphi(c)\varphi(b)^{-1}\varphi(d)^{-1} = \psi(a/b)\psi(c/d),\end{aligned}$$

as required. □

Exercise 4.33. Prove similarly that $\psi(x + y) = \psi(x) + \psi(y)$ for $x, y \in F(R)$ and that $\psi(1) = 1$.

Exercise 4.34. Let $\varphi : F \to R$ be a homomorphism from a field F to a ring $R \neq 0$. Show that φ is injective.

Theorem 4.14 does express the minimality of $F(R)$: if we consider any embedding $\varphi : R \to F$ of R in a field, then the Theorem delivers an injective map ψ from $F(R)$ onto a subfield of F, and this map sends the copy $\iota(R)$ of R inside $F(R)$ to its copy $\varphi(R)$ in F (why?). So

after identifying isomorphic rings/fields, the situation that emerges is "$R \subseteq F(R) \subseteq F$."

There is no analog of the *field of fractions* construction for non-commutative rings. The following exercises explore this theme in more detail.

Exercise 4.35. Let M be a commutative monoid with the *cancellation property:* if $ab = ac$, then $b = c$ (recall that the multiplicative monoid $(R\backslash 0, \cdot)$ of a domain has this property). Show that M can be embedded in an abelian group G. *Comments:* (1) So you want to construct a group G and an injective (monoid) homomorphism $\varphi : M \to G$; (2) to actually do this problem, just run a version of the field of fractions construction (this is easier than what we did above).

Exercise 4.36. Consider the free monoid $FM(a_1, \ldots, b_4)$ on the eight generators a_j, b_j, $1 \leq j \leq 4$, and then the relations

$$a_1a_2 = a_3a_4, \quad a_1b_2 = a_3b_4, \quad b_1a_2 = b_3a_4. \tag{4.4}$$

For $x, y \in FM$, say that $x \equiv y$ precisely if there is a finite sequence of substitutions using (4.4) that gets you from x to y.

Show that $\equiv$ is a congruence on FM.

Exercise 4.37. Let $M = (FM/\equiv)$ be the (non-commutative) quotient monoid considered in the previous Exercise. Show that M has the cancellation property: $xy = xz$ or $yx = zx$ implies that $y = z$.

Exercise 4.38. Show that $M = (FM/\equiv)$ can *not* be embedded in a group G. *Hint:* Show that if G is a group containing eight elements a_j, b_j that satisfy (4.4), then $b_1b_2 = b_3b_4$. Then show that this last relation does *not* hold in M.

It is then possible to build a (non-commutative) domain D from this M that can not be embedded in a division ring, but maybe we stop here.

4.3. **Polynomial rings.** Let R be a subring of R'. Then, for any subset $S \subseteq R'$, we define $R[S]$ as the subring of R' that is generated by R and S. Equivalently, this is the smallest subring of R' that contains $R \cup S$. We can establish the existence of this object in the usual way, by taking the intersection of all subrings of R' that contain $R \cup S$. If $S = \{s_1, \ldots, s_n\}$, we usually write $R[s_1, \ldots, s_n]$ instead of $R[\{s_1, \ldots, s_n\}]$.

Exercise 4.39. Show that $R[S \cup T] = (R[S])[T]$.

As usual, we can also build $R[S]$ from the bottom up. Let's look at this in the case $R[u]$, where we adjoin just one element $u \in R'$. Since

$u \in R[u]$ and addition and multiplication will not take us outside $R[u]$, all *polynomials* in u with coefficients from R

$$a_0 + a_1 u + a_2 u^2 + \ldots + a_n u^n, \qquad n \geq 0, a_j \in R,$$

must be in $R[u]$. Since these form a subring that contains R and u, they are exactly all of $R[u]$:

$$R[u] = \left\{ a_0 + a_1 u + a_2 u^2 + \ldots + a_n u^n : n \geq 0, a_j \in R \right\} \tag{4.5}$$

Exercise 4.40. Prove (4.5) more explicitly please.

Now if we take two distinct polynomials (that is, two sets of coefficients from R that are not identical), then there is of course no guarantee that the corresponding elements of $R[u]$ will also be distinct. For example, if $R = \mathbb{Z}$, $R' = \mathbb{Q}$, $u = 1/2$, then $2u = 1$, even though $2x$ and 1 are distinct as polynomials.

Given R and a formal symbol x, I now want to build a new ring $R[x]$ that has no such relations between distinct polynomials in x. This is similar in spirit to (but easier than) the construction of the free group. $R[x]$ will just consist of all polynomials in x with coefficients from R, and I add and multiply these in the obvious way. Two polynomials will be declared distinct unless they have exactly the same sequence of coefficients.

Now if I right away wrote a typical element of $R[x]$ as $f(x) = a_0 + a_1 x + \ldots + a_n x^n$ (we *will* do this very soon), you could complain that these operations $+, \cdot$ are undefined. Recall that $x \notin R$ is just a formal symbol, I am not operating inside a bigger ring R' now.

So, to obtain a clean formal definition of the polynomial ring, we tentatively set

$$R[x] = \{(a_0, a_1, a_2, \ldots) : a_j \in R, a_j = 0 \text{ for } j > n \text{ for some } n \geq 0\};$$

of course, our intention is to identify such a sequence with the polynomial with these coefficients as soon as possible. We then define addition and multiplication on $R[x]$ as follows:

$$(a_0, a_1, \ldots) + (b_0, b_1, \ldots) = (a_0 + b_0, a_1 + b_1, \ldots),$$
$$(a_0, a_1, \ldots)(b_0, b_1, \ldots) = (a_0 b_0, a_0 b_1 + a_1 b_0, a_0 b_2 + a_1 b_1 + a_2 b_0, \ldots).$$

These definitions are of course motivated by the fact that they give you what you would have obtained if you had added or multiplied (formal) polynomials. It is now straightforward, but mildly tedious to check that $R[x]$ with these operations becomes a ring with neutral elements $(0), (1)$; here, we agree that coefficients beyond those that are listed are equal to zero.

Exercise 4.41. Show that with these definitions, $R[x]$ becomes a ring.

From a purely formal point of view, $R[x]$ is not directly related to R; the only connection so far is that R was one of the ingredients in its construction. However, the map $R \to R[x]$, $a \mapsto (a)$, is an injective homomorphism, so an isomorphic copy of R is embedded in $R[x]$. We usually identify R with this subring of $R[x]$. Next, we introduce $x := (0,1)$; this is of course motivated by the fact that x as a polynomial has these coefficients $x = 0 + 1 \cdot x$. Then $x^2 = xx = (0,1)(0,1) = (0,0,1)$, $x^3 = (0,0,0,1)$, and, more generally, $x^n = (0,\dots,0,1)$, where the 1 is in the $(n+1)$st slot (which corresponds to index n). Also, observe that if $a \in R$, then $ax^n = (a)(0,0,\dots,0,1) = (0,0,\dots,0,a)$. By combining these facts, we obtain the formula

$$(a_0, a_1, \dots, a_n) = a_0 + a_1 x + a_2 x^2 + \dots a_n x^n.$$

Everything makes perfect sense now; the algebraic operations are performed in the ring $R[x]$. We note with some relief that the ring $R[x]$ *can* (and will) be thought of as the ring of formal polynomials in the new symbol x, with coefficients from R. The qualification "formal" is crucial here: you can also build a ring P of polynomial *functions:* its elements are the functions $f : R \to R$ of the form

$$f(x) = a_0 + a_1 x + \dots + a_n x^n, \qquad n \geq 0, a_j \in R.$$

Functions with values in a ring can be added and multiplied pointwise, for example $(f+g)(x) := f(x) + g(x)$. It's now easy to check that P becomes a ring with these operations. Now every formal polynomial $f \in R[x]$ induces a function $f(x) \in P$, given by the same expression; more formally, we can say that $f \mapsto f(x)$ defines a homomorphism $R[x] \to P$. This homomorphism is *not,* in general, injective, and P need not be isomorphic to $R[x]$. This happens for the simple reason that non-identical polynomials can define the same function. For example, if $R = \mathbb{Z}_2$, then $f(x) = 0$ and $g(x) = x^2 + x$ are the same function $\mathbb{Z}_2 \to \mathbb{Z}_2$ (why?), but of course f, g are distinct as elements of $\mathbb{Z}_2[x]$.

Exercise 4.42. (a) Show that $R[x]$ is an infinite ring if $R \neq 0$.
(b) Show that the ring of polynomial functions $R \to R$ is finite if R is finite.
(c) Show that the ring of polynomial functions $p : R \to R$ agrees with $R[x]$ for $R = \mathbb{Z}$, $\mathbb{Q}$ or $\mathbb{R}$.

Exercise 4.43. Here's an abstract version of the construction we just ran, if you like such things. Let M be a commutative monoid, R a ring. Define $R[M]$ as the collection of functions $f : M \to R$ with $f(m) \neq 0$ for only finitely many $m \in M$. It is much more suggestive

to represent these as formal linear combinations $\sum a_m m$, with $m \in M$, $a_m = f(m) \in R$, and the (formal) sum contains finitely many terms. Then introduce addition and multiplication on $R[M]$ as follows:

$$\sum a_m m + \sum b_m m := \sum (a_m + b_m) m, \quad \sum a_m m \sum b_m m := \sum c_m m,$$

and here $c_m = \sum_{pq=m} a_p b_q$; this sum is an actual, not formal, sum in R. In other words, just add and multiply formal sums formally.

Show that $R[M]$ is a ring, and that $R[\mathbb{N}_0] \cong R[x]$. (If you take $M = \mathbb{N}_0 \times \ldots \times \mathbb{N}_0$, then you obtain the polynomial ring over R in several variables.)

Exercise 4.44. Show that $R[x]$ is generated by R and x (so the notation is consistent with our earlier use of $R[u]$).

The quick summary of the construction of $R[x]$ is: a ring generated by R and x that does not have unnecessary relations. This should give a universal mapping property into arbitrary rings generated by R and one element. Here's a slightly more general version of this:

Theorem 4.15. *Let $\psi : R \to S$ be a homomorphism between rings, and let $u \in S$. Then there exists a unique homomorphism $\varphi : R[x] \to S$ with $\varphi(a) = \psi(a)$ for $a \in R$ and $\varphi(x) = u$.*

Proof. If there is such a homomorphism φ, then it must map

$$a_0 + a_1 x + \ldots + a_n x^n \mapsto \psi(a_0) + \psi(a_1) u + \ldots + \psi(a_n) u^n.$$

Conversely, it's straightforward to check that this works. □

In particular, $R \subseteq S$ could be a subring of S, and $\psi(a) = a$, $a \in R$. This is perhaps the situation we will encounter most frequently when applying Theorem 4.15.

In this case, the homomorphism φ is really just a fancy way of saying: plug u into $f(x) \in R[x]$. Or we could say that $\varphi(f)$ evaluates f at u. We do want to remember, though, that $f \in R[x]$ is a formal polynomial, not a function.

Corollary 4.16. *Let R be a subring of S and $u \in S$. Then there is an ideal $I \subseteq R[x]$, $I \cap R = 0$, so that $R[u] \cong R[x]/I$.*

This can be viewed as the ring analog of Corollary 3.12.

Proof. The evaluation homomorphism $\varphi : R[x] \to R[u]$, $a \mapsto a$, $a \in R$, $x \mapsto u$, is surjective by (4.5), so Theorem 4.7 shows that $R[u] \cong R[x]/I$, with $I = \ker(\varphi)$. Since $\varphi(a) = a$ for $a \in R$, it then also follows that $I \cap R = 0$. □

4.4. **Properties of polynomial rings.** Let $f = a_0 + a_1 x + \ldots + a_n x^n \in R[x]$, $a_n \neq 0$. Then we call $n = \deg f$ the *degree* of f. It will also be convenient to put $\deg 0 = -\infty$. If $a \in R \setminus 0$ and we think of a as an element of $R[x]$, then of course $\deg a = 0$.

Theorem 4.17. *Let R be a domain. Then $R[x]$ is a domain, and $U(R[x]) = U(R)$.*

Proof. Suppose that $fg = 0$. If neither of these is the zero polynomial, then $m = \deg f \geq 0$, $n = \deg g \geq 0$, and thus the highest coefficients a_m, b_n are both $\neq 0$. But then $fg = \ldots + a_m b_n x^{m+n} \neq 0$. So $f = 0$ or $g = 0$, and this says that $R[x]$ is a domain.

If $\deg f \geq 1$, then we see by again focusing on the highest power of x that $fg \neq 1$ for arbitrary $g \in R[x]$. This also says that if $f \in R$, then the inverse in $R[x]$, if it exists, can only be an element of R. However, if $f \in U(R)$, then of course the same inverse as before works. □

Let $f, g \in R[x]$, $g \neq 0$. Then we can divide f by g with remainder, in the following sense: if $\deg g = n$ and we denote the leading coefficient of g by b_n, then there exist $h, r \in R[x]$ and $k \geq 0$ with $\deg r < n$ and

$$b_n^k f(x) = g(x)h(x) + r(x). \tag{4.6}$$

We can do this by running the familiar long division style procedure. Let's prove the existence of k, h, r as in (4.6) along these lines. We organize the argument as an induction on $m = \deg f$.

Exercise 4.45. Convince yourself that k, h, r as in (4.6) can be found if $m = 0$.

The polynomial f is of the form $f(x) = a_m x^m + \text{lower order terms}$, $a_m \neq 0$. First of all, if $n = \deg g > m$, then we can just take $k = 0$, $h = 0$, $r = f$. If $n \leq m$, let

$$f_1(x) = b_n f(x) - a_m x^{m-n} g(x) \in R[x]. \tag{4.7}$$

Then $\deg f_1 < m$, so $b_m^j f_1(x) = k(x)g(x) + r(x)$ by the induction hypothesis for suitable $j \geq 0$, $k, r \in R[x]$, $\deg r < n$. Plug this into (4.7) to obtain that

$$b_n^{j+1} f(x) = g(x)(k(x) + a_m b_n^j x^{m-n}) + r(x),$$

and this is (4.6).

If b_n is invertible, then we can multiply (4.6) by b_n^{-k} and we obtain the slightly simpler version

$$f(x) = g(x)h(x) + r(x), \qquad \deg r < \deg g.$$

In particular, this will always work if R is a field.

Theorem 4.18. *Given $f \in R[x]$, $a \in R$, there exists a unique polynomial $h \in R[x]$ such that*

$$f(x) = (x-a)h(x) + f(a). \tag{4.8}$$

Here, the (suggestive) notation $f(a)$ refers to $\varphi_a(f)$, where $\varphi_a : R[x] \to R$ is the evaluation homomorphism from Theorem 4.15 that sends $x \mapsto a$.

Proof. Divide f by $g = x - a$ with remainder, as in (4.6). Notice that $b_1 = 1$. This gives that $f(x) = (x-a)h(x)+r(x)$, with $\deg r < \deg g = 1$. In other words, $r \in R$. Now evaluate at $x = a$ (more precisely, apply φ_a to both sides) to see that $r = f(a)$.

This also gives uniqueness because if h_1, h_2 both work in (4.8), then $(x-a)(h_2-h_1) = 0$, but this shows that $h_2 - h_1 = 0$ (if not, then look at the highest coefficient of the product to obtain a contradiction). □

Exercise 4.46. Show that in $\mathbb{Z}_6[x]$,

$$2x^2 + 4x = 2x(x+2) = 2x(4x+5), \text{ but } x + 2 \neq 4x + 5,$$

and take another look at the last part of the proof of Theorem 4.18.

Corollary 4.19. *Let $f \in R[x]$, $a \in R$. If $f(a) = 0$, then $(x-a)|f(x)$.*

We define *divisors* and *divisibility* in a general ring R in the same way as in $\mathbb{Z}$ (see Section 1.1 for this): $a|b$ means that $b = ac$ for some $c \in R$.

Theorem 4.20. *If F is a field, then $F[x]$ is a PID.*

This includes the claim that $F[x]$ is a domain, which we established earlier, in Theorem 4.17.

Proof. Although the setting looks quite different, this is essentially the same proof as the one for $\mathbb{Z}$. In both cases, division with remainder is the key tool that makes things work. In fact, a completely general version of this argument works; this is explored in Exercises 4.48, 4.49 below.

So let $I \subseteq F[x]$ be an ideal. Clearly $I = 0 = (0)$ is a principal ideal. If $I \neq 0$, fix a $g \in I$, $g \neq 0$, such that $\deg g$ is minimal among all such polynomials. Now if f is any element of I, divide f by g with remainder: $f = gh + r$, $\deg r < \deg g$. Then $r \in I$ as well, so $r = 0$ by the choice of g. We have shown that $I \subseteq (g) = \{gh : h \in R[x]\}$, and the reverse inclusion is obvious, so $I = (g)$. □

Exercise 4.47. Is $\mathbb{Z}[x]$ a PID?

Exercise 4.48. A domain R is called a *Euclidean domain* if there exists a function $\nu : R \to \mathbb{N}_0$ such that if $a, b \in R$, $b \neq 0$, then there exist $q, r \in R$ with $a = qb + r$, $\nu(r) < \nu(b)$. (This is an abstract version of division with remainder.) Show that $\mathbb{Z}$ and $F[x]$, F a field, are Euclidean domains. *Suggestion:* For $F[x]$, try $\nu(f) = 2^{\deg f}$, with $2^{-\infty} := 0$.

Exercise 4.49. Show that a Euclidean domain is a PID.

Exercise 4.50. Show that $\mathbb{Z}[\sqrt{-1}]$, the ring of Gaussian integers, is a Euclidean domain.

Exercise 4.51. If R is not a domain, then it can of course happen that $fg = 0$ for $f, g \in R[x]$, $f, g \neq 0$. However, show that if such an $f \in R[x]$ is given, then already $cf = 0$ for some $c \in R$, $c \neq 0$. *Hint:* Write $f = \sum a_j x^j$, and consider separately the case where $a_j g = 0$ for all j. In the other case, try to replace g by another polynomial of smaller degree.

Now suppose that $F \subseteq R$, where F is a field and R is a ring, and F is a subring of R. Let $u \in R$, and consider the subring $F[u] \subseteq R$. By Corollary 4.16, $F[u] \cong F[x]/I$. More precisely, $I = \ker(\varphi)$, where $\varphi : F[x] \to F[u]$ sends $a \mapsto a$, $a \in F$, and $x \mapsto u$. In other words, $g \in I$ precisely if $g(u) = 0$.

By Theorem 4.20, $I = (f)$ for suitable $f(x) \in F[x]$; in fact, from the proof, we also know that if $I \neq 0$, then any non-zero polynomial of minimal degree from I will work as f. In this case, $\deg f \geq 1$: we cannot have $f = a \in F^\times = F \setminus 0$ because then $f(u) \neq 0$. Let's summarize: $F[u] \cong F[x]/(f)$, and here any $f \in F[x]$ with $f(u) = 0$ and minimal degree works.

In the other case, $I = 0$, so $R[x] \cong R[u]$ and $g(u) \neq 0$ for all polynomials $g \in F[x]$, $g \neq 0$.

Exercise 4.52. Show that in the first case there is a *unique* monic $f \in F[x]$ of smallest possible degree with $f(u) = 0$; we call a polynomial *monic* if its leading coefficient equals 1.

We make a few definitions that are motivated by these considerations.

Definition 4.21. Suppose that $F \subseteq R$, R is a ring and F is a field (and also a subring of R). Then we say that $u \in R$ is *transcendental* over F if $f(u) \neq 0$ for all $f \in F[x]$, $f \neq 0$. An *algebraic* element is one that is not transcendental. If $u \in R$ is algebraic, then there exists a unique monic polynomial $f \in F[x]$ of smallest possible degree with $f(u) = 0$. We call f the *minimal polynomial* of u over F.

We summarize what we found above:

Proposition 4.22. *Suppose that $F \subseteq R$, F is a field and a subring of the ring R. If $u \in R$ is transcendental over F, then $F[u] \cong F[x]$. If $u \in R$ is algebraic over F with minimal polynomial $f \in F[x]$, then $F[u] \cong F[x]/(f)$.*

Let's now take another look at the second case.

Theorem 4.23. *Let $u \in R$ be algebraic over F, with minimal polynomial $f \in F[x]$. If f is* irreducible *in the sense that if $f = gh$, then one of the factors is in $F^\times$, then $F[u]$ is a field. On the other hand, if f is reducible, then $F[u]$ is not a domain.*

Proof. We know that $F[u] \cong F[x]/(f)$, and this will be a field precisely if there are no ideals other than 0 and the whole ring; see Theorem 4.9. We also know, from Theorem 4.12, that the ideals of $F[x]/(f)$ correspond to those ideals $I \subseteq F[x]$ with $I \supseteq (f)$. What do such ideals I look like? First of all, $I = (g)$, since $F[x]$ is a PID, and then we will have that $(g) \supseteq (f)$ precisely if $f \in (g)$, and this happens if and only if $f = gh$ for some $h \in F[x]$. Now if f is irreducible, then this can only work if $g \in F^\times$ or $h \in F^\times$, but then $(g) = F[x]$ or $(g) = (f)$. So in this case, $F[x]/(f)$ has no non-trivial ideals and thus is a field, as claimed.

On the other hand, if f is reducible, say $f = gh$, $\deg g, \deg h < \deg f$, then $g(u), h(u) \neq 0$ (why?), but $g(u)h(u) = f(u) = 0$. □

Of course, $F[u] \subseteq R$ will automatically be a domain if R is. In this case, the Theorem implies that minimal polynomials of algebraic elements are irreducible. In particular, this follows if R is a field itself.

Example 4.5. Let's now use these ideas to find the rings R with 4 elements. This could of course be done entirely by hand, but, as we will see, the machinery just developed will be useful.

Since the prime ring $P \subseteq R$ is in particular a subgroup of R, the characteristic of R can only be 2 or 4. If $\mathrm{char}(R) = 4$, then no room is left for other elements, so $R \cong \mathbb{Z}_4$ in this case.

If $\mathrm{char}(R) = 2$, then R contains (the field) $\mathbb{Z}_2$ as an embedded subring. Let's now first see how far we can get with a hands-on approach. In addition to $0, 1 \in \mathbb{Z}_2$, there are two more elements in R, and let's call these a, b, so $R = \{0, 1, a, b\}$. We definitely have no choice as far as the additive structure is concerned: $a + a = b + b = 0$, and $a + 1 = b$ since this is the only value that is left for this sum ($a+0 = a$, $a+a = 0$, and $a + 1 = 1$ would give $a = 0$). Similarly, $b + 1 = a$. This could also be summed up by saying that $(R, +) \cong \mathbb{Z}_2 \times \mathbb{Z}_2$.

Now the multiplication will be determined as soon as we know what a^2 is equal to because $b = 1 + a$ and of course it's clear what products involving 0 or 1 are equal to. In principle, we could have $a^2 = 0$, 1, a, or b, and then the remaining products not involving $0, 1$ would have the following values:

	R_1	R_2	R_3	R_4
a^2	a	b	1	0
b^2	b	a	0	1
ab	0	1	b	a

We don't know, at this point, if these structures are really rings. This could of course be checked directly, but we'll switch to the abstract approach now. Before we do this systematically, here's one more idea:

Definition 4.24. Let $R_1, \ldots, R_n$ be rings. Then the *direct sum* $R_1 \oplus \ldots \oplus R_n$ is defined as the set of all $(r_1, \ldots, r_n)$, $r_j \in R_j$, with componentwise addition and multiplication.

Exercise 4.53. Show that the direct sum of rings is a ring, with $0 = (0, 0, \ldots, 0)$ and $1 = (1, 1, \ldots, 1)$. Also, show that a direct sum of rings $R_j \neq 0$ is never a domain.

This gives one more ring with 4 elements, namely $\mathbb{Z}_2 \oplus \mathbb{Z}_2$.

Exercise 4.54. Show that $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ has the same multiplication table as R_1, if we put $a = (1, 0)$.

This Exercise shows that R_1 is indeed a ring, and $R_1 \cong \mathbb{Z}_2 \oplus \mathbb{Z}_2$.

Now suppose we have any ring R with $|R| = 4$, $\mathrm{char}(R) = 2$. As we discussed, if we identify $\mathbb{Z}_2$ with the prime ring of R, then the situation becomes $\mathbb{Z}_2 \subseteq R$. If we take any $a \in R \setminus \mathbb{Z}_2$, then $R = \mathbb{Z}_2[a]$ (why?). Proposition 4.22 now shows that $R \cong \mathbb{Z}_2[x]/(f)$, where $f \in \mathbb{Z}_2[x]$ is the minimal polynomial of a.

Proposition 4.25. *Suppose that* $f(x) = x^n + a_{n-1}x^{n-1} + \ldots + a_0 \in \mathbb{Z}_k[x]$ *is monic. Then* $|\mathbb{Z}_k[x]/(f)| = k^n$.

Exercise 4.55. Prove Proposition 4.25.

This shows that in our situation, the minimal polynomial f of a must be of degree 2; conversely, for any monic $f \in \mathbb{Z}_2[x]$ of degree 2, we obtain a ring $\mathbb{Z}_2[x]/(f)$ of characteristic 2 with 4 elements. There are four such polynomials $f(x) = x^2 + cx + d$, $c, d = 0, 1$. Of these, only $f(x) = x^2 + x + 1$ is irreducible.

Exercise 4.56. Show this please.

So, by Theorem 4.23, we obtain one field with 4 elements and three more rings that are not domains. To match these with the ones from the table above, let $a = x + (f) \in \mathbb{Z}_2[x]/(f)$. Consider, for example, $f(x) = x^2 + x$. Then $a^2 = -a = a$ (because $f(a) = 0$). This identifies $\mathbb{Z}_2[x]/(x^2+x) \cong R_1$, and it also shows one more time that R_1 is indeed a ring. Note that while $\mathbb{Z}_2[x]/(f)$ seems like a somewhat abstract description of a ring, it is really rather concrete and simple to use; in particular here, when f is monic of degree 2, then the elements of the ring are $s + ta$, $s, t \in \mathbb{Z}_2$, where $a = x + (f)$ satisfies $f(a) = 0$, and this relation is used to bring products again to the form $s + ta$. We are really almost exactly back to the procedure we started out with: invent a new element, call it a, and introduce a relation that clarifies how a gets multiplied.

In the same way, we can show that R_2, R_3, R_4 are (rings and) isomorphic to $\mathbb{Z}_2[x]/(f)$ for the other three choices of f: $f(x) = x^2 + x + 1$, $f(x) = x^2 + 1$, $f(x) = x^2$, in this order.

Finally, observe that if we map $\mathbb{Z}_2[x] \to \mathbb{Z}_2[x]/(x^2)$ by first sending $t \mapsto t$, $t \in \mathbb{Z}_2$, $x \mapsto x + 1$ and then applying the quotient map $\mathbb{Z}_2[x] \to \mathbb{Z}_2[x]/(x^2)$ then $x^2 + 1 \mapsto (x + 1)^2 + 1 + (x^2) = x^2 + (x^2) = 0$, so the kernel of this map includes $(x^2 + 1)$. Thus we obtain an induced homomorphism

$$\begin{array}{ccc} \mathbb{Z}_2[x] & \longrightarrow & \mathbb{Z}_2[x]/(x^2)\ . \\ \downarrow & \nearrow & \\ \mathbb{Z}_2[x]/(x^2+1) & & \end{array}$$

This map is surjective and, since both rings have 4 elements, also injective; it is an isomorphism. We have shown that $R_3 \cong R_4$.

Exercise 4.57. Show this directly, from the addition/multiplication tables of R_3, R_4. Then show that except for this pair, no two rings of order 4 are isomorphic. So there are four non-isomorphic rings of order 4, and exactly one of them is a field.

Exercise 4.58. Suppose that $(m, n) = 1$. Show that then $\mathbb{Z}_m \oplus \mathbb{Z}_n \cong \mathbb{Z}_{mn}$.

Exercise 4.59. Show that there is only one ring of order 6. Show that a ring of order 12 satisfies $\mathrm{char}(R) = 6$ or $= 12$, and show that both values occur.

Exercise 4.60. In view of Proposition 4.25, something has to change in our analysis above if we now want to find rings of order 12 and characteristic 6. Can you elaborate on this (revisit the discussion that

led to Proposition 4.22 perhaps)? Please try to find the three (non-isomorphic) rings with these data.

Exercise 4.61. What is the order of the smallest non-commutative ring?

Theorem 4.26. *Let D be a domain, $f \in D[x]$, $\deg f = n \geq 0$. Then f has at most n distinct roots (= zeros).*

Proof. Suppose that $a_1, \ldots, a_k$ are distinct roots of f. I claim that then $\prod(x-a_j)|f(x)$, and we can prove this by induction on k. The case $k = 1$ is handled by Corollary 4.19. Now consider $k \geq 2$. By the induction hypothesis, $f(x) = g(x)\prod_{j=1}^{k-1}(x - a_j)$. Then $g(a_k)\prod(a_k - a_j) = 0$, and since $a_k - a_j \neq 0$ by assumption for $j \leq k - 1$, this implies that $g(a_k) = 0$. Thus $g(x) = (x - a_k)h(x)$ by Corollary 4.19 again, and this gives the desired factorization.

Now $\deg \prod(x - a_j) = k$ if there are k factors. This shows that there cannot be more than n roots. □

In particular, this applies to polynomials with coefficients in a field, and this is the case we will be most interested in.

Exercise 4.62. Is the statement also true for $f \in R[x]$ when R is an arbitrary (commutative) ring?

Exercise 4.63. Show that there are infinitely many $x \in \mathbb{H}$ with $x^2 + 1 = 0$.

Theorem 4.27. *Let F be a field. Any finite subgroup of the multiplicative group $F^\times$ is cyclic.*

In particular, this applies to $F^\times$ itself if F is finite: $(\mathbb{Z}_p^\times, \cdot)$ is a cyclic group of order $p - 1$.

Proof. Let $G \subseteq F^\times$ be such a finite subgroup. Recall that $n = \exp(G)$ was defined as the smallest $n \geq 1$ with $x^n = 1$ for all $x \in G$. Now consider the polynomial $f(x) = x^n - 1$. By this definition of the exponent of a group, $f(a) = 0$ for all $a \in G$. So f has $|G|$ zeros, and now Theorem 4.26 implies that $|G| \leq n = \exp(G)$. Since always $\exp(G) \leq |G|$, we have that $|G| = \exp(G)$. Theorem 2.13 shows that G is cyclic. □

Again, the non-commutative version of this fails: Q is a non-cyclic finite group that can be realized as a subgroup of the (non-abelian) multiplicative group of $\mathbb{H}$.

Exercise 4.64. Show that $(\mathbb{Q}^\times, \cdot)$ is not cyclic.

Exercise 4.65. (a) Let $F = \{0, a_1, a_2, \dots, a_n\}$ be a finite field. Prove the following generalization of Wilson's Theorem (see Exercise 2.36):

$$a_1 a_2 \cdots a_n = -1$$

(b) Is this also true for $F = \mathbb{Z}_2$?

4.5. **Divisibility.** Throughout this section, D will be a domain. We want to give an abstract version, in D, of the fundamental theorem of arithmetic. We start out with some relevant definitions.

Recall that we define divisors in the expected way: we say that $a|b$ if $b = ac$ for some $c \in D$. Notice that $u \in D$ is a unit precisely if $u|1$. If $a|b$ and also $b|a$, then we say that a, b are *associates,* and we write $a \sim b$.

Proposition 4.28. *$a \sim b$ if and only if $a = ub$ for some $u \in U(D)$.*

Proof. If $a = ub$ for some unit u, then clearly $b|a$, but also $a|b$ since $b = u^{-1}a$. Thus $a \sim b$.

Conversely, if $a \sim b$, then $b = ac$ and $a = bd$, so $a = acd$. If $a \neq 0$, then this implies that $cd = 1$, so d is a unit, as desired. If $a = 0$, then $b = 0$ also, and then we can just write $a = 1b$. □

Exercise 4.66. Show that $\sim$ is an equivalence relation.

We call a a *proper divisor* of b if $a|b$, but $b \nmid a$. Equivalently, a is a proper divisor of $b \neq 0$ if $b = ac$, and c is not a unit.

Definition 4.29. We call $a \in D$ *irreducible* if $a \neq 0$ and a is not a unit, and a has no proper divisors other than units.

Exercise 4.67. Show that $a \in D$, $a \neq 0$ and not a unit, is irreducible if and only if the only factorizations of a into two factors are $a = u(u^{-1}a)$, u a unit (equivalently, if $a = bc$, then $b \sim 1$, $c \sim a$ or the other way around).

The units of $\mathbb{Z}$ are ± 1, so the irreducible elements of $\mathbb{Z}$ are $\pm p$, p prime. Put differently, the irreducible elements are exactly the associates of the primes.

The irreducible elements look like the proper substitute for primes in a general domain, so we will now be looking for factorizations into irreducible elements. As for uniqueness, observe that units can be introduced at will. More precisely, if $a = p_1 p_2 \cdots p_n$, then also $a = (u_1 p_1)(u_2 p_2) \cdots (u_n p_n)$ for any n units $u_1, \dots, u_n$ with $u_1 u_2 \cdots u_n = 1$. This puts a limit on how much uniqueness can be expected in such factorizations.

Definition 4.30. A domain D is called a *unique factorization domain* (UFD) if: (1) If $a \in D$, $a \neq 0$, is not a unit, then there are irreducible elements $p_1, \ldots, p_n$, $n \geq 1$, such that

$$a = p_1 p_2 \cdots p_n. \tag{4.9}$$

(2) If $a = p_1' p_2' \cdots p_m'$ is another factorization of this type, then $m = n$ and, after relabeling, $p_j' \sim p_j$.

In other words, a UFD is a domain in which an analog of the fundamental theorem of arithmetic holds.

We would now like to identify conditions that ensure that a given domain is a UFD. We turn to our proof of the fundamental theorem of arithmetic for inspiration. The existence of the factorization (4.9) was obtained by simply factoring a and then its factors, and then the factors of these elements etc. until the process stops. In a general domain, there is no guarantee that it will actually stop, so we introduce our first condition: we say that D satisfies the *divisor chain condition* (DCC) if no element has an infinite sequence of proper divisors. More precisely, if we have a sequence a_n with $a_{n+1} | a_n$ for all n, then there exists an N with $a_n \sim a_N$ for $n \geq N$.

Lemma 4.31. *If D satisfies the DCC, then every $a \neq 0$, a not a unit, has a factorization* (4.9) *into irreducible elements.*

Proof. We already know in outline what we want to do: we just keep factoring until this is no longer possible. To start the formal argument, I first claim that if $a \neq 0$ is not a unit, then a has an irreducible factor. To see this, look for factorizations of $a = bc$ into two non-units b, c. If that isn't possible, then a itself is irreducible and we are done (with the proof of my claim). Otherwise, write $a = a_1 b_1$, and then try to factor a_1 into non-units. Again, if this isn't possible, then a_1 is irreducible, and $a_1 | a$, as desired. If a_1 can be factored, say $a_1 = a_2 b_2$, then try to factor a_2 in the next step, and so on. We obtain a sequence of proper divisors $a_1 | a$, $a_2 | a_1$, $a_3 | a_2, \ldots$. By the DCC, this will stop at some point: a_N for some N has no proper non-unit divisors, so is irreducible, and $a_N | a$, as claimed.

Now we can use this as the basic step in a second attempt at (4.9): Given a as in the Lemma, pick an irreducible factor p_1, so $a = p_1 b_1$. If b_1 is not a unit, then we pick an irreducible factor p_2 of b_1, and then we can write $a = p_1 p_2 b_2$. We continue in this style. This produces a sequence of proper divisors $b_1 | a$, $b_2 | b_1, \ldots$, which can't continue forever, by the DCC; some b_n is a unit. □

Exercise 4.68. (a) Consider the formal generalized polynomials $f(x) = \sum a_q x^q$, with exponents $q \in \mathbb{Q}$, $q \geq 0$, and coefficients $a_q \in \mathbb{Z}$. These

form a ring R when added and multiplied in the obvious way. (If you did Exercise 4.43, then maybe you now recognize this ring as $R = \mathbb{Z}[M]$ for the monoid $M = (\mathbb{Q}_+, +)$.) Show that R is a domain that doesn't satisfy the DCC.

(b) Show that $f(x) = x$ is not a unit, but f can not be factored into irreducible elements.

Exercise 4.69. An *algebraic integer* is a number $a \in \mathbb{C}$ with $f(a) = 0$ for some monic polynomial $f \in \mathbb{Z}[x]$. These form a subring $A \subset \mathbb{C}$ (you can use this fact, you don't have to prove it here).
(a) Show that $A \cap \mathbb{Q} = \mathbb{Z}$ and conclude that A has (many) non-units.
(b) Show that A does not have any irreducible elements; in other words, any non-unit $a \neq 0$ can be factored, $a = bc$, into non-units $b, c \in A$.

Next, we turn to uniqueness (= condition (2) from the definition of a UFD). We will need more than just the DCC for this.

Example 4.6. Let $D = \mathbb{Z}[\sqrt{-5}] = \{a + b\sqrt{-5} : a, b \in \mathbb{Z}\}$. This is a domain since it is a subring of the domain $\mathbb{C}$. A key tool for discussing the arithmetic of such rings is the absolute value of a number $x \in R$. More precisely, we introduce $N(x) = |x|^2 = a^2 + 5b^2$ (N as in *norm*). This last expression shows that $N(x)$ is a non-negative integer. Also, $N(xy) = N(x)N(y)$, and this trivial identity greatly simplifies the search for divisors.

First of all, we deduce that the units of R are ± 1: indeed, if $uv = 1$, then $N(u)N(v) = 1$, so $N(u) = N(v) = 1$, but the only elements of norm 1 are ± 1. Next, we conclude that R satisfies the DCC: if $x \neq 0$ is factored, $x = yz$, then $N(y), N(z) \leq N(x)$, and the inequalities are strict unless $N(y) = 1$ or $N(z) = 1$, which makes that element a unit. So the norm of a proper factor of x is strictly smaller than $N(x)$, and thus there cannot be an infinite chain of proper divisors.

By Lemma 4.31, elements $x \neq 0, \pm 1$ have factorizations into irreducible elements. However, some of these factorizations are not unique; for example,

$$9 = 3 \cdot 3 = \left(2 + \sqrt{-5}\right)\left(2 - \sqrt{-5}\right).$$

Exercise 4.70. Show that $3, 2 \pm \sqrt{-5}$ are irreducible and $3 \not\sim 2 \pm \sqrt{-5}$.

If we reexamine the proof of the uniqueness part of the fundamental theorem of arithmetic, then we find that a key role was played by the following property of primes in $\mathbb{Z}$ (compare Proposition 1.5(b)): if $p|ab$, then $p|a$ or $p|b$.

Definition 4.32. Let D be a domain. A non-unit $p \in D$, $p \neq 0$ is called a *prime* if $p|ab$ implies that $p|a$ or $p|b$.

Exercise 4.71. Show that $3, 2 \pm \sqrt{-5} \in \mathbb{Z}[\sqrt{-5}]$ are not primes.

Lemma 4.33. *A prime is irreducible.*

Proof. Let p be a prime and suppose that $p = ab$. Then $p|ab$, so $p|a$ or $p|b$. Let's say $a = pc$. This gives that $p = p(bc)$, so b is a unit. We have seen that there are no factorizations of p that don't contain a unit, so p is irreducible, as claimed. □

Theorem 4.34. *A domain D is a UFD if and only if: (a) the DCC holds; (b) every irreducible element of D is a prime.*

Proof. We first show that conditions (a), (b) imply that D is a UFD. We already established the existence of factorizations (4.9) in Lemma 4.31 (from (a) alone). We must now show that these are also unique in the sense of condition (2) from Definition 4.30. So suppose that $a = p_1 \cdots p_n = p'_1 \cdots p'_m$. The irreducible element p'_1 is a prime, by assumption, and $p'_1 | p_1 \cdots p_n$. By repeatedly applying the defining condition for primes, we see that p'_1 divides some p_j, let's say $p_1 = up'_1$; here u must indeed be a unit, as suggested by the notation, because the irreducible element p_1 doesn't have non-unit proper factors. We now obtain that

$$up_2 \cdots p_n = p'_2 \cdots p'_m. \tag{4.10}$$

We now proceed by induction on $n \geq 1$. If $n = 1$ (= basis of our induction), then $p'_2 \cdots p'_m = u$, but irreducible elements are not units, so $m = 1$ and $p'_1 \sim p_1$.

If $n > 1$, then we apply the induction hypothesis to (4.10); observe that u doesn't interfere with this because, for example, up_2 is still irreducible. It follows that $m = n$ and $p'_j \sim p_j$ for $j \geq 2$, after relabeling, and we saw earlier that also $p'_1 \sim p_1$.

We now show that (a), (b) hold in a UFD. Suppose that $b|a$, so $a = bc$, and consider the factorizations

$$a = p_1 p_2 \cdots p_n, b = q_1 \cdots q_k, c = r_1 \cdots r_j$$

of these elements into irreducible factors. This gives the alternative factorization $a = q_1 \cdots q_k r_1 \cdots r_j$, so we now see from uniqueness that the q's were really drawn from $p_1, \ldots, p_n$, up to unit multipliers. So the divisors of a are exactly the associates of partial products $p_{j_1} \cdots p_{j_m}$ of some of the irreducible factors of a. Moreover, this will be a proper divisor precisely if at least one factor is actually dropped. This establishes DCC in a UFD; for a as above, a sequence of proper divisors terminates after at most n steps.

Finally, let p be an irreducible element of a UFD, and suppose that $p|ab$. By what we just established, this says that p must be associated

to one of the irreducible factors from the factorization of a or b (or both). But then p divides this element, as required. □

In a UFD, we obtain further benefits from the existence and uniqueness of factorizations into irreducible factors, which are analogs of the corresponding properties of $\mathbb{Z}$.

Definition 4.35. Let $a_1, \ldots, a_n$ be non-zero elements of a domain D. We say that d is a *greatest common divisor* of $a_1, \ldots, a_n$, and we write $d = (a_1, \ldots, a_n)$, if $d|a_j$, $j = 1, 2, \ldots, n$, and if also $c|a_j$ for all j, then $c|d$.

Similarly, we call m a *least common multiple* and we write $m = \operatorname{lcm}(a_1, \ldots, a_n)$ if $a_j|m$ for all j, and if k is another element with $a_j|k$ for all j, then $m|k$.

In general, gcd's and lcm's need not exist; if they do and d is a gcd, then another element e is a gcd (of the same a_j's) precisely if $e \sim d$. Similarly, if m is a lcm, then exactly the associates of m will work, too.

Exercise 4.72. Prove these remarks. Also, show that any n elements will have a gcd if any 2 elements have a gcd; in this case, $((a, b), c) = (a, b, c)$.

Theorem 4.36. *In a UFD, gcd's and lcm's always exist. Moreover, if in the factorizations of a, b we pair off associates until this is no longer possible,*

$$a = p_1 \cdots p_n q_1 \cdots q_j, \quad b = p_1 \cdots p_n r_1 \cdots r_k,$$

then

$$(a, b) = p_1 \cdots p_n, \quad \operatorname{lcm}(a, b) = p_1 \cdots p_n q_1 \cdots q_j r_1 \cdots r_k.$$

Proof. I'll prove the claim about the gcd and leave the lcm part to the reader. Clearly $d = p_1 \cdots p_n$ divides both a and b. Now suppose that e is any element with $e|a$, $e|b$. Then, as we saw above, to make $e|a$ happen, we must draw the irreducible factors of e from those of a, up to associates. Since we want both $e|a$ and $e|b$, the q's and r's are off-limits; more precisely, if a q_m, say, is an associate of an irreducible factor of b, then that factor can only be a p_j because by their definition, no q_m is associated to any r_l. So we might as well pick this p_j instead of q_m. It then follows that $e|d$, as required. □

Our next goal is to show that a PID is a UFD. It is useful to relate divisibility notions to ideals. Recall that a principal ideal can be described as $(a) = \{ax : x \in D\}$.

Lemma 4.37. *Let D be a domain. Then $a|b$ if and only if $(a) \supseteq (b)$, and a is a proper divisor of b precisely if $(a) \supsetneq (b)$. Two elements a, b are associates precisely if $(a) = (b)$.*

Exercise 4.73. Prove Lemma 4.37.

These simple observations give a very neat description of the gcd in a PID. If we say it in terms of ideals, then the conditions that $d|a$, $d|b$ become $(d) \supseteq (a)$, $(d) \supseteq (b)$. But if $a, b \in (d)$, then (d) must also contain the ideal that is generated by a and b, which (as before) we write as (a, b). So $d|a$, $d|b$ if and only if $(d) \supseteq (a, b)$. In particular, d is a gcd precisely if (d) is the smallest principal ideal that contains (a, b), if there actually is such a principal ideal.

Now in a PID any ideal is a principal ideal, so $(a, b) = (d)$ for some d, and such a d is a gcd of a, b. We have shown that the gcd exists in a PID (we will obtain a second proof of this, when we show below that a PID is a UFD); moreover, the gcd's of a, b are the elements d with $(d) = (a, b)$. So the apparent ambiguity of the notation (namely, (a, b) can refer to a gcd or to the ideal generated by a, b) actually works to our advantage here.

Theorem 4.38. *A PID D is a UFD. Moreover, if $d = (a, b)$, then there are $x, y \in D$ with $ax + by = d$.*

Proof. We'll check (a), (b) from Theorem 4.34. We can now rephrase DCC as follows: if $(a) \subseteq (a_1) \subseteq (a_2) \subseteq \ldots$ is an increasing chain of principal ideals, then $(a_n) = (a_N)$ for all $n \geq N$ for some N. To prove that this holds, consider $I = \bigcup(a_n)$.

Exercise 4.74. Show that I is an ideal. Also, show that if the sets don't increase, then a union of ideals need not be an ideal.

By the Exercise and since our domain is a PID, we have that $I = (b)$ for some b. But then $b \in I$, so $b \in (a_N)$ for some N and thus $I = (b) \subseteq (a_N)$ and $(a_n) = (a_N)$ for all $n \geq N$.

To verify condition (b) from Theorem 4.34, let p be an irreducible element, and suppose that $p|ab$ and $p \nmid a$. Consider the ideal (p, a). Since (a) is not contained in (p) under our current assumptions, we have that $a \notin (p)$, so $(p, a) \supsetneq (p)$. Since p is irreducible and (p, a), like any ideal, is a principal ideal, this is only possible if $(p, a) = D$.

The elements of (p, a) are the linear combinations $px + ay$, so it now follows that we can find x, y so that $px+ay = 1$ and hence $pxb+aby = b$. This implies that $p|b$.

Finally, if d is a gcd of a, b, then, as we saw above, $(d) = (a, b)$, the ideal generated by a, b. In particular, $d \in (a, b)$, and this says that there are $x, y \in D$ such that $ax + by = d$. □

The statement on the gcd can fail in UFDs that are not PIDs. We will show in the next section that $\mathbb{Z}[x]$ is a UFD.

Exercise 4.75. Find two polynomials $f, g \in \mathbb{Z}[x]$ (easy ones, if you can) with $(f, g) = 1$, but there are no $p, q \in \mathbb{Z}[x]$ such that $fp + gq = 1$.

Example 4.7. Let's now discuss factorization in the ring $\mathbb{Z}[\sqrt{-1}] = \mathbb{Z}[i]$ of Gaussian integers. You showed in Exercise 4.50 that $\mathbb{Z}[i]$ is Euclidean, which implies that $\mathbb{Z}[i]$ is a PID and thus also a UFD. As in our discussion of $\mathbb{Z}[\sqrt{-5}]$, a key tool will be $N(x) = |x|^2$; if $x = a + ib$, then $N(x) = a^2 + b^2$, so $N(x)$ is a non-negative integer for $x \in \mathbb{Z}[i]$.

If u is a unit, then $uv = 1$, so $N(u)N(v) = N(1) = 1$, and this implies that $N(u) = 1$. The only elements of norm 1 are ± 1, $\pm i$. Conversely, it is clear that these are units, so $U = \{\pm 1, \pm i\} = \{x \in \mathbb{Z}[i] : N(x) = 1\}$.

What are the irreducible elements of $\mathbb{Z}[i]$? One class of suspects is given by the $\mathbb{Z}$-primes $2, 3, 5, 7, 11, \ldots$ and their associates. However, it will turn out that not all of these are also primes (equivalently: irreducible) in $\mathbb{Z}[i]$.

Let's first discuss those $\mathbb{Z}$-primes p that satisfy $p \equiv 1 \mod 4$. So $p = 4n + 1$ for some $n \geq 1$. Consider the group $\mathbb{Z}_p^\times$. By Theorem 4.27, this is a cyclic group, so $\mathbb{Z}_p^\times = \langle a \rangle$ for some $a \in \mathbb{Z}_p^\times$. Consider $t = a^n$: we have that $t^4 = a^{p-1} = 1$, so $t^2 \in \mathbb{Z}_p$ is a zero of the polynomial $x^2 - 1 \in \mathbb{Z}_p[x]$. This polynomial has degree 2, so has at most two zeros, and clearly these are given by ± 1. Since $t^2 = a^{(p-1)/2} \neq 1$, it follows that $t^2 = -1$. This is an equality in $\mathbb{Z}_p$, so when written out, it says that $t^2 \equiv -1 \mod p$ or $t^2 + 1 = kp$ for some $k \in \mathbb{Z}$.

This can be rewritten as $kp = (t + i)(t - i)$. If p were a prime in $\mathbb{Z}[i]$ also, then it would now follow that $p | t \pm i$ for at least one choice of the sign, but this can't work. We have shown that p is not irreducible in $\mathbb{Z}[i]$. In fact, we can be more specific. Since p is not irreducible, we can factor it: $p = xy$, and here x, y are not units. This means that $N(x), N(y) > 1$, and since $N(p) = p^2$, we must have that $N(x) = N(y) = p$. In other words, if $x = a + ib$, then $a^2 + b^2 = p$. This means that we can factor p as $p = (a + ib)(a - ib)$. We have also established one direction of the following beautiful equivalence:

Theorem 4.39. *Let $p \geq 2$ be a prime in $\mathbb{Z}$. Then p can be written as the sum of two squares, $p = a^2 + b^2$, $a, b \in \mathbb{Z}$, if and only if $p \equiv 1 \mod 4$ or $p = 2$.*

Exercise 4.76. Finish the proof by showing that if p is a sum of two squares, then $p = 2$ or $p \equiv 1 \mod 4$.

So, to return to our main business of finding the irreducible elements of $\mathbb{Z}[i]$, we have seen that if $p \equiv 1 \mod 4$ or $p = 2$, then $p = a^2 + b^2$,

and this lets us factor p as $p = (a+ib)(a-ib)$. These factors $a \pm ib$ are irreducible because they have norm p, and this rules out factorizations that don't involve a unit.

The $\mathbb{Z}$-primes $p \equiv 3 \mod 4$ stay irreducible in $\mathbb{Z}[i]$, by the argument from above: if we had $p = xy$ with x, y not units, then it would follow that $N(x) = N(y) = p$, but this gives a representation of p as a sum of two squares.

Have we now found all irreducible elements of $\mathbb{Z}[i]$? To answer this, suppose that $\rho \in \mathbb{Z}[i]$ is irreducible. Then $\rho\overline{\rho} > 0$, and by factoring this positive integer into $\mathbb{Z}$-primes and using the fact that ρ is a prime, we see that $\rho | p$ for some $\mathbb{Z}$-prime p. So write $p = \rho x$. Then $N(\rho)N(x) = p^2$, and thus either $N(\rho) = p^2$ or $N(\rho) = p$. In the first case, we must then have that $N(x) = 1$, but this makes x a unit, so $\rho \sim p$, and these irreducible elements we found earlier (and we also see that we must have $p \equiv 3 \mod 4$ in this case).

If $N(\rho) = p$, then $\rho = a + ib$, where $a, b \in \mathbb{Z}$ with $a^2 + b^2 = p$. Again, we are back in a case already dealt with; it also follows that $p = 2$ or $p \equiv 1 \mod 4$.

Now the situation is clear. The irreducible elements of $\mathbb{Z}[i]$ are (unsurprisingly) closely related to the $\mathbb{Z}$-primes; more precisely, a $\mathbb{Z}$-prime either stays irreducible in $\mathbb{Z}[i]$, or it admits a factorization of the form $p = (a+ib)(a-ib)$, and then these factors $a \pm ib$ are irreducible. This accounts for all irreducible elements of $\mathbb{Z}[i]$. We summarize:

Theorem 4.40. *The irreducible elements of $\mathbb{Z}[i]$ are exactly the ones listed below and their associates: (1) $\mathbb{Z}$-primes p with $p \equiv 3 \mod 4$; (2) elements of the form $a + ib$, with $a^2 + b^2 = p$ for some $\mathbb{Z}$-prime p*

Integers a, b as in (2) exist precisely if $p = 2$ or $p \equiv 1 \mod 4$; given such a p, there are unique a, b with $a^2 + b^2 = p$ and $1 \leq a \leq b$, and the irreducible elements corresponding to this p are $a \pm ib$ and their associates.

Exercise 4.77. Establish the claims made in the last paragraph.

Exercise 4.78. Factor $15 - 3i \in \mathbb{Z}[i]$ into irreducible elements.

Exercise 4.79. Complete the work begun in Theorem 4.39 by proving the following: an integer $n \geq 2$ can be written as the sum of two squares if and only if the primes $p \equiv 3 \mod 4$ that occur in the factorization $n = p_1^{e_1} \cdots p_k^{e_k}$ (in $\mathbb{Z}$) all have even exponents e. *Suggestion:* Observe that n is a sum of two squares precisely if $n = (a + ib)(a - ib)$; now factor $a \pm ib$ into primes in $\mathbb{Z}[i]$.

Perhaps two classical number theoretic results are worth mentioning in this context: (a) Any arithmetic progression $an + b$, $n \geq 0$, with

$(a, b) = 1$ contains infinitely many primes (Dirichlet); (b) Any non-negative integer can be written as the sum of four squares (Lagrange).

4.6. **Unique factorization in polynomial rings.** Recall that $F[x]$ is a PID if F is a field, so we have unique factorization in polynomial rings over a field. The general situation is less clear, though; recall also that $\mathbb{Z}[x]$, for example, is not a PID.

Exercise 4.80. Prove the converse of Theorem 4.20: if $R[x]$ is a PID, then R is a field. *Suggestion:* Use the homomorphism $R[x] \to R$, $x \mapsto 0$, $a \mapsto a$ ($a \in R$), to represent $R \cong R[x]/I$, and then investigate the ideals of this ring.

Our main goal in this section is to prove

Theorem 4.41 (Gauß)**.** *If D is a UFD, then so is $D[x]$.*

For $f = a_0 + \ldots + a_n x^n \in D[x]$, $f \neq 0$, and D a UFD, we define the *content* of f as $c(f) = \gcd(a_0, \ldots, a_n)$. Here we use the fact that gcd's exist in a UFD. Notice, however, that $c(f)$ is only determined up to associates (since gcd's are). We'll be cavalier about this point; for example, we will often write $c(f) = a$ to indicate that $c(f)$ can be taken to be a or any associate of a.

If we factor out the (better: a) content of f, we can write $f(x) = c(f)f_1(x)$, with $f_1 \in D[x]$ and $c(f_1) = 1$. Such a polynomial, with content 1, is called *primitive.* Observe that this factorization of f is unique in the sense that if also $f = dg$ with $d \in D$, $g \in D[x]$, g primitive, then $d \sim c(f)$ and $g = uf_1$, $u \in D$ a unit. To see this, simply note that if $g = b_0 + \ldots + b_n x^n$, then

$$c(f) = \gcd(db_0, \ldots, db_n) = d \gcd(b_0, \ldots, b_n) = dc(g) = d,$$

so indeed $c(f) \sim d$ or $c(f) = ud$, and thus also $uf_1 = g$.

Lemma 4.42 (Gauß)**.** *Let $f, g \in D[x]$ be primitive, D a UFD. Then fg is primitive.*

Proof. A direct computational approach that looks at the coefficients of fg and their divisors would work fine, but here's a slick argument that saves some work: write $h = fg$ and suppose that $c(h) \not\sim 1$. Then there is some irreducible element $p \in D$ that divides all coefficients of h and thus also h itself. Now let $\overline{D} = D/(p)$. This is a domain because if we had $(a + (p))(b + (p)) = 0 = (p)$, then $ab \in (p)$, so $p|ab$, but then $p|a$ or $p|b$ and hence a or b represents the zero element of $\overline{D}$.

Apply the homomorphism $D[x] \to \overline{D}[x]$, $x \mapsto x$, $a \mapsto \overline{a}$ to $fg = h$ to obtain that $\overline{f}\,\overline{g} = \overline{h}$ and $\overline{h} = 0$ since, by assumption, all coefficients of h represent the zero element of $\overline{D}$. Since $\overline{D}[x]$ is a domain by Theorem

4.17, it follows that $\overline{f} = 0$ or $\overline{g} = 0$, but then all coefficients of this polynomial (let's say it was $\overline{f}$) must be zero (in $\overline{D}$). It follows that $p|c(f)$, contrary to our assumption that f is primitive. This contradiction shows that $c(h) \sim 1$. □

We will now compare factorizations in $D[x]$ with those in $F[x]$, where $F = F(D)$ is the field of fractions of D (see Section 4.2 for this). First of all, we observe that the essentially unique representation $f = cf_1$ with f_1 primitive extends to polynomials from $F[x]$.

Lemma 4.43. *Every polynomial $f \in F[x]$, $f \neq 0$, can be written as $f = \gamma f_1$, with $\gamma \in F^\times$ and $f_1 \in D[x]$ primitive. Moreover, if $f = \delta g$ is another such factorization, then $\delta = u\gamma$, $f_1 = ug$ for some unit $u \in D$.*

Proof. Write $f = \alpha_0 + \ldots + \alpha_n x^n$, with $\alpha_j \in F$, so $\alpha_j = a_j b_j^{-1}$, $a_j, b_j \in D$, $b_j \neq 0$. We get the coefficients into D by the obvious method of multiplying through by the denominators (compare this with what we would have done in the case $D = \mathbb{Z}$, so $F = \mathbb{Q}$). Let $b = b_0 b_1 \cdots b_n \in D$, $b \neq 0$. Then $h := bf \in D[x]$, and we can write $h = c(h)f_1$, with $f_1 \in D[x]$ primitive. We obtain that $bf = cf_1$ or $f = \gamma f_1$, with $\gamma = c/b$, as desired.

If also $f = \delta g$ and $\delta = d/e$, $d, e \in D$, then $cef_1 = bdg$. Since f_1, g are both primitive, we must have that $be \sim cd$ in D, that is, $uce = bd$ for some unit $u \in D$. Now our claims follow by rearranging. □

This has the following useful consequence:

Theorem 4.44. *Let D be a UFD, $f \in D[x]$. If $f = g_1 g_2$ in $F[x]$, $g_j \in F[x]$, then there exists $\alpha \in F^\times$ such that $\alpha g_1, \alpha^{-1} g_2 \in D[x]$.*

As a consequence, if $f \in D[x]$, $\deg f \geq 1$, is irreducible in $D[x]$, then f is also irreducible in $F[x]$.

In general, irreducible elements can easily fail to be irreducible in a larger ring because new elements become available for potential factorizations. For example, we saw earlier that primes $p \in \mathbb{Z}$, $p \equiv 1 \mod 4$, are not irreducible in $\mathbb{Z}[i] \supset \mathbb{Z}$. Observe also that the irreducible elements of $D \subset D[x]$ become units in $F[x]$, so the assumption that $\deg f \geq 1$ is necessary.

Proof. We focus on the case $f \neq 0$. Use Lemma 4.43 to write $g_j = \gamma_j h_j$, with $h_j \in D[x]$ primitive and $\gamma_j \in F^\times$. Then $f = \gamma_1 \gamma_2 h_1 h_2$, and $h_1 h_2 \in D[x]$ is primitive by Gauß's Lemma. Now the uniqueness part of Lemma 4.43 shows that $\gamma_1 \gamma_2 = uc(f)$ for some unit $u \in D$ from D. In particular, it follows that $\gamma_1 \gamma_2 \in D$, and thus we can take $\alpha = \gamma_1^{-1}$. □

We are now finally ready for the

Proof of Theorem 4.41. We will verify the conditions from Theorem 4.34. Let $f \in D[x]$, and suppose that $g \in D[x]$ is a proper divisor of f. If $\deg g = \deg f$, then the corresponding factorization reads $f = ag$, with $a \in D$ and a is not a unit. Then $ac(g) = c(f)$, so $c(g)$ is a proper divisor of $c(f)$.

This observation guarantees that $f \in D[x]$ cannot have an infinite chain of proper divisors: when we pass to a proper divisor, either the degree goes down, but this can happen at most $\deg f$ times, or the content of the divisor is a proper divisor of $c(f)$, but again this cannot go on forever because D is a UFD and thus satisfies the DCC. We have verified that DCC holds in $D[x]$.

Next, we show that every irreducible element of $D[x]$ is a prime. So let $f \in D[x]$ be irreducible, and suppose that $f|gh$. If $\deg f \geq 1$, then f is also irreducible in $F[x]$, by Theorem 4.44. We do know that $F[x]$ is a UFD, so $f|g$, say, in $F[x]$. In other words, $fk = g$ for some $k \in F[x]$. Use Lemma 4.43 to write $k = \gamma k_1$, with $k_1 \in D[x]$ primitive. Then $g = \gamma f k_1$, and here $fk_1 \in D[x]$ is primitive by Gauß's Lemma, so the uniqueness part of Lemma 4.43 now shows that $\gamma = uc(g) \in D$, and thus $f|g$ in $D[x]$ also.

Exercise 4.81. Let $g, h \in D[x]$. Show that $c(gh) = c(g)c(h)$.

If $f = a \in D$ is irreducible and $a|gh$, then $a|c(gh) = c(g)c(h)$, by the Exercise, so $a|c(g)$ or $a|c(h)$ because D is a UFD and thus the irreducible element a is a prime in D. It follows that a also divides the corresponding polynomial, g or h, in $D[x]$. □

Example 4.8. You probably know the classical argument for the irrationality of $\sqrt{2}$: assume that p/q solves $x^2 = 2$, derive a contradiction from the additional assumption that p/q is in reduced form. Theorem 4.44 can be used to streamline and generalize such arguments.

Let $f \in \mathbb{Z}[x]$ be a monic (= leading coefficient 1) polynomial. Then I claim that every root of f in $\mathbb{R}$ is either an integer or irrational. To see this, suppose that $f(a) = 0$ for an $a \in \mathbb{Q}$. Then $f(x) = (x-a)g(x)$ in $\mathbb{Q}[x]$. Now Theorem 4.44 shows that this factorization also works in $\mathbb{Z}[x]$, after adjusting by a factor from $\mathbb{Q}$ to get $x - a$, $g(x)$ into $\mathbb{Z}[x]$. However, since f is monic, any factor in $\mathbb{Z}[x]$ must be monic itself, after adjusting an overall sign if needed (just keep track of the highest coefficient to see this). We conclude that $x - a \in \mathbb{Z}[x]$, and this says that $a \in \mathbb{Z}$. (This can also be seen directly, with no tools, and in fact you perhaps did just that when solving Exercise 4.69.)

If you apply this to $f(x) = x^n - a$, with a a positive integer, you see that $\sqrt[n]{a}$ is irrational unless $a = N^n$ for some $N \in \mathbb{N}$. (The classical

argument alluded to above makes one wonder if perhaps $\sqrt{10}$ or $\sqrt[3]{36}$ could be rational.)

Exercise 4.82. Show that $\sqrt{2}+\sqrt{3}$ is irrational.

Exercise 4.83. Show that the polynomial $f(x) = 2x - 1$ is irreducible in $\mathbb{Z}[x]$. However, in $\mathbb{Q}[x]$ it can be factored as $f(x) = 2 \cdot (x - 1/2)$. Please explain.

Here's a useful criterion for checking irreducibility:

Theorem 4.45 (Eisenstein)**.** *Consider $f(x) = a_0 + \ldots + a_n x^n \in D[x]$, D a UFD. If there exists a prime $p \in D$ such that $p|a_j$, $0 \leq j \leq n-1$, $p \nmid a_n$, $p^2 \nmid a_0$, then f is irreducible in $F[x]$.*

The special case $D = \mathbb{Z}$, so $F = \mathbb{Q}$, is of particular interest. Notice also that the conditions from Eisenstein's criterion do not prevent the a_j from having a common divisor in D, so f need not be irreducible in $D[x]$.

Proof. As in the proof of Gauß's Lemma, consider the coefficients modulo p. More precisely, apply the homomorphism $D[x] \to \overline{D}[x]$, $x \mapsto x$, $a \mapsto \overline{a} = a + (p)$. Recall that $\overline{D} = D/(p)$ is a domain (show it again please if you are not sure). Our assumptions give that $\overline{f}(x) = \overline{a_n}x^n$, so if we have a factorization $f = gh$, then $\overline{g} = \overline{b}x^k$, $\overline{h} = \overline{c}x^{n-k}$ for some $0 \leq k \leq n$ and $\overline{b}, \overline{c} \neq 0$. If $k = 0$ or $k = n$, then one of the (original) polynomials g, h is constant and thus a unit of $F[x]$. If we had $1 \leq k \leq n-1$, then it would follow that p divides the constant terms of both g and h, but this would imply that $p^2|a_0$, so this is impossible. $\square$