16

More rings

This chapter develops a number of more advanced concepts concerning rings. These concepts will play important roles later in the text, and we prefer to discuss them now, so as to avoid too many interruptions of the flow of subsequent discussions.

16.1 Algebras

Throughout this section, R denotes a ring (i.e., a commutative ring with unity).

Sometimes, a ring may also be naturally viewed as an R-module, in which case, both the theory of rings and the theory of modules may be brought to bear to study its properties.

Definition 16.1. *An R-**algebra** is a set E, together with addition and multiplication operations on E, and a function $\mu : R \times E \to E$, such that*

(i) with respect to addition and multiplication, E forms a ring;

(ii) with respect to addition and the scalar multiplication map μ, E forms an R-module;

(iii) for all $c \in R$, and $\alpha, \beta \in E$, we have

$$\mu(c, \alpha)\beta = \mu(c, \alpha\beta) = \alpha\mu(c, \beta).$$

An R-algebra E may also be called an **algebra over** R. As we usually do for R-modules, we shall write $c\alpha$ (or $c \cdot \alpha$) instead of $\mu(c, \alpha)$. When we do this, part (iii) of the definition states that

$$(c\alpha)\beta = c(\alpha\beta) = \alpha(c\beta)$$

for all $c \in R$ and $\alpha, \beta \in E$. In particular, we may write $c\alpha\beta$ without any ambiguity. Note that there are two multiplication operations at play here: scalar multiplication

(such as $c\alpha$), and ring multiplication (such as $\alpha\beta$). Also note that since we are assuming E is commutative, the second equality in part (iii) is already implied by the first. A simple consequence of the definition is that for all $c, d \in R$ and $\alpha, \beta \in E$, we have $(c\alpha)(d\beta) = (cd)(\alpha\beta)$. From this, it follows that for all $c \in R$, $\alpha \in E$, and $k \geq 0$, we have $(c\alpha)^k = c^k \alpha^k$.

Example 16.1. Suppose E is a ring and $\tau : R \to E$ is a ring homomorphism. With scalar multiplication defined by $c\alpha := \tau(c)\alpha$ for $c \in R$ and $\alpha \in E$, one may easily check that E is indeed an R-algebra. In this case, we say that E is an R-algebra **via the map** τ. □

Example 16.2. If R is a subring of E, then with $\tau : R \to E$ being the inclusion map, we can view E as an R-algebra as in the previous example. In this case, we say that E is an R-algebra **via inclusion**. □

Example 16.3. If $\tau : R \to E$ is a natural embedding of rings, then by a slight abuse of terminology, just as we sometimes say that R is a subring of E, we shall also say that E is an R-algebra via inclusion. □

In fact, all R-algebras can be viewed as special cases of Example 16.1:

Theorem 16.2. *If E is an R-algebra, then the map*

$$\begin{aligned} \tau : \quad & R \to E \\ & c \mapsto c \cdot 1_E, \end{aligned}$$

is a ring homomorphism, and $c\alpha = \tau(c)\alpha$ for all $c \in R$ and $\alpha \in E$.

Proof. Exercise. □

In the special situation where R is a *field*, we can say even more. In this situation, and with τ as in the above theorem, then either E is trivial or τ is injective (see Exercise 7.47). In the latter case, E contains an isomorphic copy of R as a subring. To summarize:

Theorem 16.3. *If R is a field, then an R-algebra is either the trivial ring or contains an isomorphic copy of R as a subring.*

The following examples give further important constructions of R-algebras.

Example 16.4. If $E_1, \ldots, E_k$ are R-algebras, then their direct product $E_1 \times \cdots \times E_k$ is an R-algebra as well, where addition, multiplication, and scalar multiplication are defined component-wise. As usual, if $E = E_1 = \cdots = E_k$, we write this as $E^{\times k}$. □

Example 16.5. If I is an arbitrary set, and E is an R-algebra, then $\mathrm{Map}(I, E)$, which is the set of all functions $f : I \to E$, may be naturally viewed as an R-algebra, with addition, multiplication, and scalar multiplication defined pointwise. □

Example 16.6. Let E be an R-algebra and let I be an ideal of E. Then it is easily verified that I is also a submodule of E. This means that the quotient ring E/I may also be viewed as an R-module, and indeed, it is an R-algebra, called the **quotient algebra (over R) of E modulo I**. For $\alpha, \beta \in E$ and $c \in R$, addition, multiplication, and scalar multiplication in E are defined as follows:

$$[\alpha]_I + [\beta]_I := [\alpha + \beta]_I, \quad [\alpha]_I \cdot [\beta]_I := [\alpha \cdot \beta]_I, \quad c \cdot [\alpha]_I := [c \cdot \alpha]_I. \quad \square$$

Example 16.7. The ring of polynomials $R[X]$ is an R-algebra via inclusion. Let $f \in R[X]$ be a non-zero polynomial with $\mathrm{lc}(f) \in R^*$. We may form the quotient ring $E := R[X]/(f)$, which may naturally be viewed as an R-algebra, as in the previous example. If $\deg(f) = 0$, then E is trivial; so assume $\deg(f) > 0$, and consider the map

$$\begin{aligned} \tau : \quad & R \to E \\ & c \mapsto c \cdot 1_E \end{aligned}$$

from Theorem 16.2. By definition, $\tau(c) = [c]_f$. As discussed in Example 7.55, the map τ is a natural embedding of rings, and so by identifying R with its image in E under τ, we can view R as a subring of E; therefore, we can also view E as an R-algebra via inclusion. □

Subalgebras

Let E be an R-algebra. A subset S of E is called a **subalgebra (over R) of E** if it is both a subring of E and a submodule of E. This means that S contains 1_E, and is closed under addition, multiplication, and scalar multiplication; restricting these operations to S, we may view S as an R-algebra in its own right.

The following theorem gives a simple but useful characterization of subalgebras, in relation to subrings:

Theorem 16.4. *If E is an R-algebra via inclusion, and S is a subring of E, then S is a subalgebra if and only if S contains R. More generally, if E is an arbitrary R-algebra, and S is a subring of E, then S is a subalgebra of E if and only if S contains $c \cdot 1_E$ for all $c \in R$.*

Proof. Exercise. □

R-algebra homomorphisms

Let E and E' be R-algebras. A function $\rho : E \to E'$ is called an **R-algebra homomorphism** if ρ is both a ring homomorphism and an R-linear map. This means that $\rho(1_E) = 1_{E'}$, and

$$\rho(\alpha + \beta) = \rho(\alpha) + \rho(\beta),\ \rho(\alpha\beta) = \rho(\alpha)\rho(\beta),\ \text{and}\ \rho(c\alpha) = c\rho(\alpha)$$

for all $\alpha, \beta \in E$ and all $c \in R$. As usual, if ρ is bijective, then it is called an **R-algebra isomorphism**, and if, in addition, $E = E'$, it is called an **R-algebra automorphism**.

The following theorem gives a simple but useful characterization of R-algebra homomorphisms, in relation to ring homomorphisms:

Theorem 16.5. *If E and E' are R-algebras via inclusion, and $\rho : E \to E'$ is a ring homomorphism, then ρ is an R-algebra homomorphism if and only if the restriction of ρ to R is the identity map. More generally, if E and E' are arbitrary R-algebras and $\rho : E \to E'$ is a ring homomorphism, then ρ is an R-algebra homomorphism if and only if $\rho(c \cdot 1_E) = c \cdot 1_{E'}$ for all $c \in R$.*

Proof. Exercise. □

Example 16.8. If E is an R-algebra and I is an ideal of E, then as observed in Example 16.6, I is also a submodule of E, and we may form the quotient algebra E/I. The natural map

$$\begin{aligned} \rho : \quad & E \to E/I \\ & \alpha \mapsto [\alpha]_I \end{aligned}$$

is both a ring homomorphism and an R-linear map, and hence is an R-algebra homomorphism. □

Example 16.9. Since $\mathbb{C}$ contains $\mathbb{R}$ as a subring, we may naturally view $\mathbb{C}$ as an $\mathbb{R}$-algebra via inclusion. The complex conjugation map on $\mathbb{C}$ that sends $a + bi$ to $a - bi$, for $a, b \in \mathbb{R}$, is an $\mathbb{R}$-algebra automorphism on $\mathbb{C}$ (see Example 7.5). □

Many simple facts about R-algebra homomorphisms can be obtained by combining corresponding facts for ring and R-module homomorphisms. For example, the composition of two R-algebra homomorphisms is again an R-algebra homomorphism, since the composition is both a ring homomorphism and an R-linear map (Theorems 7.22 and 13.6). As another example, if $\rho : E \to E'$ is an R-algebra homomorphism, then its image S' is both a subring and a submodule of E', and hence, S' is a subalgebra of E'. The kernel K of ρ is an ideal of E, and we may form the quotient algebra E/K. The first isomorphism theorems for rings and modules (Theorems 7.26 and 13.9) tell us that E/K and S' are isomorphic

both as rings and as R-modules, and hence, they are isomorphic as R-algebras. Specifically, the map

$$\begin{aligned} \bar{\rho}: \quad & E/K \to E' \\ & [\alpha]_K \mapsto \rho(\alpha) \end{aligned}$$

is an injective R-algebra homomorphism whose image is S'.

The following theorem isolates an important subalgebra associated with any R-algebra homomorphism $\rho : E \to E$.

Theorem 16.6. *Let E be an R-algebra, and let $\rho : E \to E$ be an R-algebra homomorphism. Then the set $S := \{\alpha \in E : \rho(\alpha) = \alpha\}$ is a subalgebra of E, called the* ***subalgebra of E fixed by*** *ρ. Moreover, if E is a field, then so is S.*

Proof. Let us verify that S is closed under addition. If $\alpha, \beta \in S$, then we have

$$\begin{aligned} \rho(\alpha + \beta) &= \rho(\alpha) + \rho(\beta) \ \text{ (since } \rho \text{ is a group homomorphism)} \\ &= \alpha + \beta \ \text{ (since } \alpha, \beta \in S). \end{aligned}$$

Using the fact that ρ is a ring homomorphism, one can similarly show that S is closed under multiplication, and that $1_E \in S$. Likewise, using the fact that ρ is an R-linear map, one can also show that S is closed under scalar multiplication.

This shows that S is a subalgebra, proving the first statement. For the second statement, suppose that E is a field. Let α be a non-zero element of S, and suppose $\beta \in E$ is its multiplicative inverse, so that $\alpha\beta = 1_E$. We want to show that β lies in S. Again, using the fact that ρ is a ring homomorphism, we have

$$\alpha\beta = 1_E = \rho(1_E) = \rho(\alpha\beta) = \rho(\alpha)\rho(\beta) = \alpha\rho(\beta),$$

and hence $\alpha\beta = \alpha\rho(\beta)$; canceling α, we obtain $\beta = \rho(\beta)$, and so $\beta \in S$. □

Example 16.10. The subalgebra of $\mathbb{C}$ fixed by the complex conjugation map is $\mathbb{R}$. □

Polynomial evaluation

Let E be an R-algebra. Consider the ring of polynomials $R[X]$ (which is an R-algebra via inclusion). Any polynomial $g \in R[X]$ naturally defines a function on E: if $g = \sum_i a_i X^i$, with each $a_i \in R$, and $\alpha \in E$, then

$$g(\alpha) := \sum_i a_i \alpha^i.$$

Just as for rings, we say that α is a **root** of g if $g(\alpha) = 0_E$.

For fixed $\alpha \in E$, the **polynomial evaluation map**

$$\begin{aligned} \rho : \quad R[X] &\to E \\ g &\mapsto g(\alpha) \end{aligned}$$

is easily seen to be an R-algebra homomorphism. The image of ρ is denoted $R[\alpha]$, and is a subalgebra of E. Indeed, $R[\alpha]$ is the smallest subalgebra of E containing α, and is called the **subalgebra (over R) generated by** α. Note that if E is an R-algebra via inclusion, then the notation $R[\alpha]$ has the same meaning as that introduced in Example 7.44.

We next state a very simple, but extremely useful, fact:

Theorem 16.7. *Let $\rho : E \to E'$ be an R-algebra homomorphism. Then for all $g \in R[X]$ and $\alpha \in E$, we have*

$$\rho(g(\alpha)) = g(\rho(\alpha)).$$

Proof. Let $g = \sum_i a_i X^i \in R[X]$. Then we have

$$\begin{aligned} \rho(g(\alpha)) &= \rho\Big(\sum_i a_i \alpha^i\Big) = \sum_i \rho(a_i \alpha^i) = \sum_i a_i \rho(\alpha^i) = \sum_i a_i \rho(\alpha)^i \\ &= g(\rho(\alpha)). \quad \square \end{aligned}$$

As a special case of Theorem 16.7, if $E = R[\alpha]$ for some $\alpha \in E$, then every element of E can be expressed as $g(\alpha)$ for some $g \in R[X]$, and $\rho(g(\alpha)) = g(\rho(\alpha))$; hence, the action of ρ is completely determined by its action on α.

Example 16.11. Let $f \in R[X]$ be a non-zero polynomial with $\mathrm{lc}(f) \in R^*$. As in Example 16.7, we may form the quotient algebra $E := R[X]/(f)$.

Let $\xi := [X]_f \in E$. Then $E = R[\xi]$, and moreover, every element of E can be expressed uniquely as $g(\xi)$, where $g \in R[X]$ and $\deg(g) < \deg(f)$. In addition, ξ is a root of f. If $\deg(f) > 0$, these facts were already observed in Example 7.55, and otherwise, they are trivial.

Now let E' be any R-algebra, and suppose that $\rho : E \to E'$ is an R-algebra homomorphism, and let $\xi' := \rho(\xi)$. By the previous theorem, ρ sends $g(\xi)$ to $g(\xi')$, for each $g \in R[X]$. Thus, the image of ρ is $R[\xi']$. Also, we have $f(\xi') = f(\rho(\xi)) = \rho(f(\xi)) = \rho(0_E) = 0_{E'}$. Therefore, ξ' must be a root of f.

Conversely, suppose that $\xi' \in E'$ is a root of f. Then the polynomial evaluation map from $R[X]$ to E' that sends $g \in R[X]$ to $g(\xi') \in E'$ is an R-algebra homomorphism whose kernel contains f. Using the generalized versions of the first isomorphism theorems for rings and R-modules (Theorems 7.27 and 13.10),

we obtain the R-algebra homomorphism

$$\begin{aligned} \rho : \quad & E \to E' \\ & g(\xi) \mapsto g(\xi'). \end{aligned}$$

One sees that complex conjugation is just a special case of this construction (see Example 7.57). □

EXERCISE 16.1. Let E be an R-algebra. For $\alpha \in E$, consider the α-multiplication map on E, which sends $\beta \in E$ to $\alpha\beta \in E$. Show that this map is an R-linear map.

EXERCISE 16.2. Show that every ring may be viewed in a unique way as a $\mathbb{Z}$-algebra, and that subrings are subalgebras, and ring homomorphisms are $\mathbb{Z}$-algebra homomorphisms.

EXERCISE 16.3. Show that the only $\mathbb{R}$-algebra homomorphisms from $\mathbb{C}$ into itself are the identity map and the complex conjugation map.

16.2 The field of fractions of an integral domain

Let D be an integral domain. Just as we can construct the field of rational numbers by forming fractions involving integers, we can construct a field consisting of fractions whose numerators and denominators are elements of D. This construction is quite straightforward, though a bit tedious.

To begin with, let S be the set of all pairs of the form (a, b), with $a, b \in D$ and $b \neq 0_D$. Intuitively, such a pair (a, b) is a "formal fraction," with numerator a and denominator b. We define a binary relation $\sim$ on S as follows: for $(a_1, b_1), (a_2, b_2) \in S$, we say $(a_1, b_1) \sim (a_2, b_2)$ if and only if $a_1 b_2 = a_2 b_1$. Our first task is to show that this is an equivalence relation:

Lemma 16.8. *For all $(a_1, b_1), (a_2, b_2), (a_3, b_3) \in S$, we have*

(i) $(a_1, b_1) \sim (a_1, b_1)$;

(ii) $(a_1, b_1) \sim (a_2, b_2)$ *implies* $(a_2, b_2) \sim (a_1, b_1)$;

(iii) $(a_1, b_1) \sim (a_2, b_2)$ *and* $(a_2, b_2) \sim (a_3, b_3)$ *implies* $(a_1, b_1) \sim (a_3, b_3)$.

Proof. (i) and (ii) are rather trivial, and we do not comment on these any further. As for (iii), assume that $a_1 b_2 = a_2 b_1$ and $a_2 b_3 = a_3 b_2$. Multiplying the first equation by b_3, we obtain $a_1 b_2 b_3 = a_2 b_1 b_3$ and substituting $a_3 b_2$ for $a_2 b_3$ on the right-hand side of this last equation, we obtain $a_1 b_2 b_3 = a_3 b_2 b_1$. Now, using the fact that b_2 is non-zero and that D is an integral domain, we may cancel b_2 from both sides, obtaining $a_1 b_3 = a_3 b_1$. □

Since $\sim$ is an equivalence relation, it partitions S into equivalence classes, and for $(a, b) \in S$, we denote by $[a, b]$ the equivalence class containing (a, b), and we denote by K the set of all such equivalence classes. Our next task is to define addition and multiplication operations on equivalence classes, mimicking the usual rules of arithmetic with fractions. We want to define the sum of $[a_1, b_1]$ and $[a_2, b_2]$ to be $[a_1b_2 + a_2b_1, b_1b_2]$, and the product of $[a_1, b_1]$ and $[a_2, b_2]$ to be $[a_1a_2, b_1b_2]$. Note that since D is an integral domain, if b_1 and b_2 are non-zero, then so is the product b_1b_2, and therefore $[a_1b_2 + a_2b_1, b_1b_2]$ and $[a_1a_2, b_1b_2]$ are indeed equivalence classes. However, to ensure that this definition is unambiguous, and does not depend on the particular choice of representatives of the equivalence classes $[a_1, b_1]$ and $[a_2, b_2]$, we need the following lemma.

Lemma 16.9. *Let $(a_1, b_1), (a'_1, b'_1), (a_2, b_2), (a'_2, b'_2) \in S$, where $(a_1, b_1) \sim (a'_1, b'_1)$ and $(a_2, b_2) \sim (a'_2, b'_2)$. Then we have*

$$(a_1b_2 + a_2b_1, b_1b_2) \sim (a'_1b'_2 + a'_2b'_1, b'_1b'_2)$$

and

$$(a_1a_2, b_1b_2) \sim (a'_1a'_2, b'_1b'_2).$$

Proof. This is a straightforward calculation. Since $a_1b'_1 = a'_1b_1$ and $a_2b'_2 = a'_2b_2$, we have

$$\begin{aligned}(a_1b_2 + a_2b_1)b'_1b'_2 &= a_1b_2b'_1b'_2 + a_2b_1b'_1b'_2 = a'_1b_2b_1b'_2 + a'_2b_1b'_1b_2 \\ &= (a'_1b'_2 + a'_2b'_1)b_1b_2\end{aligned}$$

and

$$a_1a_2b'_1b'_2 = a'_1a_2b_1b'_2 = a'_1a'_2b_1b_2. \quad \square$$

In light of this lemma, we may unambiguously define addition and multiplication on K as follows: for $[a_1, b_1], [a_2, b_2] \in K$, we define

$$[a_1, b_1] + [a_2, b_2] := [a_1b_2 + a_2b_1, b_1b_2]$$

and

$$[a_1, b_1] \cdot [a_2, b_2] := [a_1a_2, b_1b_2].$$

The next task is to show that K is a ring — we leave the details of this (which are quite straightforward) to the reader.

Lemma 16.10. *With addition and multiplication as defined above, K is a ring, with additive identity $[0_D, 1_D]$ and multiplicative identity $[1_D, 1_D]$.*

Proof. Exercise. $\square$

Finally, we observe that K is in fact a field: it is clear that $[a, b]$ is a non-zero element of K if and only if $a \neq 0_D$, and hence any non-zero element $[a, b]$ of K has a multiplicative inverse, namely, $[b, a]$.

The field K is called the **field of fractions of** D. Consider the map $\tau : D \to K$ that sends $a \in D$ to $[a, 1_D] \in K$. It is easy to see that this map is a ring homomorphism, and one can also easily verify that it is injective. So, starting from D, we can synthesize "out of thin air" its field of fractions K, which essentially contains D as a subring, via the natural embedding $\tau : D \to K$.

Now suppose that we are given a field L that contains D as a subring. Consider the set K' consisting of all elements of L of the form ab^{-1}, where $a, b \in D$ and $b \neq 0_D$ — note that here, the arithmetic operations are performed using the rules for arithmetic in L. One may easily verify that K' is a subfield of L that contains D, and it is easy to see that this is the smallest subfield of L that contains D. The subfield K' of L may be referred to as the **field of fractions of** D **within** L. One may easily verify that the map $\rho : K \to L$ that sends $[a, b] \in K$ to $ab^{-1} \in L$ is an unambiguously defined ring homomorphism that maps K injectively onto K'. If we view K and L as D-algebras via inclusion, and we see that the map ρ is in fact a D-algebra homomorphism. Thus, K and K' are isomorphic as D-algebras. It is in this sense that the field of fractions K is the smallest field that contains D as a subring.

From now on, we shall simply write an element $[a, b]$ of K as the fraction a/b. In this notation, the above rules for addition, multiplication, and testing equality in K now look quite familiar:

$$\frac{a_1}{b_1} + \frac{a_2}{b_2} = \frac{a_1 b_2 + a_2 b_1}{b_1 b_2}, \quad \frac{a_1}{b_1} \cdot \frac{a_2}{b_2} = \frac{a_1 a_2}{b_1 b_2}, \quad \frac{a_1}{b_1} = \frac{a_2}{b_2} \iff a_1 b_2 = a_2 b_1.$$

Function fields

An important special case of the above construction for the field of fractions of D is when $D = F[X]$, where F is a field. In this case, the field of fractions is denoted $F(X)$, and is called the **field of rational functions (over** F**)**. This terminology is a bit unfortunate, since just as with polynomials, although the elements of $F(X)$ define functions, they are not (in general) in one-to-one correspondence with these functions.

Since $F[X]$ is a subring of $F(X)$, and since F is a subring of $F[X]$, we see that F is a subfield of $F(X)$.

More generally, we may apply the above construction to $D = F[X_1, \ldots, X_n]$, the ring of multi-variate polynomials over the field F, in which case the field of

fractions is denoted $F(X_1, \ldots, X_n)$, and is also called the field of rational functions (over F, in the variables $X_1, \ldots, X_n$).

EXERCISE 16.4. Let F be a field of characteristic zero. Show that F contains an isomorphic copy of $\mathbb{Q}$.

EXERCISE 16.5. Show that the field of fractions of $\mathbb{Z}[i]$ within $\mathbb{C}$ is $\mathbb{Q}[i]$. (See Example 7.25 and Exercise 7.14.)

16.3 Unique factorization of polynomials

Throughout this section, F denotes a field.

Like the ring $\mathbb{Z}$, the ring $F[X]$ of polynomials is an integral domain, and because of the division with remainder property for polynomials, $F[X]$ has many other properties in common with $\mathbb{Z}$. Indeed, essentially all the ideas and results from Chapter 1 can be carried over almost verbatim from $\mathbb{Z}$ to $F[X]$, and in this section, we shall do just that.

Recall that the units of $F[X]$ are precisely the units F^* of F, that is, the non-zero constants. We call two polynomials $g, h \in F[X]$ **associate** if $g = ch$ for some $c \in F^*$. It is easy to see that g and h are associate if and only if $g \mid h$ and $h \mid g$ — indeed, this follows as a special case of part (i) of Theorem 7.4. Clearly, any non-zero polynomial g is associate to a unique monic polynomial (i.e., a polynomial with leading coefficient 1), called the **monic associate** of g; indeed, the monic associate of g is $\mathrm{lc}(g)^{-1} \cdot g$ (where, as usual, $\mathrm{lc}(g)$ denotes the leading coefficient of g).

We call a polynomial $f \in F[X]$ **irreducible** if it is non-constant and all divisors of f are associate to 1 or f. Conversely, we call f **reducible** if it is non-constant and is not irreducible. Equivalently, a non-constant polynomial f is reducible if and only if there exist polynomials $g, h \in F[X]$ of degree strictly less than that of f such that $f = gh$.

Clearly, if g and h are associate polynomials, then g is irreducible if and only if h is irreducible.

The irreducible polynomials play a role similar to that of the prime numbers. Just as it is convenient to work with only *positive* prime numbers, it is also convenient to restrict attention to *monic* irreducible polynomials.

Corresponding to Theorem 1.3, every non-zero polynomial can be expressed as a unit times a product of monic irreducibles in an essentially unique way:

Theorem 16.11. *Every non-zero polynomial $f \in F[X]$ can be expressed as*

$$f = c \cdot p_1^{e_1} \cdots p_r^{e_r},$$

where $c \in F^*$, $p_1, \ldots, p_r$ are distinct monic irreducible polynomials, and $e_1, \ldots, e_r$ are positive integers. Moreover, this expression is unique, up to a reordering of the irreducible polynomials.

To prove this theorem, we may assume that f is monic, since the non-monic case trivially reduces to the monic case.

The proof of the existence part of Theorem 16.11 is just as for Theorem 1.3. If f is 1 or a monic irreducible, we are done. Otherwise, there exist $g, h \in F[X]$ of degree strictly less than that of f such that $f = gh$, and again, we may assume that g and h are monic. By induction on degree, both g and h can be expressed as a product of monic irreducible polynomials, and hence, so can f.

The proof of the uniqueness part of Theorem 16.11 is almost identical to that of Theorem 1.3. The key to the proof is the division with remainder property, Theorem 7.10, from which we can easily derive the following analog of Theorem 1.6:

Theorem 16.12. *Let I be an ideal of $F[X]$. Then there exists a unique polynomial $d \in F[X]$ such that $I = dF[X]$ and d is either zero or monic.*

Proof. We first prove the existence part of the theorem. If $I = \{0\}$, then $d = 0$ does the job, so let us assume that $I \neq \{0\}$. Since I contains non-zero polynomials, it must contain monic polynomials, since if g is a non-zero polynomial in I, then its monic associate $\mathrm{lc}(g)^{-1}g$ is also in I. Let d be a monic polynomial of minimal degree in I. We want to show that $I = dF[X]$.

We first show that $I \subseteq dF[X]$. To this end, let g be any element in I. It suffices to show that $d \mid g$. Using Theorem 7.10, we may write $g = dq + r$, where $\deg(r) < \deg(d)$. Then by the closure properties of ideals, one sees that $r = g - dq$ is also an element of I, and by the minimality of the degree of d, we must have $r = 0$. Thus, $d \mid g$.

We next show that $dF[X] \subseteq I$. This follows immediately from the fact that $d \in I$ and the closure properties of ideals.

That proves the existence part of the theorem. As for uniqueness, note that if $dF[X] = eF[X]$, we have $d \mid e$ and $e \mid d$, from which it follows that d and e are associate, and so if d and e are both either monic or zero, they must be equal. □

For $g, h \in F[X]$, we call $d \in F[X]$ a **common divisor** of g and h if $d \mid g$ and $d \mid h$; moreover, we call such a d a **greatest common divisor** of g and h if d is monic or zero, and all other common divisors of g and h divide d. Analogous to Theorem 1.7, we have:

Theorem 16.13. *For all $g, h \in F[X]$, there exists a unique greatest common divisor d of g and h, and moreover, $gF[X] + hF[X] = dF[X]$.*

Proof. We apply the previous theorem to the ideal $I := gF[X] + hF[X]$. Let

$d \in F[X]$ with $I = dF[X]$, as in that theorem. Note that $g, h, d \in I$ and d is monic or zero.

It is clear that d is a common divisor of g and h. Moreover, there exist $s, t \in F[X]$ such that $gs + ht = d$. If $d' \mid g$ and $d' \mid h$, then clearly $d' \mid (gs + ht)$, and hence $d' \mid d$.

Finally, for uniqueness, if e is a greatest common divisor of g and h, then $d \mid e$ and $e \mid d$, and hence e is associate to d, and the requirement that e is monic or zero implies that $e = d$. □

For $g, h \in F[X]$, we denote by $\gcd(g, h)$ the greatest common divisor of g and h. Note that as we have defined it, $\mathrm{lc}(g)\gcd(g, 0) = g$. Also note that when at least one of g or h are non-zero, $\gcd(g, h)$ is the unique monic polynomial of maximal degree that divides both g and h.

An immediate consequence of Theorem 16.13 is that for all $g, h \in F[X]$, there exist $s, t \in F[X]$ such that $gs + ht = \gcd(g, h)$, and that when at least one of g or h are non-zero, $\gcd(g, h)$ is the unique monic polynomial of minimal degree that can be expressed as $gs + ht$ for some $s, t \in F[X]$.

We say that $g, h \in F[X]$ are **relatively prime** if $\gcd(g, h) = 1$, which is the same as saying that the only common divisors of g and h are units. It is immediate from Theorem 16.13 that g and h are relatively prime if and only if $gF[X] + hF[X] = F[X]$, which holds if and only if there exist $s, t \in F[X]$ such that $gs + ht = 1$.

Analogous to Theorem 1.9, we have:

Theorem 16.14. *For $f, g, h \in F[X]$ such that $f \mid gh$ and $\gcd(f, g) = 1$, we have $f \mid h$.*

Proof. Suppose that $f \mid gh$ and $\gcd(f, g) = 1$. Then since $\gcd(f, g) = 1$, by Theorem 16.13 we have $fs + gt = 1$ for some $s, t \in F[X]$. Multiplying this equation by h, we obtain $fhs + ght = h$. Since $f \mid f$ by definition, and $f \mid gh$ by hypothesis, it follows that $f \mid h$. □

Analogous to Theorem 1.10, we have:

Theorem 16.15. *Let $p \in F[X]$ be irreducible, and let $g, h \in F[X]$. Then $p \mid gh$ implies that $p \mid g$ or $p \mid h$.*

Proof. Assume that $p \mid gh$. The only divisors of p are associate to 1 or p. Thus, $\gcd(p, g)$ is either 1 or the monic associate of p. If $p \mid g$, we are done; otherwise, if $p \nmid g$, we must have $\gcd(p, g) = 1$, and by the previous theorem, we conclude that $p \mid h$. □

Now to prove the uniqueness part of Theorem 16.11. Suppose we have

$$p_1 \cdots p_r = q_1 \cdots q_s,$$

where $p_1, \ldots, p_r$ and $q_1, \ldots, q_s$ are monic irreducible polynomials (with duplicates allowed among the p_i's and among the q_j's). If $r = 0$, we must have $s = 0$ and we are done. Otherwise, as p_1 divides the right-hand side, by inductively applying Theorem 16.15, one sees that p_1 is equal to q_j for some j. We can cancel these terms and proceed inductively (on r).

That completes the proof of Theorem 16.11.

Analogous to Theorem 1.11, we have:

Theorem 16.16. *There are infinitely many monic irreducible polynomials in $F[X]$.*

If F is infinite, then this theorem is true simply because there are infinitely many monic, linear polynomials; in any case, one can easily prove this theorem by mimicking the proof of Theorem 1.11 (as the reader may verify).

For a monic irreducible polynomial p, we may define the function ν_p, mapping non-zero polynomials to non-negative integers, as follows: for every polynomial $f \neq 0$, if $f = p^e g$, where $p \nmid g$, then $\nu_p(f) := e$. We may then write the factorization of f into irreducibles as

$$f = c \prod_p p^{\nu_p(f)},$$

where the product is over all monic irreducible polynomials p, with all but finitely many of the terms in the product equal to 1.

Just as for integers, we may extend the domain of definition of ν_p to include 0, defining $\nu_p(0) := \infty$. For all polynomials g, h, we have

$$\nu_p(g \cdot h) = \nu_p(g) + \nu_p(h) \text{ for all } p. \tag{16.1}$$

From this, it follows that for all polynomials g, h, we have

$$h \mid g \iff \nu_p(h) \leq \nu_p(g) \text{ for all } p, \tag{16.2}$$

and

$$\nu_p(\gcd(g, h)) = \min(\nu_p(g), \nu_p(h)) \text{ for all } p. \tag{16.3}$$

For $g, h \in F[X]$, a **common multiple** of g and h is a polynomial m such that $g \mid m$ and $h \mid m$; moreover, such an m is the **least common multiple** of g and h if m is monic or zero, and m divides all common multiples of g and h. In light of Theorem 16.11, it is clear that the least common multiple exists and is unique, and we denote the least common multiple of g and h by $\operatorname{lcm}(a, b)$. Note that as we have

defined it, $\mathrm{lcm}(g, 0) = 0$, and that when both g and h are non-zero, $\mathrm{lcm}(g, h)$ is the unique monic polynomial of minimal degree that is divisible by both g and h. Also, for all $g, h \in F[X]$, we have

$$\nu_p(\mathrm{lcm}(g, h)) = \max(\nu_p(g), \nu_p(h)) \text{ for all } p. \tag{16.4}$$

Just as in §1.3, the notions of greatest common divisor and least common multiple generalize naturally from two to any number of polynomials. We also say that a family of polynomials $\{g_i\}_{i=1}^k$ is **pairwise relatively prime** if $\gcd(g_i, g_j) = 1$ for all indices i, j with $i \neq j$.

Also just as in §1.3, any rational function $g/h \in F(X)$ can be expressed as a fraction g_0/h_0 in **lowest terms**—that is, $g/h = g_0/h_0$ and $\gcd(g_0, h_0) = 1$—and this representation is unique up to multiplication by units.

Many of the exercises in Chapter 1 carry over naturally to polynomials—the reader is encouraged to look over all of the exercises in that chapter, determining which have natural polynomial analogs, and work some of these out.

Example 16.12. Let $f \in F[X]$ be a polynomial of degree 2 or 3. Then it is easy to see that f is irreducible if and only if f has no roots in F. Indeed, if f is reducible, then it must have a factor of degree 1, which we can assume is monic; thus, we can write $f = (X - x)g$, where $x \in F$ and $g \in F[X]$, and so $f(x) = (x - x)g(x) = 0$. Conversely, if $x \in F$ is a root of f, then $X - x$ divides f (see Theorem 7.12), and so f is reducible. □

Example 16.13. As a special case of the previous example, consider the polynomials $f := X^2 - 2 \in \mathbb{Q}[X]$ and $g := X^3 - 2 \in \mathbb{Q}[X]$. We claim that as polynomials over $\mathbb{Q}$, f and g are irreducible. Indeed, neither of them have integer roots, and so neither of them have rational roots (see Exercise 1.26); therefore, they are irreducible. □

Example 16.14. In discussing the factorization of polynomials, one must be clear about the coefficient domain. Indeed, if we view f and g in the previous example as polynomials over $\mathbb{R}$, then they factor into irreducibles as

$$f = (X - \sqrt{2})(X + \sqrt{2}), \quad g = (X - \sqrt[3]{2})(X^2 + \sqrt[3]{2}\,X + \sqrt[3]{4}),$$

and over $\mathbb{C}$, g factors even further, as

$$g = (X - \sqrt[3]{2})\big(X - \sqrt[3]{2}(1 + i\sqrt{3})/2\big)\big(X - \sqrt[3]{2}(1 - i\sqrt{3})/2\big). \quad \square$$

EXERCISE 16.6. Suppose $f = \sum_{i=0}^{\ell} c_i X^i$ is an irreducible polynomial over F, where $c_0 \neq 0$ and $c_\ell \neq 0$. Show that the "reverse" polynomial $\tilde{f} := \sum_{i=0}^{\ell} c_{\ell-i} X^i$ is also irreducible.

16.4 Polynomial congruences

Throughout this section, F denotes a field.

Many of the results from Chapter 2 on congruences modulo a positive integer n carry over almost verbatim to congruences modulo a non-zero polynomial $f \in F[X]$. We state these results here — the proofs of these results are essentially the same as in the integer case, and as such, are omitted for the most part.

Because of the division with remainder property for polynomials, we have the analog of Theorem 2.4:

Theorem 16.17. *Let $g, f \in F[X]$, where $f \neq 0$. Then there exists a unique $z \in F[X]$ such that $z \equiv g \pmod{f}$ and $\deg(z) < \deg(f)$, namely, $z := g \bmod f$.*

Corresponding to Theorem 2.5, we have:

Theorem 16.18. *Let $g, f \in F[X]$ with $f \neq 0$, and let $d := \gcd(g, f)$.*

(i) For every $h \in F[X]$, the congruence $gz \equiv h \pmod{f}$ has a solution $z \in F[X]$ if and only if $d \mid h$.

(ii) For every $z \in F[X]$, we have $gz \equiv 0 \pmod{f}$ if and only if $z \equiv 0 \pmod{f/d}$.

(iii) For all $z, z' \in F[X]$, we have $gz \equiv gz' \pmod{f}$ if and only if $z \equiv z' \pmod{f/d}$.

Let $g, f \in F[X]$ with $f \neq 0$. Part (iii) of Theorem 16.18 gives us a **cancellation law** for polynomial congruences:

$$\text{if } \gcd(g, f) = 1 \text{ and } gz \equiv gz' \pmod{f}, \text{ then } z \equiv z' \pmod{f}.$$

We say that $z \in F[X]$ is a **multiplicative inverse of g modulo f** if $gz \equiv 1 \pmod{f}$. Part (i) of Theorem 16.18 says that g has a multiplicative inverse modulo f if and only if $\gcd(g, f) = 1$. Moreover, part (iii) of Theorem 16.18 says that the multiplicative inverse of g, if it exists, is uniquely determined modulo f.

As for integers, we may generalize the "mod" operation as follows. Suppose $g, h, f \in F[X]$, with $f \neq 0$, $g \neq 0$, and $\gcd(g, f) = 1$. If s is the rational function $h/g \in F(X)$, then we define $s \bmod f$ to be the unique polynomial $z \in F[X]$ satisfying

$$gz \equiv h \pmod{f} \quad \text{and} \quad \deg(z) < \deg(f).$$

With this notation, we can simply write $g^{-1} \bmod f$ to denote the unique multiplicative inverse of g modulo f of degree less than $\deg(f)$.

Corresponding to Theorem 2.6, we have:

Theorem 16.19 (Chinese remainder theorem). *Let $\{f_i\}_{i=1}^k$ be a pairwise relatively prime family of non-zero polynomials in $F[X]$, and let $g_1, \ldots, g_k$ be arbitrary polynomials in $F[X]$. Then there exists a solution $g \in F[X]$ to the system of congruences*

$$g \equiv g_i \pmod{f_i} \quad (i = 1, \ldots, k).$$

Moreover, any $g' \in F[X]$ is a solution to this system of congruences if and only if $g \equiv g' \pmod{f}$, where $f := \prod_{i=1}^k f_i$.

Let us recall the formula for the solution g (see proof of Theorem 2.6). We have

$$g := \sum_{i=1}^{k} g_i e_i,$$

where

$$e_i := f_i^* t_i, \quad f_i^* := f/f_i, \quad t_i := (f_i^*)^{-1} \bmod f_i \quad (i = 1, \ldots, k).$$

Now, let us consider the special case of the Chinese remainder theorem where $f_i = X - x_i$ with $x_i \in F$, and $g_i = y_i \in F$, for $i = 1, \ldots, k$. The condition that $\{f_i\}_{i=1}^k$ is pairwise relatively prime is equivalent to the condition that the x_i's are distinct. Observe that a polynomial $g \in F[X]$ satisfies the system of congruences

$$g \equiv g_i \pmod{f_i} \quad (i = 1, \ldots, k)$$

if and only if

$$g(x_i) = y_i \quad (i = 1, \ldots, k).$$

Moreover, we have $f_i^* = \prod_{j \neq i}(X - x_j)$ and $t_i = 1/\prod_{j \neq i}(x_i - x_j) \in F$. So we get

$$g = \sum_{i=1}^{k} y_i \frac{\prod_{j \neq i}(X - x_j)}{\prod_{j \neq i}(x_i - x_j)}.$$

The reader will recognize this as the usual *Lagrange interpolation formula* (see Theorem 7.15). Thus, the Chinese remainder theorem for polynomials includes Lagrange interpolation as a special case.

Polynomial quotient algebras. Let $f \in F[X]$ be a polynomial of degree $\ell \geq 0$, and consider the quotient ring $E := F[X]/(f)$. As discussed in Example 16.7, we may naturally view E as an F-algebra. Moreover, if we set $\xi := [X]_f \in E$, then $E = F[\xi]$, and viewing E as a vector space over F, we see that $\{\xi^{i-1}\}_{i=1}^{\ell}$ is a basis for E.

Now suppose $\alpha \in E$. We have $\alpha = [g]_f = g(\xi)$ for some $g \in F[X]$, and from

the above discussion about polynomial congruences, we see that α is a unit if and only if $\gcd(g, f) = 1$.

If $\ell = 0$, then E is trivial. If f is irreducible, then E is a field, since $g \not\equiv 0 \pmod{f}$ implies $\gcd(g, f) = 1$. If f is reducible, then E is not a field, and indeed, not even an integral domain: for any non-trivial factor $g \in F[X]$ of f, $[g]_f \in E$ is a zero divisor.

The Chinese remainder theorem for polynomials also has a more algebraic interpretation. Namely, if $\{f_i\}_{i=1}^k$ is a pairwise relatively prime family of non-zero polynomials in $F[X]$, and $f := \prod_{i=1}^k f_i$, then the map

$$\begin{aligned} \theta : \quad F[X]/(f) &\to F[X]/(f_1) \times \cdots \times F[X]/(f_k) \\ [g]_f &\mapsto ([g]_{f_1}, \ldots, [g]_{f_k}) \end{aligned}$$

is unambiguously defined, and is in fact an F-algebra isomorphism. This map may be seen as a generalization of the ring isomorphism $\bar{\rho}$ discussed in Example 7.54.

Example 16.15. The polynomial X^2+1 is irreducible over $\mathbb{R}$, since if it were not, it would have a root in $\mathbb{R}$ (see Example 16.12), which is clearly impossible, since -1 is not the square of any real number. It follows immediately that $\mathbb{C} = \mathbb{R}[X]/(X^2+1)$ is a field, without having to explicitly calculate a formula for the inverse of a non-zero complex number. □

Example 16.16. Consider the polynomial $f := X^4 + X^3 + 1$ over $\mathbb{Z}_2$. We claim that f is irreducible. It suffices to show that f has no irreducible factors of degree 1 or 2.

If f had a factor of degree 1, then it would have a root; however, $f(0) = 0+0+1 = 1$ and $f(1) = 1+1+1 = 1$. So f has no factors of degree 1.

Does f have a factor of degree 2? The polynomials of degree 2 are X^2, $X^2 + X$, $X^2 + 1$, and $X^2 + X + 1$. The first and second of these polynomials are divisible by X, and hence not irreducible, while the third has a 1 as a root, and hence is also not irreducible. The last polynomial, $X^2 + X + 1$, has no roots, and hence is the only irreducible polynomial of degree 2 over $\mathbb{Z}_2$. So now we may conclude that if f were not irreducible, it would have to be equal to

$$(X^2 + X + 1)^2 = X^4 + 2X^3 + 3X^2 + 2X + 1 = X^4 + X^2 + 1,$$

which it is not.

Thus, $E := \mathbb{Z}_2[X]/(f)$ is a field with $2^4 = 16$ elements. We may think of elements E as bit strings of length 4, where the rule for addition is bit-wise "exclusive-or." The rule for multiplication is more complicated: to multiply two given bit strings, we interpret the bits as coefficients of polynomials (with the left-most bit the coefficient of X^3), multiply the polynomials, reduce the product modulo f, and

write down the bit string corresponding to the reduced product polynomial. For example, to multiply 1001 and 0011, we compute

$$(X^3 + 1)(X + 1) = X^4 + X^3 + X + 1,$$

and

$$(X^4 + X^3 + X + 1) \bmod (X^4 + X^3 + 1) = X.$$

Hence, the product of 1001 and 0011 is 0010.

Theorem 7.29 says that E^* is a cyclic group. Indeed, the element $\xi :=$ 0010 (i.e., $\xi = [X]_f$) is a generator for E^*, as the following table of powers shows:

i	ξ^i	i	ξ^i
1	0010	8	1110
2	0100	9	0101
3	1000	10	1010
4	1001	11	1101
5	1011	12	0011
6	1111	13	0110
7	0111	14	1100
		15	0001

Such a table of powers is sometimes useful for computations in small finite fields such as this one. Given $\alpha, \beta \in E^*$, we can compute $\alpha\beta$ by obtaining (by table lookup) i, j such that $\alpha = \xi^i$ and $\beta = \xi^j$, computing $k := (i + j) \bmod 15$, and then obtaining $\alpha\beta = \xi^k$ (again by table lookup). □

16.5 Minimal polynomials

Throughout this section, F denotes a field.

Suppose that E is an arbitrary F-algebra, and let α be an element of E. Consider the polynomial evaluation map

$$\begin{aligned} \rho : \quad F[X] &\to E \\ g &\mapsto g(\alpha), \end{aligned}$$

which is an F-algebra homomorphism. By definition, the image of ρ is $F[\alpha]$. The kernel of ρ is an ideal of $F[X]$, and since every ideal of $F[X]$ is principal, it follows that $\operatorname{Ker} \rho = \phi F[X]$ for some polynomial $\phi \in F[X]$; moreover, we can make the choice of ϕ unique by insisting that it is monic or zero. The polynomial ϕ is called the **minimal polynomial of α (over F)**.

On the one hand, suppose $\phi \neq 0$. Since any polynomial that is zero at α is a polynomial multiple of ϕ, we see that ϕ is the unique monic polynomial of smallest

degree that vanishes at α. Moreover, the first isomorphism theorems for rings and modules tell us that $F[\alpha]$ is isomorphic (as an F-algebra) to $F[X]/(\phi)$, via the isomorphism

$$\begin{aligned} \bar{\rho}: \quad F[X]/(\phi) &\to F[\alpha] \\ [g]_\phi &\mapsto g(\alpha). \end{aligned}$$

Under this isomorphism, $[X]_\phi \in F[X]/(\phi)$ corresponds to $\alpha \in F[\alpha]$, and we see that $\{\alpha^{i-1}\}_{i=1}^m$ is a basis for $F[\alpha]$ over F, where $m = \deg(\phi)$. In particular, every element of $F[\alpha]$ can be written uniquely as $\sum_{i=1}^m c_i \alpha^{i-1}$, where $c_1, \ldots, c_m \in F$.

On the other hand, suppose $\phi = 0$. This means that no non-zero polynomial vanishes at α. Also, it means that the map ρ is injective, and hence $F[\alpha]$ is isomorphic (as an F-algebra) to $F[X]$; in particular, $F[\alpha]$ is not finitely generated as a vector space over F.

Note that if $\alpha \in E$ has a minimal polynomial $\phi \neq 0$, then $\deg(\phi) > 0$, unless E is trivial (i.e., $1_E = 0_E$), in which case $\phi = 1$.

Example 16.17. Consider the real numbers $\sqrt{2}$ and $\sqrt[3]{2}$.

We claim that $X^2 - 2$ is the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$. To see this, first observe that $\sqrt{2}$ is a root of $X^2 - 2$. Thus, the minimal polynomial of $\sqrt{2}$ divides $X^2 - 2$. However, as we saw in Example 16.13, the polynomial $X^2 - 2$ is irreducible over $\mathbb{Q}$, and hence must be equal to the minimal polynomial of $\sqrt{2}$ over $\mathbb{Q}$.

A similar argument shows that $X^3 - 2$ is the minimal polynomial of $\sqrt[3]{2}$ over $\mathbb{Q}$.

We also see that $\mathbb{Q}[\sqrt{2}]$ is isomorphic (as a $\mathbb{Q}$-algebra) to $\mathbb{Q}[X]/(X^2 - 2)$, and since $X^2 - 2$ is irreducible, it follows that the ring $\mathbb{Q}[\sqrt{2}]$ is actually a field. As a vector space over $\mathbb{Q}$, $\mathbb{Q}[\sqrt{2}]$ has dimension 2, and every element of $\mathbb{Q}[\sqrt{2}]$ may be written uniquely as $a + b\sqrt{2}$ for $a, b \in \mathbb{Q}$. Indeed, for all $a, b \in \mathbb{Q}$, not both zero, the multiplicative inverse of $a + b\sqrt{2}$ is $(a/c) + (b/c)\sqrt{2}$, where $c := a^2 - 2b^2$.

Similarly, $\mathbb{Q}[\sqrt[3]{2}]$ is a field and has dimension 3 as a vector space over $\mathbb{Q}$, and every element of $\mathbb{Q}[\sqrt[3]{2}]$ may be written uniquely as $a + b\sqrt[3]{2} + c\sqrt[3]{4}$ for $a, b, c \in \mathbb{Q}$. □

A simple but important fact is the following:

Theorem 16.20. *Suppose E is an F-algebra, and that as an F-vector space, E has finite dimension n. Then every $\alpha \in E$ has a non-zero minimal polynomial of degree at most n.*

Proof. Indeed, the family of elements

$$1_E, \alpha, \ldots, \alpha^n$$

must be linearly dependent (as must any family of $n+1$ elements of a vector space

of dimension n), and hence there exist $c_0, \ldots, c_n \in F$, not all zero, such that

$$c_0 1_E + c_1 \alpha + \cdots + c_n \alpha^n = 0_E,$$

and therefore, the non-zero polynomial $f := \sum_i c_i X^i$ vanishes at α. □

Example 16.18. Let $f \in F[X]$ be a monic polynomial of degree ℓ, and consider the F-algebra $E := F[X]/(f) = F[\xi]$, where $\xi := [X]_f \in E$. Clearly, the minimal polynomial of ξ over F is f. Moreover, as a vector space over F, E has dimension ℓ, with $\{\xi^{i-1}\}_{i=1}^{\ell}$ being a basis. Therefore, every $\alpha \in E$ has a non-zero minimal polynomial of degree at most ℓ. □

EXERCISE 16.7. In the field E in Example 16.16, what is the minimal polynomial of $\mathtt{1011}$ over $\mathbb{Z}_2$?

EXERCISE 16.8. Let $\rho : E \to E'$ be an F-algebra homomorphism, let $\alpha \in E$, let ϕ be the minimal polynomial of α over F, and let ϕ' be the minimal polynomial of $\rho(\alpha)$ over F. Show that $\phi' \mid \phi$, and that $\phi' = \phi$ if ρ is injective.

EXERCISE 16.9. Show that if the factorization of f over $F[X]$ into monic irreducibles is $f = f_1^{e_1} \cdots f_r^{e_r}$, and if $\alpha = [h]_f \in F[X]/(f)$, then the minimal polynomial ϕ of α over F is $\mathrm{lcm}(\phi_1, \ldots, \phi_r)$, where each ϕ_i is the minimal polynomial of $[h]_{f_i^{e_i}} \in F[X]/(f_i^{e_i})$ over F.

16.6 General properties of extension fields

We now discuss a few general notions related to extension fields. These are all quite simple applications of the theory developed so far. Recall that if F and E are fields, with F being a subring of E, then F is called a subfield of E, and E is called an extension field of F. As usual, we shall blur the distinction between a subring and a natural embedding; that is, if $\tau : F \to E$ is a natural embedding, we shall simply identify elements of F with their images in E under τ, and in so doing, we may view E as an extension field of F. Usually, the map τ will be clear from context; for example, if $E = F[X]/(f)$ for some irreducible polynomial $f \in F[X]$, then we shall simply say that E is an extension field of F, although strictly speaking, F is embedded in E via the map that sends $c \in F$ to $[c]_f \in E$.

We start with some definitions. Let E be an extension field of a field F. Then E is an F-algebra via inclusion, and in particular, an F-vector space. If E is a finite dimensional F-vector space, then we say that E is a **finite extension of** F, and $\dim_F(E)$ is called the **degree (over F)** of the extension, and is denoted $(E : F)$; otherwise, we say that E is an **infinite extension of** F.

An element $\alpha \in E$ is called **algebraic over** F if there exists a non-zero polynomial $g \in F[X]$ such that $g(\alpha) = 0$, and in this case, we define the **degree of** α **(over** F**)** to be the degree of its minimal polynomial over F (see §16.5); otherwise, α is called **transcendental over** F. If all elements of E are algebraic over F, then we call E an **algebraic extension of** F.

Suppose E is an extension field of a field F. For $\alpha \in E$, we define

$$F(\alpha) := \{g(\alpha)/h(\alpha) : g, h \in F[X],\ h(\alpha) \neq 0\}.$$

It is easy to see that $F(\alpha)$ is a subfield of E, and indeed, it is the smallest subfield of E containing F and α. Clearly, the ring $F[\alpha] = \{g(\alpha) : g \in F[X]\}$, which is the smallest subring of E containing F and α, is a subring of $F(\alpha)$. We derive some basic properties of $F(\alpha)$ and $F[\alpha]$. The analysis naturally breaks down into two cases, depending on whether α is algebraic or transcendental over F.

On the one hand, suppose α is algebraic over F. Let ϕ be the minimal polynomial of α over F, so that $\deg(\phi) > 0$, and the quotient ring $F[X]/(\phi)$ is isomorphic (as an F-algebra) to the ring $F[\alpha]$ (see §16.5). Since $F[\alpha]$ is a subring of a field, it must be an integral domain, which implies that $F[X]/(\phi)$ is an integral domain, and so ϕ is irreducible. This in turn implies that $F[X]/(\phi)$ is a field, and so $F[\alpha]$ is not just a subring of E, it is a *subfield* of E. Since $F[\alpha]$ is itself already a subfield of E containing F and α, it follows that $F(\alpha) = F[\alpha]$. Moreover, $F[\alpha]$ is a finite extension of F; indeed $(F[\alpha] : F) = \deg(\phi)$ = the degree of α over F, and the elements $1, \alpha, \ldots, \alpha^{m-1}$, where $m := \deg(\phi)$, form a basis for $F[\alpha]$ over F.

On the other hand, suppose that α is transcendental over F. In this case, the minimal polynomial of α over F is the zero polynomial, and the ring $F[\alpha]$ is isomorphic (as an F-algebra) to the ring $F[X]$ (see §16.5), which is definitely not a field. But consider the "rational function evaluation map" that sends $g/h \in F(X)$ to $g(\alpha)/h(\alpha) \in F(\alpha)$. Since no non-zero polynomial over F vanishes at α, it is easy to see that this map is well defined, and is in fact an F-algebra isomorphism. Thus, we see that $F(\alpha)$ is isomorphic (as an F-algebra) to $F(X)$. It is also clear that $F(\alpha)$ is an infinite extension of F.

Let us summarize the above discussion in the following theorem:

Theorem 16.21. *Let E be an extension field of a field F.*

(i) If $\alpha \in E$ is algebraic over F, then $F(\alpha) = F[\alpha]$, and $F[\alpha]$ is isomorphic (as an F-algebra) to $F[X]/(\phi)$, where ϕ is the minimal polynomial of α over F, which is irreducible; moreover, $F[\alpha]$ is a finite extension of F, and $(F[\alpha] : F) = \deg(\phi)$ = the degree of α over F, and the elements $1, \alpha, \ldots, \alpha^{m-1}$, where $m := \deg(\phi)$, form a basis for $F[\alpha]$ over F.

(ii) If $\alpha \in E$ is transcendental over F, then $F(\alpha)$ is isomorphic (as an F-algebra) to the rational function field $F(X)$, while the subring $F[\alpha]$ is

isomorphic (as an F-algebra) to the ring of polynomials $F[X]$; moreover, $F(\alpha)$ is an infinite extension of F.

Suppose E is an extension field of a field K, which itself is an extension of a field F. Then E is also an extension field of F. The following theorem examines the relation between the degrees of these extensions, in the case where E is a finite extension of K, and K is a finite extension of F. The proof is a simple calculation, which we leave to the reader to verify.

Theorem 16.22. *Suppose E is a finite extension of a field K, with a basis $\{\beta_j\}_{j=1}^m$ over K, and K is a finite extension of F, with a basis $\{\alpha_i\}_{i=1}^n$ over F. Then the elements*

$$\alpha_i\beta_j \quad (i = 1, \ldots, n;\ j = 1, \ldots, m)$$

form a basis for E over F. In particular, E is a finite extension of F and

$$(E : F) = (E : K)(K : F).$$

Now suppose that E is a finite extension of a field F. Let K be an intermediate field, that is, a subfield of E containing F. Then evidently, E is a finite extension of K (since any basis for E over F also spans E over K), and K is a finite extension of F (since as F-vector spaces, K is a subspace of E). The previous theorem then implies that $(E : F) = (E : K)(K : F)$. We have proved:

Theorem 16.23. *If E is a finite extension of a field F, and K is a subfield of E containing F, then E is a finite extension of K, K is a finite extension of F, and $(E : F) = (E : K)(K : F)$.*

Again, suppose that E is a finite extension of a field F. Theorem 16.20 implies that E is algebraic over F, and indeed, that each element of E has degree over F bounded by $(E : F)$. However, we can say a bit more about these degrees. Suppose $\alpha \in E$. Then the degree of α over F is equal to $(F[\alpha] : F)$, and by the previous theorem, applied to $K := F[\alpha]$, we have $(E : F) = (E : F[\alpha])(F[\alpha] : F)$. In particular, the degree of α over F divides $(E : F)$. We have proved:

Theorem 16.24. *If E is a finite extension of a field F, then it is an algebraic extension, and for each $\alpha \in E$, the degree of α over F divides $(E : F)$.*

Example 16.19. Continuing with Example 16.17, we see that the real numbers $\sqrt{2}$ and $\sqrt[3]{2}$ are algebraic over $\mathbb{Q}$. The fields $\mathbb{Q}[\sqrt{2}]$ and $\mathbb{Q}[\sqrt[3]{2}]$ are extension fields of $\mathbb{Q}$, where $(\mathbb{Q}[\sqrt{2}] : \mathbb{Q}) = 2 =$ the degree of $\sqrt{2}$ over $\mathbb{Q}$, and $(\mathbb{Q}[\sqrt[3]{2}] : \mathbb{Q}) = 3 =$ the degree of $\sqrt[3]{2}$ over $\mathbb{Q}$. As both of these fields are finite extensions of $\mathbb{Q}$, they are algebraic extensions as well. Since their degrees over $\mathbb{Q}$ are prime numbers, it follows that they have no subfields other than themselves and $\mathbb{Q}$. In particular,

if $\alpha \in \mathbb{Q}[\sqrt{2}] \setminus \mathbb{Q}$, then $\mathbb{Q}[\alpha] = \mathbb{Q}[\sqrt{2}]$. Similarly, if $\alpha \in \mathbb{Q}[\sqrt[3]{2}] \setminus \mathbb{Q}$, then $\mathbb{Q}[\alpha] = \mathbb{Q}[\sqrt[3]{2}]$. □

Example 16.20. Continuing with Example 16.18, suppose $f \in F[X]$ is a monic *irreducible* polynomial of degree ℓ, so that $E := F[X]/(f) = F[\xi]$, where $\xi := [X]_f \in E$, is an extension field of F. The element ξ is algebraic of degree ℓ over F. Moreover, E is a finite extension of F, with $(E : F) = \ell$; in particular, E is an algebraic extension of F, and for each $\alpha \in E$, the degree of α over F divides ℓ. □

As we have seen in Example 16.14, an irreducible polynomial over a field may be reducible when viewed as a polynomial over an extension field. A **splitting field** is a finite extension of the coefficient field in which a given polynomial splits completely into linear factors. As the next theorem shows, splitting fields always exist.

Theorem 16.25. *Let F be a field, and $f \in F[X]$ a non-zero polynomial of degree n. Then there exists a finite extension E of F over which f factors as*

$$f = c(X - \alpha_1)(X - \alpha_2) \cdots (X - \alpha_n),$$

where $c \in F$ and $\alpha_1, \ldots, \alpha_n \in E$.

Proof. We may assume that f is monic. We prove the existence of E by induction on the degree n of f. If $n = 0$, then the theorem is trivially true. Otherwise, let h be an irreducible factor of f, and set $K := F[X]/(h)$, so that $\xi := [X]_h \in K$ is a root of h, and hence of f. So over K, which is a finite extension of F, the polynomial f factors as

$$f = (X - \xi)g,$$

where $g \in K[X]$ is a monic polynomial of degree $n - 1$. Applying the induction hypothesis, there exists a finite extension E of K over which g splits into linear factors. Thus, over E, f splits into linear factors, and by Theorem 16.22, E is a finite extension of F. □

EXERCISE 16.10. In the field E in Example 16.16, find all the elements of degree 2 over $\mathbb{Z}_2$.

EXERCISE 16.11. Let E be an extension field of a field F, and let $\alpha_1, \ldots, \alpha_n \in E$ be algebraic over F. Show that the ring $F[\alpha_1, \ldots, \alpha_n]$ (see Example 7.45) is in fact a field, and that $F[\alpha_1, \ldots, \alpha_n]$ is a finite (and hence algebraic) extension of F.

EXERCISE 16.12. Consider the real numbers $\sqrt{2}$ and $\sqrt[3]{2}$. Show that

$$(\mathbb{Q}[\sqrt{2}, \sqrt[3]{2}] : \mathbb{Q}) = (\mathbb{Q}[\sqrt{2} + \sqrt[3]{2}] : \mathbb{Q}) = 6.$$

EXERCISE 16.13. Consider the real numbers $\sqrt{2}$ and $\sqrt{3}$. Show that

$$(\mathbb{Q}[\sqrt{2}, \sqrt{3}] : \mathbb{Q}) = (\mathbb{Q}[\sqrt{2} + \sqrt{3}] : \mathbb{Q}) = 4.$$

EXERCISE 16.14. Show that if E is an algebraic extension of K, and K is an algebraic extension of F, then E is an algebraic extension of F.

EXERCISE 16.15. Let E be an extension of F. Show that the set of all elements of E that are algebraic over F is a subfield of E containing F.

EXERCISE 16.16. Consider a field F and its field of rational functions $F(X)$. Let $\alpha \in F(X) \setminus F$. Show that X is algebraic over $F(\alpha)$, and that α is transcendental over F.

EXERCISE 16.17. Let E be an extension field of a field F. Suppose $\alpha \in E$ is transcendental over F, and that E is algebraic over $F(\alpha)$. Show that for every $\beta \in E$, β is transcendental over F if and only if E is algebraic over $F(\beta)$.

16.7 Formal derivatives

Throughout this section, R denotes a ring.

Consider a polynomial $g \in R[X]$. If Y is another indeterminate, we may evaluate g at $X + Y$, and collecting monomials of like degree in Y, we may write

$$g\big(X + Y\big) = g_0 + g_1 Y + g_2 Y^2 + \cdots \tag{16.5}$$

where $g_i \in R[X]$ for $i = 0, 1, 2, \ldots$. Evidently, $g_0 = g$ (just substitute 0 for Y in (16.5)), and we may write

$$g\big(X + Y\big) \equiv g + g_1 Y \pmod{Y^2}. \tag{16.6}$$

We define the **formal derivative** of g, denoted $\mathbf{D}(g)$, to be the unique polynomial $g_1 \in R[X]$ satisfying (16.6). We stress that unlike the "analytical" notion of derivative from calculus, which is defined in terms of limits, this definition is purely "symbolic." Nevertheless, some of the usual rules for derivatives still hold:

Theorem 16.26. *We have:*

(i) $\mathbf{D}(c) = 0$ *for all* $c \in R$;

(ii) $\mathbf{D}(X) = 1$;

(iii) $\mathbf{D}(g + h) = \mathbf{D}(g) + \mathbf{D}(h)$ *for all* $g, h \in R[X]$;

(iv) $\mathbf{D}(gh) = \mathbf{D}(g)h + g\mathbf{D}(h)$ *for all* $g, h \in R[X]$.

Proof. Parts (i) and (ii) are immediate from the definition. Parts (iii) and (iv) follow from the definition by a simple calculation. Suppose

$$g\big(X + Y\big) \equiv g + g_1 Y \pmod{Y^2} \quad \text{and} \quad h\big(X + Y\big) \equiv h + h_1 Y \pmod{Y^2}$$

where $g_1 = \mathbf{D}(g)$ and $h_1 = \mathbf{D}(h)$. Then

$$(g + h)\big(X + Y\big) \equiv g\big(X + Y\big) + h\big(X + Y\big) \equiv (g + h) + (g_1 + h_1)Y \pmod{Y^2},$$

and

$$(gh)\big(X + Y\big) \equiv g\big(X + Y\big)h\big(X + Y\big) \equiv gh + (g_1 h + g h_1)Y \pmod{Y^2}. \quad \square$$

Combining parts (i) and (iv) of this theorem, we see that $\mathbf{D}(cg) = c\mathbf{D}(g)$ for all $c \in R$ and $g \in R[X]$. This fact can also be easily derived directly from the definition of the derivative.

Combining parts (ii) and (iv) of this theorem, together with a simple induction argument, we see that $\mathbf{D}(X^n) = nX^{n-1}$ for all positive integers n. This fact can also be easily derived directly from the definition of the derivative by considering the binomial expansion of $(X + Y)^n$.

Combining part (iii) of this theorem and the observations in the previous two paragraphs, we see that for any polynomial $g = \sum_{i=0}^{k} a_i X^i \in R[X]$, we have

$$\mathbf{D}(g) = \sum_{i=1}^{k} i a_i X^{i-1}, \tag{16.7}$$

which agrees with the usual formula for the derivative of a polynomial.

The notion of a formal derivative can be generalized to multi-variate polynomials. Let $g \in R[X_1, \ldots, X_n]$. For any $i = 1, \ldots, n$, we can view g as a polynomial in the variable X_i, whose coefficients are elements of $R[X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n]$. Then if we formally differentiate with respect to the variable X_i, we obtain the formal "partial" derivative $\mathbf{D}_{X_i}(g)$.

EXERCISE 16.18. Show that for $g_1, \ldots, g_n \in R[X]$, we have

$$\mathbf{D}\Big(\prod_i g_i\Big) = \sum_i \mathbf{D}(g_i) \prod_{j \neq i} g_j$$

and that for $g \in R[X]$, and $n \geq 1$, we have

$$\mathbf{D}(g^n) = n g^{n-1} \mathbf{D}(g).$$

EXERCISE 16.19. Prove the "chain rule" for formal derivatives: if $g, h \in R[X]$

and $f = g(h) \in R[X]$, then $\mathbf{D}(f) = \mathbf{D}(g)(h) \cdot \mathbf{D}(h)$; more generally, if $g \in R[X_1, ..., X_n]$, and $h_1, ..., h_n \in R[X]$, and $f = g(h_1, ..., h_n) \in R[X]$, then

$$\mathbf{D}_X(f) = \sum_{i=1}^{n} \mathbf{D}_{X_i}(g)(h_1, ..., h_n)\mathbf{D}_X(h_i).$$

EXERCISE 16.20. Let $g \in R[X]$, and let $x \in R$ be a root of g. Show that x is a multiple root of g if and only if x is also a root of $\mathbf{D}(g)$ (see Exercise 7.18).

EXERCISE 16.21. Let $g \in R[X]$ with $\deg(g) = k \geq 0$, and let $x \in R$. Show that if we evaluate g at $X + x$, writing

$$g\big(X + x\big) = \sum_{i=0}^{k} b_i X^i,$$

with $b_0, \ldots, b_k \in R$, then we have

$$i! \cdot b_i = (\mathbf{D}^i(g))(x) \text{ for } i = 0, \ldots, k.$$

EXERCISE 16.22. Suppose p is a prime, $g \in \mathbb{Z}[X]$, and $x \in \mathbb{Z}$, such that $g(x) \equiv 0 \pmod{p}$ and $\mathbf{D}(g)(x) \not\equiv 0 \pmod{p}$. Show that for every positive integer e, there exists an integer $\hat{x}$ such that $g(\hat{x}) \equiv 0 \pmod{p^e}$, and give an efficient procedure to compute such an $\hat{x}$, given p, g, x, and e. Hint: mimic the "lifting" procedure discussed in §12.5.2.

16.8 Formal power series and Laurent series

We discuss generalizations of polynomials that allow an infinite number of non-zero coefficients. Although we are mainly interested in the case where the coefficients come from a field F, we develop the basic theory for general rings R.

16.8.1 Formal power series

The ring $R[\![X]\!]$ of **formal power series over** R consists of all formal expressions of the form

$$g = a_0 + a_1 X + a_2 X^2 + \cdots,$$

where $a_0, a_1, a_2, \ldots \in R$. Unlike ordinary polynomials, we allow an infinite number of non-zero coefficients. We may write such a formal power series as

$$g = \sum_{i=0}^{\infty} a_i X^i.$$

Formally, such a formal power series is an infinite sequence $\{a_i\}_{i=0}^{\infty}$, and the rules for addition and multiplication are *exactly* the same as for polynomials. Indeed, the formulas (7.2) and (7.3) in §7.2 for addition and multiplication may be applied directly — all of the relevant sums are finite, and so everything is well defined. We leave it to the reader to verify that with addition and multiplication so defined, $R[\![X]\!]$ indeed forms a ring. We shall not attempt to interpret a formal power series as a function, and therefore, "convergence" issues shall simply not arise.

Clearly, $R[\![X]\!]$ contains $R[X]$ as a subring. Let us consider the group of units of $R[\![X]\!]$.

Theorem 16.27. *Let $g = \sum_{i=0}^{\infty} a_i X^i \in R[\![X]\!]$. Then $g \in (R[\![X]\!])^*$ if and only if $a_0 \in R^*$.*

Proof. If a_0 is not a unit, then it is clear that g is not a unit, since the constant term of a product of formal power series is equal to the product of the constant terms.

Conversely, if a_0 is a unit, we show how to define the coefficients of the inverse $h = \sum_{i=0}^{\infty} b_i X^i$ of g. Let $f = gh = \sum_{i=0}^{\infty} c_i X^i$. We want $f = 1$, which means that $c_0 = 1$ and $c_i = 0$ for all $i > 0$. Now, $c_0 = a_0 b_0$, so we set $b_0 := a_0^{-1}$. Next, we have $c_1 = a_0 b_1 + a_1 b_0$, so we set $b_1 := -a_1 b_0 \cdot a_0^{-1}$. Next, we have $c_2 = a_0 b_2 + a_1 b_1 + a_2 b_0$, so we set $b_2 := -(a_1 b_1 + a_2 b_0) \cdot a_0^{-1}$. Continuing in this way, we see that if we define $b_i := -(a_1 b_{i-1} + \cdots + a_i b_0) \cdot a_0^{-1}$ for $i \geq 1$, then $gh = 1$. □

Example 16.21. In the ring $R[\![X]\!]$, the multiplicative inverse of $1 - X$ is $\sum_{i=0}^{\infty} X^i$. □

EXERCISE 16.23. Let F be a field. Show that every non-zero ideal of $F[\![X]\!]$ is of the form (X^m) for some uniquely determined integer $m \geq 0$.

16.8.2 Formal Laurent series

One may generalize formal power series to allow a finite number of negative powers of X. The ring $R((X))$ of **formal Laurent series over** R consists of all formal expressions of the form

$$g = a_m X^m + a_{m+1} X^{m+1} + \cdots,$$

where m is allowed to be any integer (possibly negative), and $a_m, a_{m+1}, \ldots \in R$. Thus, elements of $R((X))$ may have an infinite number of terms involving positive powers of X, but only a finite number of terms involving negative powers of X. We may write such a formal Laurent series as

$$g = \sum_{i=m}^{\infty} a_i X^i.$$

Formally, such a formal Laurent series is a doubly infinite sequence $\{a_i\}_{i=-\infty}^{\infty}$, with the restriction that for some integer m, we have $a_i = 0$ for all $i < m$. We may again use the usual formulas (7.2) and (7.3) to define addition and multiplication (where the indices i, j, and k now range over all integers, not just the non-negative integers). Note that while the sum in (7.3) has an infinite number of terms, only finitely many of them are non-zero.

One may naturally view $R[\![X]\!]$ as a subring of $R((X))$, and of course, $R[X]$ is a subring of $R[\![X]\!]$ and so also a subring of $R((X))$.

Theorem 16.28. *If D is an integral domain, then $D((X))$ is an integral domain.*

Proof. Let $g = \sum_{i=m}^{\infty} a_i X^i$ and $h = \sum_{i=n}^{\infty} b_i X^i$, where $a_m \neq 0$ and $b_n \neq 0$. Then $gh = \sum_{i=m+n}^{\infty} c_i X^i$, where $c_{m+n} = a_m b_n \neq 0$. □

Theorem 16.29. *Let $g \in R((X))$, and suppose that $g \neq 0$ and $g = \sum_{i=m}^{\infty} a_i X^i$ with $a_m \in R^*$. Then g has a multiplicative inverse in $R((X))$.*

Proof. We can write $g = X^m g'$, where g' is a formal power series whose constant term is a unit, and hence there is a formal power series h such that $g'h = 1$. Thus, $X^{-m}h$ is the multiplicative inverse of g in $R((X))$. □

As an immediate corollary, we have:

Theorem 16.30. *If F is a field, then $F((X))$ is a field.*

EXERCISE 16.24. Let F be a field. Show that $F((X))$ is the field of fractions of $F[\![X]\!]$; that is, there is no subfield $E \subsetneq F((X))$ that contains $F[\![X]\!]$.

16.8.3 Reversed Laurent series

While formal Laurent series are useful in some situations, in many others, it is more useful and natural to consider **reversed Laurent series over** R. These are formal expressions of the form

$$g = \sum_{i=-\infty}^{m} a_i X^i,$$

where $a_m, a_{m-1}, \ldots \in R$. Thus, in a reversed Laurent series, we allow an infinite number of terms involving negative powers of X, but only a finite number of terms involving positive powers of X. Formally, such a reversed Laurent series is a doubly infinite sequence $\{a_i\}_{i=-\infty}^{\infty}$, with the restriction that for some integer m, we have $a_i = 0$ for all $i > m$. We may again use the usual formulas (7.2) and (7.3) to define

addition and multiplication — and again, the sum in (7.3) has only finitely many non-zero terms.

The ring of all reversed Laurent series is denoted $R((X^{-1}))$, and as the notation suggests, the map that sends X to X^{-1} (and acts as the identity on R) is an R-algebra isomorphism of $R((X))$ with $R((X^{-1}))$. Also, one may naturally view $R[X]$ as a subring of $R((X^{-1}))$.

For $g = \sum_{i=-\infty}^{m} a_i X^i \in R((X^{-1}))$ with $a_m \neq 0$, let us define the **degree of** g, denoted $\deg(g)$, to be the value m, and the **leading coefficient of** g, denoted $\mathrm{lc}(g)$, to be the value a_m. As for ordinary polynomials, we define the degree of 0 to be $-\infty$, and the leading coefficient of 0 to be 0. Note that if g happens to be a polynomial, then these definitions of degree and leading coefficient agree with that for ordinary polynomials.

Theorem 16.31. *For $g, h \in R((X^{-1}))$, we have $\deg(gh) \leq \deg(g) + \deg(h)$, where equality holds unless both $\mathrm{lc}(g)$ and $\mathrm{lc}(h)$ are zero divisors. Furthermore, if $h \neq 0$ and $\mathrm{lc}(h)$ is a unit, then h is a unit, and we have $\deg(gh^{-1}) = \deg(g) - \deg(h)$.*

Proof. Exercise. □

It is also natural to define a **floor function** for reversed Laurent series: for $g \in R((X^{-1}))$ with $g = \sum_{i=-\infty}^{m} a_i X^i$, we define

$$\lfloor g \rfloor := \sum_{i=0}^{m} a_i X^i \in R[X];$$

that is, we compute the floor function by simply throwing away all terms involving negative powers of X.

Theorem 16.32. *Let $g, h \in R[X]$ with $h \neq 0$ and $\mathrm{lc}(h) \in R^*$, and using the usual division with remainder property for polynomials, write $g = hq + r$, where $q, r \in R[X]$ with $\deg(r) < \deg(h)$. Let h^{-1} denote the multiplicative inverse of h in $R((X^{-1}))$. Then $q = \lfloor gh^{-1} \rfloor$.*

Proof. Multiplying the equation $g = hq + r$ by h^{-1}, we obtain $gh^{-1} = q + rh^{-1}$, and $\deg(rh^{-1}) < 0$, from which it follows that $\lfloor gh^{-1} \rfloor = q$. □

Let F be a field, so that $F((X^{-1}))$ is also field (this is immediate from Theorem 16.31). Now, $F((X^{-1}))$ contains $F[X]$ as a subring, and hence contains (an isomorphic copy of) the rational function field $F(X)$. Just as $F(X)$ corresponds to the field of rational numbers, $F((X^{-1}))$ corresponds to the field real numbers. Indeed, we can think of real numbers as decimal numbers with a finite number of digits to the left of the decimal point and an infinite number to the right, and reversed Laurent series have a similar "syntactic" structure. In many ways, this

syntactic similarity between the real numbers and reversed Laurent series is more than just superficial.

EXERCISE 16.25. Write down the rule for determining the multiplicative inverse of an element of $R((X^{-1}))$ whose leading coefficient is a unit in R.

EXERCISE 16.26. Let F be a field of characteristic other than 2. Show that a non-zero $g \in F((X^{-1}))$ has a square-root in $F((X^{-1}))$ if and only if $\deg(g)$ is even and $\mathrm{lc}(g)$ has a square-root in F.

EXERCISE 16.27. Let R be a ring, and let $a \in R$. Show that the multiplicative inverse of $X - a$ in $R((X^{-1}))$ is $\sum_{j=1}^{\infty} a^{j-1} X^{-j}$.

EXERCISE 16.28. Let R be an arbitrary ring, let $a_1, \ldots, a_\ell \in R$, and let

$$f := (X - a_1)(X - a_2) \cdots (X - a_\ell) \in R[X].$$

For $j \geq 0$, define the "power sum"

$$s_j := \sum_{i=1}^{\ell} a_i^j.$$

Show that in the ring $R((X^{-1}))$, we have

$$\frac{\mathbf{D}(f)}{f} = \sum_{i=1}^{\ell} \frac{1}{(X - a_i)} = \sum_{j=1}^{\infty} s_{j-1} X^{-j},$$

where $\mathbf{D}(f)$ is the formal derivative of f.

EXERCISE 16.29. Continuing with the previous exercise, derive **Newton's identities**, which state that if $f = X^\ell + c_1 X^{\ell-1} + \cdots + c_\ell$, with $c_1, \ldots, c_\ell \in R$, then

$$\begin{aligned}
s_1 + c_1 &= 0 \\
s_2 + c_1 s_1 + 2c_2 &= 0 \\
s_3 + c_1 s_2 + c_2 s_1 + 3c_3 &= 0 \\
&\vdots \\
s_\ell + c_1 s_{\ell-1} + \cdots + c_{\ell-1} s_1 + \ell c_\ell &= 0 \\
s_{j+\ell} + c_1 s_{j+\ell-1} + \cdots + c_{\ell-1} s_{j+1} + c_\ell s_j &= 0 \quad (j \geq 1).
\end{aligned}$$

16.9 Unique factorization domains (∗)

As we have seen, both the ring of integers and the ring of polynomials over a field enjoy a unique factorization property. These are special cases of a more general phenomenon, which we explore here.

Throughout this section, D denotes an integral domain.

We call $a, b \in D$ **associate** if $a = ub$ for some $u \in D^*$. Equivalently, a and b are associate if and only if $a \mid b$ and $b \mid a$ (see part (i) of Theorem 7.4). A non-zero element $p \in D$ is called **irreducible** if it is not a unit, and all divisors of p are associate to 1 or p. Equivalently, a non-zero, non-unit $p \in D$ is irreducible if and only if it cannot be expressed as $p = ab$ where neither a nor b are units.

Definition 16.33. *We call D a **unique factorization domain (UFD)** if*

(i) every non-zero element of D that is not a unit can be written as a product of irreducibles in D, and

(ii) such a factorization into irreducibles is unique up to associates and the order in which the factors appear.

Another way to state part (ii) of the above definition is that if $p_1 \cdots p_r$ and $p'_1 \cdots p'_s$ are two factorizations of some element as a product of irreducibles, then $r = s$, and there exists a permutation π on the indices $\{1, \ldots, r\}$ such that p_i and $p'_{\pi(i)}$ are associate.

As we have seen, both $\mathbb{Z}$ and $F[X]$ are UFDs. In both of those cases, we chose to single out a distinguished irreducible element among all those associate to any given irreducible: for $\mathbb{Z}$, we always chose positive primes, and for $F[X]$, we chose monic irreducible polynomials. For any specific unique factorization domain D, there may be such a natural choice, but in the general case, there will not be (but see Exercise 16.30 below).

Example 16.22. Having already seen two examples of UFDs, it is perhaps a good idea to look at an example of an integral domain that is not a UFD. Consider the subring $\mathbb{Z}[\sqrt{-3}]$ of the complex numbers, which consists of all complex numbers of the form $a + b\sqrt{-3}$, where $a, b \in \mathbb{Z}$. As this is a subring of the field $\mathbb{C}$, it is an integral domain (one may also view $\mathbb{Z}[\sqrt{-3}]$ as the quotient ring $\mathbb{Z}[X]/(X^2 + 3)$).

Let us first determine the units in $\mathbb{Z}[\sqrt{-3}]$. For $a, b \in \mathbb{Z}$, we have $N(a+b\sqrt{-3}) = a^2 + 3b^2$, where N is the usual norm map on $\mathbb{C}$ (see Example 7.5). If $\alpha \in \mathbb{Z}[\sqrt{-3}]$ is a unit, then there exists $\alpha' \in \mathbb{Z}[\sqrt{-3}]$ such that $\alpha\alpha' = 1$. Taking norms, we obtain

$$1 = N(1) = N(\alpha\alpha') = N(\alpha)N(\alpha').$$

Since the norm of an element of $\mathbb{Z}[\sqrt{-3}]$ is a non-negative integer, this implies that $N(\alpha) = 1$. If $\alpha = a + b\sqrt{-3}$, with $a, b \in \mathbb{Z}$, then $N(\alpha) = a^2 + 3b^2$, and it is clear

that $N(\alpha) = 1$ if and only if $\alpha = \pm 1$. We conclude that the only units in $\mathbb{Z}[\sqrt{-3}]$ are ± 1.

Now consider the following two factorizations of 4 in $\mathbb{Z}[\sqrt{-3}]$:

$$4 = 2 \cdot 2 = (1 + \sqrt{-3})(1 - \sqrt{-3}). \tag{16.8}$$

We claim that 2 is irreducible. For suppose, say, that $2 = \alpha\alpha'$, for $\alpha, \alpha' \in \mathbb{Z}[\sqrt{-3}]$, with neither a unit. Taking norms, we have $4 = N(2) = N(\alpha)N(\alpha')$, and therefore, $N(\alpha) = N(\alpha') = 2$—but this is impossible, since there are no integers a and b such that $a^2 + 3b^2 = 2$. By the same reasoning, since $N(1 + \sqrt{-3}) = N(1 - \sqrt{-3}) = 4$, we see that $1 + \sqrt{-3}$ and $1 - \sqrt{-3}$ are both irreducible. Further, it is clear that 2 is not associate to either $1 + \sqrt{-3}$ or $1 - \sqrt{-3}$, and so the two factorizations of 4 in (16.8) are fundamentally different. □

For $a, b \in D$, we call $d \in D$ a **common divisor** of a and b if $d \mid a$ and $d \mid b$; moreover, we call such a d a **greatest common divisor** of a and b if all other common divisors of a and b divide d. We say that a and b are **relatively prime** if the only common divisors of a and b are units. It is immediate from the definition of a greatest common divisor that it is unique, up to multiplication by units, if it exists at all. Unlike in the case of $\mathbb{Z}$ and $F[X]$, in the general setting, greatest common divisors need not exist; moreover, even when they do, we shall not attempt to "normalize" greatest common divisors, and we shall speak only of "a" greatest common divisor, rather than "the" greatest common divisor.

Just as for integers and polynomials, we can generalize the notion of a greatest common divisor in an arbitrary integral domain D from two to any number of elements of D, and we can also define a **least common multiple** of any number of elements as well.

Although these greatest common divisors and least common multiples need not exist in an arbitrary integral domain D, if D is a UFD, they will always exist. The existence question easily reduces to the question of the existence of a greatest common divisor and least common multiple of a and b, where a and b are non-zero elements of D. So assuming that D is a UFD, we may write

$$a = u \prod_{i=1}^{r} p_i^{e_i} \quad \text{and} \quad b = v \prod_{i=1}^{r} p_i^{f_i},$$

where u and v are units, $p_1, \ldots, p_r$ are non-associate irreducibles, and $e_1, \ldots, e_r$ and $f_1, \ldots, f_r$ are non-negative integers, and it is easily seen that

$$\prod_{i=1}^{r} p_i^{\min(e_i, f_i)}$$

is a greatest common divisor of a and b, while

$$\prod_{i=1}^{r} p_i^{\max(e_i, f_i)}$$

is a least common multiple of a and b.

It is also evident that in a UFD D, if $c \mid ab$ and c and a are relatively prime, then $c \mid b$. In particular, if p is irreducible and $p \mid ab$, then $p \mid a$ or $p \mid b$. This is equivalent to saying that if p is irreducible, then the quotient ring D/pD is an integral domain (and the ideal pD is a prime ideal — see Exercise 7.38). The converse also holds:

Theorem 16.34. *Suppose D satisfies part (i) of Definition 16.33, and that D/pD is an integral domain for every irreducible $p \in D$. Then D is a UFD.*

Proof. Exercise. □

EXERCISE 16.30. (a) Show that the "is associate to" relation is an equivalence relation.

(b) Consider an equivalence class C induced by the "is associate to" relation. Show that if C contains an irreducible element, then all elements of C are irreducible.

(c) Suppose that for every equivalence class C that contains irreducibles, we choose one element of C, and call it a **distinguished irreducible**. Show that D is a UFD if and only if every non-zero element of D can be expressed as $up_1^{e_1} \cdots p_r^{e_r}$, where u is a unit, $p_1, \ldots, p_r$ are distinguished irreducibles, and this expression is unique up to a reordering of the p_i's.

EXERCISE 16.31. Show that the ring $\mathbb{Z}[\sqrt{-5}]$ is not a UFD.

EXERCISE 16.32. Let D be a UFD and F its field of fractions. Show that

(a) every element $x \in F$ can be expressed as $x = a/b$, where $a, b \in D$ are relatively prime, and

(b) that if $x = a/b$ for $a, b \in D$ relatively prime, then for any other $a', b' \in D$ with $x = a'/b'$, we have $a' = ca$ and $b' = cb$ for some $c \in D$.

EXERCISE 16.33. Let D be a UFD and let $p \in D$ be irreducible. Show that there is no prime ideal Q of D with $\{0_D\} \subsetneq Q \subsetneq pD$ (see Exercise 7.38).

16.9.1 Unique factorization in Euclidean and principal ideal domains

Our proofs of the unique factorization property in both $\mathbb{Z}$ and $F[X]$ hinged on the division with remainder property for these rings. This notion can be generalized, as follows.

Definition 16.35. *We say D is a* ***Euclidean domain*** *if there is a "size function" S mapping the non-zero elements of D to the set of non-negative integers, such that for all $a, b \in D$ with $b \neq 0$, there exist $q, r \in D$, with the property that $a = bq + r$ and either $r = 0$ or $S(r) < S(b)$.*

Example 16.23. Both $\mathbb{Z}$ and $F[X]$ are Euclidean domains. In $\mathbb{Z}$, we can take the ordinary absolute value function $|\cdot|$ as a size function, and for $F[X]$, the function $\deg(\cdot)$ will do. □

Example 16.24. Recall again the ring

$$\mathbb{Z}[i] = \{a + bi : a, b \in \mathbb{Z}\}$$

of Gaussian integers from Example 7.25. Let us show that this is a Euclidean domain, using the usual norm map N on complex numbers (see Example 7.5) for the size function. Let $\alpha, \beta \in \mathbb{Z}[i]$, with $\beta \neq 0$. We want to show the existence of $\kappa, \rho \in \mathbb{Z}[i]$ such that $\alpha = \beta\kappa + \rho$, where $N(\rho) < N(\beta)$. Suppose that in the field $\mathbb{C}$, we compute $\alpha\beta^{-1} = r + si$, where $r, s \in \mathbb{Q}$. Let m, n be integers such that $|m - r| \leq 1/2$ and $|n - s| \leq 1/2$—such integers m and n always exist, but may not be uniquely determined. Set $\kappa := m + ni \in \mathbb{Z}[i]$ and $\rho := \alpha - \beta\kappa$. Then we have

$$\alpha\beta^{-1} = \kappa + \delta,$$

where $\delta \in \mathbb{C}$ with $N(\delta) \leq 1/4 + 1/4 = 1/2$, and

$$\rho = \alpha - \beta\kappa = \alpha - \beta(\alpha\beta^{-1} - \delta) = \delta\beta,$$

and hence

$$N(\rho) = N(\delta\beta) = N(\delta)N(\beta) \leq \frac{1}{2}N(\beta). \quad \square$$

Theorem 16.36. *If D is a Euclidean domain and I is an ideal of D, then there exists $d \in D$ such that $I = dD$.*

Proof. If $I = \{0\}$, then $d = 0$ does the job, so let us assume that $I \neq \{0\}$. Let d be any non-zero element of I such that $S(d)$ is minimal, where S is a size function that makes D into a Euclidean domain. We claim that $I = dD$.

It will suffice to show that for all $c \in I$, we have $d \mid c$. Now, we know that there exists $q, r \in D$ such that $c = dq + r$, where either $r = 0$ or $S(r) < S(d)$. If $r = 0$, we are done; otherwise, r is a non-zero element of I with $S(r) < S(d)$, contradicting the minimality of $S(d)$. □

Recall that an ideal of the form $I = dD$ is called a principal ideal. If all ideals of D are principal, then D is called a **principal ideal domain (PID)**. Theorem 16.36 says that every Euclidean domain is a PID.

PIDs enjoy many nice properties, including:

Theorem 16.37. *If D is a PID, then D is a UFD.*

For the rings $\mathbb{Z}$ and $F[X]$, the proof of part (i) of Definition 16.33 was a quite straightforward induction argument (as it also would be for any Euclidean domain). For a general PID, however, this requires a different sort of argument. We begin with the following fact:

Theorem 16.38. *If D is a PID, and $I_1 \subseteq I_2 \subseteq \cdots$ are ideals of D, then there exists an integer k such that $I_k = I_{k+1} = \cdots$.*

Proof. Let $I := \bigcup_{i=1}^{\infty} I_i$, which is an ideal of D (see Exercise 7.37). Thus, $I = dD$ for some $d \in D$. But $d \in \bigcup_{i=1}^{\infty} I_i$ implies that $d \in I_k$ for some k, which shows that $I = dD \subseteq I_k$. It follows that $I = I_k = I_{k+1} = \cdots$. □

We can now prove the existence part of Theorem 16.37:

Theorem 16.39. *If D is a PID, then every non-zero, non-unit element of D can be expressed as a product of irreducibles in D.*

Proof. Let $c \in D$, $c \neq 0$, and c not a unit. If c is irreducible, we are done. Otherwise, we can write $c = ab$, where neither a nor b are units. As ideals, we have $cD \subsetneq aD$ and $cD \subsetneq bD$. If we continue this process recursively, building up a "factorization tree" where c is at the root, a and b are the children of c, and so on, then the recursion must stop, since any infinite path in the tree would give rise to ideals

$$cD = I_1 \subsetneq I_2 \subsetneq \cdots,$$

contradicting Theorem 16.38. □

The proof of the uniqueness part of Theorem 16.37 is essentially the same as for proofs we gave for $\mathbb{Z}$ and $F[X]$.

Analogous to Theorems 1.7 and 16.13, we have:

Theorem 16.40. *Let D be a PID. For all $a, b \in D$, there exists a greatest common divisor d of a and b, and moreover, $aD + bD = dD$.*

Proof. Exercise. □

As an immediate consequence of the previous theorem, we see that in a PID D, for all $a, b \in D$ with greatest common divisor d, there exist $s, t \in D$ such that

$as + bt = d$; moreover, $a, b \in D$ are relatively prime if and only if there exist $s, t \in D$ such that $as + bt = 1$.

Analogous to Theorems 1.9 and 16.14, we have:

Theorem 16.41. *Let D be a PID. For all $a, b, c \in D$ such that $c \mid ab$ and a and c are relatively prime, we have $c \mid b$.*

Proof. Exercise. □

Analogous to Theorems 1.10 and 16.15, we have:

Theorem 16.42. *Let D be a PID. Let $p \in D$ be irreducible, and let $a, b \in D$. Then $p \mid ab$ implies that $p \mid a$ or $p \mid b$.*

Proof. Exercise. □

Theorem 16.37 now follows immediately from Theorems 16.39, 16.42, and 16.34.

EXERCISE 16.34. Show that $\mathbb{Z}[\sqrt{-2}]$ is a Euclidean domain.

EXERCISE 16.35. Consider the polynomial

$$X^3 - 1 = (X - 1)(X^2 + X + 1).$$

Over $\mathbb{C}$, the roots of $X^3 - 1$ are $1, (-1 \pm \sqrt{-3})/2$. Let $\omega := (-1 + \sqrt{-3})/2$, and note that $\omega^2 = -1 - \omega = (-1 - \sqrt{-3})/2$, and $\omega^3 = 1$.

(a) Show that the ring $\mathbb{Z}[\omega]$ consists of all elements of the form $a + b\omega$, where $a, b \in \mathbb{Z}$, and is an integral domain. This ring is called the ring of **Eisenstein integers**.

(b) Show that the only units in $\mathbb{Z}[\omega]$ are ± 1, $\pm\omega$, and $\pm\omega^2$.

(c) Show that $\mathbb{Z}[\omega]$ is a Euclidean domain.

EXERCISE 16.36. Show that in a PID, all non-zero prime ideals are maximal (see Exercise 7.38).

Recall that for a complex number $\alpha = a + bi$, with $a, b \in \mathbb{R}$, the norm of α was defined as $N(\alpha) = \alpha\overline{\alpha} = a^2 + b^2$ (see Example 7.5). There are other measures of the "size" of a complex number that are useful. The **absolute value** of α is defined as $|\alpha| := \sqrt{N(\alpha)} = \sqrt{a^2 + b^2}$. The **max norm** of α is defined as $M(\alpha) := \max\{|a|, |b|\}$.

EXERCISE 16.37. Let $\alpha, \beta \in \mathbb{C}$. Prove the following statements:

(a) $|\alpha\beta| = |\alpha||\beta|$;

(b) $|\alpha + \beta| \leq |\alpha| + |\beta|$;

(c) $N(\alpha + \beta) \leq 2(N(\alpha) + N(\beta))$;

(d) $M(\alpha) \leq |\alpha| \leq \sqrt{2}M(\alpha)$.

The following exercises develop algorithms for computing with Gaussian integers. For computational purposes, we assume that a Gaussian integer $\alpha = a + bi$, with $a, b \in \mathbb{Z}$, is represented as the pair of integers (a, b).

EXERCISE 16.38. Let $\alpha, \beta \in \mathbb{Z}[i]$.

(a) Show how to compute $M(\alpha)$ in time $O(\text{len}(M(\alpha)))$ and $N(\alpha)$ in time $O(\text{len}(M(\alpha))^2)$.

(b) Show how to compute $\alpha + \beta$ in time $O(\text{len}(M(\alpha)) + \text{len}(M(\beta)))$.

(c) Show how to compute $\alpha \cdot \beta$ in time $O(\text{len}(M(\alpha)) \cdot \text{len}(M(\beta)))$.

(d) Assuming $\beta \neq 0$, show how to compute $\kappa, \rho \in \mathbb{Z}[i]$ such that $\alpha = \beta\kappa + \rho$, $N(\rho) \leq \frac{1}{2}N(\beta)$, and $N(\kappa) \leq 4N(\alpha)/N(\beta)$. Your algorithm should run in time $O(\text{len}(M(\alpha)) \cdot \text{len}(M(\beta)))$. Hint: see Example 16.24; also, to achieve the stated running time bound, your algorithm should first test if $M(\beta) \geq 2M(\alpha)$.

EXERCISE 16.39. Using the division with remainder algorithm from part (d) of the previous exercise, adapt the Euclidean algorithm for (ordinary) integers to work with Gaussian integers. On inputs $\alpha, \beta \in \mathbb{Z}[i]$, your algorithm should compute a greatest common divisor $\delta \in \mathbb{Z}[i]$ of α and β in time $O(\ell^3)$, where $\ell := \max\{\text{len}(M(\alpha)), \text{len}(M(\beta))\}$.

EXERCISE 16.40. Extend the algorithm of the previous exercise, so that it computes $\sigma, \tau \in \mathbb{Z}[i]$ such that $\alpha\sigma + \beta\tau = \delta$. Your algorithm should run in time $O(\ell^3)$, and it should also be the case that $\text{len}(M(\sigma))$ and $\text{len}(M(\tau))$ are $O(\ell)$.

The algorithms in the previous two exercises for computing greatest common divisors in $\mathbb{Z}[i]$ run in time cubic in the length of their input, whereas the corresponding algorithms for $\mathbb{Z}$ run in time quadratic in the length of their input. This is essentially because the running time of the algorithm for division with remainder discussed in Exercise 16.38 is insensitive to the size of the quotient.

To get a quadratic-time algorithm for computing greatest common divisors in $\mathbb{Z}[i]$, in the following exercises we shall develop an analog of the binary gcd algorithm for $\mathbb{Z}$.

EXERCISE 16.41. Let $\pi := 1 + i \in \mathbb{Z}[i]$.

(a) Show that $2 = \pi\bar{\pi} = -i\pi^2$, that $N(\pi) = 2$, and that π is irreducible in $\mathbb{Z}[i]$.

(b) Let $\alpha \in \mathbb{Z}[i]$, with $\alpha = a + bi$ for $a, b \in \mathbb{Z}$. Show that $\pi \mid \alpha$ if and only if $a - b$ is even, in which case

$$\frac{\alpha}{\pi} = \frac{a+b}{2} + \frac{b-a}{2}i.$$

(c) Show that for all $\alpha \in \mathbb{Z}[i]$, we have $\alpha \equiv 0 \pmod{\pi}$ or $\alpha \equiv 1 \pmod{\pi}$.

(d) Show that the quotient ring $\mathbb{Z}[i]/\pi\mathbb{Z}[i]$ is isomorphic to the ring $\mathbb{Z}_2$.

(e) Show that for all $\alpha \in \mathbb{Z}[i]$ with $\alpha \equiv 1 \pmod{\pi}$, there exists a unique $\varepsilon \in \{\pm 1, \pm i\}$ such that $\alpha \equiv \varepsilon \pmod{2\pi}$.

(f) Show that for all $\alpha, \beta \in \mathbb{Z}[i]$ with $\alpha \equiv \beta \equiv 1 \pmod{\pi}$, there exists a unique $\varepsilon \in \{\pm 1, \pm i\}$ such that $\alpha \equiv \varepsilon\beta \pmod{2\pi}$.

EXERCISE 16.42. We now present a "$(1 + i)$-ary gcd algorithm" for Gaussian integers. Let $\pi := 1 + i \in \mathbb{Z}[i]$. The algorithm takes non-zero $\alpha, \beta \in \mathbb{Z}[i]$ as input, and runs as follows:

$\rho \leftarrow \alpha,\ \rho' \leftarrow \beta,\ e \leftarrow 0$
while $\pi \mid \rho$ and $\pi \mid \rho'$ do $\rho \leftarrow \rho/\pi,\ \rho' \leftarrow \rho'/\pi,\ e \leftarrow e + 1$
repeat
    while $\pi \mid \rho$ do $\rho \leftarrow \rho/\pi$
    while $\pi \mid \rho'$ do $\rho' \leftarrow \rho'/\pi$
    if $M(\rho') < M(\rho)$ then $(\rho, \rho') \leftarrow (\rho', \rho)$
    determine $\varepsilon \in \{\pm 1, \pm i\}$ such that $\rho' \equiv \varepsilon\rho \pmod{2\pi}$
(∗) $\rho' \leftarrow \rho' - \varepsilon\rho$
until $\rho' = 0$
$\delta \leftarrow \pi^e \cdot \rho$
output δ

Show that this algorithm correctly computes a greatest common divisor of α and β, and that it can be implemented so as to run in time $O(\ell^2)$, where $\ell := \max(\mathrm{len}(M(\alpha)), \mathrm{len}(M(\beta)))$. Hint: to analyze the running time, for $i = 1, 2, \ldots$, let v_i (respectively, v_i') denote the value of $|\rho\rho'|$ just before (respectively, after) the execution of the line marked $(*)$ in loop iteration i, and show that

$$v_i' \leq (1 + \sqrt{2})v_i \ \text{ and } \ v_{i+1} \leq v_i'/2\sqrt{2}.$$

EXERCISE 16.43. Extend the algorithm of the previous exercise, so that it computes $\sigma, \tau \in \mathbb{Z}[i]$ such that $\alpha\sigma + \beta\tau = \delta$. Your algorithm should run in time $O(\ell^2)$, and it should also be the case that $\mathrm{len}(M(\sigma))$ and $\mathrm{len}(M(\tau))$ are $O(\ell)$. Hint: adapt the algorithm in Exercise 4.10.

EXERCISE 16.44. In Exercise 16.41, we saw that 2 factors as $-i(1 + i)^2$ in $\mathbb{Z}[i]$,

where $1+i$ is irreducible. This exercise examines the factorization in $\mathbb{Z}[i]$ of prime numbers $p > 2$. Show that:

(a) for every irreducible $\pi \in \mathbb{Z}[i]$, there exists a unique prime number p such that π divides p;

(b) for all prime numbers $p \equiv 1 \pmod 4$, we have $p = \pi\bar{\pi}$, where $\pi \in \mathbb{Z}[i]$ is irreducible, and the complex conjugate $\bar{\pi}$ of π is also irreducible and not associate to π;

(c) all prime numbers $p \equiv 3 \pmod 4$ are irreducible in $\mathbb{Z}[i]$.

Hint: for parts (b) and (c), use Theorem 2.34.

16.9.2 Unique factorization in $D[X]$

In this section, we prove the following:

Theorem 16.43. *If D is a UFD, then so is $D[X]$.*

This theorem implies, for example, that $\mathbb{Z}[X]$ is a UFD. Applying the theorem inductively, one also sees that $\mathbb{Z}[X_1, \ldots, X_n]$ is a UFD, as is $F[X_1, \ldots, X_n]$ for every field F.

We begin with some simple observations. First, recall that for an integral domain D, $D[X]$ is an integral domain, and the units in $D[X]$ are precisely the units in D. Second, it is easy to see that an element of D is irreducible in D if and only if it is irreducible in $D[X]$. Third, for $c \in D$ and $f = \sum_i c_i X^i \in D[X]$, we have $c \mid f$ if and only if $c \mid c_i$ for all i.

We call a non-zero polynomial $f \in D[X]$ **primitive** if the only elements of D that divide f are units. If D is a UFD, then given any non-zero polynomial $f \in D[X]$, we can write it as $f = cf'$, where $c \in D$ and $f' \in D[X]$ is a primitive polynomial: just take c to be a greatest common divisor of all the coefficients of f.

Example 16.25. In $\mathbb{Z}[X]$, the polynomial $f = 4X^2 + 6X + 20$ is not primitive, but we can write $f = 2f'$, where $f' = 2X^2 + 3X + 10$ is primitive. □

It is easy to prove the existence part of Theorem 16.43:

Theorem 16.44. *Let D be a UFD. Every non-zero, non-unit element of $D[X]$ can be expressed as a product of irreducibles in $D[X]$.*

Proof. Let f be a non-zero, non-unit polynomial in $D[X]$. If f is a constant, then because D is a UFD, f factors into irreducibles in D. So assume f is not constant. If f is not primitive, we can write $f = cf'$, where c is a non-zero, non-unit in D, and f' is a primitive, non-constant polynomial in $D[X]$. Again, as D is a UFD, c factors into irreducibles in D.

From the above discussion, it suffices to prove the theorem for non-constant, primitive polynomials $f \in D[X]$. If f is itself irreducible, we are done. Otherwise, we can write $f = gh$, where $g, h \in D[X]$ and neither g nor h are units. Further, by the assumption that f is a primitive, non-constant polynomial, both g and h must also be primitive, non-constant polynomials; in particular, both g and h have degree strictly less than $\deg(f)$, and the theorem follows by induction on degree. □

The uniqueness part of Theorem 16.43 is (as usual) more difficult. We begin with the following fact:

Theorem 16.45. *Let D be a UFD, let p be an irreducible in D, and let $g, h \in D[X]$. Then $p \mid gh$ implies $p \mid g$ or $p \mid h$.*

Proof. Consider the quotient ring D/pD, which is an integral domain (because D is a UFD), and the corresponding ring of polynomials $(D/pD)[X]$, which is also an integral domain. Also consider the natural map that sends $a \in D$ to $\bar{a} := [a]_p \in D/pD$, which we can extend coefficient-wise to a ring homomorphism from $D[X]$ to $(D/pD)[X]$ (see Example 7.46). If $p \mid gh$, then we have

$$0 = \overline{gh} = \bar{g}\bar{h},$$

and since $(D/pD)[X]$ is an integral domain, it follows that $\bar{g} = 0$ or $\bar{h} = 0$, which means that $p \mid g$ or $p \mid h$. □

Theorem 16.46. *Let D be a UFD. The product of two primitive polynomials in $D[X]$ is also primitive.*

Proof. Let $g, h \in D[X]$ be primitive polynomials, and let $f := gh$. If f is not primitive, then $c \mid f$ for some non-zero, non-unit $c \in D$, and as D is a UFD, there is some irreducible element $p \in D$ that divides c, and therefore, divides f as well. By Theorem 16.45, it follows that $p \mid g$ or $p \mid h$, which implies that either g is not primitive or h is not primitive. □

Suppose that D is a UFD and that F is its field of fractions. Any non-zero polynomial $f \in F[X]$ can always be written as $f = (c/d)f'$, where $c, d \in D$, with $d \neq 0$, and $f' \in D[X]$ is primitive. To see this, clear the denominators of the coefficients of f, writing $df = f''$, where $0 \neq d \in D$ and $f'' \in D[X]$. Then take c to be a greatest common divisor of the coefficients of f'', so that $f'' = cf'$, where $f' \in D[X]$ is primitive. Then we have $f = (c/d)f'$, as required. Of course, we may assume that c and d are relatively prime—if not, we may divide c and d by a greatest common divisor.

Example 16.26. Let $f = (3/5)X^2 + 9X + 3/2 \in \mathbb{Q}[X]$. Then we can write $f = (3/10)f'$, where $f' = 2X^2 + 30X + 5 \in \mathbb{Z}[X]$ is primitive. □

As a consequence of the previous theorem, we have:

Theorem 16.47. *Let D be a UFD and let F be its field of fractions. Suppose that $f, g \in D[X]$ and $h \in F[X]$ are non-zero polynomials such that $f = gh$ and g is primitive. Then $h \in D[X]$.*

Proof. Write $h = (c/d)h'$, where $c, d \in D$ and $h' \in D[X]$ is primitive. Let us assume that c and d are relatively prime. Then we have

$$d \cdot f = c \cdot gh'. \tag{16.9}$$

We claim that $d \in D^*$. To see this, note that (16.9) implies that $d \mid (c \cdot gh')$, and the assumption that c and d are relatively prime implies that $d \mid gh'$. But by Theorem 16.46, gh' is primitive, from which it follows that d is a unit. That proves the claim.

It follows that $c/d \in D$, and hence $h = (c/d)h' \in D[X]$. □

Theorem 16.48. *Let D be a UFD and F its field of fractions. If $f \in D[X]$ with $\deg(f) > 0$ is irreducible, then f is also irreducible in $F[X]$.*

Proof. Suppose that f is not irreducible in $F[X]$, so that $f = gh$ for non-constant polynomials $g, h \in F[X]$, both of degree strictly less than that of f. We may write $g = (c/d)g'$, where $c, d \in D$ and $g' \in D[X]$ is primitive. Set $h' := (c/d)h$, so that $f = gh = g'h'$. By Theorem 16.47, we have $h' \in D[X]$, and this shows that f is not irreducible in $D[X]$. □

Theorem 16.49. *Let D be a UFD. Let $f \in D[X]$ with $\deg(f) > 0$ be irreducible, and let $g, h \in D[X]$. If f divides gh in $D[X]$, then f divides either g or h in $D[X]$.*

Proof. Suppose that $f \in D[X]$ with $\deg(f) > 0$ is irreducible. This implies that f is a primitive polynomial. By Theorem 16.48, f is irreducible in $F[X]$, where F is the field of fractions of D. Suppose f divides gh in $D[X]$. Then because $F[X]$ is a UFD, f divides either g or h in $F[X]$. But Theorem 16.47 implies that f divides either g or h in $D[X]$. □

Theorem 16.43 now follows immediately from Theorems 16.44, 16.45, and 16.49, together with Theorem 16.34.

In the proof of Theorem 16.43, there is a clear connection between factorization in $D[X]$ and $F[X]$, where F is the field of fractions of D. We should perhaps make this connection more explicit. Let $f \in D[X]$ be a non-zero polynomial. We may write f as

$$f = up_1^{a_1} \cdots p_r^{a_r} f_1^{b_1} \cdots f_s^{b_s}.$$

where $u \in D^*$, the p_i's are non-associate, irreducible elements of D, and the f_j's are non-associate, irreducible, non-constant polynomials over D (and in particular, primitive). For $j = 1, \ldots, s$, let $g_j := \mathrm{lc}(f_j)^{-1} f_j$ be the monic associate of f_j in $F[X]$. Then in $F[X]$, f factors as

$$f = c g_1^{b_1} \cdots g_s^{b_s},$$

where

$$c := u \cdot \prod_i p_i^{a_i} \cdot \prod_j \mathrm{lc}(f_j)^{b_j} \in F,$$

and the g_j's are distinct, irreducible, monic polynomials over F.

Example 16.27. Consider the polynomial $f = 4X^2 + 2X - 2 \in \mathbb{Z}[X]$. Over $\mathbb{Z}[X]$, f factors as $2(2X - 1)(X + 1)$, where each of these three factors is irreducible in $\mathbb{Z}[X]$. However, over $\mathbb{Q}[X]$, f factors as $4(X - 1/2)(X + 1)$, where 4 is a unit, and the other two factors are irreducible. □

The following theorem provides a useful criterion for establishing that a polynomial is irreducible.

Theorem 16.50 (Eisenstein's criterion). *Let D be a UFD and F its field of fractions. Let $f = c_n X^n + c_{n-1} X^{n-1} + \cdots + c_0 \in D[X]$. If there exists an irreducible $p \in D$ such that*

$$p \nmid c_n,\ p \mid c_{n-1},\ \cdots,\ p \mid c_0,\ p^2 \nmid c_0,$$

then f is irreducible over F.

Proof. Let f be as above, and suppose it were not irreducible in $F[X]$. Then by Theorem 16.48, we could write $f = gh$, where $g, h \in D[X]$, both of degree strictly less than that of f. Let us write

$$g = a_k X^k + \cdots + a_0 \ \text{ and } \ h = b_\ell X^\ell + \cdots + b_0,$$

where $a_k \neq 0$ and $b_\ell \neq 0$, so that $0 < k < n$ and $0 < \ell < n$. Now, since $c_n = a_k b_\ell$, and $p \nmid c_n$, it follows that $p \nmid a_k$ and $p \nmid b_\ell$. Further, since $c_0 = a_0 b_0$, and $p \mid c_0$ but $p^2 \nmid c_0$, it follows that p divides one of a_0 or b_0, but not both—for concreteness, let us assume that $p \mid a_0$ but $p \nmid b_0$. Also, let m be the smallest positive integer such that $p \nmid a_m$—note that $0 < m \leq k < n$.

Now consider the natural map that sends $a \in D$ to $\bar{a} := [a]_p \in D/pD$, which we can extend coefficient-wise to a ring homomorphism from $D[X]$ to $(D/pD)[X]$ (see Example 7.46). Because D is a UFD and p is irreducible, D/pD is an integral domain. Since $f = gh$, we have

$$\bar{c}_n X^n = \bar{f} = \bar{g}\bar{h} = (\bar{a}_k X^k + \cdots + \bar{a}_m X^m)(\bar{b}_\ell X^\ell + \cdots + \bar{b}_0). \tag{16.10}$$

But notice that when we multiply out the two polynomials on the right-hand side of (16.10), the coefficient of X^m is $\bar{a}_m \bar{b}_0 \neq 0$, and as $m < n$, this clearly contradicts the fact that the coefficient of X^m in the polynomial on the left-hand side of (16.10) is zero. □

As an application of Eisenstein's criterion, we have:

Theorem 16.51. *For every prime number q, the qth cyclotomic polynomial*

$$\Phi_q := \frac{X^q - 1}{X - 1} = X^{q-1} + X^{q-2} + \cdots + 1$$

is irreducible over $\mathbb{Q}$.

Proof. Let

$$f := \Phi_q\big(X + 1\big) = \frac{(X+1)^q - 1}{(X+1) - 1}.$$

It is easy to see that

$$f = \sum_{i=0}^{q-1} c_i X^i, \quad \text{where } c_i = \binom{q}{i+1} \ (i = 0, \ldots, q-1).$$

Thus, $c_{q-1} = 1$, $c_0 = q$, and for $0 < i < q - 1$, we have $q \mid c_i$ (see Exercise 1.14). Theorem 16.50 therefore applies, and we conclude that f is irreducible over $\mathbb{Q}$. It follows that Φ_q is irreducible over $\mathbb{Q}$, since if $\Phi_q = gh$ were a non-trivial factorization of Φ_q, then $f = \Phi_q\big(X + 1\big) = g\big(X + 1\big) \cdot h\big(X + 1\big)$ would be a non-trivial factorization of f. □

EXERCISE 16.45. Show that neither $\mathbb{Z}[X]$ nor $F[X, Y]$ (where F is a field) are PIDs (even though they are UFDs).

EXERCISE 16.46. Let $f \in \mathbb{Z}[X]$ be a monic polynomial. Show that if f has a root $x \in \mathbb{Q}$, then $x \in \mathbb{Z}$, and x divides the constant term of f.

EXERCISE 16.47. Let D be a UFD, let p be an irreducible element of D, and consider the natural map that sends $a \in D$ to $\bar{a} := [a]_p \in D/pD$, which we extend coefficient-wise to a ring homomorphism from $D[X]$ to $(D/pD)[X]$ (see Example 7.46). Show that if $f \in D[X]$ is a primitive polynomial such that $p \nmid \mathrm{lc}(f)$ and $\bar{f} \in (D/pD)[X]$ is irreducible, then f is irreducible.

EXERCISE 16.48. Let a be a non-zero, square-free integer, with $a \notin \{\pm 1\}$, and let n be a positive integer. Show that the polynomial $X^n - a$ is irreducible in $\mathbb{Q}[X]$.

EXERCISE 16.49. Show that the polynomial $X^4 + 1$ is irreducible in $\mathbb{Q}[X]$.

EXERCISE 16.50. Let F be a field, and consider the ring of bivariate polynomials $F[X, Y]$. Show that in this ring, the polynomial $X^2 + Y^2 - 1$ is irreducible, provided F does not have characteristic 2. What happens if F has characteristic 2?

EXERCISE 16.51. Design and analyze an efficient algorithm for the following problem. The input is a pair of polynomials $g, h \in \mathbb{Z}[X]$, along with their greatest common divisor d in the ring $\mathbb{Q}[X]$. The output is the greatest common divisor of g and h in the ring $\mathbb{Z}[X]$.

EXERCISE 16.52. Let $g, h \in \mathbb{Z}[X]$ be non-zero polynomials with $d := \gcd(g, h) \in \mathbb{Z}[X]$. Show that for every prime p not dividing $\mathrm{lc}(g)\,\mathrm{lc}(h)$, we have $\bar{d} \mid \gcd(\bar{g}, \bar{h})$, and except for finitely many primes p, we have $\bar{d} = \gcd(\bar{g}, \bar{h})$. Here, $\bar{d}$, $\bar{g}$, and $\bar{h}$ denote the images of d, g, and h in $\mathbb{Z}_p[X]$ under the coefficient-wise extension of the natural map from $\mathbb{Z}$ to $\mathbb{Z}_p$ (see Example 7.47).

EXERCISE 16.53. Let F be a field, and let $g, h \in F[X, Y]$. Define $V(g, h) := \{(x, y) \in F \times F : g(x, y) = h(x, y) = 0\}$. Show that if g and h are relatively prime, then $V(g, h)$ is a finite set. Hint: consider the rings $F(X)[Y]$ and $F(Y)[X]$.

16.10 Notes

The "$(1 + i)$-ary gcd algorithm" in Exercise 16.42 for computing greatest common divisors of Gaussian integers is based on algorithms in Weilert [106] and Damgård and Frandsen [31]. The latter paper also develops a corresponding algorithm for Eisenstein integers (see Exercise 16.35). Weilert [107] presents an asymptotically fast algorithm that computes the greatest common divisor of Gaussian integers of length at most ℓ in time $O(\ell^{1+o(1)})$.