Algebra

Willem A. de Graaf

Contents

Chapter 1

Introduction

1.1 What is algebra

It is a fruitless effort to try to define precisely what the mathematical area of algebra is. The whole of science is divided into an ever increasing number of areas, but there are no precise borders between them, as the same piece of knowledge can play an important role in many different areas. However, roughly, it is possible to say that algebra is concerned with the study of algebraic structures. An algebraic structure is a set with one or more operations. An operation on a set S is a map $S \times S \to S$, i.e., a function that takes a pair of elements of S and associates a third element of S to it. Some examples of algebraic structures immediately come to mind:

- the set $\mathbb{N} = \{0, 1, 2, \ldots\}$ of non-negative integers with the operations $+$ (addition) and $\cdot$ (multiplication),
- the set of integers $\mathbb{Z}$ with the same operations as $\mathbb{N}$,
- the set of rational numbers $\mathbb{Q}$ again with addition and multiplication,
- the set $F = \{f : M \to M\}$ of functions from a set M to itself, with the operation of composition of functions $(f, g) \mapsto f \circ g$, where $f \circ g(m) = f(g(m))$.

Not all algebraic structures that one can think of are equally interesting. However, there are algebraic structes that are defined in a seemingly weird way, on the face of it just as pastime for academics with nothing better to do, but which do turn out to have important applications. As an example we consider the set $\mathcal{N} = \{m \in \mathbb{Z} \mid m \geq 0\}$. For a subset $S \subset \mathcal{N}$ we define

$$\operatorname{mex}(S) = \min\{m \in \mathcal{N} \mid m \notin S\}$$

(the *m*inimal *ex*cluded value of S). Now for $a, b \in \mathcal{N}$ we set

$$a \oplus b = \operatorname{mex}(\{a' \oplus b \mid 0 \leq a' < a\} \cup \{a \oplus b' \mid 0 \leq b' < b\}).$$

So $0 \oplus 0 = \operatorname{mex}(\emptyset) = 0$, $1 \oplus 0 = 0 \oplus 1 = \operatorname{mex}(\{0\}) = 1$, $1 \oplus 1 = \operatorname{mex}(\{1\}) = 0$. This seems to be a rather odd operation, but it has interesting applications in combinatorial game theory (see Section 4.8.5 or [Sie13]).

An important concept in this context is the one of *isomorphism*. Roughly speaking, two algebraic structures are isomorphic if there is a bijective map between them that preserves the operations. For example, if we consider structures with one operation, then we say that the structure S with operation $\cdot$ is isomorphic to the structure T with operation $\circ$ if there is a bijection $\sigma : S \to T$ such that $\sigma(s_1 \cdot s_2) = \sigma(s_1) \circ \sigma(s_2)$ for all $s_1, s_2 \in S$. Similarly, if we consider structures with two operations (such as $\mathbb{Z}$, $\mathbb{Q}$,...) then an isomorphism is required to preserve both operations.

Example 1.1.1 Consider $\mathbb{Z}$ with just the addition, and the set of 2×2-matrices

$$A = \{\left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right) \mid m \in \mathbb{Z}\},$$

with the operation of matrix multiplication. Then a small calculation shows that $\sigma : \mathbb{Z} \to A$, $\sigma(m) = \left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)$ is an isomorphism.

Two isomorphic structures are equal as far as their algebraic structure is concerned. However, in order to solve a particular problem, it may be advantageous to work with a particular structure, and not with another isomorphic one. In the example above, when working with the set A one may use the fact that its elements can be seen as linear maps and bring the machinery of linear algebra to bear on a particular problem.

Historically the interest in algebra has come from attempts to solve certain problems. One such problem was what is called Fermat's last theorem, stating that for an integer $n \geq 3$ there are no integers a, b, c, all of them nonzero, such that $a^n + b^n = c^n$. This was proved by Fermat for $n = 4$, and later Euler had published a proof for $n = 3$. Later investigations showed the connection with the problem of unique factorization. In order to study the latter it was necessary to develop the theory of certain algebraic structures called rings. These are the subject of Chapter 2.

A second problem has been the search for a formula to solve polynomial equations of degree 5. We all know how to solve an equation of degree 2, $x^2 + bx + c = 0$, namely

$$x = \frac{-b \pm \sqrt{b^2 - 4c}}{2}.$$

Similar formulas for degree 3 and 4 have been worked out by many mathematicians (the most famous probably being Cardano (1501-1576), who published such formulas in his book Ars Magna). So the problem was to find a formula for polynomials of degree 5, until Galois (1811-1832) showed that no such formula could exist. He introduced a completely new idea (in fact, it was so revolutionary that his contemporaries did not understand it). Galois attached an algebraic structure, called a group, to a polynomial. By investigating this type of algebraic structure he could show that in general there cannot be a formula for the roots of a polynomial of degree 5 (in terms of arithmetic operations and k-th roots $\sqrt[k]{\ }$). The group of Galois encodes the symmetries that exist between the roots of a given polynomial. Similar ideas have subsequently arisen in many areas of science; it has turned out that looking at a group of symmetries can lead to surprising and spectacular results (for an overview, see the book by Stewart ([Ste07])). In Chapter 3 we will study symmetries and groups.

1.2 Mathematical induction

In algebra, the proof technique of induction plays a particularly important role. Therefore we briefly have a look at it.

Consider the set $\mathbb{N}$, and let $P(n)$ be a property that an element $n \in \mathbb{N}$ can have. Then the *ordinary principle of induction* states that if

$$P(0) \text{ holds and} \tag{1.2.1}$$
$$P(n) \Rightarrow P(n+1) \text{ for all } n \geq 0 \tag{1.2.2}$$

then $P(n)$ holds for all $n \in \mathbb{N}$. Here (1.2.1) is the *base case*, and the hypothesis $P(n)$ in (1.2.2) is called the *induction hypothesis*.

This is intuitively clear, as this makes it possible to write down a proof for $P(n_0)$ for each $n_0 \in \mathbb{N}$. Indeed, let $n_0 = 3$. We have $P(0)$ and as $P(0) \Rightarrow P(1)$, also $P(1)$ holds. Since $P(1) \Rightarrow P(2)$, also $P(2)$ holds. Finally, $P(2) \Rightarrow P(3)$ proves $P(3)$. However, it must be clear that having a proof of $P(n_0)$ for each n_0 separately, is not the same as having a proof that $P(n)$ holds for all $n \in \mathbb{N}$. (The proofs become longer and longer, and cannot be combined to a single proof.) So we must accept the principle of induction as a kind of axioma.

Some proofs by induction do not use the ordinary principle, but rather the *strong principle of induction*. This states that if for all $n \geq 0$

$$(P(k) \text{ for all } k \text{ with } 0 \leq k < n) \Rightarrow P(n), \tag{1.2.3}$$

then $P(n)$ holds for all $n \in \mathbb{N}$. Here the base case ($P(0)$) is included in (1.2.3): the statement "$P(k)$ for all k with $0 \leq k < n$" is trivially true for $n = 0$ (there are no such k), so if we have shown (1.2.3) for all $n \geq 0$ then $P(0)$ follows.

Theorem 1.2.1 *The* well ordering principle *for* $\mathbb{N}$ *states that every non-empty subset of* $\mathbb{N}$ *has a smallest element. The ordinary principle of induction holds for* $\mathbb{N}$ *if and only if the strong principle induction holds for* $\mathbb{N}$ *if and only if the well ordering principle holds for* $\mathbb{N}$.

Proof. Suppose that the principle of strong induction holds for $\mathbb{N}$. Suppose also that $P(0)$ and $P(n) \Rightarrow P(n+1)$ for all $n \geq 0$. We show that $P(n)$ for all $n \geq 0$ by strong induction. Consider the statement $S(n)$ which asserts that $P(k)$ holds for all k with $0 \leq k < n$. Let $n \in \mathbb{N}$, $n \geq 0$. If $n = 0$ then $S(n) \Rightarrow P(0)$ as $P(0)$ is true. Let $n > 0$ and suppose $S(n)$. Then in particular, $P(n-1)$. So because $P(l) \Rightarrow P(l+1)$ for all $l \geq 0$, it follows that $P(n)$. We see that for all n we have $S(n) \Rightarrow P(n)$, which is (1.2.3). So by strong induction we see that $P(n)$ holds for all $n \geq 0$. We conclude that the ordinary principle of induction holds.

Suppose that the ordinary principle of induction holds. Consider the statement $P(n)$ which asserts that if a set $S \subseteq \mathbb{N}$ contains an $x \in \mathbb{N}$ with $x \leq n$, then S has a smallest element. We show $P(n)$ by induction. Firstly, if an $S \subset \mathbb{N}$ contains an x with $x \leq 0$, then $x = 0$. As 0 is the smallest element of $\mathbb{N}$, we see that 0 is also the smallest element of S. Secondly, suppose $P(n)$ and let $S \subset \mathbb{N}$ contain an x with $x \leq n+1$. If S has no element y with $y < n+1$ then $x = n+1$ is the smallest element of S. On the other hand, if S contains such a y then it contains a $y \leq n$. By $P(n)$ it follows that S has a smallest element. By induction we conclude that $P(n)$ holds for all $n \in \mathbb{N}$. Now let $S \subset \mathbb{N}$ be non-empty. Hence it contains some $n \in \mathbb{N}$. Therefore, by $P(n)$ it follows that S has a smallest element.

Finally, suppose that the well ordering principle holds. Suppose (1.2.3). Let $S = \{n \in \mathbb{N} \mid P(n) \text{ does not hold}\}$. If S is non-empty, then S has a smallest element m. So for $0 \leq k < m$ we have $P(k)$. Therefore, we also have $P(m)$, which is a contradiction. We conclude $S = \emptyset$ and $P(n)$ holds for all $n \in \mathbb{N}$. □

Of course, it is also possible to use induction to prove statements for integers ranging over different sets, like $\{k \in \mathbb{Z} \mid k \geq k_0\}$. In these cases, when using ordinary induction, the base case is $P(k_0)$ and the induction step is $P(n) \Rightarrow P(n+1)$ for all $n \geq k_0$. When using strong induction, the induction hypothesis is $P(k)$ for all k with $k_0 \leq k < n$ (from which then $P(n)$ has to be seen to follow).

More in general, it is possible to use induction to prove statements for the elements of any set where the well ordering principle holds. For example, we can consider the set $\mathbb{N}^m$ consisting of m-tuples $(i_1, \ldots, i_m)$, with $i_k \in \mathbb{N}$ for $1 \leq k \leq m$. This set can be ordered by the *lexicographical order*: $(i_1, \ldots, i_m) < (j_1, \ldots, j_m)$ if $i_k < j_k$ where k is minimal with the property that $i_k \neq j_k$. It is a non trivial fact that this is a well ordering (a fact that can be proved by induction on m!). It can be used to prove statements for elements of the set $\mathbb{N}^m$ by induction.

Because of the equivalences of Theorem 1.2.1 it is clear that any proof using induction can be rewritten to a proof using the well-ordering principle. On some occasions the use of induction leads to a more elegant proof, on other occasions the use of the well-ordering principle can be preferred.

When doing a proof by induction it is a good idea to state clearly what the set of elements is for which the statement is supposed to hold, and to precisely formulate the induction hypothesis.

1.3 Associativity

Frequently the operations of an algebraic structure are associative. Here we briefly go into a consequence of that property.

Let S be a set with operation $\cdot$. Then we say that $\cdot$ is *associative* if $s \cdot (t \cdot u) = (s \cdot t) \cdot u$ for all $s, t, u \in S$.

So let S be a set with an associative operation $\cdot$. Let $s_1, \ldots, s_n \in S$ which we multiply together in that order. There are many ways to do this. For example, for $n = 4$ we have the possibilities

$$s_1 \cdot (s_2 \cdot (s_3 \cdot s_4)),\ s_1 \cdot ((s_2 \cdot s_3) \cdot s_4),\ (s_1 \cdot s_2) \cdot (s_3 \cdot s_4),\ ((s_1 \cdot s_2) \cdot s_3) \cdot s_4,\ (s_1 \cdot (s_2 \cdot s_3)) \cdot s_4.$$

So when writing the expression we want $s_1, \ldots, s_n$ to appear in that order. Only the way of bracketing the expression can be different. We claim that all such bracketings lead to the same result. We show that by induction on n. The induction hypothesis is that for all expressions with m elements of S, where $1 \leq m < n$, the result is independent of the bracketing. If n is 1 or 2, then the result is trivially independent on the bracketing, so suppose that $n \geq 3$. Consider two different bracketings where the elements $s_1, \ldots, s_n$ appear in that order, so that we get two elements $v_1 \cdot v_2$ and $w_1 \cdot w_2$, where v_1 involves $s_1, \ldots, s_k$, v_2 involves $v_{k+1}, \ldots, v_n$, w_1 involves $s_1, \ldots, s_l$, w_2 involves $s_{l+1}, \cdots, s_n$. By induction on n we have that $v_1 = s_1 \cdot x_1$, $w_1 = s_1 \cdot y_1$, where x_1 involves $s_2, \ldots, s_k$, and y_1 involves $s_2, \ldots, s_l$. Then by the associative property, $v_1 \cdot v_2 = (s_1 \cdot x_1) \cdot v_2 = s_1 \cdot (x_1 \cdot v_2)$ and similarly, $w_1 \cdot w_2 = s_1 \cdot (y_1 \cdot w_2)$. By induction $x_1 \cdot v_2 = y_1 \cdot w_2$. Hence $v_1 \cdot v_2 = w_1 \cdot w_2$.

For this reason, when dealing with an associative operation, we do not write the bracketing. For example, we write $s_1 \cdot s_2 \cdot s_3 \cdot s_4$ (or even simpler, $s_1 s_2 s_3 s_4$) and not one of the bracketed expressions above.

Using the associativity of the operation we can also define an exponentiation. Let $s \in S$ and $n \in \mathbb{Z}$, $n \geq 1$. Then we let s^n denote the n-fold product of s with itself. Because of what we have seen above, it does not matter how we bracket this product. Therefore we have the following fundamental property

$$s^m \cdot s^n = s^{m+n} \text{ for all } m, n \in \mathbb{Z},\ m, n \geq 1.$$

Now suppose that additionally S has a *neutral element*, that is, there is a $1 \in S$ with $1 \cdot s = s \cdot 1 = s$ for all $s \in S$. Let $s \in S$ have an *inverse*, that is, an $s^{-1} \in S$ such that $s \cdot s^{-1} = s^{-1} \cdot s = 1$. Then we define $s^0 = 1$ and for $n \leq -1$ we let s^n be the $(-n)$-fold product of s^{-1} with itself. Then we have the following extension of the above formula:

$$s^m \cdot s^n = s^{m+n} \text{ for all } m, n \in \mathbb{Z}. \tag{1.3.1}$$

Indeed, if $m, n \geq 1$ then this has already been observed. If either m or n is zero then it is obvious. If both $m, n \leq -1$ then this again follows from the above. So suppose $m \leq -1$, $n \geq 1$. Then

$$s^m \cdot s^n = s^{m+1} \cdot (s^{-1} \cdot s) \cdot s^{n-1} = s^{m+1} \cdot s^{n-1}.$$

We see that we can finish the proof by induction on n.

1.4 Equivalence relations

Here we have a short look at equivalence relations. The main usage for us of these relations lies in the construction of quotient structures. The general pattern of these is as follows. Let S be an algebraic structure (so a set with one or more operations). Then, roughly speaking, a substructure is a subset $T \subset S$ such that the operations of S restrict to T. For example, the set of even integers $2\mathbb{Z} = \{2m \mid m \in \mathbb{Z}\}$ is a substructure of $\mathbb{Z}$. We can use T to define an equivalence relation on S. In the example the relation can be defined as follows: $a \sim b$ if and only if $a - b \in 2\mathbb{Z}$. Now if the given substructure is *distinguished* (the definition of this depends on the kind of structure), then the operations of S induce operations on the set of equivalence classes of the defined relation. This set of equivalence classes thus gets an algebraic structure itself, which is called the quotient of S by T.

Let A be a non-empty set. A *relation* on A is a subset R of $A \times A$. Instead of $(a, b) \in R$ we write aRb to express that $a, b \in A$ are in the relation R.

The relation $R \subset A \times A$ is called an *equivalence relation* if it

- is *reflexive*, that is, aRa for all $a \in A$,
- is *symmetric*, that is, $aRb \Rightarrow bRa$ for all $a, b \in A$,
- is *transitive*, that is, aRb and $bRc \Rightarrow aRc$ for all $a, b, c \in A$.

Consider an equivalence relation R on the set A. Then for $a \in A$ we define

$$[a] = \{b \in A \mid aRb\},$$

which is called the *equivalence class* of a (it is the set of all $b \in A$ that are equivalent to a in the relation R). Note that $a \in [a]$ because R is reflexive.

Example 1.4.1 Let $A = \mathbb{Z}$ and $R_2 = \{(a, b) \in \mathbb{Z}^2 \mid a - b \text{ is even } \}$. Then R_2 is an equivalence relation. For $a \in \mathbb{Z}$ we have that $[a]$ is the set of even integers if a is even, and it is the set of odd integers if a is odd.

One problem with equivalence classes is that we can have $[a] = [b]$, also if $a \neq b$. The next proposition has some characterizations of this situation.

Proposition 1.4.2 *Let R be an equivalence relation on the set A. Let $a, b \in A$. The following are equivalent:*

(i) aRb,

(ii) $a \in [b]$,

(iii) $[a] \subseteq [b]$,

(iv) $[a] = [b]$,

(v) $[a] \cap [b] \neq \emptyset$.

Proof. If aRb then also bRa so that $a \in [b]$, showing (i) $\Rightarrow$ (ii).

Suppose that $a \in [b]$, so that bRa. Let $c \in [a]$, that is, aRc. By transitivity we see that bRc and $c \in [b]$. We conclude that (ii) implies (iii).

Suppose (iii). Let $c \in [b]$, so that bRc. As $a \in [a]$ also $a \in [b]$, and hence bRa, and by symmetry aRb. Now by transitivity it follows that aRc and $c \in [a]$. We conclude that $[a] = [b]$.

The implication (iv) $\Rightarrow$ (v) is obvious.

Finally, suppose (v). Then there exists a $c \in A$ lying in $[a]$ and in $[b]$. So aRc and bRc. The latter implies cRb, so by transitivity it follows aRb. □

In particular it follows that two equivalence classes $[a]$, $[b]$ are either equal ($[a] = [b]$), or they are disjoint ($[a] \cap [b] = \emptyset$).

A *partition* of a set A is defined to be a collection of non-empty, pairwise disjoint subsets of A, whose union is A.

We see that the collection of equivalence classes of an equivalence relation R on A is a partition of A.

Example 1.4.3 Consider the equivalence relation of Example 1.4.1. The equivalence classes of R_2 are exactly the classes [0] (all even integers) and [1] (all odd integers).

Chapter 2

Rings

In this chapter we look at algebraic structures called rings. Historically, one of the main problems in the theory of rings has been to establish whether a given ring satisfies a property analogous to the unique prime factorization in $\mathbb{Z}$. This will be the main theme of this chapter. The interest in this problem has mainly come from attempts to prove Fermat's last theorem: if $n \geq 3$ then there are no nonzero $x, y, z \in \mathbb{Z}$ with $x^n + y^n = z^n$. In mysterious ways this theorem relates to the problem of unique factorization.

Also we look at so-called congruences, which at first sight have nothing to do with unique factorization. We will see some applications of the arithmetic of congruences (also called modular arithmetic), the most famous of which is the RSA cryptosystem. Generalizing the concept of congruence to rings other than $\mathbb{Z}$ leads to the concept of an ideal. It turns out that with this concept we can say something new about the problem of unique factorization. So congruences have something to do with unique factorization after all. The final section has a few applications of unique factorization, among which is a proof of Fermat's last theorem for $n = 3$.

2.1 Definition and first examples

Definition 2.1.1 *A* ring *is a nonempty set R with two operations, $+$ (addition) and $\cdot$ (multiplication) with the following properties.*

1. *(a) $+$ is* commutative*, i.e., $r + s = s + r$ for all $r, s \in R$.*
   (b) $+$ is associative*, i.e., $r + (s + t) = (r + s) + t$ for all $r, s, t \in R$.*
   (c) There exists a $0 \in R$ (called zero) such that $0 + r = r + 0 = r$ for all $r \in R$.
   (d) For each $r \in R$ there exists an $s \in R$ with $r + s = 0$. (This s is called the opposite *of r and denoted $-r$.)*
2. *$\cdot$ is associative, i.e., $r \cdot (s \cdot t) = (r \cdot s) \cdot t$ for all $s, r, t \in R$.*
3. *The* distributive laws *hold: $r \cdot (s + t) = r \cdot s + r \cdot t$, $(s + t) \cdot r = s \cdot r + t \cdot r$ for all $r, s, t \in R$.*

Usually (except in this section!) we omit the $\cdot$ when doing a multiplication, that is, we write rs instead of $r \cdot s$.

Some examples immediately come to mind: $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$. Note that $\mathbb{N}$ (with the usual addition and multiplication) is not a ring as not every element has an opposite.

The zoo of algebraic objects has many rings, and in order to divide them into certain categories, a number of properties that a ring can have are used. Some of these are as follows.

- A ring R is said to be *commutative* if the multiplication is commutative, i.e., $s \cdot t = t \cdot s$ for all $s, t \in R$.
- A ring R is said to have a *unity* (or *one*) if there is an element $1 \in R$ with $1 \cdot r = r \cdot 1 = r$ for all $r \in R$.

- An $r \in R$ with $r \neq 0$ such that there exists an $s \in R$, $s \neq 0$ with $r \cdot s = 0$ or $s \cdot r = 0$ is called a *zero divisor*. A commutative ring with unity 1, with $1 \neq 0$, without zero divisors is called a *domain*.

- Let R be a ring with unity 1, with $1 \neq 0$. An element $r \in R$ is said to be *invertible* if there is an $s \in R$ with $r \cdot s = s \cdot r = 1$. (This s is called the *inverse* of r and denoted r^{-1}.) A commutative ring with unity 1, such that $1 \neq 0$, such that every nonzero element is invertible is called a *field*.

We have that $\mathbb{Z}$ is a domain, but not a field. The set of even integers $2\mathbb{Z} = \{2m \mid m \in \mathbb{Z}\}$ is a ring, but has no unity, so it is not a domain. The set of 2×2-matrices

$$M_2(\mathbb{Z}) = \{\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \mid a, b, c, d \in \mathbb{Z}\}$$

together with matrix addition and matrix multiplication, is a ring with unity. But it is not commutative, and has zero divisors. So it is not a domain. The rings $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are fields.

There are many immediate consequences of the definitions. Let R be a ring, $r, s \in R$. Then:

- The zero is unique. (Let $0'$ have the same property; then $0 + 0' = 0'$ but also $0 + 0' = 0$.)

- The opposite of r is unique. (Let $s_1, s_2 \in R$ have the property that $r + s_i = 0$, $i = 1, 2$. Then $s_2 = s_2 + 0 = s_2 + (r + s_1) = (s_2 + r) + s_1 = 0 + s_1 = s_1$.)

- Analogously one shows that the unity of R (if R has a unity), and the inverse of $r \in R$ (if r has an inverse) are unique.

- $r \cdot 0 = 0 \cdot r = 0$ ($r \cdot 0 = r \cdot (0 + 0) = r \cdot 0 + r \cdot 0$, and now add $-(r \cdot 0)$ to both sides).

- $r \cdot (-s) = (-r) \cdot s = -(r \cdot s)$ ($0 = r \cdot 0 = r \cdot (s + (-s)) = r \cdot s + r \cdot (-s)$, and add $-(r \cdot s)$ to both sides).

- $(-r) \cdot (-s) = r \cdot s$ (using the previous point: $(-r) \cdot (-s) = -(r \cdot (-s)) = -(-(r \cdot s)) = r \cdot s$).

One word about the condition $1 \neq 0$ that we have put in several places. If $1 = 0$ then using the fourth point above, we see $r = 1 \cdot r = 0 \cdot r = 0$, so that $R = \{0\}$, a rather trivial ring. We see that if R has more than one element, then a unity is automatically different from zero.

We note that a field is automatically a domain. Indeed, let R be a field and $r \in R$, $r \neq 0$. If $r \cdot s = 0$ then $0 = r^{-1} \cdot 0 = r^{-1} \cdot (r \cdot s) = (r^{-1} \cdot r) \cdot s = 1 \cdot s = s$. So s is necessarily zero and therefore r is not a zero divisor.

Also we remark that in a domain the *cancellation law* holds: if $t \neq 0$ and $t \cdot r = t \cdot s$ then $r = s$. Indeed, $t \cdot r = t \cdot s$ implies that $0 = t \cdot r + (-(t \cdot s)) = t \cdot r + t \cdot (-s) = t \cdot (r + (-s))$. So, since there are no zero divisors, $r + (-s) = 0$, implying $r = s$.

I have more than once been asked by non mathematicians why it is that "minus times minus is plus". One answer is that there is no reason for this. One could define a set with two operations (addition and multiplication), where this does not hold. However, if we want the operations to satisfy the (in general rather desirable) properties of Definition 2.1.1, then we necessarily have that minus times minus is plus. (See the last point above.)

2.2 Factorization in $\mathbb{Z}$

2.2.1 Divisibility

Let $a, b \in \mathbb{Z}$. We say that a *divides* b if there is a $c \in \mathbb{Z}$ with $b = ca$. Notation: $a|b$. If a does not divide b then we write $a \not| b$. Examples: $2|6$, $13|143$, $12 \not| 26$. If a divides b we also say that a is a *divisor* of b, or that b is a *multiple* of a, or that b is *divisible* by a.

Note that this definition only uses the ring operations of $\mathbb{Z}$. An alternative definition could read: "a divides b if $\frac{b}{a}$ lies in $\mathbb{Z}$". Firstly, this definition creates some problems when $a = 0$. Secondly, and

more importantly, it tacitly uses the inverse of a, which lies in $\mathbb{Q}$ and not in $\mathbb{Z}$. Therefore we prefer the first definition given above.

We have some immediate properties:

- $a|0$ for all $a \in \mathbb{Z}$ (since $0 = 0 \cdot a$), and $0|a$ if and only if $a = 0$.
- If $a|b$ and $b|c$ then $a|c$.
- $a|b$ and $b|a$ if and only if $a = \pm b$. (Indeed, if $b = c_1 a$ and $a = c_2 b$, then $b = c_1 c_2 b$ so that, if $b \neq 0$, we get $c_1 c_2 = 1$ and c_1, c_2 are both ± 1. If $b = 0$ then a has to be 0 as well.)

Proposition 2.2.1 (Division with remainder) *Let $a, b \in \mathbb{Z}$ with $b \neq 0$. Then there are unique $q, r \in \mathbb{Z}$ with $a = qb + r$ and $0 \leq r < |b|$.*

Proof. First we show existence. For that consider the set

$$M = \{a - qb \mid q \in \mathbb{Z}\} \cap \mathbb{Z}_{\geq 0}.$$

We note that M is not the empty set because $b \neq 0$. Let r be the minimal element of M. If $r > |b|$ then $r - |b|$ also lies in M, contrary to the choice of r. Hence $0 \leq r < |b|$. Moreover, as $r \in M$ there is a $q \in \mathbb{Z}$ with $r = a - qb$ and hence $a = qb + r$.

In order to show uniqueness, suppose that $a = qb + r = q'b + r'$ where $0 \leq r, r' < |b|$. Then $0 = (q - q')b + r - r'$ so that $r - r'$ is a multiple of b. As $|r - r'| < |b|$, it follows that $r - r' = 0$ and $q - q' = 0$ as well. □

The r of the proposition is called the *remainder* of a upon division by b. Example: $73 = 3 \cdot 21 + 10$, so 10 is the remainder of 73 upon division by 21.

2.2.2 Greatest common divisor

The next concept we look at is the greatest common divisor. Also here the question arises how to define this so that the definition makes sense also for a larger class of rings. The first try, also motivating the name, is to define the greatest common divisor of integers a, b as the largest integer dividing both a and b. However, this uses the natural ordering on $\mathbb{Z}$, something that other rings may not have. It turns out that the following definition does carry over nicely to many other rings.

Definition 2.2.2 *Let $a, b \in \mathbb{Z}$. An element d of $\mathbb{Z}$ is called a* greatest common divisor *of a, b if*

1. *d divides a and b,*
2. *if $c \in \mathbb{Z}$ divides a and b then c divides d.*

The disadvantage of this definition is that it is not immediately clear that every $a, b \in \mathbb{Z}$ have a greatest common divisor. This is shown in the next theorem.

Theorem 2.2.3 *Let $a, b \in \mathbb{Z}$. They have a greatest common divisor d. Moreover, there exist $s, t \in \mathbb{Z}$ such that $sa + tb = d$.*

Proof. If $a = b = 0$ then we take $d = 0$, and $s = t = 0$. Otherwise we set

$$I = \{xa + yb \mid x, y \in \mathbb{Z}\}.$$

Then I contains positive integers. Let d be the minimal positive element of I. We first show that d divides both a and b. Perform a division with remainder (Proposition 2.2.1), and obtain $q, r \in \mathbb{Z}$ with $a = qd + r$ and $0 \leq r < d$. Now $a, d \in I$ and hence also $r = a - qd$ lies in I. By the choice of d it follows that $r = 0$ and we see that d divides a. In the same way it is shown that d divides b.

Let $c \in \mathbb{Z}$ be such that c divides both a and b. Then $a = a_0 c$, $b = b_0 c$ for certain $a_0, b_0 \in \mathbb{Z}$. As $d \in I$ there are $s, t \in \mathbb{Z}$ with $d = sa + tb$, so that $d = (sa_0 + tb_0)c$ and we see that c divides d.

The last statement follows immediately from the fact that d lies in I. □

The previous proof does not yield a method for finding a greatest common divisor of given $a, b \in \mathbb{Z}$. A good method for that is called the *Euclidean algorithm* as a description of it is contained in Euclid's Elements. It is based on the following lemma.

Lemma 2.2.4 *Let $a, b \in \mathbb{Z}$. Let $q, r \in \mathbb{Z}$ be such that $a = qb + r$. Then $d \in \mathbb{Z}$ is a greatest common divisor of a, b if and only if it is a greatest common divisor of b, r.*

Proof. Suppose that $d \in \mathbb{Z}$ is a greatest common divisor of a, b. Then in particular $d|a$, $d|b$ so that $a = a_0 d$, $b = b_0 d$ for certain $a_0, b_0 \in \mathbb{Z}$. So $r = a - qb = (a_0 - qb_0)d$ and we see that d divides r as well. Suppose that $c \in \mathbb{Z}$ divides b and r. Then by a similar argument we see that c divides a because $a = qb + r$. As d is a greatest common divisor of a, b it follows that c divides d. We conclude that d is also a greatest common divisor of b, r.

Now suppose that $d \in \mathbb{Z}$ is a greatest common divisor of b, r. Then $d|b$ and $d|r$ and hence d also divides a as $a = qb + r$. Let $c \in \mathbb{Z}$ divide both a, b. Then c divides r as $r = a - qb$. Hence c divides d. The conclusion is that d is a greatest common divisor of a, b. □

In order to describe the algorithm we first note that a greatest common divisor of a, b is also a greatest common divisor of $-a, b$, $a, -b$, $-a, -b$, b, a. Hence we may assume $a \geq b \geq 0$. We define a sequence of integers $r_i \geq 0$ as follows. First of all, $r_0 = a$, $r_1 = b$. Secondly, if r_{i-1}, r_i are defined we perform a division with remainder to obtain $q_i, r_{i+1} \in \mathbb{Z}$ with $r_{i-1} = q_i r_i + r_{i+1}$ and $0 \leq r_{i+1} < r_i$. The procedure terminates when we find a k with $r_{k+1} = 0$.

Note that $r_0 \geq r_1 > r_2 > r_3 \cdots$, so the procedure has to terminate by the well-ordering principle.

Now, d is a greatest common divisor of r_{i-1}, r_i if and only if it is a greatest common divisor of r_i, r_{i+1} (by Lemma 2.2.4). Hence d being a greatest common divisor of a, b is equivalent to it being a greatest common divisor of r_{i-1}, r_i for $1 \leq i \leq k+1$. But as $r_{k+1} = 0$, a greatest common divisor of r_k, r_{k+1} is simply r_k. We conclude that r_k is a greatest common divisor of a, b as well.

Example 2.2.5 Let $a = 966$, $b = 294$. Then we perform the following divisions

$$\begin{aligned} 966 &= 3 \cdot 294 + 84 \Rightarrow r_2 = 84 \\ 294 &= 3 \cdot 84 + 42 \Rightarrow r_3 = 42 \\ 84 &= 2 \cdot 42 + 0 \Rightarrow r_4 = 0, \end{aligned}$$

so that 42 is a greatest common divisor of a, b.

Theorem 2.2.3 also promises us the existence of $s, t \in \mathbb{Z}$ such that $sa + tb = d$. The Euclidean algorithm can be extended to find those s, t as well. Fittingly, the resulting algorithm is called the *extended Euclidean algorithm.*

We consider the sequence r_i as defined before. We also define sequences s_i, t_i with the property that $s_i a + t_i b = r_i$. This is done as follows. Firstly, $s_0 = 1$, $t_0 = 0$ and $s_1 = 0$, $t_1 = 1$. Secondly, suppose that s_{i-1}, t_{i-1}, s_i, t_i are defined such that $s_{i-1}a + t_{i-1}b = r_{i-1}$, $s_i a + t_i b = r_i$. Recall that $r_{i-1} = q_i r_i + r_{i+1}$. Hence

$$r_{i+1} = s_{i-1}a + t_{i-1}b - q_i(s_i a + t_i b) = (s_{i-1} - q_i s_i)a + (t_{i-1} - q_i t_i)b,$$

so we set $s_{i+1} = s_{i-1} - q_i s_i$, $t_{i+1} = t_{i-1} - q_i t_i$. When we found k with $r_{k+1} = 0$ then $d = r_k$ is a greatest common divisor of a, b, and with $s = s_k$, $t = t_k$ we have $d = sa + tb$.

Example 2.2.6 We perform the computation by writing the equations $r_i = s_i a + t_i b$ one above the other. The $(i+1)$-st equation is obtained by subtracting q_i times the i-th equation from the $(i-1)$-st

equation. Again let $a = 966$, $b = 294$. Then

$$\begin{aligned} 966 &= 1 \cdot a + 0 \cdot b \\ 294 &= 0 \cdot a + 1 \cdot b \\ 84 &= 1 \cdot a - 3 \cdot b \\ 42 &= -3 \cdot a + 10 \cdot b. \end{aligned}$$

Remark 2.2.7

- Let $a, b \in \mathbb{Z}$ and d, d' greatest common divisors of a, b. The Definition 2.2.2 immediately implies that $d|d'$ and $d'|d$. As seen in Section 2.2.1, this entails $d' = \pm d$. On the other hand, it is clear that if d is a greatest common divisor of a, b, then so is $-d$. It follows that a greatest common divisor is uniquely determined up to its sign. So by requiring that a greatest common divisor be non-negative, it is uniquely determined. With that convention we also speak about *the* gratest common divisor d of $a, b \in \mathbb{Z}$ and write $d = \gcd(a, b)$.

- Let $a, b \in \mathbb{Z}$ with $\gcd(a, b) = 1$. Then we say that a, b are *coprime*.

- Consider the sequences r_i, q_i as defined in the Euclidean algorithm. Then $r_{i-1} = q_i r_i + r_{i+1} \geq r_i + r_{i+1} \geq 2r_{i+1}$ (note that $q_i \geq 1$ as $r_{i-1} \geq r_i$). So we have $r_2 \leq \frac{1}{2} r_0$, $r_4 \leq \frac{1}{2} r_2 \leq \frac{1}{4} r_0$ and so on. Hence $r_{2j} \leq \frac{1}{2^j} r_0$. Furthermore, if $r_{2j} < 1$, then the algorithm has terminated on or somewhere before the $2j$-th step. So also if $\frac{1}{2^j} r_0 < 1$ then the algorithm has terminated on or somewhere before the $2j$-th step. But this is equivalent to $2^j > r_0 = a$ or to $j > \log_2(a)$. The conclusion is that the algorithm terminates in at most $2 \log_2(a)$ steps. This is expressed by saying that the algorithm has linear complexity. (As $\log_2(a)$ is proportional to the number of digits of a, this means that the number of steps is bounded by a linear function in the number of digits of a.)

2.2.3 Factorization

Here we show that every integer greater than 1 can uniquely be written as a product of primes. In our definition of the concept of prime we do not adhere to the principle that such a definition should carry over to a wider class of rings (as we did when defining the greatest common divisor). The reason for this is that around primes and factorization of integers there is a lot of folklore and standard conventions. Using more generally applicable definitions and terminology would lead us too far away from these.

An integer $p > 1$ is said to be *prime* if its only positive divisors are 1 and p. Examples of primes are: 2, 3, 5, $\ldots$, 290835625986031209641, $\ldots$.

Lemma 2.2.8 (i) *Let $a, b, c \in \mathbb{Z}$ with a, b coprime. If $a|bc$ then $a|c$.*

(ii) *Let $a_1, \ldots, a_n \in \mathbb{Z}$ and $p \in \mathbb{Z}$ a prime. If p divides the product $a_1 \cdots a_n$, then there exists a_i with $p|a_i$.*

Proof. The greatest common divisor of a, b is 1, so there exist $s, t \in \mathbb{Z}$ with $sa + tb = 1$ (Theorem 2.2.3). Multiplying by c we get $csa + ctb = c$. As $a|bc$, there is an $m \in \mathbb{Z}$ with $bc = ma$. Hence $c = csa + tbc = (cs + tm)a$ and a divides c.

For the second part we use induction on n. The case $n = 1$ is obvious. Suppose that $n \geq 1$ and the claim holds for all sets of $n - 1$ integers. If $p|a_1$ then there is nothing to prove. Otherwise p and a_1 are coprime. By Lemma 2.2.8 it follows that p divides $a_2 \cdots a_n$. By induction we now conclude that there is an i with $2 \leq i \leq n$ such that $p|a_i$. □

Theorem 2.2.9 *Let $a \in \mathbb{Z}$, $a > 1$. Then there are primes $p_1, \ldots, p_m$ (not necessarily different) in $\mathbb{Z}$ such that $a = p_1 \cdots p_m$. The primes p_i are uniquely determined up to their order.*

Proof. We use strong induction. The induction hypothesis is that all integers b with $2 \leq b < a$ can be written as a product of a unique sequence of primes. If a itself is prime, then there is nothing to prove. Otherwise, $a = bc$ with $2 \leq b, c < a$. By the induction hypothesis b and c can be written as a product of primes. Hence so can a. Furthermore, if we have two factorizations $a = p_1 \cdots p_m = q_1 \cdots q_r$, where the p_i, q_j are primes, then p_1 divides $q_1 \cdots q_r$. So by second part of the previous lemma there is an i such that $p_1 | q_i$. But as both are prime, it follows that $p_1 = q_i$. Hence $p_2 \cdots p_m = q_1 \cdots q_{i-1} q_{i+1} \cdots q_r$. Again by the induction hypothesis both products contain the same primes, and each prime occurs the same number of times. Therefore the same holds for the original factorizations of a. □

Note that the proof says nothing on how to find the prime factors of a given $a \in \mathbb{Z}$. In fact, this is a very difficult problem, and since the 60's a lot of research has gone into the development of algorithms for factorising integers. Nowadays many algorithms are known, but none of these have a polynomial time complexity (this means that the number of steps taken by the algorithm is bounded by a polynomial function in the number of digits of a).

2.3 Factorization in polynomial rings

2.3.1 Polynomial rings

In analysis a polynomial (say over $\mathbb{R}$) is a function $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ for all $x \in \mathbb{R}$. Algebra focuses on the arithmetic properties of polynomials. So they are not primarily functions, but objects for which we have an addition and a multiplication. But, if polynomials are not functions, what are they? A first answer could be to say that polynomials are expressions of the form $2 + 3x + x^3 - 7x^8$. However, a few things here are not so clear: what is x?, when are two polynomials equal?, what is precisely meant by "expression"? For this reason we will give a formal definition of the term polynomial. In practice we will not work with the formal definition, but just with expressions like the one above. However, we always keep in the back of our minds that by such an expression we mean something slightly different.

Definition 2.3.1 *Let R be a ring. A* polynomial *with coefficients in R is an infinite sequence $(a_0, a_1, \ldots, a_n, \ldots)$ with $a_i \in R$, such that there is an $N \in \mathbb{N}$ with $a_k = 0$ for all $k > N$.*

We will represent a polynomial $(a_0, a_1, \ldots)$ by the formal sum

$$\sum_{n=0}^{\infty} a_n x^n.$$

Here x is just a symbol, called an *indeterminate*. It could also be any other letter. We say that the sum is formal because we do not perform the addition and then obtain something else. For example, if we consider the infinite sequence of integers $(m_0, m_1, \ldots)$ where $m_i = i$ for $0 \leq i \leq 7$ and $m_i = 0$ for $i \geq 8$, then $\sum_{i=0}^{\infty} m_i = 28$. Here we perform the addition a number of times, and get an answer. This is not so with the formal sum: there the summation is just another notation for the infinite sequence.

When we use the indeterminate x, the set of all polynomials with coefficients in R is denoted $R[x]$. It is called the *ring of polynomials in the indeterminate x with coefficients in R.* Now we justify this terminology by introducing an addition and a multiplication on $R[x]$, making it into a ring.

Let $f = \sum_{n=0}^{\infty} a_n x^n$, $g = \sum_{n=0}^{\infty} b_n x^n$ be elements of $R[x]$. Then we define

$$f + g = \sum_{n=0}^{\infty} (a_n + b_n) x^n$$

$$fg = \sum_{n=0}^{\infty} c_n x^n \text{ where } c_n = \sum_{i=0}^{n} a_i b_{n-i}.$$

It is straightforward to see that both $f + g$ and fg are elements of $R[x]$.

Proposition 2.3.2 *With the operations defined above, $R[x]$ is a ring.*

Proof. We need to check the requirements of Definition 2.1.1. The commutativity and associativity of $+$ follows from the respective properties of $+$ in R. The zero is the sequence consisting of only $0 \in R$. Finally, $-f = \sum_{n=0}^{\infty} -a_n x^n$.

The distributive laws are verified by brute force, for example:

$$\begin{aligned}\sum_{n=0}^{\infty} a_n x^n \left(\sum_{n=0}^{\infty} b_n x^n + \sum_{n=0}^{\infty} c_n x^n \right) &= \sum_{n=0}^{\infty} a_n x^n \left(\sum_{n=0}^{\infty} (b_n + c_n) x^n \right) \\ &= \sum_{n=0}^{\infty} \left(\sum_{i=0}^{n} a_i (b_{n-i} + c_{n-i}) \right) x^n \\ &= \sum_{n=0}^{\infty} \left(\sum_{i=0}^{n} a_i b_{n-i} + \sum_{i=0}^{n} a_i c_{n-i} \right) x^n \\ &= \left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{n=0}^{\infty} b_n x^n \right) + \left(\sum_{n=0}^{\infty} a_n x^n \right) \left(\sum_{n=0}^{\infty} c_n x^n \right).\end{aligned}$$

The associativity of the multiplication can also be verified by brute force. An alternative argument is the following. Let f_1, f_2, g_1, g_2, h_1, h_2 be elements of $R[x]$ and suppose that $s(tu) = (st)u$ where s is one of the f_i, t is one of the g_i and u is one of the h_i. Then using the distributive laws we see that

$$(f_1 + f_2)\left((g_1 + g_2)(h_1 + h_2)\right) = \left((f_1 + f_2)(g_1 + g_2)\right)(h_1 + h_2).$$

From this it follows that it suffices to check associativity for f, g, h when $f = a_r x^r$, $g = b_s x^s$, $h = c_t x^t$ (this means that $f = \sum_{n=0}^{\infty} a_n x^n$, where all $a_n = 0$ except that a_r, and similarly for f, g). But for those we have $f(gh) = a_r b_s c_t x^{r+s+t} = (fg)h$. □

We also have that $R[x]$ is commutative if and only if R is commutative. Furthermore, $R[x]$ has a unity if and only if R has one (namely, if the unity of R is 1, then the unity of $R[x]$ is $(1, 0, 0, 0, \ldots)$).

When working with polynomials we do not write them as infinite sequences. We write them in the normal way, and for that we use the following conventions. Let $f = \sum_{n=0}^{\infty} a_n x^n$ and let n_0 be maximal such that $a_{n_0} \neq 0$. Then we write $f = a_0 + a_1 x + \cdots + a_{n_0} x^{n_0}$. Second, if $a_i = 0$ then we omit the term $a_i x^i$. Third, if R has a unity and $a_i = 1$ then we write x^i instead of $1x^i$.

Let $f \in R[x]$ and write $f = a_0 + a_1 x + \cdots + a_{n_0} x^{n_0}$ with $a_{n_0} \neq 0$. If $a_{n_0} = 1$ (of course, only in case R has a unity), then f is said to be *monic*. The integer n_0 is called the *degree* of f, and denoted $\deg(f)$. We will use the convention that $\deg(0) = -\infty$. With this convention we have the following fundamental property.

Lemma 2.3.3 *Let R be a ring without zero divisors and $f, g \in R[x]$. Then $\deg(fg) = \deg(f) + \deg(g)$. Secondly, $R[x]$ has no zero divisors.*

Proof. If one of f, g is 0 then this holds because of our convention (and the rule $-\infty + a = -\infty$ for all $a \in \mathbb{Z}$). So suppose that both are nonzero and write $f = a_0 + \cdots + a_m x^m$, $g = b_0 + \cdots + b_n x^n$ with $m, n \geq 0$ and $a_m, b_n \neq 0$. Then in fg there is only one term of degree $m + n$, namely $a_m b_n x^{m+n}$, and all other terms are of smaller degree. So it cannot cancel against other terms. Furthermore, as R has no zero divisors $a_m b_n \neq 0$. Hence $\deg(fg) = m + n$. In particular, $fg \neq 0$. It follows that $R[x]$ has no zero divisors. □

Corollary 2.3.4 *Let R be a domain. Then $R[x]$ is a domain.*

Remark 2.3.5 Instead of the sequences defined in Definition 2.3.1 we could simply consider infinite sequences without the requirement that the coefficients from a certain point onwards are all zero. Then we can define addition and multiplication in the same way, and again get a ring. This ring is called the ring of *formal power series* in the indeterminate x, and denoted $R[[x]]$.

Remark 2.3.6 We can iterate the definition of polynomial ring, taking $R[x]$ instead of R, and y instead of x. Then we have the ring $R[x][y]$ which is the polynomial ring in two indeterminates x, y. Instead of $R[x][y]$ we write $R[x, y]$. We can repeat this any number of times to obtain the polynomial ring $R[x_1, \ldots, x_n]$ in n indeterminates.

2.3.2 Divisibility

Now we consider polynomials over fields. Let F be a field, and $F[x]$ the polynomial ring over F in the indeterminate x. By Corollary 2.2.1 this is a domain. As seen in Section 2.1 this implies that the calcellation law holds: if $fh = gh$ for certain $f, g, h \in F[x]$, $h \neq 0$ then $f = g$.

Divisibility in $F[x]$ is defined in exactly the same way as in $\mathbb{Z}$ (Section 2.2.1): for $f, g \in F[x]$ we say that f divides g if there is an $h \in F[x]$ with $g = hf$. Notation: $f|g$.

Lemma 2.3.7 *Let $f, g \in F[x]$ with $f|g$ and $g|f$. Then $f = ag$ for a certain $a \in F$.*

Proof. If one of f, g is zero, then the other is zero as well, and we can take $a = 1$ (or any other element of F). If both are non-zero, then there are non-zero $h_1, h_2 \in F[x]$ with $f = h_1 g$, $g = h_2 f$, so that $f = h_1 h_2 f$ and hence $h_1 h_2 = 1$. Then $\deg(h_1) + \deg(h_2) = \deg(1) = 0$ so that $\deg(h_1) = \deg(h_2) = 0$ and $h_1, h_2 \in F$. □

For the next proposition we recall our convention that $\deg(0) = -\infty$.

Proposition 2.3.8 (Division with remainder) *Let $f, g \in F[x]$, $g \neq 0$. Then there are unique $q, r \in F[x]$ with $f = qg + r$ and $\deg(r) < \deg(g)$.*

Proof. We define sequences r_i, q_i in the following way. First of all we set $q_0 = 0$, $r_0 = f$. Then $f = q_0 g + r_0$. Suppose $i \geq 0$ and q_i, r_i are defined with $f = q_i g + r_i$. If $\deg(r_i) < \deg(g)$ we stop and set $q = q_i$, $r = r_i$. If $\deg(r_i) \geq \deg(g)$ then write

$$g = a_0 + a_1 x + \cdots + a_m x^m \text{ with } a_m \neq 0$$
$$r_i = b_0 + b_1 x + \cdots + b_n x^n \text{ with } a_n \neq 0.$$

Then $n \geq m$ and we set $r_{i+1} = r_i - \frac{b_n}{a_m} x^{n-m} g$, $q_{i+1} = q_i + \frac{b_n}{a_m} x^{n-m}$. Then $f = q_{i+1} g + r_{i+1}$ and $\deg(r_{i+1}) < \deg(r_i)$. The latter condition implies that after a finite number of steps the procedure terminates, and we find q, r.

In order to show uniqueness, suppose that $f = qg + r$ and $f = q'g + r'$ with $\deg(r), \deg(r') < \deg(g)$. By subtracting we see that $r' - r = (q - q')g$. If $q \neq q'$ then $\deg((q - q')g) \geq \deg(g) > \deg(r' - r)$ and we get a contradiction. It follows that $q = q'$ and hence also $r = r'$. □

Example 2.3.9 The proof tells us how to find q, r. For example, let $f = x^4 - 3x^3 + x + 1$, $g = x^2 - x - 1$. Then $r_0 = f$, $q_0 = 0$. Since $\deg(r_0) \geq \deg(g)$ we set $r_1 = r_0 - x^2 g = -2x^3 + x^2 + x + 1$, $q_1 = q_0 + x^2 = x^2$. Continuing we get $r_2 = r_1 + 2xg = -x^2 - x + 1$, $q_2 = q_1 - 2x = x^2 - 2x$, $r_3 = r_2 + g = -2x$, $q_3 = q_2 - 1 = x^2 - 2x - 1$. Now $\deg(r_3) < \deg(g)$, so we stop with $q = x^2 - 2x - 1$, $r = -2x$.

2.3.3 Greatest common divisor

The definition of the concept of greatest common divisor of two polynomials is a transcription of the one for integers (Definition 2.2.2). We also have an immediate analogue of Theorem 2.2.3, with an almost identical proof.

Definition 2.3.10 *Let F be a field, and $f, g \in F[x]$. A $d \in F[x]$ is called a* greatest common divisor *of f, g if*

1. *d divides both f and g,*

2. if $h \in F[x]$ divides both f, g then h divides d.

Theorem 2.3.11 *Let F be a field and $f, g \in F[x]$. They have a greatest common divisor d. Moreover, there exist $h_1, h_2 \in F[x]$ such that $h_1 f + h_2 g = d$.*

Proof. If $f = g = 0$ then we can take $d = 0$. Otherwise we set

$$I = \{h_1 f + h_2 g \mid h_1, h_2 \in F[x]\}.$$

Then I contains nonzero polynomials. The set $\{\deg(h) \mid h \in I, h \neq 0\}$ has a minimum $m_0 \geq 0$. Let $d \in I$, $d \neq 0$, be of degree m_0. Write $f = qd + r$ with $\deg(r) < \deg(d)$ (Proposition 2.3.8). As $d \in I$, also $qd \in I$ and since $f \in I$, also $f - qd \in I$. It follows that $r \in I$ and by the choice of d we must have $r = 0$ and hence $d|f$. In the same way it is seen that $d|g$.

Let $h \in F[x]$ be such that $h|f$, $h|g$. Then $f = uh$, $g = vh$ for certain $u, v \in F[x]$. As $d \in I$, there are $h_1, h_2 \in F[x]$ with $d = h_1 f + h_2 g$. But the latter is equal to $(h_1 u + h_2 v)h$. It follows that $h|d$.

As already observed, the last statement follows directly from $d \in I$. □

In order to compute a greatest common divisor of two polynomials we have a Euclidean algorithm that is the analogue of the Euclidean algorithm for integers. First we have an analogue of Lemma 2.2.4. The proof can almost be copied verbatim; therefore we omit it here.

Lemma 2.3.12 *Let $f, g \in F[x]$. Let $q, r \in F[x]$ be such that $f = qg + r$. Then $d \in F[x]$ is a greatest common divisor of f, g if and only if it is a greatest common divisor of g, r.*

The algorithm based on this runs as follows. Set $r_0 = f$, $r_1 = g$. Let $i \geq 1$. Supposing that r_{i-1}, r_i have been defined we do the following. If $r_i = 0$ then we stop, and the result is r_{i-1}. If $r_i \neq 0$ then we perform a division with remainder to find q_i, r_{i+1} with $r_{i-1} = q_i r_i + r_{i+1}$ and $\deg(r_{i+1}) < \deg(r_i)$.

Note that the degrees of the r_i form a strictly decreasing sequence of non-negative integers. So this procedure terminates after a finite number of steps. By Lemma 2.3.12, d is a greatest common divisor of f, g if and only if it is a greatest common divisor of r_{i-1}, r_i for $i \geq 1$. So if $r_i = 0$ then r_{i-1} is a greatest common divisor of f, g.

Example 2.3.13 Let $f = x^5 - x^3 + x - 1$, $g = x^3 - 1$ be elements of $\mathbb{Q}[x]$. So $r_0 = f$, $r_1 = g$. We have $r_0 = q_1 r_1 + r_2$ with $q_1 = x^2 - 1$, $r_2 = x^2 + x + 2$. Continuing, $r_1 = q_2 r_2 + r_3$ with $q_2 = x - 1$, $r_3 = 3x - 3$, and $r_2 = q_3 r_3 + r_4$ with $q_3 = \frac{1}{3}x + \frac{2}{3}$, $r_4 = 0$. It follows that a greatest common divisor of f, g is $3x - 3$.

Also here we have an extended Euclidean algorithm for computing polynomials h_1, h_2 such that $h_1 f + h_2 g$ is the greatest common divisor of f, g. It works exactly in the same way as for integers. Therefore we omit the details.

Remark 2.3.14 Let $f, g \in F[x]$ not both zero. Let d, d' be greatest common divisors of f, g. Then they are non-zero, and by applying Definition 2.3.10 twice we see that $d|d'$ and $d'|d$. By Lemma 2.3.7 this implies that $d' = ad$ for some $a \in F$ and as $d, d' \neq 0$ we have $a \neq 0$. Conversely, if d is a greatest common divisor of f, g then it is clear that ad for $a \in F$, $a \neq 0$ is also a greatest common divisor. We see that a greatest common divisor is uniquely determined up to a non-zero scalar multiple. So if we require that a greatest common divisor be monic, then it is uniquely determined. In that case we write $d = \gcd(f, g)$. So in Example 2.3.13 we have $\gcd(f, g) = x - 1$.

2.3.4 Factorization

In this section we look at the analogue of prime factorization in $\mathbb{Z}$. One confusing thing is that here the relevant polynomials are not called prime but irreducible.

Definition 2.3.15 *Let F be a field. An $f \in F[x]$ is called* irreducible *if $f \notin F$ and for every factorization $f = gh$ with $g, h \in F[x]$ we have $g \in F$, or $h \in F$.*

Example 2.3.16 It is more difficult to give examples of irreducible polynomials than of primes in $\mathbb{Z}$. Polynomials of degree 1 are trivially irreducible. For another example: $x^6 + x^3 + 1$ is irreducible in $\mathbb{Q}[x]$, but this is not immediately obvious.

The proof of the next lemma is extremely similar to the one of Lemma 2.2.8. Therefore we omit it. Subsequently we have an analogue of Theorem 2.2.9. As polynomials are not the same as integers, the formulation is a bit more cumbersome. But the proof follows the same line of argument.

Lemma 2.3.17 (i) *Let $f, g, h \in F[x]$ with $\gcd(f, g) = 1$. If $f|gh$ then $f|h$.*

(ii) *Let $g_1, \ldots, g_n \in F[x]$ and $f \in F[x]$ irreducible. If f divides the product $g_1 \cdots g_n$, then there exists g_i with $f|g_i$.*

Theorem 2.3.18 *Let $f \in F[x]$, $f \notin F$, be a monic polynomial. Then there exist monic irreducible polynomials $g_1, \ldots, g_s \in F[x]$ with $f = g_1 \cdots g_s$. Furthermore, if $h_1, \ldots, h_t \in F[x]$ are monic and irreducible with $f = h_1 \cdots h_t$ then $s = t$ and there are indices $j_1, \ldots, j_s$ such that $\{j_1, \ldots, j_s\} = \{1, \ldots, s\}$ and $g_i = h_{j_i}$.*

Proof. We use strong induction on the degree of f, the induction hypothesis being that the theorem holds for all polynomials of degree k with $1 \leq k < \deg(f)$. First we show the existence of a factorization in irreducibles. If f is irreducible then we take $s = 1$, $g_1 = f$. Otherwise $f = f_1 f_2$ with f_i polynomials of degrees between 1 and $\deg(f) - 1$. Hence by induction the f_i can be written as a product of irreducibles, and therefore so can f.

Now suppose that we have two factorizations as in the theorem. By Lemma 2.3.17(ii) we see that there is a j_1 such that $g_1|h_{j_1}$. Hence there is a $u \in F[x]$ with $h_{j_1} = ug_1$. As h_{j_1} is irreducible we have that $u \in F$ or $g_1 \in F$. But the latter is impossible as g_1 is irreducible. It follows that $u \in F$ and since both g_1 and h_{j_1} are monic it follows that $u = 1$. By cancelling the factor g_1 from both sides we have $g_2 \cdots g_s = h_1 \cdots h_{j_1-1} h_{j_1+1} \cdots h_m$. Now the induction hypothesis immediately finishes the proof. □

As with integers, the question poses itself how to find the factorization into irreducibles of a given polynomial in $F[x]$. Here it seems to be even more difficult. Indeed, while it is straightforward to find the factorization of small integers by hand (for example, after a few trial divisions we find that $403 = 13 \cdot 31$), we have no immediate idea how to find the factorization of small degree polynomials by hand. Consider for example the factorization $x^5 - x^4 + x^3 - x - 2 = (x^2 - x + 2)(x^3 - x - 1)$ in $\mathbb{Q}[x]$. We cannot immitate the obvious method for the integers and divide the polynomial on the left hand side by all irreducible polynomials of degree 2, as there is an infinite number of them. For polynomial factorization a number of algorithms have been developed, and the way they work depends very much on the base field F. For $F = \mathbb{Q}$ there exists a polynomial time algorithm for factorizing polynomials (based on the LLL algorithm, named after its creators, Lenstra, Lenstra, Lovasz). So it seems that, at least from a complexity point of view, factorizing polynomials is easier than factorizing integers.

2.3.5 Roots of polynomials

In Section 2.3.1 we said that polynomials are not functions. However, to a polynomial $f \in F[x]$ we can associate a function $F \to F$, by $a \mapsto f(a)$, where $f(a) = a_0 + a_1 a + \cdots + a_n a^n$ if $f = a_0 + a_1 x + \cdots + a_n x^n$. (Note that we do *not* write $x \mapsto f(x)$ as x is the indeterminate, and not an element of F.)

For a fixed $a \in F$ this yields a map $\varphi_a : F[x] \to F$, $\varphi_a(f) = f(a)$. A first observation is that φ_a respects addition and multiplication, that is, $\varphi_a(f + g) = \varphi_a(f) + \varphi_a(g)$, $\varphi_a(fg) = \varphi_a(f)\varphi_a(g)$. This is immediate from the definition of addition and multiplication of polynomials.

An $a \in F$ is said to be a *root* of f if $f(a) = 0$. We can connect the existence of a root of f to a certain type of divisor of f.

Proposition 2.3.19 *Let $f \in F[x]$. Then $a \in F$ is a root of f if and only if $x - a$ divides f.*

Proof. If $x - a$ divides f then $f = (x - a)g$ and $\varphi_a(f) = \varphi_a(x - a)\varphi_a(g) = 0 \cdot g(a) = 0$. Conversely, if $f(a) = 0$ then we perform a division with remainder to obtain $q, r \in F[x]$ with $f = q(x - a) + r$ with

$\deg(r) < 1$, that is, $r \in F$. Then $0 = \varphi_a(f) = \varphi_a(q) \cdot 0 + \varphi_a(r) = r$. It follows that $x - a$ divides f. □

An immediate consequence is that polynomials of degree 2 or 3 in $F[x]$ without roots in F are irreducible. Indeed, if such a polynomial is not irreducible, then it has a factor of degree 1.

2.4 Factorization in domains

In his attempt to prove Fermat's last theorem for $n = 3$, Euler formulated the following statement. Let p, q be coprime integers with $p^2 + 3q^2 = x^3$ for some $x \in \mathbb{Z}$. Then there exist $a, b \in \mathbb{Z}$ with $p = a^3 - 9ab^2$, $q = 3a^2b - 3b^3$. (L. Euler, *Vollständige Anleitung zur Algebra*, St Petersburg, 1770.)

One way of trying to prove it, found in his work[1], runs as follows. Consider complex numbers of the form $u + v\sqrt{-3}$, with $u, v \in \mathbb{Z}$. We have $p^2 + 3q^2 = (p + q\sqrt{-3})(p - q\sqrt{-3})$. As p, q are coprime, the same holds for $p+q\sqrt{-3}$ and $p-q\sqrt{-3}$. So both of the latter are cubes. Hence there exist $a, b \in \mathbb{Z}$ with $p + q\sqrt{-3} = (a + b\sqrt{-3})^3$. Now

$$(a + b\sqrt{-3})^3 = a^3 - 9ab^2 + (3a^2b - 3b^3)\sqrt{-3},$$

from which the statement follows.

One step in this argument is: if AB is a cube and A, B are coprime then both A and B are cubes. However, in order that this step be valid we need that the numbers of the form $u+v\sqrt{-3}$ allow unique factorization. But they do not.

Of course, at this point it is not clear at all what is meant by unique factorization of numbers of the form $u + v\sqrt{-3}$. Secondly, it is not clear why they do not have that property.

After Euler the next person to achieve progress on proving Fermat's last theorem was Sophie Germain (1776-1831). She considered proving the theorem for prime exponent p (it is easy to see that this is enough) and divided the proof into two cases. The first case occurs when p does not divide x, y, z. The second case carries the assumption that p divides exactly one of x, y, z. Germain proved a theorem which establishes the first case for all primes < 100.

In 1847 the French mathematician Lamé, considering Fermat's last theorem for general odd prime exponent n, had the idea to write

$$x^n + y^n = (x + y)(x + \zeta y)(x + \zeta^2 y) \cdots (x + \zeta^{n-1} y), \tag{2.4.1}$$

where $\zeta = e^{\frac{2\pi i}{n}}$. (To see this, note that $T^n - 1 = (T - 1)(T - \zeta) \cdots (T - \zeta^{n-1})$, set $T = -x/y$ and multiply by $-y^n$.) If the factors $x + y, \ldots, x + \zeta^{n-1}y$ are relatively prime, then Lamé's idea was to show that $x^n + y^n = z^n$ implies that each of these factors is an n-th power, and derive a contradiction from that. However, nothing came of this. Firstly, Liouville pointed out that for this conclusion to be justified, the property of unique factorization was necessary. Secondly, Kummer had a few years earlier already shown that the unique factorization property does not hold for the kind of numbers considered by Lamé. (See [Edw77], §4.1, for a very interesting historical overview.)

In this section we look at factorization in domains in general, using a class of examples that includes the numbers $a + b\sqrt{-3}$ considered by Euler. It is customary to restrict to domains when looking at properties of unique factorization. However, it is possible to also include rings with zero divisors; here we do not go into that, but refer to [Gal78].

2.4.1 A class of examples

Here we describe a class of rings that will serve as the main source of examples in this section. We start with a polynomial $f = x^2 + a_1x + a_0$ in $\mathbb{Q}[x]$ with $a_1, a_0 \in \mathbb{Z}$. *We assume that f is irreducible.* In Section 2.3.5 we have seen that this is equivalent to f not having roots in $\mathbb{Q}$. Let $\alpha \in \mathbb{C}$ be a root

[1] Although, note that Edwards ([Edw77], p. 43) writes: "Euler very seriously confuses necessary and sufficient conditions in this part of his Algebra [E9] and it is very difficult to determine what he meant to say."

of f, and denote the second root of f by $\bar{\alpha}$ (in fact, if $\alpha \notin \mathbb{R}$, then $\bar{\alpha}$ is the complex conjugate of α). Then as $f = (x-\alpha)(x-\bar{\alpha})$ we have

$$\bar{\alpha} = -a_1 - \alpha \text{ and } \alpha\bar{\alpha} = a_0. \tag{2.4.2}$$

Set

$$\mathbb{Z}[\alpha] = \{a + b\alpha \mid a, b \in \mathbb{Z}\}.$$

Then $\mathbb{Z}[\alpha]$ is a subset of $\mathbb{C}$. For $a, b, c, d \in \mathbb{Z}$ we have

- $a + b\alpha = 0$ if and only if $a = b = 0$. Indeed, $a + b\alpha = 0$ with $b \neq 0$ implies $\alpha = -\frac{a}{b}$ and by Proposition 2.3.19, $x + \frac{a}{b}$ divides f. But because f is irreducible, this is not possible.
- $a + b\alpha = c + d\alpha$ if and only if $a = c$ and $b = d$. This follows immediately from the previous point.
- $(a + b\alpha) + (c + d\alpha) = (a + c) + (b + d)\alpha$, showing that $\mathbb{Z}[\alpha]$ is closed under summation.
- As $f(\alpha) = 0$ we have $\alpha^2 = -a_1\alpha - a_0$. Hence $(a + b\alpha)(c + d\alpha) = ac + (ad + bc)\alpha + bd\alpha^2 = (ac - a_0bd) + (ad + bc - a_1bd)\alpha$, so that $\mathbb{Z}[\alpha]$ is also closed under multiplication. (Note that here we need that the coefficients of f lie in $\mathbb{Z}$.)

It follows that with the addition and multiplication from $\mathbb{C}$, the set $\mathbb{Z}[\alpha]$ becomes a commutative ring with unity (which is $1 = 1 + 0\alpha$). Indeed, all the properties that we need for the operations in $\mathbb{Z}[\alpha]$ follow at once because they hold in $\mathbb{C}$. Furthermore, since $\mathbb{Z}[\alpha] \subset \mathbb{C}$, which is a field, it has no zero divisors. It follows that $\mathbb{Z}[\alpha]$ is a domain.

Example 2.4.1 Let $f = x^2 + 3$. Then f has no roots in $\mathbb{Q}$ and hence is irreducible. We take $\alpha = \sqrt{-3} = i\sqrt{3}$, so that $\bar{\alpha} = -\sqrt{-3}$. In this case $\mathbb{Z}[\sqrt{-3}] = \{a + b\sqrt{-3} \mid a, b \in \mathbb{Z}\}$ is the ring that played a role in Euler's attempt to prove Fermat's last theorem for $n = 3$.

For working with the rings $\mathbb{Z}[\alpha]$ it is very useful to consider the function

$$N : \mathbb{Z}[\alpha] \to \mathbb{Z}_{\geq 0}, \quad N(a + b\alpha) = |(a + b\alpha)(a + b\bar{\alpha})| = |a^2 - a_1ab + a_0b^2|,$$

which is called the *norm* of $\mathbb{Z}[\alpha]$. (Note that the last equality shows that the image of N is contained in $\mathbb{Z}_{\geq 0}$.) It has the following properties:

- $N(\eta) = 0$ if and only if $\eta = 0$,
- $N(\eta\xi) = N(\eta)N(\xi)$ for $\eta, \xi \in \mathbb{Z}[\alpha]$,
- $N(\eta) = 1$ if and only if η is invertible in $\mathbb{Z}[\alpha]$.

Proof. Throughout we use the fact that $\bar{\alpha} \in \mathbb{Z}[\alpha]$, which is immediate from (2.4.2).

Write $\eta = a + b\alpha$. If $N(\eta) = 0$ then $(a + b\alpha)(a + b\bar{\alpha}) = 0$. Since $\mathbb{Z}[\alpha]$ has no zero divisors it follows that $a + b\alpha = 0$ or $a + b\bar{\alpha} = 0$. In both cases $a = b = 0$ and $\eta = 0$.

For the second point we can write $\eta = a + b\alpha$, $\xi = c + d\alpha$ and verify the identity by brute force. More elegantly, we can define the map $\sigma : \mathbb{Z}[\alpha] \to \mathbb{Z}[\alpha]$ by $\sigma(a + b\alpha) = a + b\bar{\alpha}$. Then σ respects the multiplication: $\sigma(\eta\xi) = \sigma(\eta)\sigma(\xi)$. This is obvious, because for multiplying η, ξ we just use the equality $\alpha^2 = -a_1\alpha - a_0$; and $\bar{\alpha}$ satisfies the same equality. Then $N(\eta\xi) = |\eta\xi\sigma(\eta\xi)| = |\eta\xi\sigma(\eta)\sigma(\xi)| = |\eta\sigma(\eta)||\xi\sigma(\xi)| = N(\eta)N(\xi)$.

If η is invertible, then there is a $\xi \in \mathbb{Z}[\alpha]$ with $\eta\xi = 1$. So using the second point, $N(\eta)N(\xi) = 1$. Since both lie in $\mathbb{Z}_{\geq 0}$, both have to be 1. Conversely, write $\eta = a + b\alpha$. If $N(\eta) = 1$ then $(a + b\alpha)(a + b\bar{\alpha}) = \pm 1$. If it is 1, then $a + b\bar{\alpha}$ is the inverse of η. If it is -1, then $-a - b\bar{\alpha}$ is the inverse of η. We conclude that η is invertible. □

Example 2.4.2 Consider $\mathbb{Z}[\sqrt{-3}]$ as in Example 2.4.1. Then $N(a+b\sqrt{-3}) = |a^2+3b^2| = a^2+3b^2$. The invertible elements of $\mathbb{Z}[\alpha]$ are those with norm equal to 1. It immediately follows that these are ± 1.

Now let $f = x^2-2$, and $\alpha = \sqrt{2}$. Here $N(a+b\sqrt{2}) = |a^2-2b^2|$. Hence $a+b\sqrt{2}$ is invertible if and only if $a^2-2b^2 = \pm 1$. It follows that $1+\sqrt{2}$ is invertible. It can be shown that all invertible elements of $\mathbb{Z}[\sqrt{2}]$ are of the form $\pm(1+\sqrt{2})^m$ for $m \in \mathbb{Z}$. So in this case the ring has many invertible elements.

More in general we can take $k > 0$, not a square, and consider $f = x^2-k$. Then $a+b\sqrt{k} \in \mathbb{Z}[\sqrt{k}]$ is invertible if and only if $a^2-kb^2 = \pm 1$. This is known as Pell's equation, which has a long history in its own right. It can be shown that there is a fundamental solution $\eta = a_0 + b_0\sqrt{k}$ and that all other invertible elements are of the form $\pm\eta^m$ for $m \in \mathbb{Z}$. But this solution is not necessarily easy to find. For example, when $k = 109$ then a fundamental solution is $8890182 + 851525\sqrt{109}$.

Remark 2.4.3 Set $\mathbb{Q}(\alpha) = \{a+b\alpha \mid a,b \in \mathbb{Q}\}$. (Here we can also write $\mathbb{Q}[\alpha]$, but we use the other parentheses to emphasize that it is a field.) Then in the same way as for $\mathbb{Z}[\alpha]$ it is seen that $\mathbb{Q}(\alpha)$ is a domain. It contains $\mathbb{Z}[\alpha]$. Let $a+b\alpha \in \mathbb{Q}(\alpha)$ be non-zero. Using (2.4.2) we see that $\bar{\alpha} \in \mathbb{Q}(\alpha)$ and $(a+b\alpha)(a+b\bar{\alpha}) = a^2 - aba_1 + b^2a_0$ lies in $\mathbb{Q}$. It is not zero as $\mathbb{Q}(\alpha)$ (being a subset of $\mathbb{C}$) has no zero divisors. It follows that

$$\frac{1}{a^2 - aba_1 + b^2a_0}(a+b\bar{\alpha})$$

is the inverse of $a+b\alpha$. The conclusion is that $\mathbb{Q}(\alpha)$ is a field containing $\mathbb{Z}[\alpha]$.

In the same way as for $\mathbb{Z}[\alpha]$ we define the norm of $\mathbb{Q}(\alpha)$, i.e., $N(a+b\alpha) = |(a+b\alpha)(a+b\bar{\alpha})| = |a^2 - a_1ab + a_0b^2|$ for $a,b \in \mathbb{Q}$. In the same way as for $\mathbb{Z}[\alpha]$ we see that for $\eta, \xi \in \mathbb{Q}(\alpha)$ we have $N(\eta) = 0$ if and only if $\eta = 0$ and $N(\eta\xi) = N(\eta)N(\xi)$. Furthermore, for $\eta \in \mathbb{Q}(\alpha)$, $\eta \neq 0$ we see that

$$N(\eta)N(\eta^{-1}) = N(\eta\eta^{-1}) = N(1) = 1$$

so that $N(\eta^{-1}) = \frac{1}{N(\eta)}$.

2.4.2 Some terminology

In the previous sections we have seen two factorization theorems (Theorems 2.2.9, 2.3.18). We want to see in which rings similar theorems hold. For that we first need to make the terminology a bit clearer and more general. The basic building blocks for factorizations are called prime numbers for the integers and irreducible polynomials for the polynomials. Both adjectives (prime and irreducible) are used in this more general context, but they are made to denote somewhat different properties. Secondly, in the theorem for polynomials, the uniqueness part is obtained by only considering monic polynomials. However, that is a bit unsatisfactory: it is obvious that the factorizations $6x^2-18x+12 = (2x-2)(3x-6) = (3x-3)(2x-4)$ are essentially the same, as the factors in the second factorization are scalar multiples of the factors in the first factorization. So we need to state more clearly what is meant by "unique factorization".

Let D be a domain (that is, a commutative ring with unity $1 \neq 0$, without zero divisors). First of all, as for the integers and polynomials we say for $a,b \in D$ that a *divides* b if there is a $c \in D$ with $b = ca$. Notation: $a|b$.

Definition 2.4.4 *Let $a \in D$. Then*

- *a is* irreducible *if $a \neq 0$, a is not invertible, and if $a = bc$ with $b,c \in D$ then b or c is invertible.*
- *a is* prime *if $a \neq 0$ a is not invertible, and if $a|bc$ with $b,c \in D$, then $a|b$ or $a|c$.*

It is straightforward to see that a prime element is also irreducible. (Indeed, suppose $a \in D$ is prime and $a = bc$ for certain $b,c \in A$. Then $a|bc$ and, as a is prime, $a|b$ or $a|c$. Suppose that $a|b$; then $b = da$ for a certain $d \in D$. But then $a = bc = adc$, and because in D the cancellation law holds (see Section 2.1), we have $dc = 1$ and c is invertible. In the same way $a|c$ implies that b is invertible.) The converse does not always hold; we will see many examples of this, for a first example see Example 2.4.7 below.

Example 2.4.5 We see that the irreducible elements in $\mathbb{Z}$ are exactly $\pm p$, where p is a prime according to the definition given in Section 2.2.3. In $\mathbb{Z}$ we have that an irreducible is also prime (we leave this as an exercise).

Example 2.4.6 Let F be a field. The irreducibles in $F[x]$ are exactly the irreducible polynomials as defined in Definition 2.3.15. In this ring an irreducible is also prime (and again we leave the verification as an exercise).

Example 2.4.7 Let F be a field and $A = \{a_0 + a_2x^2 + \cdots + a_nx^n \mid n \geq 0\} \subset F[x]$ (so A consists of the polynomials with zero linear term). The sum and product of two elements of A is again in A. So because $F[x]$ is a domain, A is a domain as well. We note that the element $x^2 \in A$ is irreducible. (Indeed, if $fg = x^2$ for certain $f, g \in A$, then $\deg(f) + \deg(g) = 2$ (Lemma 2.3.3) and we see that f or g has to be in F as A does not have elements of degree 1.) But x^2 is not prime: $x^2|x^6$ but $x^6 = x^3 \cdot x^3$ and $x^2 \not| x^3$!

Definition 2.4.8 *Two elements $a, b \in D$ are said to be* associated *if there is an invertible element $\epsilon \in D$ with $a = \epsilon b$. In this case a is called an* associate *of b.*

Example 2.4.9 In $\mathbb{Z}$ the invertible elements are ± 1, so the associates of $a \in \mathbb{Z}$ are $\pm a$. The invertible elements of $F[x]$ are the elements of $F^* = \{a \in F \mid a \neq 0\}$. Hence the associates of $f \in F[x]$ are af with $a \in F^*$.

Next we have to decide what to use as the "atoms" for our factorizations (we have to choose between irreducibles and primes). It appears that it is best to choose the irreducibles. Then we argue as follows for an $a \in D$. If $a = 0$ or if a is invertible, then it makes not much sense to look for a unique factorization of a, so suppose that $a \neq 0$ and that a is not invertible. If a cannot be written as a product of two non-invertible elements, then a is irreducible. Otherwise $a = bc$, where both b and c are non-invertible. Now we continue by finding factorizations of b and c. Looking at the proofs of Theorems 2.2.9, 2.3.18 we see that this is how the argument goes. In general it is not always guaranteed that this process always terminates (it can happen that we write a as products of more and more non-invertibles, without end). However, in many domains it does terminate and every (non-zero, non-invertible) element can be written as a product of irreducibles. The main problem is the uniqueness of the factorization. In order to express that we have two definitions.

Definition 2.4.10 *Let D be a domain in which every non-zero and non-invertible element can be written as a product of irreducible elements. Then we say that the factorization in D is* essentially unique *if for each $a \in A$ with $a \neq 0$ and a non-invertible we have the following. If $a = p_1 \cdots p_s$, $a = q_1 \cdots q_t$ are two ways of writing a as product of irreducible elements p_i, q_j, then $s = t$ and possibly after having changed the order of the q_j, we have that p_i, q_i are associates for $1 \leq i \leq s$.*

Definition 2.4.11 *Let D be a domain. If every non-zero and non-invertible element of D can be written as an essentially unique product of irreducible elements, then we say that D is a* unique factorization domain.

2.4.3 A criterion for unique factorization

From the discussion at the end of the previous section two questions emerge on a given domain D:

1. Is factorization possible in D? (That is, can every $a \in D$, $a \neq 0$, a not invertible, be written as a product of irreducibles?)
2. Is factorization unique, in the sense of Definition 2.4.10?

Here we give a sufficient criterion for the first question, and a necessary and sufficient criterion for the second.

Proposition 2.4.12 *Let D be a domain, and suppose there is a function $\nu : D \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ such that $\nu(ab) > \nu(b)$ for all non-zero and non-invertible $a \in D$ and all non-zero $b \in D$. Then every non-zero and non-invertible element of D is a product of irreducible elements of D.*

Proof. Let $S \subset D$ be the set of all $a \in D$ that are non-zero, non-invertible, and cannot be written as a product of irreducible elements. Suppose that S is not empty. Then $\nu(S) = \{\nu(a) \mid a \in S\}$ is a non-empty subset of $\mathbb{Z}_{\geq 0}$ and hence has a minimal element n_0. Let $a_0 \in S$ be such that $\nu(a_0) = n_0$. As $a_0 \in S$ it cannot be irreducible. So there are $b, c \in D$, both non-invertible, with $a_0 = bc$. Then $\nu(a_0) = \nu(bc)$ which is bigger than both $\nu(b)$ and $\nu(c)$. Hence b, c do not lie in S and can therefore be written as a product of irreducibles. Hence so can a_0, and we have a contradiction to our assumption that $S \neq \emptyset$. □

Remark 2.4.13 This proof is actually a proof by induction, disguised as a proof based on the well-ordering principle. See Section 1.2.

Corollary 2.4.14 *Let $\mathbb{Z}[\alpha]$ be as in Section 2.4.1. Every non-zero and non-invertible element of $\mathbb{Z}[\alpha]$ can be written as a product of irreducible elements.*

Proof. Use the previous proposition with the norm N as the function ν. The properties of N shown in Section 2.4.1 immediately imply that ν has the property required in Proposition 2.4.12. □

Theorem 2.4.15 *Let D be a domain in which every non-zero and non-invertible element can be written as a product of irreducible elements. Then D is a unique factorization domain if and only if every irreducible element of D is prime.*

Remark 2.4.16 In Section 2.4.2 we have seen that every prime is also irreducible. Hence the factorization is essentially unique if and only if the sets of irreducibles and primes in D coincide.

Proof. Suppose that D is a unique factorization domain. Let $a \in D$ be irreducible. We have to show that it is prime. Suppose that $a|bc$ for certain $b, c \in D$. Then $bc = da$ for a certain $d \in D$. Let $b = b_1 \cdots b_m$, $c = c_1 \cdots c_n$, $d = d_1 \cdots d_r$ be the factorizations of b, c, and d as products of irreducibles. Then

$$b_1 \cdots b_m c_1 \cdots c_n = d_1 \cdots d_r a$$

is an equality between two products of irreducibles. By Definition 2.4.10 a is associated with a b_i or with a c_j. Suppose that a is associated with b_i. Then $b_i = ua$, where u is invertible. But then $a|b$. In the same way we see that when a and c_j are associated, then $a|c$. The conclusion is that a is prime.

Suppose that every irreducible is prime. Let $a \in D$ and let $a = p_1 \cdots p_m = q_1 \cdots q_n$ be two factorizations in irreducibles. We may assume that $n \geq m$. Because p_1 is prime, it divides one of the q_i. After reordering we may assume that $i = 1$. Hence $q_1 = u_1 p_1$. But as q_1 is irreducible this implies that u_1 is invertible (because p_1 is not invertible, as it is irreducible). We substitute and cancel the factor p_1 from both sides and obtain $p_2 \cdots p_m = u_1 q_2 \cdots q_n$. Note that p_2 cannot divide u_1 as u_1 is invertible. Hence we can continue in the same way, and after m steps we obtain the equality $1 = u_1 \cdots u_m q_{m+1} \cdots q_n$. If $n > m$ then $q_n|1$, implying that q_n is invertible. That is not the case, so that $n = m$ and $q_i = u_i p_i$, which means that p_i and q_i are associated. □

Example 2.4.17 Consider $\mathbb{Z}[\sqrt{-3}]$ as in Examples 2.4.1, 2.4.2. We recall that $N(a + b\sqrt{-3}) = a^2 + 3b^2$. We have the following equality

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

We show that $1 + \sqrt{-3}$ is irreducible. Suppose that $1 + \sqrt{-3} = \eta\xi$ for certain $\eta, \xi \in \mathbb{Z}[\sqrt{-3}]$. Then $4 = N(1 + \sqrt{-3}) = N(\eta)N(\xi)$. Note that there are no elements in $\mathbb{Z}[\sqrt{-3}]$ of norm 2. Hence we may assume that $N(\eta) = 1$, which implies that η is invertible. (See the properties of the norm in Section

2.4.1). It follows that $1+\sqrt{-3}$ is irreducible. In exactly the same way we see that $1-\sqrt{-3}$ and 2 are irreducible. The invertible elements of $\mathbb{Z}[\sqrt{-3}]$ are ± 1. So $1+\sqrt{-3}$ is not associated with 2. We see that the above equality presents two factorizations of the same element that are essentially different. So the factorization in $\mathbb{Z}[\sqrt{-3}]$ is not essentially unique.

A slight variation is to say that $1+\sqrt{-3}$ divides $2 \cdot 2$, but it does not divide 2 (indeed, if $2 = \eta(1+\sqrt{-3})$ then $N(\eta) = 1$ and η is invertible, implying $\eta = \pm 1$, a contradiction). Hence $1+\sqrt{-3}$ is an irreducible which is not prime. Again it follows that the factorization in $\mathbb{Z}[\sqrt{-3}]$ is not essentially unique.

2.4.4 Euclidean domains

To investigate divisibility and factorization in the integers and in polynomial rings, we made good use of division with remainder. This procedure can be copied for a wider class of rings, which are called Euclidean (because there is a Euclidean algorithm for finding a greatest common divisor).

Definition 2.4.18 *A domain D is said to be* Euclidean *if there is a function $\nu : D \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ with*

1. *$\nu(ab) \geq \nu(b)$ for all non-zero $a, b \in D$,*
2. *for all $a, b \in D$ with $b \neq 0$ there are $q, r \in D$ with $a = qb + r$ and $r = 0$ or $\nu(r) < \nu(b)$.*

Remark 2.4.19 Let D be a Euclidean domain, and $a, b \in D$ non-zero. Suppose further that a is not invertible. Let $q, r \in D$ be such that $b = qab + r$ with $r = 0$ or $\nu(r) < \nu(ab)$. If $r = 0$ then since the cancellation law holds in D, we have $1 = qa$ and a is invertible, contrary to our hypothesis. So $\nu(ab) > \nu(r) = \nu((1-qa)b) \geq \nu(b)$. Now we see that Proposition 2.4.12 implies that every non-zero and non-invertible element of D can be written as a product of irreducibles.

Remark 2.4.20 Let D be a Euclidean domain, and ν as in Definition 2.4.18. Let $a \in D$ be non-zero. Then $\nu(a) = \nu(a \cdot 1) \geq \nu(1)$. So $\nu(1)$ is the minimal value attained by ν. Now suppose that a is non-invertible. Then if we take $b = 1$ in Remark 2.4.19 we see that $\nu(a) = \nu(a \cdot 1) > \nu(1)$. Conversely, suppose that a is invertible and let $b \in D$ be such that $ba = 1$. Then $\nu(1) = \nu(ba) \geq \nu(a)$. But $\nu(1)$ is the minimal vallue of ν and therefore $\nu(a) = \nu(1)$. The conclusion is that $\nu(1)$ is the minimal value of ν and $\nu(a) = \nu(1)$ if and only if a is invertible.

Remark 2.4.21 Let $D = \mathbb{Z}[\alpha]$ as in Section 2.4.1. From the properties of the norm N we see that with $\nu = N$ we automatically have the first part of the requirements for ν in Definition 2.4.18. The problem is to decide whether or not the second part holds. If it does, then we say that $\mathbb{Z}[\alpha]$ is *norm-Euclidean*. It is possible that a $\mathbb{Z}[\alpha]$ is Euclidean, but not norm-Euclidean. For an example, consider $f = x^2 - x - 17$, with root $\alpha = \frac{1+\sqrt{69}}{2}$. We have that $\mathbb{Z}[\frac{1+\sqrt{69}}{2}]$ is Euclidean, but not norm-Euclidean (this is rather non-trivial; for the computer assisted proof see [Cla94]).

Example 2.4.22

1. $D = \mathbb{Z}$ with $\nu(a) = |a|$.
2. Let F be a field and $D = F[x]$. Then D is Euclidean with $\nu(f) = \deg(f)$.
3. As usual we write $i = \sqrt{-1}$, and consider $D = \mathbb{Z}[i]$. We have $N(a+bi) = a^2 + b^2$. We show that $\mathbb{Z}[i]$ is norm-Euclidean. According to Remark 2.4.21, we need to show that for given $\eta, \xi \in \mathbb{Z}[i]$, $\xi \neq 0$, there are $q, r \in \mathbb{Z}[i]$ with $\eta = q\xi + r$ and $r = 0$ or $N(r) < N(\xi)$. For that we work in the field $\mathbb{Q}(i)$ (see Remark 2.4.3; note that the norm extends to this field). In this field ξ has an inverse, so that $\eta = q\xi + r$ is equivalent to $\eta\xi^{-1} = q + r\xi^{-1}$. Here $\eta\xi^{-1}$, $r\xi^{-1}$ lie in $\mathbb{Q}(i)$ and $q \in \mathbb{Z}[i]$. The idea now is to choose a q such that $r\xi^{-1}$ has coefficients that are as small as possible, and hopefully r will have small norm. More precisely, write $\eta\xi^{-1} = u + vi$ with $u, v \in \mathbb{Q}$. Let $u_0, v_0 \in \mathbb{Z}$ be such that $u = u_0 + x_0$, $v = v_0 + y_0$ with $|x_0|, |y_0| \leq \frac{1}{2}$. Set $q = u_0 + v_0 i$ and $r = \xi(x_0 + y_0 i)$. Then $\eta\xi^{-1} = q + x_0 + y_0 i$ so that $\eta = q\xi + r$ and therefore also $r \in \mathbb{Z}[i]$. Furthermore, $N(r) = N(\xi)(x_0^2 + y_0^2) \leq N(\xi)(\frac{1}{2}^2 + \frac{1}{2}^2) = \frac{1}{2}N(\xi) < N(\xi)$. We conclude that $\mathbb{Z}[i]$ is norm-Euclidean.

Now we give an example of the division with remainder. Let $\eta = 2+3i$, $\xi = 1+i$. Note that the inverse of $a+bi$ is $\frac{1}{a^2+b^2}(a-bi)$, so $\xi^{-1} = \frac{1}{2}(1-i)$. We have $\eta\xi^{-1} = \frac{5}{2} + \frac{1}{2}i$. We can take $q = 2$ and $r = \xi(\frac{1}{2} + \frac{1}{2}i) = i$. Indeed, $2+3i = 2(1+i)+i$ and $N(i) = 1 < N(1+i) = 2$. Note also that here the choice of q is not unique. We could also have taken $q = 3+i$ and $r = \xi(-\frac{1}{2} - \frac{1}{2}i) = -i$, and get the equality $2+3i = (3+i)(1+i) - i$. However, note also that the number of q's that can be taken is always finite.

Many things that we have seen for the integers and polynomial rings can be copied for Euclidean rings. The first of these is the greatest common divisor and the Euclidean algorithm. We immediately have the analogues of Definition 2.2.2, Theorem 2.2.3 and Lemma 2.2.4.

Definition 2.4.23 *Let D be a Euclidean domain. Let $a, b \in D$. An element d of D is called a* greatest common divisor *of a, b if*

1. *d divides a and b,*
2. *if $c \in D$ divides a and b then c divides d.*

Theorem 2.4.24 *Let D be a Euclidean domain. Let $a, b \in D$. They have a greatest common divisor d. Moreover, there exist $s, t \in D$ such that $sa + tb = d$.*

Proof. Let $\nu : D \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ be as in Definition 2.4.18.

If $a = b = 0$ then we take $d = 0$ and $s = t = 0$. Otherwise we set

$$I = \{xa + yb \mid x, y \in D\}.$$

Then I contains nonzero elements. Let $M = \{\nu(c) \mid c \in I, c \neq 0\}$. Then M has a minimal element, m_0, and let $d \in I$ be such that $d \neq 0$ and $\nu(d) = m_0$. There are $q, r \in D$ with $a = qd + r$ and $r = 0$ or $\nu(r) < \nu(d)$. Now $a, d \in I$ and hence also $r = a - qd$ lies in I. By the choice of d it follows that $r = 0$ and we see that d divides a. In the same way it is shown that d divides b.

Let $c \in D$ be such that c divides both a and b. Then $a = a_0c$, $b = b_0c$ for certain $a_0, b_0 \in D$. As $d \in I$ there are $s, t \in D$ with $d = sa + tb$, so that $d = (sa_0 + tb_0)c$ and we see that c divides d.

The last statement follows immediately from $d \in I$. □

We also have a *Euclidean algorithm* for finding a greatest common divisor of $a, b \in D$. It is based on the following lemma. The proof of Lemma 2.2.4 can be copied almost verbatim, so we omit it here.

Lemma 2.4.25 *Let D be a Euclidean domain. Let $a, b \in D$. Let $q, r \in D$ be such that $a = qb + r$. Then $d \in D$ is a greatest common divisor of a, b if and only if it is a greatest common divisor of b, r.*

The algorithm for finding the greatest common divisor is now the exact analogue of the algorithms for integers and for polynomials. Here we get a sequence $r_0, r_1, \ldots$ with $\nu(r_0) > \nu(r_1) > \cdots$ so again the algorithm has to terminate. We do not describe the details of this algorithm, but rather illustrate it with an example.

Example 2.4.26 Let $D = \mathbb{Z}[i]$. In Example 2.4.22 we have seen that it is Euclidean with respect to the norm N (and $N(a+bi) = a^2 + b^2$). Let $\zeta = 3+i$, $\eta = 2$. We want to find a greatest common divisor of ζ, η. We set $r_0 = \zeta$, $r_1 = \eta$. Next we compute q_1, r_2 with $r_0 = q_1r_1 + r_2$ and $N(r_2) < N(r_1)$. We have $r_0r_1^{-1} = \frac{3}{2} + \frac{1}{2}i$ so we can take $q_1 = 1$, $r_2 = r_1(\frac{1}{2} + \frac{1}{2}i) = 1+i$. Next we compute q_2, r_3 such that $r_1 = q_2r_2 + r_3$ and $N(r_3) < N(r_2)$. We have $r_1r_2^{-1} = 1-i$, so that $q_2 = 1-i$ and $r_3 = 0$. So now we stop, and a greatest common divisor is the last non-zero remainder that we obtained, that is $1+i$.

Remark 2.4.27 There is also the question as to how unique a greatest common divisor is. Let D be a Euclidean domain, $a, b \in D$ and d, d' greatest common divisors of a, b. We may assume that a, b are not both zero, as otherwise $d = d' = 0$. By applying Definition 2.4.23 twice we find $d|d'$ and $d'|d$. Hence $d' = ud$, $d = vd'$ for certain $u, v \in D$. Therefore $d = vd' = uvd$ and since in D the cancellation law holds, it follows that $uv = 1$ and u, v are invertible. Conversely, it is straightfoward to see that if d

is a greatest common divisor of a, b and $u \in D$ is invertible, then ud is also a greatest common divisor of a, b. We conclude that a greatest common divisor is unique up to multiplication by an invertible element. In other words, if d is a greatest common divisor of a, b then the set of greatest common divisors of a, b consists of all associates of d.

Next we look at factorization into irreducibles. In Remark 2.4.19 we have seen that in a Euclidean domain a factorization into irreducibles is possible for all (non-zero, non-invertible) elements. By the next theorem, we also have that the factorization is essentially unique. In other words, a Euclidean domain is a unique factorization domain.

Theorem 2.4.28 *Let D be a Euclidean domain. Then the factorization of non-zero and non-invertible elements of D as products of irreducible elements is essentially unique.*

Proof. By Theorem 2.4.15 it is enough to show that every irreducible element of D is prime. So let $a \in D$ be irreducible. Let $b, c \in D$ be such that $a|bc$. Let d be a greatest common divisor of a, b. Then $d|a$ so that $a = ud$ for a certain $u \in D$. As a is irreducible it follows that one of u, d is invertible. If u is invertible then because $d|b$ we also have $a|b$. If d is invertible, then we write $d = sa + tb$ for certain $s, t \in D$ (these exist by Theorem 2.4.24). Hence $1 = d^{-1}sa + d^{-1}tb$, and $c = d^{-1}sac + d^{-1}tbc$. From $a|bc$ we have that there exists a $v \in D$ with $bc = va$. Thus $c = d^{-1}sac + d^{-1}tva = (d^{-1}sc + d^{-1}tv)a$ and $a|c$. We conclude that a is prime. □

Example 2.4.29 In Example 2.4.22 we have seen that $\mathbb{Z}[i]$ is Euclidean. Therefore by the previous theorem, it has unique factorization into irreducibles. Furthermore, by Theorem 2.4.15 every irreducible element of $\mathbb{Z}[i]$ is prime. However, it is not immediately clear what the irreducibles (or primes) of $\mathbb{Z}[i]$ are. Since $\eta \in \mathbb{Z}[i]$ is invertible if and only if $N(\eta) = 1$, it follows that the invertibles of $\mathbb{Z}[i]$ are $\pm 1, \pm i$. But $2 = (1 + i)(1 - i)$. So since $1 \pm i$ are not invertible, it follows that 2 is not irreducible (or prime) in $\mathbb{Z}[i]$.

Example 2.4.30 Consider $\mathbb{Z}[\sqrt{-3}]$. As seen in Example 2.4.17 this ring has no unique factorization. Therefore, by the previous theorem it is not Euclidean either.

Example 2.4.31 Let $f = x^2 - x + 5$. The roots of f are $\frac{1 \pm \sqrt{-19}}{2}$. They do not lie in $\mathbb{Q}$, and therefore f is irreducible in $\mathbb{Q}[x]$ (as f has degree 2!). We set $\alpha = \frac{1+\sqrt{-19}}{2}$, and consider the ring $\mathbb{Z}[\alpha]$. Here we show that it is *not* Euclidean. We cannot do that with the method of Example 2.4.30, as it turns out that $\mathbb{Z}[\alpha]$ has unique factorization into irreducibles (see Example 2.6.28).

For the norm we have $N(a + b\alpha) = |a^2 + ab + 5b^2|$. If $ab \geq 0$ then $a^2 + ab + 5b^2 \geq 0$. If $ab < 0$ then we write $a^2 + ab + 5b^2 = (a + b)^2 - ab + 4b^2$ and see that $a^2 + ab + 5b^2 \geq 0$ as well. So $N(a + b\alpha) = a^2 + ab + 5b^2$.

As seen in Section 2.4.1 $a + b\alpha$ is invertible if and only if $1 = N(a + b\alpha) = a^2 + ab + 5b^2$. If $ab \geq 0$ then this equation implies $b = 0$, $a \pm 1$. If $ab < 0$ then again we write $a^2 + ab + 5b^2 = (a + b)^2 - ab + 4b^2$ and we see that $b = 0$ and $a = \pm 1$ as well. We conclude that ± 1 are the invertible elements of $\mathbb{Z}[\alpha]$.

We claim that there are no elements in $\mathbb{Z}[\alpha]$ with norm 2 or 3. Indeed, suppose that $N(a+b\alpha) = 2$. This amounts to $a^2 + ab + 5b^2 = 2$. If $ab \geq 0$ then this implies $b = 0$ and $a^2 = 2$ which has no solution in $\mathbb{Z}$. If $ab < 0$ then as above we get $(a + b)^2 - ab + 4b^2 = 2$ and again we have $b = 0$ and $a^2 = 2$, which is impossible. In exactly the same way we see that there are no elements of norm 3.

Now suppose that $\mathbb{Z}[\alpha]$ is Euclidean with function ν as in Definition 2.4.18. Let $S = \{\nu(\xi) \mid \xi \in \mathbb{Z}[\alpha] \text{ non-invertible}, \xi \neq 0\}$, let m_0 be the minimal element of S and let $\eta \in \mathbb{Z}[\alpha]$ be such that $\nu(\eta) = m_0$. Because we assumed $\mathbb{Z}[\alpha]$ to be Euclidean, there exist $q, r \in \mathbb{Z}[\alpha]$ with $2 = q\eta + r$ and $r = 0$ or $\nu(r) < \nu(\eta)$. In other words, $r = 0$ or r is invertible (by definition of S). Therefore, r is one of $0, 1, -1$. Now $r = 1$ implies $1 = q\eta$ which is excluded because η is not invertible. Hence $r \neq 1$.

Suppose that $r = 0$; then $q\eta = 2$. So $N(q)N(\eta) = 4$. Since there are no elements of norm 2, and $N(\eta) \geq 2$, it follows that $N(\eta) = 4$ and $N(q) = 1$. We write $\alpha = q'\eta + r'$ with $r' = 0$ or $\nu(r') < \nu(\eta)$. Again it follows that r' is one of 0, 1, -1. If $r' = 0$ then α is divisible by η. But that is not possible because $N(\alpha) = 5$ and $N(\eta) = 4$. If $r' = 1$ then $\alpha - 1$ is divisible by η, which is excluded because

$N(\alpha-1)=5$. If $r'=-1$ then $\alpha+1$ is divisible by η, but because $N(\alpha+1)=7$ this is again impossible. So we see that we cannot have $r=0$.

In exactly the same way we see that $r=-1$ is impossible. The conclusion is that $\mathbb{Z}[\alpha]$ cannot be Euclidean.

2.5 Congruences

In this section we use certain equivalence relations (called congruences) on $\mathbb{Z}$ to construct finite rings. Congruences were introduced by Carl F. Gauss in his *Disquisitiones Arithmeticae* (1801). Unlike the problem of unique factorization, the interest in this does not seem to have been driven by attempts to solve a praticular problem, but rather by human curiosity. The theory of congruences has found many applications in mathematics, of which we will see a few. Furthermore, in the second half of the 20st century a new development in cryptography, the so-called RSA system, was based on arithmetic with congruences. Towards the end of the section we will see what this system is about.

2.5.1 Definition, examples and a first application

Let $n\in\mathbb{Z}$, $n\neq 0$. We define the relation R_n on $\mathbb{Z}$ by stipulating that $(a,b)\in R_n$ if $n|(a-b)$. We leave it as an exercise to check that this indeed defines an equivalence relation. It is clear that $-n$ defines the same relation as n. Therefore we always assume that $n>0$.

The equivalence class of $a\in\mathbb{Z}$ is denoted $[a]_n$ (or just by $[a]$ if it is clear which n we are using), so

$$[a]_n=\{b\in\mathbb{Z}\mid n|(a-b)\}=\{a+kn\mid k\in\mathbb{Z}\}.$$

In this context, an equivalence class is also called a *congruence class*. As noted in Section 1.4 for $a,b\in\mathbb{Z}$ we can have $[a]_n=[b]_n$, also if $a\neq b$. From Proposition 1.4.2 we see that $[a]_n=[b]_n$ if and only if $a\in[b]_n$ if and only if n divides $a-b$. We say that a is a *representative* of the congruence class $[a]_n$. So b is also a representative of $[a]_n$ if and only if n divides $a-b$.

Instead of $(a,b)\in R_n$, or $b\in[a]_n$, we often also write $a\equiv b \bmod n$, and we say that a and b are congruent modulo n, or that a is equal to b modulo n.

Example 2.5.1 We have $15\equiv 46 \bmod 31$, $7\equiv 31 \bmod 8$.

Proposition 2.5.2 *R_n has exactly n congruence classes, namely $[0]_n,[1]_n,\ldots,[n-1]_n$.*

Proof. The listed classes are not equal because if $0\leq a<b\leq n-1$ then $0<b-a<n$ so that n does not divide $b-a$. For $a\in\mathbb{Z}$ there are $q,r\in\mathbb{Z}$ with $a=qn+r$ and $0\leq r<n$ (Proposition 2.2.1). Then n divides $a-r$ so that $[a]_n=[r]_n$. It follows that there are no more congruence classes than the listed ones. □

The set of all congruence classes is denoted $\mathbb{Z}/n\mathbb{Z}$, so

$$\mathbb{Z}/n\mathbb{Z}=\{[0]_n,\ldots,[n-1]_n\}.$$

We want to define arithmetic operations on $\mathbb{Z}/n\mathbb{Z}$ and we try to do it in the stupidest possible way:

$$[a]_n+[b]_n=[a+b]_n$$
$$[a]_n\cdot[b]_n=[ab]_n.$$

But potentially there is a problem with this. For example, we have $\mathbb{Z}/3\mathbb{Z}=\{\alpha,\beta,\gamma\}$ with $\alpha=[0]$, $\beta=[1]$, $\gamma=[2]$. So $\beta+\gamma=[1]+[2]=[3]=[0]$. But also $\beta=[7]$ and $\gamma=[-1]$. Hence $\beta+\gamma=[7-1]$, which could be different from $[0]$. Now the miracle happens: $[7-1]=[6]=[0]$. The point now is that this miracle always happens.

Lemma 2.5.3 *The operations $+$ and $\cdot$ on $\mathbb{Z}/n\mathbb{Z}$ are* well-defined*, that is, the result of these operations does not depend on the representatives of the congruence classes that we use.*

Proof. Let $a, b, c, d \in \mathbb{Z}$ be such that $[a]_n = [c]_n$, $[b]_n = [d]_n$. Then we must show that $[a+b]_n = [c+d]_n$, $[ab]_n = [cd]_n$. The equalities above are equivalent to $a = c + kn$, $b = d + ln$ for certain $k, l \in \mathbb{Z}$. So

$$a + b = c + d + (k + l)n \text{ and } ab = cd + (cl + dk + kln)n,$$

implying the desired conclusion. □

Remark 2.5.4 An important consequence of this lemma is that when we do arithmetic modulo n we can reduce any intermediate result modulo n. Stated like this it is a bit vague, so let's do some examples. Let $a, b \in \mathbb{Z}$, and let $c, d \in \mathbb{Z}$ be such that $a \equiv c \bmod n$, $b \equiv d \bmod n$ (in other words, such that $[a]_n = [c]_n$, $[b]_n = [d]_n$). Then $a + b \equiv c + d \bmod n$ and $ab \equiv cd \bmod n$. So, for example, $40 + 21 \equiv 1 + 8 \bmod 13 \equiv 9 \bmod 13$. Here we see that instead of computing $40 + 21 = 61$ and establishing that this is equal to 9 modulo 13, we can reduce 40 and 21 modulo 13 first and then do the addition. This is particularly useful when doing multiplications, for example

$$25 \cdot 103 \cdot 77 \equiv 10 \cdot -2 \cdot 2 \bmod 15 \equiv -5 \cdot 2 \bmod 15 \equiv 5 \bmod 15.$$

We can do this off the top of our head, whereas computing $25 \cdot 103 \cdot 77 = 198275$ and reducing that modulo 15 is much harder!

Proposition 2.5.5 *With the operations defined above $\mathbb{Z}/n\mathbb{Z}$ is a commutative ring with unity (the unity being $[1]_n$).*

Proof. The associativity of $+$ follows directly from the associativity of the addition in $\mathbb{Z}$:

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

Note that this computation is so simple because of Lemma 2.5.3: the result of the addition does not depend on the representatives of the classes. So we do not have to worry about them: we can do sums as if we were in $\mathbb{Z}$, and just put brackets [ ] around everything.

The associativity of $\cdot$, the commutativity of $+$ and $\cdot$ and the distributive laws follow in the same way.

The zero of $\mathbb{Z}/n\mathbb{Z}$ is $[0]$: $[0] + [a] = [0 + a] = [a]$. The opposite of an $[a]$ is $[-a]$: $[a] + [-a] = [a - a] = [0]$. The unity is $[1]$: $[1][a] = [1 \cdot a] = [a]$. □

Example 2.5.6 Since $\mathbb{Z}/n\mathbb{Z}$ is a finite ring, we can make tables of the addition and multiplication. For example, for $n = 6$ we get Table 2.1. Here, for brevity, we write a instead of $[a]$.

Table 2.1: Addition and multiplication tables of $\mathbb{Z}/6\mathbb{Z}$.

$+$	0	1	2	3	4	5
0	0	1	2	3	4	5
1	1	2	3	4	5	0
2	2	3	4	5	0	1
3	3	4	5	0	1	2
4	4	5	0	1	2	3
5	5	0	1	2	3	4

$\cdot$	0	1	2	3	4	5
0	0	0	0	0	0	0
1	0	1	2	3	4	5
2	0	2	4	0	2	4
3	0	3	0	3	0	3
4	0	4	2	0	4	2
5	0	5	4	3	2	1

We see that the addition table is not very interesting, as each row is equal to the previous row, shifted by one. From the multiplication table we can see some interesting phenomena. For example, we see that $\mathbb{Z}/6\mathbb{Z}$ contains zero divisors, namely [2], [3] and [4]. The other non-zero elements, [1] and [5], are invertible. (And we also see that the inverse of [5] is [5] itself.)

As a first application of the arithmetic in $\mathbb{Z}/n\mathbb{Z}$ we show the following curiosity.

Proposition 2.5.7 *An integer is divisible by 9 if and only if the sum of its decimal digits is divisible by 9.*

Proof. Let $m \in \mathbb{Z}$. Note that $9|m$ if and only if $[m]_9 = [0]_9$. Write $m = a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \cdots + a_s \cdot 10^s$, where $0 \leq a_i \leq 9$. That is, $a_0, \ldots, a_s$ are the decimal digits of m. Now $[10]_9 = [1]_9$, and by using the definitions of addition and multiplication in $\mathbb{Z}/9\mathbb{Z}$ repeatedly we compute

$$\begin{aligned}
[m]_9 &= [a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \cdots + a_s \cdot 10^s]_9 \\
&= [a_0]_9 + [a_1]_9 \cdot [10]_9 + [a_2]_9 \cdot [10]_9^2 + \cdots + [a_s]_9 \cdot [10]_9^s \\
&= [a_0]_9 + [a_1]_9 \cdot [1]_9 + [a_2]_9 \cdot [1]_9^2 + \cdots + [a_s]_9 \cdot [1]_9^s \\
&= [a_0]_9 + [a_1 \cdot 1]_9 + [a_2 \cdot 1^2]_9 + \cdots + [a_s \cdot 1^s]_9 \\
&= [a_0 + \cdots + a_s]_9.
\end{aligned}$$

So $[m]_9 = [0]_9$ if and only if $[a_0 + \cdots + a_s]_9 = 0$ if and only if $9|(a_0 + \cdots + a_s)$. □

Example 2.5.8 Let $m = 183380664473623892305971 45$. The sum of its decimal digits is 117. The sum of the decimal digits of the latter is 9. Hence 117 is divisible by 9, and therefore so is m.

2.5.2 Certain finite fields and Eisenstein's criterion for irreducibility

Here we show a nice application of the theory of congruences, namely a criterion due to Eisenstein by which it is sometimes possible to prove that a given polynomial is irreducible.

Theorem 2.5.9 *Let $p \in \mathbb{Z}$ be a prime. Then $\mathbb{Z}/p\mathbb{Z}$ is a field.*

Proof. By Proposition 2.5.5 $\mathbb{Z}/p\mathbb{Z}$ is a commutative ring with unity, which is $[1]_p$. So we only have to show that every non-zero element of $\mathbb{Z}/p\mathbb{Z}$ has a multiplicative inverse. Let $a \in \mathbb{Z}$ be such that $[a]_p \neq [0]_p$. This is the same as saying that p does not divide a. As p is prime this implies that $\gcd(a, p) = 1$. Hence by Theorem 2.2.3 there are $s, t \in \mathbb{Z}$ with $sa + tp = 1$. Therefore $[s]_p[a]_p = [sa]_p = [1 - tp]_p = [1]_p$ and the inverse of $[a]_p$ is $[s]_p$. We conclude that $\mathbb{Z}/p\mathbb{Z}$ is a field. □

The field $\mathbb{Z}/p\mathbb{Z}$, for p prime, is often denoted $\mathbb{F}_p$.

Now we need a small intermezzo on homomorphisms of rings. Let R, S be two rings. A map $f : R \to S$ is called a *ring homomorphism* if

$$f(r_1 + r_2) = f(r_1) + f(r_2) \text{ and } f(r_1 \cdot r_2) = f(r_1) \cdot f(r_2) \text{ for all } r_1, r_2 \in R.$$

(Here the $+$ on the left is the addition of R, whereas the $+$ on the right is the addition of S. The same goes for the multiplication.) A ring homomorphism which is bijective is called a *ring isomorphism.* If there exists a ring isomorphism $f : R \to S$ then the rings R, S are said to be *isomorphic* and we write $R \cong S$.

Lemma 2.5.10 *Let R, S be rings and $f : R \to S$ a ring homomorphism. Let $0_R, 0_S$ denote the zeros of R and S respectively. Then $f(0_R) = 0_S$ and $f(-r) = -f(r)$ for all $r \in R$. Furthermore, if R, S have unities, which are denoted $1_R, 1_S$ respectively, then $f(1_R) = 1_S$, unless $f(1_R)$ is a zero divisor in S.*

Proof. We have $f(0_R) = f(0_R + 0_R) = f(0_R) + f(0_R)$. Adding $-f(0_R)$ to both sides we obtain $f(0_R) = 0_S$. So also $0_S = f(r - r) = f(r) + f(-r)$; hence $f(-r) = -f(r)$.

In the same spirit we see that $f(1_R) = f(1_R \cdot 1_R) = f(1_R) \cdot f(1_R)$, hence $f(1_R)(1_S - f(1_R)) = 0_S$. So if $f(1_R)$ is not a zero divisor then $1_S - f(1_R) = 0_S$, or $f(1_R) = 1_S$. □

Remark 2.5.11 For all $a \in A$ we have $f(a)(1_S - f(1_R)) = f(a \cdot 1_R)(1_S - f(1_R)) = f(a)(f(1_R)(1_S - f(1_R))) = 0_S$. So if $1_S - f(1_R) \neq 0$ then $f(a)$ is either zero or a zero divisor for all $a \in A$.

Let $f : R \to S$ be a ring homomorphism. Then its *kernel* is

$$\ker(f) = \{r \in R \mid f(r) = 0_S\}.$$

By the previous lemma, $0_R \in \ker(f)$. It is straightforward to see that $\ker(f) = \{0_R\}$ if and only if f is injective. Indeed, suppose that $\ker(f) = \{0_R\}$. Now $f(r_1) = f(r_2)$ is the same as $f(r_1 - r_2) = 0_S$ and therefore $r_1 - r_2 = 0_R$, so that f is injective. Conversely, if f is injective and $f(r) = 0_S$ then $f(r) = f(0_R)$ and $r = 0_R$, so that $\ker(f) = \{0_R\}$.

Example 2.5.12 Let $n \in \mathbb{Z}$, $n > 0$ and define $\pi : \mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ by $\pi(a) = [a]_n$. The definitions of the addition and multiplication in $\mathbb{Z}/n\mathbb{Z}$ immediately imply that this is a ring homomorphism. Its kernel is $\{a \in \mathbb{Z} \mid \pi(a) = [0]_n\} = \{a \in \mathbb{Z} \mid n|a\} = \{kn \mid k \in \mathbb{Z}\}$.

Now we want to show Eisenstein's criterion for irreducibility of polynomials in $\mathbb{Q}[x]$. For this we first need a lemma which is due to Gauss.

Lemma 2.5.13 (Gauss) *Let $f \in \mathbb{Z}[x]$ and suppose that there are $g, h \in \mathbb{Q}[x]$ with $\deg(g), \deg(h) > 0$ and $f = gh$. Then there are rational numbers $\alpha, \beta \in \mathbb{Q}$ with $\alpha g, \beta h \in \mathbb{Z}[x]$ and $f = (\alpha g)(\beta h)$.*

Proof. A polynomial $k \in \mathbb{Z}[x]$ is said to be *primitive* if there is no prime $p \in \mathbb{Z}$ that divides all coefficients of k. Let $k, l \in \mathbb{Z}[x]$ be primitive. Then we claim that kl is primitive. Write $k = a_0 + a_1 x + \cdots + a_m x^m$, $l = b_0 + b_1 x + \cdots + b_n x^n$ with $a_m, b_n \neq 0$. Let p be a prime, and let s, t be minimal such that p does *not* divide a_s, b_t. The coefficient of x^{s+t} in kl is

$$a_s b_t + (a_{s-1} b_{t+1} + a_{s-2} b_{t+2} + \cdots) + (a_{s+1} b_{t-1} + a_{s+2} b_{t-2} + \cdots).$$

However p divides a_i for $0 \leq i < s$ and b_j for $0 \leq j < t$. Hence p divides all elements that appear between the brackets above. But p does not divide $a_s b_t$. Therefore, p does not divide the coefficient of x^{s+t} in kl. We conclude that kl is primitive.

For the proof of the lemma we may assume that f is primitive. There are primitive polynomials $\tilde{g}$, $\tilde{h}$ in $\mathbb{Z}[x]$ such that $g = \frac{a}{b}\tilde{g}$, $h = \frac{c}{d}\tilde{h}$ and $a, b, c, d \in \mathbb{Z}_{\geq 1}$ are such that $\gcd(a, b) = \gcd(c, d) = 1$. Let $s, t \in \mathbb{Z}_{\geq 1}$ be such that $\gcd(s, t) = 1$ and $\frac{s}{t} = \frac{ac}{bd}$. Then $f = \frac{s}{t}\tilde{g}\tilde{h}$. Moreover, by the above, $\tilde{g}\tilde{h}$ is primitive. But because $f \in \mathbb{Z}[x]$ we see that t divides every coefficient in $\tilde{g}\tilde{h}$. Hence $t = 1$. So $f = s\tilde{g}\tilde{h}$, and because $f, \tilde{g}, \tilde{h}$ are primitive we have $s = 1$ as well. The conclusion is that $\alpha = \frac{b}{a}$, $\beta = \frac{d}{c}$ do the job. □

Theorem 2.5.14 (Eisenstein) *Let $f \in \mathbb{Q}[x]$, $f = a_0 + a_1 x + \cdots + a_n x^n$ with $a_i \in \mathbb{Z}$. Suppose that there is a prime p with*

$$\begin{aligned} &p|a_i, 0 \leq i \leq n-1, \\ &p \not| a_n, \\ &p^2 \not| a_0. \end{aligned}$$

Then f is irreducible in $\mathbb{Q}[x]$.

Proof. We consider the map $\psi_p : \mathbb{Z}[x] \to \mathbb{F}_p[x]$ defined by $\psi_p(b_0 + \cdots + b_m x^m) = [b_0]_p + [b_1]_p x + \cdots + [b_m]_p x^m$. Because of the definitions of the arithmetic operations in the various rings that occur ($\mathbb{F}_p$, $\mathbb{Z}[x]$, $\mathbb{F}_p[x]$) we have that this is a ring homomorphism.

Suppose that f is not irreducible. Then by Lemma 2.5.13 there are $g, h \in \mathbb{Z}[x]$ with $f = gh$ and $\deg(g), \deg(h) \geq 1$. Because of the hypothesis on p we have $\psi_p(f) = [a_n]_p x^n$. But also $\psi_p(gh) = \psi_p(g)\psi_p(h)$. Because of unique factorization (Theorem 2.3.18) it follows that $\psi_p(g) = [b_m]_p x^m$, $\psi_p(h) = [c_r]_p x^r$. So, writing $g = b_0 + \cdots + b_m x^m$, $h = c_0 + \cdots + c_r x^r$, it follows that $p|b_0$, $p|c_0$. But that implies that p^2 divides a_0 as $a_0 = b_0 c_0$. This is excluded, and therefore f is irreducible. □

Example 2.5.15 Using the criterion it is immediate that the following polynomials are irreducible in $\mathbb{Q}[x]$: $x^n - 2$, $x^3 - 6x + 3$.

2.5.3 The Chinese remainder theorem

The Chinese remainder theorem is so called because an instance of it appeared in the third century Chinese mathematics book Sun Zi Suanjing. In there it is stated: "Now there are an unknown number of things. If we count by threes, there is a remainder 2; if we count by fives, there is a remainder 3; if we count by sevens, there is a remainder 2. Find the number of things." Here we state it in a more modern form, as an isomorphism between certain rings. For that we first need to look at a construction of rings which is called the *direct product*.

Let R, S be rings. Consider the product $R \times S = \{(r, s) \mid r \in R, s \in S\}$ on which we define the following operations

$$(r_1, s_1) + (r_2, s_2) = (r_1 + r_2, s_1 + s_2) \text{ and } (r_1, s_1) \cdot (r_2, s_2) = (r_1 \cdot r_2, s_1 \cdot s_2).$$

It is clear that with these operations $R \times S$ is a ring. It is called the *direct product* of R, S. It is also clear that if R, S are commutative, then so is $R \times S$, if R and S have unities 1_R, 1_S respectively, then $R \times S$ has unity $(1_R, 1_S)$. However, it is *not* true that if R, S are domains, that then $R \times S$ is a domain as well. Indeed, the zero of $R \times S$ is $(0_R, 0_S)$, where 0_R, 0_S are the zeroes of R, S respectively. So for example, $(1_R, 0_S) \cdot (0_R, 1_S) = (0_R, 0_S)$ and we see that $R \times S$ has zero divisors.

Theorem 2.5.16 (Chinese remainder theorem) *Let m, n be coprime positive integers. Then there is a well-defined map*

$$\sigma : \mathbb{Z}/mn\mathbb{Z} \to \mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}$$

with $\sigma([a]_{mn}) = ([a]_m, [a]_n)$. Moreover, this map is an isomorphism of rings.

Proof. If $[a]_{mn} = [b]_{mn}$ then $mn|a - b$. So each of m, n divides $a - b$ and $[a]_m = [b]_m$, $[a]_n = [b]_n$. Therefore, setting $\sigma([a]_{mn}) = ([a]_m, [a]_n)$ yields a well-defined map.

It is immediate that σ is a homomorphism of rings:

$$\sigma([a]_{mn} + [b]_{mn}) = \sigma([a + b]_{mn}) = ([a + b]_m, [a + b]_n) = ([a]_m, [a]_n) + ([b]_m, [b]_n)$$
$$\sigma([a]_{mn}[b]_{mn}) = \sigma([ab]_{mn}) = ([ab]_m, [ab]_n) = ([a]_m, [a]_n)([b]_m, [b]_n).$$

Let $a \in \mathbb{Z}$ be such that $\sigma([a]_{mn}) = ([0]_m, [0]_n)$. Since $\sigma([a]_{mn}) = ([a]_m, [a]_n)$ it follows that $[a]_m = [0]_m$ and $[a]_n = [0]_n$, or that $m|a$ and $n|a$. Lemma 2.2.8(i) now implies that $mn|a$. Hence $[a]_{mn} = [0]_{mn}$. So the kernel of σ is $\{[0]_{mn}\}$. As seen in Section 2.5.2, this means that σ is injective.

Finally we observe that

$$|\mathbb{Z}/mn\mathbb{Z}| = mn = |\mathbb{Z}/m\mathbb{Z}| \cdot |\mathbb{Z}/n\mathbb{Z}| = |\mathbb{Z}/m\mathbb{Z} \times \mathbb{Z}/n\mathbb{Z}|.$$

Furthermore, an injective map between finite sets of the same cardinality has to be surjective. It follows that σ is surjective as well, and therefore it is an isomorphism. □

We can easily state a more general version of this theorem. Let $m_1, \ldots, m_t$ be positive pairwise coprime integers (the latter means that $\gcd(m_i, m_j) = 1$ if $i \neq j$). Then there is a ring isomorphism

$$\sigma : \mathbb{Z}/(m_1 \cdots m_t)\mathbb{Z} \to \mathbb{Z}/m_1\mathbb{Z} \times \cdots \times \mathbb{Z}/m_t\mathbb{Z}$$

with $\sigma([a]_{m_1 \cdots m_t}) = ([a]_{m_1}, \ldots, [a]_{m_t})$. This can be proved in the same way as Theorem 2.5.16. Alternatively, it can be proved by induction, using Theorem 2.5.16, by considering the isomorphisms

$$\mathbb{Z}/(m_1 \cdots m_t)\mathbb{Z} \to \mathbb{Z}/m_1\mathbb{Z} \times \mathbb{Z}/(m_2 \cdots m_t)\mathbb{Z} \to \mathbb{Z}/m_1\mathbb{Z} \times (\mathbb{Z}/m_2\mathbb{Z} \times \cdots \times \mathbb{Z}/m_t\mathbb{Z}) \to \mathbb{Z}/m_1\mathbb{Z} \times \cdots \times \mathbb{Z}/m_t\mathbb{Z}.$$

One thing that Theorem 2.5.16 says is that if m, n are coprime positive integers, and a_1, a_2 are any integers then there is an $x \in \mathbb{Z}$ with

$$\begin{aligned} x &\equiv a_1 \bmod m \\ x &\equiv a_2 \bmod n. \end{aligned} \tag{2.5.1}$$

Indeed, because the map σ is surjective, it follows that there is an $x \in \mathbb{Z}$ with $\sigma([x]_{mn}) = ([a_1]_m, [a_2]_n)$. However, the proof of the theorem does not tell us how to find this x. Now we show how that is done. For this we consider a concrete example:

$$x \equiv 7 \bmod 10$$
$$x \equiv 11 \bmod 23.$$

The trick is to first write the general solution to one of the equations, with one parameter in it. And then determine a value for the parameter so that it satisfies also the other equation. Lets consider the first equation. Any solution to that can be written $x = 7 + 10s$. The second equation then says $7 + 10s \equiv 11 \bmod 23$, or $10s \equiv 4 \bmod 23$. Note that any integer s satisfying this will give us a solution of the equations. Since $\gcd(10, 23) = 1$ there are $u, v \in \mathbb{Z}$ with $u \cdot 10 + v \cdot 23 = 1$. These can be found with the extended Euclidean algorithm (see Section 2.2.2). In this case a possibility is $u = -16$, $v = 7$. So we see that $-16 \cdot 10 \equiv 1 \bmod 23$. We now multiply the equation that we got by -16 and obtain $-16 \cdot 10 \cdot s \equiv -16 \cdot 4 \bmod 23$, or equivalently, $s \equiv -64 \bmod 23$. Now $-64 \equiv 5 \bmod 23$ so we use $s = 5$ and get $x = 7 + 50 = 57$.

Now we look at the general equations (2.5.1). Write $x = a_1 + ms$. Any s that we substitute here will give an x satisfying the first equation. We substitute in the second equation and get $ms \equiv a_2 - a_1 \bmod n$. There are $u, v \in \mathbb{Z}$ with $um + vn = 1$, so that $um \equiv 1 \bmod n$. We multiply the equation by u and get $ums \equiv u(a_2 - a_1) \bmod n$, or $s \equiv u(a_2 - a_1) \bmod n$. So we can use $s = u(a_2 - a_1)$. Then $x = a_1 + mu(a_2 - a_1)$. Alternatively, by a division with remainder we find q, r such that $u(a_2 - a_1) = qn + r$ and $0 \leq r < n$. Then $u(a_2 - a_1) \equiv r \bmod n$ and we can also use $s = r$. Then we get $x = a_1 + rm$, which usually is a smaller solution.

The Chinese remainder theorem can be extended to more than two factors, and similarly the above method can be extended to more than two equations. Suppose, for example, that we have three equations

$$x \equiv a_1 \bmod m$$
$$x \equiv a_2 \bmod n$$
$$x \equiv a_3 \bmod r,$$

where m, n, r are pairwise relatively prime. Then we first pick two of the equations, say the first two, and solve them, finding a $b \in \mathbb{Z}$ with $b \equiv a_1 \bmod m$, $b \equiv a_2 \bmod n$. Next we find $x \in \mathbb{Z}$ satisfying

$$x \equiv b \bmod mn$$
$$x \equiv a_3 \bmod r,$$

then it is clear that x satisfies all three equations.

In many applications the Chinese remainder theorem is used to chop up a computation modulo a big integer into a few computations modulo smaller integers, which are much easier. We illustrate that by a method for computing $a^s \bmod n$, for given integers a, s, n. For this we first show how to compute a^s by using a small number of multiplications. We look at an example.

Example 2.5.17 We compute a^{10}. The naive way of doing that is to compute

$$a^2 = a \cdot a,\ a^3 = a \cdot a^2, a^4 = a \cdot a^3, \ldots, a^{10} = a \cdot a^9,$$

which takes 9 multiplications. But alternatively we can compute

$$a^2 = a \cdot a,\ a^4 = a^2 \cdot a^2,\ a^8 = a^4 \cdot a^4,\ a^{10} = a^2 \cdot a^8,$$

needing 4 multiplications.

In order to use the trick of the previous example as the basis of a general method, we note that every positive integer can be written as a sum of distinct powers of 2. Indeed, if this is false then there is a smallest positive integer m that cannot be written as a sum of distinct powers of 2. Then $m > 1$

as $2^0 = 1$. Let k be maximal with $2^k \leq m$. By the hypothesis on m we have $2^k < m$. Again by the hypothesis on m we have that $m - 2^k$ can be written as a sum of distinct powers of 2. If any term in the latter sum is equal to 2^k then $2 \cdot 2^k \leq m$, but $2 \cdot 2^k = 2^{k+1}$, contrary to the choice of k.

Note that this proof also gives a method to write a given m as a sum of distinct powers of 2: let k be maximal with $2^k \leq m$ and continue with $m - 2^k$. Take for example $m = 10$. Then 2^3 is the highest power of 2 not exceeding 10. But $10 - 2^3 = 2$, so $10 = 2 + 2^3$.

Now write $s = 2^{i_1} + \cdots + 2^{i_r}$ with $i_1 < i_2 < \cdots < i_r$. By repeated squaring compute

$$a^2 = a \cdot a,\ a^4 = a^2 \cdot a^2, \ldots, a^{2^{i_r}} = a^{2^{i_r - 1}} \cdot a^{2^{i_r - 1}}$$

and then

$$a^s = a^{2^{i_1}} \cdot a^{2^{i_2}} \cdots a^{2^{i_r}}.$$

This algorithm for computing a^s is called the *exponentiation by repeated squaring.*

Remark 2.5.18 This method uses a low number of multiplications, but not necessarily the lowest one. For example consider computing a^{15}. We have $15 = 1 + 2 + 2^2 + 2^3$. So we compute

$$a^2 = a \cdot a,\ a^4 = a^2 \cdot a^2,\ a^8 = a^4 \cdot a^4 \text{ and } a^{15} = a \cdot a^2 \cdot a^4 \cdot a^8,$$

needing 6 multiplications. But we can also compute

$$a^2 = a \cdot a,\ a^3 = a \cdot a^2,\ a^6 = a^3 \cdot a^3,\ a^{12} = a^6 \cdot a^6 \text{ and } a^{15} = a^3 \cdot a^{12},$$

which uses 5 multiplications.

Next we look at the modular part, that is, computing $a^s \bmod n$. Suppose that $n = pq$ with $p, q > 1$ and $\gcd(p, q) = 1$. Then by the repeated squaring mathod we find a_1, a_2 such that $a_1 \equiv a^s \bmod p$, $a_2 \equiv a^s \bmod q$. Finally we solve the equations

$$x \equiv a_1 \bmod p$$
$$x \equiv a_2 \bmod q.$$

By the Chinese remainder theorem a solution x satisfies $x \equiv a^s \bmod n$.

Example 2.5.19 We compute $17^{13} \bmod 33$. We have $13 = 1 + 2^2 + 2^3$. Secondly, $33 = 3 \cdot 11$. We first compute $17^{13} \bmod 3$. Since $17 \equiv 2 \bmod 3$ we compute

$$2^2 \bmod 3 \equiv 1,\ 2^4 \bmod 3 \equiv 1,\ 2^8 \bmod 3 \equiv 1 \text{ and } 2^{13} \bmod 3 \equiv 2 \cdot 1 \cdot 1 = 2.$$

Next we compute $17^{13} \bmod 11$. Now $17 \equiv 6 \bmod 11$ and

$$6^2 \bmod 11 \equiv 3,\ 6^4 \bmod 11 \equiv 9,\ 6^8 \bmod 11 \equiv 4 \text{ and } 6^{17} \bmod 11 \equiv 6 \cdot 9 \cdot 4 \equiv 7 \bmod 11.$$

Finally we solve

$$x \equiv 2 \bmod 3$$
$$x \equiv 7 \bmod 11.$$

Set $x = 2 + 3s$ and substitute, to obtain $2 + 3s \equiv 7 \bmod 11$, or $3s \equiv 5 \bmod 11$. We have $4 \cdot 3 - 1 \cdot 11 = 1$, so after multiplying by 4, we get $s \equiv 20 \bmod 11$, so we can take $s = 9$ and $x = 29$. We see that a potentially complicated exponentiation can be reduced to a few multiplications that can easily be done by hand.

As a second application of the Chinese remainder theorem we now have a look at *secret sharing.* The principle of secret sharing can be illustrated by the following story. Victoria and her brother Allan find a map of a hidden treasure. But before they can go and retreive it, they first need to go home to put good shoes on. However, being brother and sister, they don't trust each other, and suspect that

the one who holds the map could go and find the treasure alone. So they tear the map in two in such a way that no half can be used in itself to find the treasure, and Victoria and Allan each get one half. When they have put their shes on they get back together, reconstruct the map, and find the treasure.

A more serious application lies in the command and control of heavy weapons, such as nuclear missiles. It is clear that can lead to serious consequences if just one single person can decide to deploy them. Therefore typically a few persons get part of a code. When enough of them get together and decide that the weapons should be used, then by putting together their pieces, they can reconstruct the code.

Here is one simple way how this can be realized using the Chinese remainder theorem. First the secret code is represented by a big integer N. Let p be the number of participants, that is, the people who get a part of the code. Choose pairwise coprime integers $n_1, \dots, n_p$ such that $n_0 = n_1 \cdots n_p > N$. The piece of information given to the i-th participant is the unique integer a_i with $0 \leq a_i < n_i$ and $N \equiv a_i \bmod n_i$. When the participants come together they can reconstruct N by solving the equations

$$\begin{aligned} x &\equiv a_1 \bmod n_1 \\ &\vdots \\ x &\equiv a_p \bmod n_p. \end{aligned}$$

Then $x \equiv N \bmod n_0$. So because N is the unique integer in the interval $[0, n_0)$ that is congruent to N modulo n_0, from x we can find N.

Like this the system works no better than giving each participant a piece of the secret code, which they can then put together when they meet. However the system based on the chinese remainder theorem is more flexible than that. In practical applications it is often necessary that the secret code can already be reconstructed when t participants come together, for some $t < p$. In order to achieve that the n_i are chosen such that for each $i_1 < \cdots < i_t$ we have $n_{i_1} \cdots n_{i_t} > N$, but also for each $j_1 < \cdots < j_{t-1}$ we have $n_{j_1} \cdots n_{j_{t-1}} < N$. This makes sure that any group of t participants can reconstruct N, but no group of $t-1$ participants can. (This is called an (t, p)-threshold scheme.)

This system was proposed by Mignotte in [Mig83]. It may not be entirely secure because the knowledge of just a few of the a_i does give some information on N, namely that it is an integer of the form $a_i + sn_i$. Here we do not go into these questions.

2.5.4 The RSA cryptosystem

Cryptography is the science of secret messages. The model that is commonly used consists of three persons: Alice, Bob and Eve. Alice wants to send secret messages to Bob, while Eve tries to intercept and read them (Eve is the *eavesdropper*). Alice transforms her message into something which cannot be directly understood: we say that she encrypts the plaintext into ciphertext. For this she uses a particular method of encryption, which usually can be made to work in different ways, depending on an initial setting called the *encryption key*. Upon receipt, Bob decrypts the message using a *decryption key*. It is assumed that Eve knows the encryption and decryption methods, but does not know the keys. If she can discover what the keys are, then she can also decrypt the message.

Example 2.5.20 Let the messages be written in the alphabet, consisting of the 26 letters A to Z. (Usually one would include more symbols, such as a space, but here we just restrict to using 26 letters.) We number these from 0 to 25, i.e.,

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

Pick an integer m_0 with $0 \leq m_0 \leq 25$. The encryption method consists in writing instead of the letter numbered i the letter numbered $(i + m_0) \bmod 26$. So, if we take $m_0 = 11$, then for example

$$\text{MOUNTAIN} \mapsto 12, 14, 20, 13, 19, 0, 8, 13 \mapsto 23, 25, 5, 24, 4, 11, 19, 24 \mapsto \text{XZFYELTY}.$$

The decryption method consists in writing, instead of the letter numbered i, the letter numbered $(i - m_0) \bmod 26$.

This method is called the *Caesar cipher* because the Roman historian Suetonius wrote that it was used by Julius Caesar to communicate with his generals. This simple method was also used more recently by the mafia boss Bernardo Provenzano to encrypt his secret messages to his collaborators.

At first this method looks quite good, for who would guess that XZFYELTY means MOUNTAIN? However, there are several problems with it. First of all, there are just 26 possible keys, so Eve can just try them all. Secondly, a given letter is always encrypted in the same way. In English the most common letter is E, so by looking for the most common letter in a piece of cyphertext, Eve can get a good guess as to what the encryption of E should be. Once she finds how one letter is encrypted, she also knows the key. This way of trying to break a system is called *frequency analysis*.

Another problem is that somehow Alice must communicate the key to Bob. If Eve intercepts that message, then she knows the key immediately. (This is not only a theoretical problem. In the second world war the German army used the so-called Enigma machine to send secret messages. In 1942 some codebooks were found by the British navy on board of a German submarine. These gave the British codebreakers a huge advantage.) In order to get around this, *public key cryptosystems* have been introduced. The idea that such systems can exist is due to Diffie and Hellman in their 1976 paper [DH76]. In these systems the encryption and decryption keys are not the same, and it is very difficult to work out what the decryption key is, from the knowledge of the encryption key only. Roughly such a system works as follows. There are two keys: the *public key* used for encryption and the *private key* used for decryption. Bob computes these two keys, makes the public key public, and keeps the private key to himself. Alice encrypts her message using the public key and sends it to Bob, who can decrypt using his private key. Eve also knows the public key, but from that it is practically impossible to arrive at the private key necessary for decryption.

Although Diffie and Hellman introduced the idea of public key cryptosystems, they did not come up with one. This was first done by Rivest, Shamir and Adleman in a paper of 1978, [RSA78]. Their system is nowadays called RSA (using the first letters of the names of its inventors). We first describe how it works, then we prove its correctness, and finally we indicate how pieces of text can be encrypted.

To compute the public and private keys, Bob does the following:

- He chooses two distinct (large) primes p, q and computes $n = pq$, $\varphi(n) = (p-1)(q-1)$.
- He chooses an r with $1 < r < \varphi(n)$ and $\gcd(r, \varphi(n)) = 1$.
- He uses the extended Euclidean algorithm to compute s, t with $0 < s < \varphi(n)$ and $sr + t\varphi(n) = 1$. (Note that $sr + t\varphi(n) = (s + m\varphi(n))r + (t - mr)\varphi(n)$, so we can change s arbitrarily modulo $\varphi(n)$ and hence may assume that $0 < s < \varphi(n)$.)
- The *public key* is (r, n) and the *private key* is (s, n).

The encryption and decryption methods run as follows. An integer a with $0 \leq a < n$ is *encrypted* as the integer b with $0 \leq b < n$ and $b \equiv a^r \bmod n$. The integer b with $0 \leq b < n$ is *decrypted* as the integer a with $0 \leq a < n$ and $a \equiv b^s \bmod n$.

Example 2.5.21 Bob chooses $p = 5$, $q = 13$ and obtains $n = 65$, $\varphi(n) = 48$, and chooses $r = 11$. By the extended Euclidean algorithm he finds $-13 \cdot r + 3 \cdot \varphi(n) = 1$, so he has $s = -13 + \varphi(n) = 35$. Hence the public key is $(11, 65)$ and the private key is $(35, 65)$.

In order to encrypt messages the same correspondence between letters and integers as in Example 2.5.20 is used. So for example, to encrypt F, Alice computes

$$F \mapsto 5 \mapsto 5^{11} \bmod 65 \equiv 60 \bmod 65$$

and she sends 60 to Bob, who computes

$$60 \mapsto 60^{35} \equiv 5 \bmod 65$$

and recovers the letter F.

In order to show the correctness of this scheme we need to prove that decryption is the inverse of encryption. For this we first have the following theorem first stated by Fermat.

Theorem 2.5.22 (Fermat's little theorem) *Let p be a prime and $a \in \mathbb{Z}$. Then $a^p \equiv a \bmod p$.*

Proof. First we prove it for $a \geq 0$ by induction. The case $a = 0$ is trivial, so suppose $a \geq 0$ and that $a^p \equiv a \bmod p$. We have

$$(a+1)^p = a^p + \binom{p}{1}a^{p-1} + \binom{p}{2}a^2 + \cdots + \binom{p}{p-1}a + 1.$$

For $1 \leq i \leq p-1$ we have $p! = \binom{p}{i}(p-i)!i!$. But p divides the left-hand side, and does not divide $(p-i)!$ nor $i!$. Hence p divides $\binom{p}{i}$. It follows that $(a+1)^p \equiv a^p + 1 \bmod p \equiv a + 1 \bmod p$, where the last equality follows from the induction hypothesis. This proves the theorem for $a \geq 0$.

If $a < 0$ then $a = -b$ with $b > 0$. Then if p is odd we have

$$a^p = -b^p \equiv -b \bmod p \equiv a \bmod p.$$

For $p = 2$ we have $a^2 = b^2 \equiv b \bmod 2 \equiv a \bmod 2$. □

Corollary 2.5.23 *Let p be a prime, and $a \in \mathbb{Z}$ such that $p \not| a$. Then $a^{p-1} \equiv 1 \bmod p$.*

Proof. By the previous theorem p divides $a^p - a = a(a^{p-1} - 1)$. But p does not divide a. Hence it divides $a^{p-1} - 1$. □

Proposition 2.5.24 *Let p, q be primes, $p \neq q$. Set $n = pq$, $\varphi(n) = (p-1)(q-1)$. Let $x \in \mathbb{Z}$, $x \geq 1$. Then $a^{x\varphi(n)+1} \equiv a \bmod n$ for all $a \in \mathbb{Z}$.*

Proof. We show that p divides $a^{x\varphi(n)+1} - a$. If $p|a$ then this certainly holds. So assume that $p \not| a$. Then by the previous corollary,

$$a^{x\varphi(n)+1} - a = a((a^{p-1})^{x(q-1)} - 1) \equiv a(1^{x(q-1)} - 1) \bmod p \equiv 0 \bmod p.$$

In the same way we see that q divides $a^{x\varphi(n)+1} - a$. Since $\gcd(p, q) = 1$ we conclude that n divides $a^{x\varphi(n)+1} - a$. □

Now let $p, q, r, s, n, \varphi(n)$ be as in the RSA system. Since $0 < s < \varphi(n)$ and $sr + t\varphi(n) = 1$ we necessarily have $t < 0$. Let a be an integer. Then

$$(a^r)^s = a^{rs} = a^{-t\varphi(n)+1} \equiv a \bmod n$$

by Proposition 2.5.24. It follows that decryption is indeed the inverse of encryption.

We are however not quite done. From Example 2.5.21 we see that the encryption of a letter does not necessarily lead to another letter. Secondly, just encrypting one letter at the time will result in a system which can be broken by frequency analysis, as we have seen with the Caesar cipher. The solution is to represent a larger piece of text by an integer. It is based on the following lemma.

Lemma 2.5.25 *Let $m \geq 2$ be an integer and $k \geq 0$. Then every integer a with $0 \leq a < m^{k+1}$ can uniquely be written as $a_0 + a_1 m + a_2 m^2 + \cdots + a_k m^k$, where $0 \leq a_i < m$.*

Proof. We use induction on k. If $k = 0$ then we can take $a_0 = a$. Suppose $k \geq 1$ and that every integer b with $0 \leq b < m^k$ can uniquely be written as $b = b_0 + b_1 m + \cdots + b_{k-1} m^{k-1}$, where $0 \leq b_i < m$. By a division with remainder (Proposition 2.2.1) we find unique $q, r \in \mathbb{Z}$ with $a = qm + r$ and $0 \leq r < m$. Because $a < m^{k+1}$ we have $q < m^k$. So by induction there are $a_1, \ldots, a_k$ such that $q = a_1 + a_2 m + \cdots + a_k m^{k-1}$. Now by setting $a_0 = r$ we obtain $a = a_0 + a_1 m + \cdots + a_k m^k$. In order to see uniqueness observe that $a_0 = r$ is uniquely determined. The uniqueness of the other a_i follows

from the induction hypothesis applied to q. □

Here the expression $a_0 + a_1 m + a_2 m^2 + \cdots + a_k m^k$ is called the *representation of a in the base m*. We write $a = (a_0, a_1, \ldots, a_k)_m$. So, for example, $(1,0,2,1)_3 = 1 + 0 \cdot 3 + 2 \cdot 9 + 1 \cdot 27 = 46$ and $(1,1,0,1)_2 = 1 + 2 + 8 = 11$.

Now let N be the number of symbols used for our messages. In Example 2.5.20 we have $N = 26$. Let p, q be the primes used by Bob in the RSA system, and let $n = pq$, $\varphi(n)$, r, s be as usual. Then choose k such that $N^k < n < N^{k+1}$. Then we encrypt pieces of text of length k in one stroke. This works as follows. Consider a piece of text of length k, consisting of the letters $X_0, \ldots, X_{k-1}$. Each X_i corresponds to a unique integer a_i with $0 \leq a_i < N$. Set $a = (a_0, \ldots, a_{k-1})_N$ so that by Lemma 2.5.25, $0 \leq a < N^k$. Let b be the unique integer with $0 \leq b < n$ and $b \equiv a^r \bmod n$. Since $b < n < N^{k+1}$, we have that $b = (b_0, \ldots, b_k)_N$ where $0 \leq b_i < N$. So each b_i corresponds to a unique letter, and we can write represent b as a piece of text consisting of $k+1$ letters. This is the encrypted form of the original piece of text.

Example 2.5.26 Again we use the 26 letters of the alphabet, enumerated as in Example 2.5.20. Bob chooses $p = 13$, $q = 59$, so that $n = 767$, $\varphi(n) = 12 \cdot 58 = 696$. Furthermore, he chooses $r = 131$, so that $s = 611$ (exercise). Since $26^2 = 676$ we choose $k = 2$ so that $26^2 < n < 26^3$. This means that we encrypt blocks of two letters into blocks of three letters. Let's encrypt the word MOUNTAIN. We compute

$$\text{MO} \mapsto (12,14)_{26} = 12 + 14 \cdot 26 = 376 \mapsto 376^{131} \bmod 767 \equiv 363 = 25 + 13 \cdot 26 + 0 \cdot 26^2 = (25,13,0)_{26} \mapsto \text{ZNA}$$
$$\text{UN} \mapsto (30,13)_{26} = 30 + 13 \cdot 26 = 368 \mapsto 368^{131} \bmod 767 \equiv 673 = 23 + 25 \cdot 26 + 0 \cdot 26^2 = (23,25,0)_{26} \mapsto \text{XZA}$$
$$\text{TA} \mapsto (19,0)_{26} = 19 + 0 \cdot 26 = 19 \mapsto 19^{131} \bmod 767 \equiv 635 = 11 + 24 \cdot 26 + 0 \cdot 26^2 = (11,24,0)_{26} \mapsto \text{LYA}$$
$$\text{IN} \mapsto (8,13)_{26} = 8 + 13 \cdot 26 = 346 \mapsto 346^{131} \bmod 767 \equiv 577 = 5 + 22 \cdot 26 + 0 \cdot 26^2 = (5,22,0)_{26} \mapsto \text{FWA}.$$

So the encryption of MOUNTAIN is ZNAXZALYAFWA.

Remark 2.5.27

- In order to perform a frequency analysis on an encrypted piece of text we need to deal with frequencies of blocks of length k. However, if k is not very small, then there are too many possible blocks for this to be practical. Secondly, the frequency of each block will be rather low, so it will be difficult to draw any useful conclusion from this.
- One should note that there are many more blocks of $k+1$ letters than there are blocks of k letters. *So not every block of $k+1$ letters is the encryption of a block of k letters.*
- We see that the method of exponentiation by repeated squaring, as explained in Section 2.5.3, is very useful here. However, the trick with splitting the computation using the Chinese remainder theorem can only be used by Bob, and not by Alice, as she does not know the factorization of n.

Finally we say some words on the security of the RSA system. Eve knows the public key (r, n). But in order to decrypt she has to find s, which is not computed from n but from $\varphi(n) = (p-1)(q-1)$. So if Eve can factorize n then she can find $\varphi(n)$ and s. On the other hand, if she has a way to find $\varphi(n)$, then she also has a way of factorizing n. Indeed, knowing $\varphi(n)$ she can define the polynomial $f = x^2 + (\varphi(n) - n - 1)x + n$. A small computation shows that $f = (x-p)(x-q)$. So from f Eve can find p, q by computing the roots of f. Therefore, Bob has to choose p, q in such a way that $n = pq$ is practically impossible to factorize *with the currently available algorithms*. For this p, q have to be "big" (whatever that means), but that is not enough, as certain algorithms can sometimes also factorize big numbers if the factors have special properties. The next example illustrates this.

Example 2.5.28 Let n be a given product of two odd primes p, q, $p < q$ Set $s = \frac{q-p}{2}$, $t = \frac{q+p}{2}$. Then also $n = t^2 - s^2$. Instead of trying to find p, q directly we can try to find s, t. For this we let t_0 be the smallest integer above $\sqrt{n}$, and for $i \geq 0$ we see whether $t_i^2 - n$ is a square. If it is, then $t_i^2 - n = s^2$ and $n = t_i^2 - s^2 = (t_i - s)(t_i + s)$ and we have factorized n. Otherwise, we set $t_{i+1} = t_i + 1$. For

example, consider $n = 943$. Then $t_0 = 31$ and $t_0^2 - n = 18$, which is not a square. Then $t_1 = 32$ and $t_1^2 - n = 81 = 9^2$. So we find $n = (32 - 9)(32 + 9) = 23 \cdot 41$.

This method, which by the way is due to Fermat, works well when the final t_i is close to t_0, which means that s is small. In turn that is the case when p, q are close together. So p, q can be big, but when they are close together, then this factorization method will find them quite quickly.

2.6 Ideals in rings

After having noticed that rings like $\mathbb{Z}[\alpha]$ may fail to have unique factorization into irreducibles, Ernst Kummer introduced the concept of ideal number, and showed that these do admit unique factorization. Then using these ideal numbers he managed to show that for odd *regular* primes p the equation $x^p + y^p = z^p$ has no solution in the integers with $xyz \neq 0$. In this proof also the factorization (2.4.1) (or rather its ideal theoretic counterpart) plays an important role. Here we do not go into the definition of the concept of regular prime. We just note that the first ten irregular primes are 37, 59, 67, 101, 103, 131, 149, 157, 233, 257. So Kummer's proof was a huge advance in the theory concerning Fermat's last theorem.

In 1876 Dedekind generalized Kummer's concept of ideal number, and introduced the notion of *ideal*. Here we will study ideals of mainly commutative rings. They are used to construct quotient rings. The rings $\mathbb{Z}/n\mathbb{Z}$, seen in the previous section, are examples of these, so ideals are a way of doing congruences in rings other than $\mathbb{Z}$. We will see a more general version of the Chinese remainder theorem. Finally we will show how the theory of ideals enables us to get some more information on the property of unique factorization in rings.

2.6.1 Ideals and quotients

Definition 2.6.1 *Let R be a ring. A subset $I \subset R$ is said to be an* ideal *if*

1. $0 \in I$,
2. $a - b \in I$ *for all* $a, b \in I$,
3. *for all* $a \in I$ *and* $b \in R$ *we have* $ab, ba \in I$.

Remark 2.6.2

- The second condition is equivalent to $-a \in I$ and $a + b \in I$ for all $a, b \in I$.
- If R is commutative then in the third condition it is enough to have $ba \in I$ for all $b \in R$, $a \in I$.

Example 2.6.3 Consider the ring $\mathbb{Z}$. Let $n \in \mathbb{Z}$ and set

$$n\mathbb{Z} = \{na \mid a \in \mathbb{Z}\}.$$

Then it is straightforward to see that $n\mathbb{Z}$ is an ideal of $\mathbb{Z}$. Moreover, it is not hard to see that all ideals of $\mathbb{Z}$ are of this form. Indeed, let $I \subset \mathbb{Z}$ be an ideal. If $I = \{0\}$ then $I = 0\mathbb{Z}$. Otherwise I contains positive integers. Let n be the mimimal positive integer in I. Then trivially $n\mathbb{Z} \subset I$. On the other hand, for $a \in I$ there are $q, r \in \mathbb{Z}$ with $a = qn + r$ and $0 \leq r < n$ (Proposition 2.2.1). But then $qn \in I$ and from $r = a - qn$ we see that $r \in I$. By the choice of n we get $r = 0$ and $a = qn$. Hence $I \subset n\mathbb{Z}$ so that $I = n\mathbb{Z}$.

Let R be a *commutative* ring, let $a_1, \ldots, a_s \in R$ and set

$$I = \left\{ \sum_{i=1}^{s} b_i a_i \mid b_i \in R \right\}.$$

Then it is obvious that I is an ideal of R. It is called the *ideal generated by* $a_1, \ldots, a_s$. We write $I = \langle a_1, \ldots, a_s \rangle$. An ideal generated by one element (that is, of the form $\langle a \rangle$) is said to be *principal*. The previous example shows that all ideals in $\mathbb{Z}$ are principal. However, this does not necessarily hold for other rings.

In Section 2.5.1 the congruence relation was defined as aR_nb if and only if $n|(b-a)$. With the notation of Example 2.6.3 that is equivalent to $b-a \in n\mathbb{Z}$. Now we do the same for an ideal in any ring. Let R be a ring and $I \subset R$ an ideal. Then we define the relation $\sim_I$ in R by $a \sim_I b$ if $b-a \in I$. From Definition 2.6.1 it follows immediately that this is an equivalence relation. The equivalence class of an $a \in R$ is

$$[a]_I = \{b \in R \mid b-a \in I\} = \{a+c \mid c \in I\},$$

and we write this last set as $a+I$. If it is clear which ideal we are working with we also write $[a]$ instead of $[a]_I$.

The set of all these equivalence classes is denoted R/I. It is called the *quotient* of R by I.

We note that $[a]_I = [b]_I$ if and only if $b \in [a]_I$ if and only if $b-a \in I$ if and only if $b = a+c$ for some $c \in I$.

As in Section 2.5.1 we define an addition and multiplication on these classes by

$$[a]_I + [b]_I = [a+b]_I$$
$$[a]_I \cdot [b]_I = [ab]_I.$$

As in Section 2.5.1 we must show that these operations are well-defined. So let $a, b, c, d \in R$ be such that $[a]_I = [c]_I$, $[b]_I = [d]_I$. Then $a = c+u$, $b = d+v$ for certain $u, v \in I$, and

$$[a+b]_I = [c+d+u+v]_I = [c+d]_I, \quad [ab]_I = [cd+cv+ud+uv]_I = [cd]_I.$$

(For the last equality note that $cv+ud+uv \in I$ and that we do not assume R to be commutative here.) It follows that the above operations are indeed well-defined. The proof of the next proposition is entirely analogous to the one of Proposition 2.5.5. Therefore we omit it.

Proposition 2.6.4 *With the operations defined above R/I is a ring. If R is commutative then so is R/I. If R has a unity 1 then $[1]_I$ is the unity of R/I.*

An important class of examples of ideals is provided by the kernels of homomorphisms. Concerning these we have the following useful fact.

Proposition 2.6.5 *Let R, S be rings and $f : R \to S$ a ring homomorphism. Then $\ker(f)$ is an ideal of R. Secondly, the map $\bar{f} : R/\ker(f) \to S$ with $\bar{f}([a]) = f(a)$ is well-defined, and an injective ring homomorphism. Thirdly, if f is surjective then $\bar{f}$ is an isomorphism.*

Proof. We write $0_R, 0_S$ for the zeros of R and S. By Lemma 2.5.10 we have $0_R \in \ker(f)$. Let $a \in R$. Then $0_S = f(a+(-a)) = f(a)+f(-a)$ implying $f(-a) = -f(a)$. So for $a, b \in R$ we see that $f(a-b) = f(a+(-b)) = f(a)+f(-b) = f(a)-f(b)$. Hence if $a, b \in \ker(f)$ then also $a-b \in \ker(f)$. Let $a \in \ker(f)$ and $b \in R$. Then $f(ab) = f(a)f(b) = 0_S f(b) = 0_S$, and analogously, $f(ba) = 0_S$. Therefore $ab, ba \in \ker(f)$ and we conclude that $\ker(f)$ is an ideal of R.

We now show that $\bar{f}$ is well-defined. Let $a, b \in R$ be such that $[a] = [b]$. Then $a = b+c$ for a certain $c \in \ker(f)$. Hence $f(a) = f(b+c) = f(b)+f(c) = f(b)+0_S = f(b)$.

The fact that $\bar{f}$ is a ring homomorphism now follows immediately from the fact that f is one. For example, $\bar{f}([a]+[b]) = \bar{f}([a+b]) = f(a+b) = f(a)+f(b) = \bar{f}([a])+\bar{f}([b])$. Similarly we see that $\bar{f}([a][b]) = \bar{f}([a])\bar{f}([b])$.

If $\bar{f}([a]) = \bar{f}([b])$ then $f(a) = f(b)$ and $f(a-b) = 0_S$, meaning that $a-b \in \ker f$ and therefore $[a] = [b]$. So $\bar{f}$ is injective.

If f is surjective, then obviously so is $\bar{f}$ (they have the same images). So in that case $\bar{f}$ is an isomorphism. □

Remark 2.6.6 The previous proposition is sometimes called the First Isomorphism Theorem. This suggests that there are more theorems of this kind, and indeed we have the following:

- (Second Isomorphism Theorem) Let R be a ring with ideals I, J. Then $I \cap J$ is an ideal of I and J is an ideal of $I+J$ and $(I+J)/J \cong I/(I \cap J)$.

- (Third Isomorphism Theorem) Let R be a ring with ideals I, J with $I \subset J$. Then J/I is an ideal in R/I and $(R/I)/(J/I) \cong R/J$.

(Compare Remark 3.4.8.)

2.6.2 What does a quotient look like?

In the previous section we defined the quotient of a ring R by an ideal I. It was rather straightforward to see that this quotient itself is a ring, with the operations induced by the operations of R. However, this construction is rather abstract and it is by no means clear what R/I looks like. (This is a general problem with quotients of algebraic structures.) We now look at an example that we already know well.

Example 2.6.7 Let $R = \mathbb{Z}$ and $I = n\mathbb{Z}$ for some $n \geq 1$. Example 2.6.3 shows that every ideal in $\mathbb{Z}$ is of this form. In Section 2.5 we have seen that for the quotient we have

$$\mathbb{Z}/n\mathbb{Z} = \{[0], \ldots, [n-1]\}.$$

Furthermore, we can compute addition and multiplication tables of $\mathbb{Z}/n\mathbb{Z}$ using this list of elements.

This example makes clear what we are after: we want an overview of the elements of R/I together with a way of adding and multiplying these elements. More precisely, we want a easily computable bijection $\psi : S \to R/I$, where S is some set, such that the inverse ψ^{-1} is also easily computable. Here we require that the set S be significantly more explicit than the set R/I itself. (We take the view that we know a more explicit set when we see one, and do not go into what we could mean by "more explicit".) If we have this, then we can just work with the set S, and the operations of addition and multiplication are performed using ψ, i.e., for $s, t \in S$ we set $s + t = \psi^{-1}(\psi(s) + \psi(t))$, $st = \psi^{-1}(\psi(s)\psi(t))$.

Example 2.6.8 In the context of the previous example we can set $S = \{0, 1, \ldots, n-1\}$, and define ψ by $\psi(i) = [i]$. Then $\psi^{-1}([a]) = a \bmod n$, were the latter denotes the unique integer in S congruent to a modulo n.

Unfortunately, it is not always easy to find a good candidate for S. When R is a polynomial ring we can do something similar to the previous example. So let F be a field, and consider the ring $F[x]$ of polynomials in the indeterminate x with coefficients in F. Let $f \in F[x]$ and set $I = \langle f \rangle$, that is

$$I = \{gf \mid g \in F[x]\}.$$

(In a completely analogous way as for $\mathbb{Z}$ it can be shown that every ideal of $F[x]$ is of this form; see also Proposition 2.6.16). Let $n = \deg(f)$ and set

$$S = \{h \in F[x] \mid \deg(h) < n\}$$

and define $\psi : S \to F[x]/I$ by $\psi(h) = [h]_I$. We first show that ψ is bijective. Firstly, if $\psi(h_1) = \psi(h_2)$ then $[h_1]_I = [h_2]_I$ implying that f divides $h_1 - h_2$, but because the degree of the h_i are at most $n-1$ that means $h_1 - h_2 = 0$. Hence ψ is injective. Secondly, let $[g]_I \in F[x]/I$. There are $q, r \in F[x]$ with $g = qf + r$ and $\deg(r) < \deg(f)$ (Proposition 2.3.8). But then $r \in S$ and $[g]_I = [r]_I$ so that $[g]_I = \psi(r)$. This also immediately gives us a construction of the inverse of ψ: let $[g]_I \in F[x]/I$ and q, r as above. Then $\psi^{-1}([g]_I) = r$.

Example 2.6.9 Let $F = \mathbb{R}$ and $f = x^2 + 1$, $I = \langle f \rangle$. Then $S = \{a + bx \mid a, b \in \mathbb{R}\}$. Addition and multiplication on S are defined using ψ, as shown above. This immediately gives $(a+bx)+(c+dx) = (a+c)+(b+d)x$. For the multiplication we note that $[a + bx]_I[c + dx]_I = [ac + (ad + bc)x + bdx^2]_I$. The latter polynomial we have to divide by f. But that is easy in this case, we have $ac + (ad + bc)x + bdx^2 = bd(x^2 + 1) + ac - bd + (ad + bc)x$. It follows that elements of S are multiplied like $(a+bx)(c+dx) = (ac-bd)+(ad+bc)x$. With these definitions of addition and multiplication S is a

ring isomorphic to R/I. In this case, the operations remind us strongly of the operations of $\mathbb{C}$. And indeed, if we define $f : S \to \mathbb{C}$, $f(a+bx) = a+bi$ then it is easily seen that f is an isomorphism. It follows that $\mathbb{R}[x]/\langle f\rangle$ is a ring isomorphic to $\mathbb{C}$ (so, in fact, it is a field).

We can also show the latter isomorphism a bit differently. Define $\sigma : \mathbb{R}[x] \to \mathbb{C}$ by $\sigma(h) = h(i)$. Then it is obvious that σ is a ring homomorphism. (In the notation of Section 2.3.5 we have $\sigma = \varphi_i$.) It is also obvious that σ is surjective. Let $h \in \ker(\sigma)$. Then $0 = \sigma(h) = h(i)$. There are $q, r \in \mathbb{R}[x]$ with $h = q(x^2+1) + r$ and $\deg(r) < 2$. Then $0 = h(i) = q(i)(i^2+1) + r(i)$, so $r(i) = 0$. But because $\deg(r) \leq 1$ this means that $r = 0$. This shows that $\ker(\sigma) = \langle x^2+1\rangle$. Hence by Proposition 2.6.5 we conclude that $\mathbb{R}[x]/I$ is isomorphic to $\mathbb{C}$.

Example 2.6.10 Consider the ring $\mathbb{Z}[i]$ (Section 2.4.1). Let $I = \langle 1+2i\rangle$ be the ideal generated by $1+2i$. What does $\mathbb{Z}[i]/I$ look like? In order to get some information on this we first compute some more elements of I. For example:

$$i(1+2i) = -2+i, \quad 1+2i-2(-2+i) = 5.$$

So $-2+i$ and 5 lie in I. It follows that $[a+bi]_I = [k]_I$, where $k = 0,1,2,3,4$. Indeed, as $a+bi = a+2b+b(-2+i)$ we see that $[a+bi]_I = [a+2b]_I$. Furthermore, $a+2b = q\cdot 5 + k$ where $0 \leq k \leq 4$, so that $[a+2b]_I = [k]_I$. On the other hand, the classes $[k]_I$ for $0 \leq k \leq 4$ are all different. Indeed, if two of them were equal then we would have $l \in I$ for some l with $1 \leq l \leq 4$. Then, as $\gcd(5,l) = 1$ there are integers u, v with $ul + 5v = 1$ and we see $1 \in I$. But the latter would mean that $1+2i$ is invertible, which it is not. So we see that for our set S we can take $S = \{0,1,2,3,4\}$. The map ψ is $\psi(k) = [k]_I$ and $\psi^{-1}([a+bi]_I) = (a+2b) \bmod 5$. Hence $\mathbb{Z}[i]/I \cong \mathbb{Z}/5\mathbb{Z}$.

This argument also suggests another approach, analogous to the second one of the previous example. Define $\sigma : \mathbb{Z}[i] \to \mathbb{Z}/5\mathbb{Z}$ by $\sigma(a+bi) = [a+2b]_5$. A small computation shows that this is a ring homomorphism. As $1+2i \in \ker(\sigma)$ we have $I \subset \ker(\sigma)$ (indeed, any element of I can be written $\eta(1+2i)$ and $\sigma(\eta(1+2i)) = \sigma(\eta)\sigma(1+2i) = \sigma(\eta)\cdot 0 = 0$). Conversely, if $a+bi \in \ker(\sigma)$ then $[a+2b]_5 = [0]_5$, or $a+2b = 5l$ for some $l \in \mathbb{Z}$ and $a+bi = -2b+5l+bi = b(-2+i) + l\cdot 5$ which lies in I. It follows that $\ker(\sigma) = I$ and by Proposition 2.6.5, $\mathbb{Z}[i]/I \cong \mathbb{Z}/5\mathbb{Z}$.

2.6.3 Sums and products of ideals and the Chinese remainder theorem

The notion of ideal started as "ideal number"; so it is natural to have arithmetic operations, sum and product, on ideals. Let R be a ring and I, J ideals of R. Then the sum, respectively the product, of I, J is defined to be the smallest ideal containing $a+b$ for all $a \in I$, $b \in J$, respectively the smallest ideal containing ab for all $a \in I$, $b \in J$. For the sum this is easy because the set $\{a+b \mid a \in I, b \in J\}$ is easily seen to be an ideal of R. So we define

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

For the product the situation is a bit more tricky because in general the set $\{ab \mid a \in I, b \in J\}$ is not an ideal of R. The problem is that it is not closed under addition. We get around this problem by taking the set consisting of all finite sums of elements of the form ab for $a \in I$, $b \in J$, that is

$$IJ = \{\sum_{i=1}^{m} a_i b_i \mid m \geq 1, a_i \in I, b_i \in J\}.$$

It is immediate that this again is an ideal of R.

Example 2.6.11 In Example 2.6.3 we have seen that all ideals of $\mathbb{Z}$ are of the form $n\mathbb{Z}$ for some $n \in \mathbb{Z}$. So let $m\mathbb{Z}$, $n\mathbb{Z}$ be two ideals of $\mathbb{Z}$. We claim that $m\mathbb{Z} + n\mathbb{Z} = d\mathbb{Z}$ where $d = \gcd(m,n)$. Indeed, since $d|m$ we have $m\mathbb{Z} \subset d\mathbb{Z}$, and similarly, $n\mathbb{Z} \subset d\mathbb{Z}$. Hence $m\mathbb{Z} + n\mathbb{Z} \subset d\mathbb{Z}$. Conversely, there are $u, v \in \mathbb{Z}$ with $um + vn = d$ (Theorem 2.2.3). So for $a \in \mathbb{Z}$ we have $da = mua + nva \in m\mathbb{Z} + n\mathbb{Z}$. Hence we have $d\mathbb{Z} \subset m\mathbb{Z} + n\mathbb{Z}$ as well.

For the product we have $(m\mathbb{Z})(n\mathbb{Z}) = mn\mathbb{Z}$. This is straightforward to see. Let ma_i, nb_i be elements of $m\mathbb{Z}$, $n\mathbb{Z}$ respectively. Then $\sum_i (ma_i)(nb_i) = mn(\sum_i a_i b_i) \in mn\mathbb{Z}$. Hence $(m\mathbb{Z})(n\mathbb{Z}) \subset mn\mathbb{Z}$. Conversely, let $mna \in mn\mathbb{Z}$. We write $mna = (m)(na)$ showing that $mna \in (m\mathbb{Z})(n\mathbb{Z})$.

Definition 2.6.12 *Let R be a ring and I, J ideals of R. Then I, J are said to be* coprime *if $I+J = R$.*

Example 2.6.13 Using Example 2.6.11 we see that the ideals $m\mathbb{Z}$, $n\mathbb{Z}$ are coprime if and only if $\gcd(m, n) = 1$. But that is the same as saying that m, n are coprime.

Theorem 2.6.14 *Let R be a commutative ring with unity 1. Let $I, J \subset R$ be coprime ideals. Then $IJ = I \cap J$ and the map*

$$\sigma : R/IJ \to R/I \times R/J$$

with $\sigma([a]_{IJ}) = ([a]_I, [a]_J)$ is well-defined and an isomorphism of rings.

Proof. Since $I + J = R$ there are $x_0 \in I$, $y_0 \in J$ with $x_0 + y_0 = 1$.

We show that $IJ = I \cap J$. Let $a \in I$, $b \in J$; then $ab \in I$ (as $a \in I$) and $ab \in J$ (as $b \in J$). Hence $ab \in I \cap J$ and therefore $IJ \subset I \cap J$. Let $a \in I \cap J$. Then $a = a \cdot 1 = ax_0 + ay_0$. Now $ax_0 = x_0a$ and hence it lies in IJ. Secondly, $ay_0 \in IJ$. It follows that $a \in IJ$ and we have $I \cap J \subset IJ$; showing that we have equality.

Let $a, b \in R$ be such that $[a]_{IJ} = [b]_{IJ}$. Then $b - a \in IJ$ so that $b - a \in I \cap J$ and therefore $[a]_I = [b]_I$, $[a]_J = [b]_J$. We have shown that σ is well-defined.

The fact that σ is a homomorphism of rings follows directly from the definitions of the arithmetic operations in the rings involved.

Now we show that σ is surjective. We note that $[x_0]_I = [0]_I$ and $[x_0]_J = [1 - y_0]_J = [1]_J$ and similarly, $[y_0]_I = [1]_I$, $[y_0]_J = [0]_J$. Let $a, b \in R$ then

$$\sigma([ay_0 + bx_0]_{IJ}) = ([ay_0 + bx_0]_I, [ay_0 + bx_0]_J) = ([a]_I[y_0]_I + [b]_I[x_0]_I, [a]_J[y_0]_J + [b]_J[x_0]_J) = ([a]_I, [b]_J).$$

We determine the kernel of σ. Suppose that $\sigma([a]_{IJ}) = ([0]_I, [0]_J)$. Then $[a]_I = [0]_I$ and $[a]_J = [0]_J$ and thus $a \in I$, $a \in J$. Hence $a \in I \cap J = IJ$. It follows that $\ker(\sigma) = \{[0]_{IJ}\}$. Therefore σ is injective as well. □

Example 2.6.15 Let $R = \mathbb{R}[x]$ and $I = \langle x^3 + x\rangle$. Set $J_1 = \langle x\rangle$ and $J_2 = \langle x^2 + 1\rangle$. We leave it as an exercise to show that $I = J_1J_2$. Since $x^2 + 1 + (-x)x = 1$ we see that $J_1 + J_2 = \mathbb{R}[x]$. So by the Chinese remainder theorem

$$\mathbb{R}[x]/I \cong \mathbb{R}[x]/\langle x\rangle \times \mathbb{R}[x]/\langle x^2 + 1\rangle.$$

Using Example 2.6.9 we see that the latter is isomorphic to $\mathbb{R} \times \mathbb{C}$.

2.6.4 Principal ideal domains and prime and maximal ideals

A domain D is said to be a *principal ideal domain* if all of its ideals are principal, that is, every ideal is of the form

$$\langle a\rangle = \{ba \mid b \in D\}.$$

In Example 2.6.3 we have seen that $\mathbb{Z}$ is a principal ideal domain. We will first generalize the proof in that example to Euclidean domains. Then we will study prime and maximal ideals. Finally we establish a link between the theory of ideals in rings and the question of unique factorization: we show that a principal ideal domain has unique factorization into irreducibles.

Proposition 2.6.16 *Let D be a Euclidean domain. Then D is a principal ideal domain.*

Proof. Let $\nu : D \setminus \{0\} \to \mathbb{Z}_{\geq 0}$ be as in Definition 2.4.18. Let $I \subset D$ be an ideal of D. If $I = \{0\}$ then $I = \langle 0\rangle$. Otherwise I contains non-zero elements. Let $S = \{\nu(a) \mid a \in I \setminus \{0\}\}$. Then S has a minimal element m_0 and let $b \in I$, $b \neq 0$ be such that $\nu(b) = m_0$. Let $a \in I$, $a \neq 0$. Then there are $q, r \in D$ with $a = qb + r$ and $r = 0$ or $\nu(r) < \nu(b)$. Since $r = a - qb$ we have $r \in I$. So by the choice of b it follows that $r = 0$ and $a = qb$. Hence $I \subset \langle b\rangle$. But as $b \in I$ we also have $\langle b\rangle \subset I$ and hence $I = \langle b\rangle$. □

Example 2.6.17 Consider the ring $\mathbb{Z}[\sqrt{-3}]$. As seen in Example 2.4.30 this ring is not Euclidean. It is not a principal ideal domain either. To see that define

$$I = \langle 2, 1+\sqrt{-3}\rangle = \{\alpha \cdot 2 + \beta \cdot (1+\sqrt{-3}) \mid \alpha, \beta \in \mathbb{Z}[\sqrt{-3}]\}.$$

Suppose that there exists a $\gamma \in \mathbb{Z}[\sqrt{-3}]$ with $I = \langle\gamma\rangle$. Then $2 = \eta\gamma$ for a certain $\eta \in \mathbb{Z}[\sqrt{-3}]$. But in Example 2.4.17 we have seen that 2 is irreducible. Hence η or γ is invertible. If γ is invertible then $1 \in I$. So in that case there are $\alpha, \beta \in \mathbb{Z}[\sqrt{-3}]$ with $\alpha \cdot 2 + \beta \cdot (1+\sqrt{-3}) = 1$. Writing $\alpha = a + b\sqrt{-3}$, $\beta = c + d\sqrt{-3}$ we see that this amounts to

$$\begin{aligned} 2a + c - 3d &= 1 \\ 2b + c + d &= 0. \end{aligned}$$

Subtracting the second equation from the first we get $2(a-b) - 4d = 1$. But here the left hand side is even, and the right hand side is odd. It follows that γ is not invertible. So η is invertible and hence $\eta = \pm 1$ (see Example 2.4.2). Hence $\gamma = \pm 2$. But 2 does not divide $1 + \sqrt{-3}$. So we have reached a contradiction, and we conclude that I is not principal.

In this example we see that ideals being principal or not could have something to do with unique factorization, as the elements 2 and $1+\sqrt{-3}$ are also used to show that $\mathbb{Z}[\sqrt{-3}]$ does not have unique factorization (Example 2.4.17).

Definition 2.6.18 *Let R be a commutative ring with unity 1. An ideal I of A is called* prime *if $I \neq R$ and for all $a, b \in R$ with $ab \in I$ we have $a \in I$, or $b \in I$.*

Example 2.6.19 Let $p \in \mathbb{Z}$ be prime, then $p\mathbb{Z}$ is a prime ideal of $\mathbb{Z}$. Indeed, $1 \notin p\mathbb{Z}$ so $p\mathbb{Z} \neq \mathbb{Z}$. Secondly, if $ab \in p\mathbb{Z}$ then $ab = pc$ for some $c \in R$, that is, p divides ab. As p is prime we must have $p|a$ or $p|b$. In the first case $a \in p\mathbb{Z}$. In the second case $b \in p\mathbb{Z}$.

The ideal $6\mathbb{Z}$ is not a prime ideal because $6 \in 6\mathbb{Z}$, but neither 2 nor 3 lie in $6\mathbb{Z}$.

Proposition 2.6.20 *Let R be a commutative ring with unity 1 and let $I \subset R$ be an ideal. Then I is prime if and only if R/I is a domain.*

Proof. Suppose that I is prime. By Proposition 2.6.4 we already know that R/I is a commutative ring with unity $[1]$. In order to prove that it is a domain we must show that $[1] \neq [0]$ and that R/I has no zero divisors. Firstly, $[1] = [0]$ is the same as $1 \in I$ or $I = R$; but because I is prime, this is not the case. Secondly, let $[a], [b] \in R/I$ be such that $[a][b] = [0]$. Then $ab \in I$ so that $a \in I$ or $b \in I$ as I is prime. But $a \in I$ is the same as $[a] = [0]$ and $b \in I$ is the same as $[b] = [0]$. So one of $[a], [b]$ is $[0]$. Hence there are no zero divisors in R/I.

Now suppose that R/I is a domain. In particular that means that $[1] \neq [0]$, so that $I \neq R$. Let $a, b \in R$ be such that $ab \in I$. Then $[ab] = [0]$. But $[ab] = [a][b]$ and because R/I has no zero divisors, it follows that $[a] = [0]$ or $[b] = [0]$. In other words, $a \in I$ or $b \in I$. We conclude that I is prime. □

Definition 2.6.21 *Let R be a commutative ring with unity 1. An ideal I of A is called* maximal *if $I \neq R$ and the only ideals J of R with $I \subset J \subset R$ are $J = I$ and $J = R$.*

Example 2.6.22 Let $p \in \mathbb{Z}$ be prime, then $p\mathbb{Z}$ is a maximal ideal of $\mathbb{Z}$. Indeed, $1 \notin p\mathbb{Z}$ so $p\mathbb{Z} \neq \mathbb{Z}$. Secondly, let $J \subset \mathbb{Z}$ be an ideal with $p\mathbb{Z} \subset J \subset \mathbb{Z}$. By Example 2.6.3 there is a $d \in \mathbb{Z}$ with $J = d\mathbb{Z}$. Because $p\mathbb{Z} \subset J$ we have that $p \in d\mathbb{Z}$, or $p = da$ for some $a \in \mathbb{Z}$, or $d|p$. As p is prime this can only happen if $d = \pm 1$ or $d = \pm p$. In the first case $J = \mathbb{Z}$. In the second case $J = p\mathbb{Z}$.

The ideal $6\mathbb{Z}$ is not a maximal ideal because $6\mathbb{Z} \subsetneq 2\mathbb{Z} \subsetneq \mathbb{Z}$.

Comparing this with Example 2.6.19 and Definition 2.4.4, we see that when proving that $p\mathbb{Z}$ is a prime ideal we use the fact that p is prime, whereas when proving that $p\mathbb{Z}$ is maximal we use the fact that p is irreducible.

Proposition 2.6.23 *Let R be a commutative ring with unity 1 and let $I \subset R$ be an ideal. Then I is maximal if and only if R/I is a field.*

Proof. Suppose that I is maximal. By By Proposition 2.6.4 we already know that R/I is a commutative ring with unity $[1]$. In order to prove that it is a field we must show that $[1] \neq [0]$ and that for $[a] \in R/I$ with $[a] \neq [0]$ we have that there is a $[b] \in R/I$ with $[a][b] = [1]$. As in the proof of Proposition 2.6.20 we have that $[1] \neq [0]$ follows from $I \neq R$. Furthermore, $[a] \neq [0]$ is the same as $a \notin I$. Set $J = I + \langle a \rangle = \{c + ab \mid c \in I, b \in R\}$, which is an ideal of R. It strictly contains I as $a \notin I$. Thus $J = R$, so that there is a $c \in I$ and $b \in R$ with $c + ab = 1$. But that means $[a][b] = [ab] = [1 - c] = [1]$ and we are done.

Now suppose that R/I is a field. Then $[1] \neq [0]$ so that $I \neq R$. Let J be an ideal of R with $I \subset J \subset R$ and suppose that $J \neq I$. Let $a \in J \setminus I$. As R/I is a field, there is a $b \in R$ with $[a][b] = [1]$, or $ab = 1 + c$ for some $c \in I$. But $a \in J$ implies $ab \in J$, so $ab - c \in J$ and we see that $1 \in J$. That means $J = R$. □

Corollary 2.6.24 *Let R be a commutative ring with unity. Then every maximal ideal of R is prime.*

Proof. This follows from the previous propositions, along with the observation that a field is a domain. □

Example 2.6.25 Let $I = \langle x^2 + 1 \rangle \subset \mathbb{R}[x]$. Then I is maximal as $\mathbb{R}[x]/I \cong \mathbb{C}$ is a field (Example 2.6.9).

Now we show that principal ideal domains have unique factorization. This gives another proof of the fact that Euclidean domains have unique factorization. But there are principal ideal domains that are not Euclidean. So this extends the range of domains where we know that unique factorization holds.

Lemma 2.6.26 *Let R be a principal ideal domain. Let $a \in R$, $a \neq 0$; then the following are equivalent:*

(i) *$\langle a \rangle$ is a maximal ideal of R.*

(ii) *$\langle a \rangle$ is a prime ideal of R.*

(iii) *a is prime.*

(iv) *a is irreducible.*

Proof. From Corollary 2.6.24 it immediately follows that (i) implies (ii).

Suppose that $\langle a \rangle$ is a prime ideal. Then a is not invertible as otherwise $\langle a \rangle = R$. Let $b, c \in R$ be such that $a|bc$. Then $bc \in \langle a \rangle$. Hence $b \in \langle a \rangle$ or $c \in \langle a \rangle$. The first possibility means that $a|b$, and from the second possibility we have $a|c$. We conclude that a is prime. So (ii) implies (iii).

In general we have that prime elements are irreducible (this was noted below Definition 2.4.4). So it remains to show that (iv) implies (i). We suppose that a is irreducible. Then in particular a is not invertible so that $\langle a \rangle \neq R$. Let J be an ideal of R with $\langle a \rangle \subset J \subset R$. Because R is a principal ideal domain there is a $b \in R$ with $J = \langle b \rangle$. As $a \in J$ we have $a = bc$ for a certain $c \in R$. But a is irreducible, so this means that b or c is invertible. If b is invertible then $J = R$. If c is invertible then $J = \langle a \rangle$. We conclude that $\langle a \rangle$ is maximal. □

Theorem 2.6.27 *Let R be a principal ideal domain. Then R is a unique factorization domain (see Definition 2.4.11).*

Proof. Define

$$S = \{a \in R \mid a \neq 0 \text{ and } a \text{ not invertible and } a \text{ is not a product of irreducibles}\}.$$

Suppose that $S \neq \emptyset$. For $a \in S$ we write $I_a = \langle a \rangle$. We claim that there exists an $a \in S$ such that I_a is not strictly contained in any I_b for $b \in S$. Indeed, suppose that there is no such $a \in S$. Let $a_1 \in S$. Then there exists $a_2 \in S$ with $I_{a_1} \subsetneq I_{a_2}$. But there is also an $a_3 \in S$ with $I_{a_2} \subsetneq I_{a_3}$. So there is an infinite series $a_1, a_2, \ldots$ with $I_{a_i} \subsetneq I_{a_{i+1}}$. Set

$$J = \bigcup_{n \geq 1} I_{a_n}.$$

Then J is an ideal of R. Hence there is a $b \in R$ with $J = \langle b \rangle$. By the definition of J there is an $n \geq 1$ with $b \in I_{a_n}$. But then $J \subset I_{a_n}$. Again using the definition of J we then see that $I_{a_m} \subset J \subset I_{a_n}$ for all m. But this is a contradiction, and the claim is proved.

Now let $a \in S$ be such that I_a is not strictly contained in any I_b for $b \in S$. As $a \in S$ it is not irreducible. Hence $a = bc$ with $b, c \in R$ not invertible. Therefore $I_a \subset I_b$ and $I_a \subset I_c$. If $I_a = I_b$ then $b = au$ for a certain $u \in R$. Then $a = bc = auc$ and as R is a domain, $uc = 1$ and c is invertible. We know that c is not invertible, and therefore $I_a \subsetneq I_b$. In the same way it is seen that $I_a \subsetneq I_c$. By the choice of a that implies that $b, c \notin S$. Hence b, c are products of irreducibles. Therefore so is a, and we have a contradiction showing that $S = \emptyset$.

The uniqueness of the factorization follows from Lemma 2.6.26 together with Theorem 2.4.15. □

Example 2.6.28 As in Example 2.4.31 we consider the ring $\mathbb{Z}[\alpha]$ with $\alpha = \frac{1+\sqrt{-19}}{2}$ which is a zero of $f = x^2 - x + 5$. In the mentioned example it is shown that $\mathbb{Z}[\alpha]$ is not Euclidean. Here we will show that it is a principal ideal domain. So by Theorem 2.6.27 it has unique factorization after all.

As seen in Example 2.4.31 we have $N(a + b\alpha) = a^2 + ab + 5b^2$. Furthermore, since $\alpha \notin \mathbb{R}$ we also have that N coincides with the square of the complex norm, i.e., if we write $a + b\alpha = u + vi$ then $N(a + b\alpha) = u^2 + v^2$. (This is clear because in this case $\bar{\alpha}$ is the complex conjugate of α.)

Let $I \subset \mathbb{Z}[\alpha]$ be an ideal. Set $S = \{N(\eta) \mid \eta \in I, \eta \neq 0\}$. Let m_0 be the minimal element of S, and let $\eta_0 \in I$ be such that $N(\eta_0) = m_0$. We will show that $I = \langle \eta_0 \rangle$. For that we set

$$M = \eta_0^{-1} I = \{\eta_0^{-1} \xi \mid \xi \in I\} \subset \mathbb{Q}(\alpha).$$

We prove two statements on M:

1. *Let $\theta \in M$ have the property that there exists a $\zeta \in \mathbb{Z}[\alpha]$ with $N(\theta - \zeta) < 1$. Then $\theta \in \mathbb{Z}[\alpha]$.* Write $\theta = \eta_0^{-1}\xi$ for some $\xi \in I$. Then $N(\theta - \zeta) = N(\eta_0^{-1})N(\xi - \eta_0\zeta) = \frac{N(\xi - \eta_0\zeta)}{N(\eta_0)}$. If this is < 1 then $N(\eta_0) > N(\xi - \eta_0\zeta)$. So by the choice of η_0 we see that $\xi - \eta_0\zeta = 0$.

2. *Let $\theta \in M$, $\theta \notin \mathbb{Z}[\alpha]$ and write $\theta = u + vi$ where $u \in \mathbb{Q}$, $v \in \mathbb{R}$. Then $|v - m\frac{\sqrt{19}}{2}| \geq \frac{\sqrt{3}}{2}$ for all $m \in \mathbb{Z}$.* Suppose that $|v - m\frac{\sqrt{19}}{2}| < \frac{\sqrt{3}}{2}$ for a certain $m \in \mathbb{Z}$. Choose $d \in \mathbb{Z}$ such that $|d + \frac{m}{2} - u| \leq \frac{1}{2}$. Set $\zeta = d + m\alpha$ then $\zeta - \theta = (d + \frac{m}{2} - u) + (m\frac{\sqrt{19}}{2} - v)i$. Hence $N(\zeta - \theta) = (d + \frac{m}{2} - u)^2 + (v - m\frac{\sqrt{19}}{2})^2 < \frac{1}{4} + \frac{3}{4} = 1$. By 1. it now follows $\theta \in \mathbb{Z}[\alpha]$, which is a contradiction.

Let $\zeta \in \mathbb{Z}[\alpha]$, then $\eta_0\zeta \in I$ so that $\zeta = \eta_0^{-1}(\eta_0\zeta)$ and we see that $\zeta \in M$. Hence $\mathbb{Z}[\alpha] \subset M$.

We claim that $M = \mathbb{Z}[\alpha]$. So suppose that there is a $\theta_1 \in M$, but $\theta_1 \notin \mathbb{Z}[\alpha]$. Write $\theta_1 = u_1 + v_1 i$ where $u_1 \in \mathbb{Q}$, $v_1 \in \mathbb{R}$. Let $f : \mathbb{Z} \to \mathbb{R}$ be given by $f(m) = v_1 - m\frac{\sqrt{19}}{2}$. Let $S = \{m \in \mathbb{Z} \mid f(m) \geq 0\}$ and let $m_0 \in S$ be such that $f(m) \geq f(m_0)$ for all $m \in S$ (it is clear that such an m_0 exists as $\lim_{m \to -\infty} f(m) = \infty$). Then by 2. we see that $v_1 - m_0\frac{\sqrt{19}}{2} \geq \frac{\sqrt{3}}{2}$. We also claim that $v_1 - m_0\frac{\sqrt{19}}{2} \leq \frac{\sqrt{19}}{2} - \frac{\sqrt{3}}{2}$. Indeed, suppose that $v_1 - m_0\frac{\sqrt{19}}{2} > \frac{\sqrt{19}}{2} - \frac{\sqrt{3}}{2}$. Then $v_1 - (m_0 + 1)\frac{\sqrt{19}}{2} > -\frac{\sqrt{3}}{2}$. But by 2. we have that $v_1 - (m_0 + 1)\frac{\sqrt{19}}{2}$ does not lie in the interval $(-\frac{\sqrt{3}}{2}, \frac{\sqrt{3}}{2})$. Hence $v_1 - (m_0 + 1)\frac{\sqrt{19}}{2} \geq \frac{\sqrt{3}}{2}$. So also $m_0 + 1 \in S$, but $v_1 - (m_0 + 1)\frac{\sqrt{19}}{2} < v_1 - m_0\frac{\sqrt{19}}{2}$, contrary to the choice of m_0.

Set $\theta_2 = \theta_1 - m_0\alpha$. Then $\theta_2 = u_2 + v_2 i$ with $v_2 \in [\frac{\sqrt{3}}{2}, \frac{\sqrt{19}}{2} - \frac{\sqrt{3}}{2}]$. Next, let $m_1 \in \mathbb{Z}$ be such that $u_2 + m_1 \in (-\frac{1}{2}, \frac{1}{2}]$ and set $\theta_3 = \theta_2 + m_1$. Note that also $\theta_2, \theta_3 \in M$ but both do not lie in $\mathbb{Z}[\alpha]$.

Write $\theta = \theta_3 = u + vi$. Then $\sqrt{3} \leq 2v \leq \sqrt{19} - \sqrt{3}$ and hence

$$-\frac{\sqrt{19}}{2} + \sqrt{3} \leq 2v - \frac{\sqrt{19}}{2} \leq \frac{\sqrt{19}}{2} - \sqrt{3}.$$

As $\frac{\sqrt{19}}{2} - \sqrt{3} < \frac{\sqrt{3}}{2}$ we infer from 2. that $2\theta \in \mathbb{Z}[\alpha]$. So we can write $2\theta = a + b\alpha$ with $a, b \in \mathbb{Z}$. Hence $2\theta = a + \frac{b}{2} + b\frac{\sqrt{19}}{2}i$. So $v = \frac{b}{2}\frac{\sqrt{19}}{2}$ and as seen above it lies in $[\frac{\sqrt{3}}{2}, \frac{\sqrt{19}}{2} - \frac{\sqrt{3}}{2}]$. By approximating the square roots involved, it immediately follows that $b = 1$ is the only possibility. But also $v = \frac{a}{2} + \frac{1}{4}$ lies in $(-\frac{1}{2}, \frac{1}{2}]$. Hence $a = -1$ or $a = 0$. If $a = 0$ then $2\theta = \alpha$ and hence $\frac{\alpha}{2} \in M$. If $a = -1$ then $2\theta = \bar{\alpha}$ and we reach the same conclusion.

Then also $\bar{\alpha}\frac{\alpha}{2} \in M$. But the latter is equal to $\frac{5}{2}$. So by 2. it lies in $\mathbb{Z}[\alpha]$. But it obviously does not. Therefore we have reached a contradiction, and conclude that $M = \mathbb{Z}[\alpha]$. Finally, this immediately implies that $I = \langle \eta_0 \rangle$.

2.7 Applications of unique factorization in Euclidean domains

Here we show some theorems whose proofs use the property of unique factorization in a Euclidean domain. They are all concerned with finding integer solutions to *Diophantine equations*: a polynomial $f \in \mathbb{Z}[x_1, \ldots, x_n]$ is given and the question is to find all $(k_1, \ldots, k_n) \in \mathbb{Z}^n$ such that $f(k_1, \ldots, k_n) = 0$ (or to prove that no such solutions exist).

First we have the following fundamental lemma.

Lemma 2.7.1 *Let D be a Euclidean domain. Let $\zeta, \eta \in D$ be coprime (that is, having greatest common divisor 1) and such that $\zeta\eta = \theta^k$ for a certain $\theta \in D$ and $k \geq 1$. Then there are invertible elements $u, v \in D$ and $x, y \in D$ such that $\zeta = ux^k$, $\eta = vy^k$.*

Proof. Let $\zeta = p_1 \cdots p_l$, $\eta = q_1 \cdots q_m$, $\theta = r_1 \cdots r_n$ be the factorizations of ζ, η, θ in irreducibles. Then $\zeta\eta = \theta^k$ translates to

$$p_1 \cdots p_l \cdot q_1 \cdots q_m = r_1^k \cdots r_n^k.$$

This gives two factorizations into irreducibles of the same element. So we see that each r_i is associated with a p_s or with a q_t, but not with both (as ζ and η are coprime). Furthermore, if r_i is associated with a p_s then there exist $p_{s_2}, \ldots, p_{s_k}$ such that r_i is associated with p_{s_j}, $2 \leq j \leq k$. Conversely, each p_s is associated with an r_i. It follows that we can reorder the factors of ζ such that

$$\zeta = p_1 \cdots p_l = (p_{1,1} \cdots p_{1,k}) \cdots (p_{a,1} \cdots p_{a,k}),$$

where each $p_{s,j}$ is associated with the same r_{i_s}. So there are invertible elements $\epsilon_{s,j}$ such that $p_{s,j} = \epsilon_{s,j} r_{i_s}$ for $1 \leq s \leq a$, $1 \leq j \leq k$. Let u denote the product of all $\epsilon_{s,j}$. Then

$$\zeta = u r_{i_1}^k \cdots r_{i_a}^k = u(r_{i_1} \cdots r_{i_a})^k.$$

The argument for η is completely analogous. □

Corollary 2.7.2 *Let $a, b, c \in \mathbb{Z}$ be such that $\gcd(a, b) = 1$ and $ab = c^3$. Then there are $u, v \in \mathbb{Z}$ with $a = u^3$, $b = v^3$.*

Proof. By the previous lemma there are $u', v' \in \mathbb{Z}$ such that $a = \pm(u')^3$, $b = \pm(v')^3$ (because the invertible elements of $\mathbb{Z}$ are ± 1). But if $a = -(u')^3$ then also $a = (-u')^3$ and we set $u = -u'$. We do a similar thing for b. □

We say that a triple $(a, b, c) \in \mathbb{Z}^3$ is *coprime* if $\gcd(a, b) = \gcd(a, c) = \gcd(b, c) = 1$.

2.7.1 Pythagorean triples

A triple $(a,b,c) \in \mathbb{Z}^2$ is said to be *Pythagorean* if $a^2+b^2=c^2$. (Of course the name stems from the fact that such a triple consists of the sides of a right triangle.) A Pythagorean triple is called *primitive* if (a,b,c) is coprime. Here we will show a method to obtain all primitive Pythagorean triples.

Lemma 2.7.3 *Let (a,b,c) be a primitive Pythagorean triple. Then one of a,b is even and the other is odd.*

Proof. As $\gcd(a,b)=1$, not both a,b are even. We observe the following general fact: let $m \in \mathbb{Z}$ then $m^2 \equiv 0,1 \bmod 4$ and $m^2 \equiv 0 \bmod 4$ if and only if m is even, and $m^2 \equiv 1 \bmod 4$ if and only if m is odd. (This is proved by considering the cases $m \equiv 0,1,2,3 \bmod 4$ separately.) So if both a,b were odd, then $(a^2+b^2) \equiv 2 \bmod 4$. But no square is equal to 2 modulo 4, so we obtain a contradiction with $a^2+b^2=c^2$. □

Theorem 2.7.4 *Let (a,b,c) be a primitive Pythagorean triple with a odd and $c>0$. Then there exist $s,t \in \mathbb{Z}$ with $\gcd(s,t)=1$ and $a=s^2-t^2$, $b=2st$, $c=s^2+t^2$.*

Proof. We work with the domain $\mathbb{Z}[i]$ which is Euclidean (Example 2.4.22). In $\mathbb{Z}[i]$ we have the factorization $a^2+b^2=(a-bi)(a+bi)$. Let $\zeta \in \mathbb{Z}[i]$ be irreducible and suppose that ζ divides both $a+bi$ and $a-bi$. Then ζ divdes their sum $2a$ and ζ divides c^2. As ζ is also a prime element (see Theorem 2.4.15) it divides c. But $\gcd(a,c)=1$ and c is odd by Lemma 2.7.3, so $\gcd(2a,c)=1$ as well and hence there are integers x,y with $2ax+cy=1$. It follows that ζ divides 1, and hence that ζ is invertible, which is a contradiction. We conclude that in $\mathbb{Z}[i]$ the elements $a+bi$, $a-bi$ are coprime.

Lemma 2.7.1 now shows that there exist an invertible element $u \in \mathbb{Z}[i]$ and a $\eta \in \mathbb{Z}[i]$ with $a+bi=u\eta^2$. Write $\eta=s+ti$, then $a+bi=u((s^2-t^2)+2sti)$. The invertible elements of $\mathbb{Z}[i]$ are ± 1, $\pm i$. If $u=1$ then the above equation immediately gives $a=s^2-t^2$, $b=2st$. If $u=-1$ then set $s_0=-t$, $t_0=s$ and from $a+bi=-s^2+t^2-2sti=s_0^2-t_0^2+2s_0t_0i$ we get $a=s_0^2-t_0^2$, $b=2s_0t_0$. From $u=\pm i$ it would follow that a is even, which is excluded.

Finally, note that $\gcd(s,t)$ has to be 1, as otherwise the triple is not primitive. □

Remark 2.7.5 There are a few other ways to prove this theorem, without using the ring $\mathbb{Z}[i]$. In fact, this theorem was already known in antiquity, but the ring $\mathbb{Z}[i]$ was not. One way is to show first that for $x,y \in \mathbb{Q}$ with $x^2+y^2=1$ and $(x,y) \neq (-1,0)$ there is a $z \in \mathbb{Q}$ with

$$x=\frac{1-z^2}{1+z^2}, \quad y=\frac{2z}{1+z^2}.$$

(This can be done by intersecting the circle $x^2+y^2=1$ with the line $x=-zy+1$). Secondly, if $a^2+b^2=c^2$ then setting $x=\frac{a}{c}$, $y=\frac{b}{c}$ we find $x,y \in \mathbb{Q}$ with $x^2+y^2=1$. So we obtain a z as above, and by writing $z=\frac{s}{t}$ one finishes the proof.

2.7.2 Fermat for $n=3$

Here we show that there are no non-zero integers x,y,z with $x^3+y^3=z^3$. Traditionally, the proof of Fermat's last theorem for exponent p (i.e., that there are no non-zero integers x,y,z with $x^p+y^p=z^p$) is divided into two cases. In the *first case* one assumes that p does not divide xyz. In the *second case* it is assumed that $p|xyz$. Usually the first case is much easier. For $p=3$ it is dealt with by a little lemma.

Lemma 2.7.6 *There do not exist $x,y,z \in \mathbb{Z}$ with $3 \nmid xyz$ and $x^3+y^3=z^3$.*

Proof. Here is a table of $a^3 \bmod 9$ against $a \bmod 9$:

$a \bmod 9$	0	1	2	3	4	5	6	7	8
$a^3 \bmod 9$	0	1	−1	0	1	−1	0	1	−1

We see that $3 \not| a$ implies that $a^3 \equiv \pm 1 \bmod 9$.

Let $x, y, z \in \mathbb{Z}$ be all not divisible by 3. Then $x^3 + y^3$ is equal to one of $2, 0, -2$ modulo 9. But this can then not be equal to $z^3 \bmod 9$. □

Now we turn to the much harder second case. In the proof a central role is played by the ring $\mathbb{Z}[\omega]$ where

$$\omega = \frac{-1 + \sqrt{-3}}{2}.$$

We have that ω is a root of $f = x^2 + x + 1$. As seen in Section 2.4.1 this ring has a norm defined by $N(a + b\omega) = |a^2 - ab + b^2|$. We leave it as an exercise to verify that this norm makes $\mathbb{Z}[\omega]$ into a Euclidean ring, and that

$$\pm 1, \pm\omega, \pm(1 + \omega)$$

are the invertible elements of $\mathbb{Z}[\omega]$. Let $\bar{\omega} = \frac{-1-\sqrt{-3}}{2}$ be the other root of f. Then $\bar{\omega} = \omega^2 = -1 - \omega$. For $\eta = a + b\omega$ we define $\bar{\eta} = a + b\bar{\omega}$. As observed in Section 2.4.1 we have that $\overline{\eta\zeta} = \bar{\eta}\bar{\zeta}$ for all $\eta, \zeta \in \mathbb{Z}[\omega]$.

Instead of dealing with the second case directly we first prove something slightly different.

Proposition 2.7.7 *Let $(x, y, w) \in \mathbb{Z}^3$ be coprime such that $3 \not| xyw$. Then $x^3 + y^3 \neq 3^{3m}w^3$ for all $m \geq 0$.*

Proof. The proof is by induction on m. The case $m = 0$ is dealt with by Lemma 2.7.6. So let $m \geq 1$. The induction hypothesis is that for all coprime triples $(p, q, r) \in \mathbb{Z}^3$ with $3 \not| pqr$ we have $p^3 + q^3 \neq 3^{3(m-1)}r^3$. Under this assumption we prove the statement for m. So suppose that we have a coprime triple $(x, y, w) \in \mathbb{Z}^3$ with $3 \not| xyw$ and $x^3 + y^3 = 3^{3m}w^3$. It is our aim to derive a contradiction. We divide the rest of the proof in a number of steps.

1. Consider the factorization
$$x^3 + y^3 = (x + y)(x^2 - xy + y^2).$$
By Fermat's little theorem (Theorem 2.5.22) we have $a^3 \equiv a \bmod 3$ for all $a \in \mathbb{Z}$. Therefore $x + y \equiv x^3 + y^3 \bmod 3 \equiv 3^{3m}w^3 \bmod 3 \equiv 0 \bmod 3$. It follows that $y = -x + 3k$ for some $k \in \mathbb{Z}$. But then $x^2 - xy + y^2 = 3x^2 - 9kx + 9k^2 \equiv 3x^2 \bmod 9$. From this it follows that 3 divides $x^2 - xy + y^2$. By assumption $3 \not| x$ so it also follows that 9 does not divide $x^2 - xy + y^2$. Hence
$$3|(x^2 - xy + y^2) \text{ and } 3^{3m-1}|(x + y).$$

2. We claim that $\gcd(x + y, x^2 - xy + y^2) = 3$. By 1., 3 divides both $x + y$ and $x^2 - xy + y^2$. On the other hand, $\gcd(x, y) = 1$ implies that $\gcd(x + y, xy) = 1$. Hence there are integers u, v with $u(x + y) + v(xy) = 1$. But then $3u(x + y) + v(x + y)^2 - v(x^2 - xy + y^2) = 3$. So any prime dividing both $x + y$ and $x^2 - xy + y^2$ divides 3.

3. Define $\alpha, \beta \in \mathbb{Z}$ by $x + y = 3^{3m-1}\alpha$, $x^2 - xy + y^2 = 3\beta$. Then by 2., along with $3 \not| w$, it follows that $\gcd(\alpha, \beta) = 1$. Moreover, $\alpha\beta = w^3$. So by Corollary 2.7.2 we have that $\alpha = \gamma^3$, $\beta = \delta^3$ for certain $\gamma, \delta \in \mathbb{Z}$. Furthermore, γ, δ are not divisible by 3.

4. Set $\pi = \omega - 1$. Then $N(\pi) = 3$ and therefore π is irreducible. We have $\bar{\pi} = -2 - \omega = (1 + \omega)\pi$. As $1+\omega$ is invertible, $\bar{\pi}$ is associated with π, $N(\bar{\pi}) = 3$ and $\bar{\pi}$ is irreducible as well. Furthermore, $\pi\bar{\pi} = 3$.

5. Now we consider the following factorization in $\mathbb{Z}[\omega]$
$$x^3 + y^3 = (x + y)(x + \omega y)(x + \omega^2 y).$$
We claim that there are $\zeta, \eta \in \mathbb{Z}[\omega]$ with $\zeta(x+\omega y)+\eta(x+\omega^2 y) = \pi$. Note that $\pi x = -x+x\omega = \omega(x+\omega y)-(x+\omega^2 y)$ and $\pi y = -y+y\omega = (1+\omega)(x+\omega y)-(1+\omega)(x+\omega^2 y)$. From $\gcd(x, y) = 1$ we see that there are integers u, v with $ux + vy = 1$ and therefore $u(\pi x) + v(\pi y) = \pi$. Now on the left we substitute the expressions for πx and πy just obtained, and we see that indeed ζ, η exist.

6. We claim that π divides $x+\omega y$. First of all, $\pi|3$ by 3. So since $3|x+y$ (by 1.) we see that $\pi|x+y$. Finally we note that $x+\omega y = x+y+\pi y$.

7. Because of the previous point we can define $\mu \in \mathbb{Z}[\omega]$ by $x+\omega y = \pi\mu$. We claim that $\mu = \epsilon\theta^3$ for some $\epsilon, \theta \in \mathbb{Z}[\omega]$ with ϵ invertible. In order to see this observe that $x+\omega^2 y = \overline{x+\omega y} = \bar{\pi}\bar{\mu} = (1+\omega)\pi\bar{\mu}$ (for the last equality see 4.). Let ζ, η be as in 5. After dividing by π we obtain $\zeta\mu + \eta(1+\omega)\bar{\mu} = 1$ and it follows that $\mu, \bar{\mu}$ are coprime in $\mathbb{Z}[\omega]$. Now on the one hand, $(x+\omega y)(x+\omega^2 y) = x^2 - xy + y^2 = 3\delta^3$ (see 3.). On the other hand $(x+\omega y)(x+\omega^2 y) = \pi\bar{\pi}\mu\bar{\mu} = 3\mu\bar{\mu}$. Hence $\mu\bar{\mu} = \delta^3$ and we conclude with Lemma 2.7.1.

8. We have $\epsilon = \pm 1$. First of all, $\pi\mu = x+y\omega = x+y+\pi y$. By 3., $x+y = 3^{3m-1}\alpha = 3\cdot 3^{3m-2}\alpha = \pi\bar{\pi}3^{3m-2}\alpha$. After dividing by π we have $\mu = 3^{3m-2}\bar{\pi}\alpha + y$. Hence $y \equiv \mu \bmod 3$. By hypothesis 3 does not divide y, so $y \equiv \pm 1 \bmod 3$. Therefore, as $\mu = \epsilon\theta^3$ we see that $\epsilon\theta^3 \equiv \pm 1 \bmod 3$.

   Write $\theta = a+b\omega$. Then $\theta^3 = a^3 + 3a^2b\omega + 3a(b\omega)^2 + \omega^3 b^3 \equiv a^3 + b^3 \bmod 3$. So θ^3 is congruent to an integer modulo 3. From $\epsilon\theta^3 \equiv \pm 1 \bmod 3$ we see that 3 does not divide θ^3. Hence $\theta^3 \equiv \pm 1 \bmod 3$. So also $\epsilon \equiv \pm 1 \bmod 3$. But the only invertibles that satisfy that are ± 1.

9. If $\epsilon = -1$ then we replace θ by $-\theta$ so that we have $\mu = \theta^3$ and $x+\omega y = \pi\theta^3$. Write $\theta = a+b\omega$ then we get
$$x+\omega y = (-a^3 - 3a^2b + 6ab^2 - b^3) + (a^3 - 6a^2b + 3ab^2 + b^3)\omega.$$
   So since $\gcd(x,y) = 1$ we also have $\gcd(a,b) = 1$. It also follows that $x+y = 9ab(b-a)$. With 3. this implies that $ab(b-a) = 3^{3m-3}\gamma^3 = (3^{m-1}\gamma)^3$. The triple $(a, b, b-a)$ is coprime. So by Corollary 2.7.2 there are coprime $s, t, u \in \mathbb{Z}$ with $a = s^3$, $b = t^3$, $b-a = u^3$. Then $s^3+u^3 = t^3$. If $m = 1$ then from $ab(b-a) = \gamma^3$ and the fact that $3 \not| \gamma$ (see 3.) we get that 3 does not divide stu and we have a contradiction with Lemma 2.7.6. If $m > 1$ then with $v = -t$ we get $s^3+u^3+v^3 = 0$ and $s^3u^3v^3 = -ab(b-a) = (-3^{m-1}\gamma)^3$. As (s,u,v) is coprime we see that 3^{m-1} divides one of s, u, v and the others are not divisibly by 3. Suppose that 3^{m-1} divides v. Then $v = 3^{m-1}l$, where 3 does not divide l because it does not divide γ. Then $s^3 + u^3 = (-v)^3 = 3^{3(m-1)}(-l)^3$ and we have a contradiction with the induction hypothesis. Therefore the proof is complete.

□

Theorem 2.7.8 *There are no $x, y, z \in \mathbb{Z}$ with $xyz \neq 0$ and $x^3 + y^3 = z^3$.*

Proof. Suppose that there is a coprime triple $(x,y,z) \in \mathbb{Z}^3$ with $x^3+y^3 = z^3$. We want to derive a contradiction. By Lemma 2.7.6 we see that 3 divides exactly one of x, y, z. Suppose that it divides y. Then write $y = 3^m w$ where w is not divisible by 3. Then $(x, -z, -w)$ is coprime with $x^3 + (-z)^3 = 3^{3m}(-w)^3$ and we obtain a contradiction with the previous proposition. If 3 divides x or z then we proceed in exactly the same way.

Now let $x, y, z \in \mathbb{Z}$ with $xyz \neq 0$ and $x^3+y^3 = z^3$. If there is a prime p dividing one of x, y, z then it also divides the third and by setting $x' = x/p$, $y' = y/p$, $z' = z/p$ we have $(x')^3 + (y')^3 = (z')^3$. Continuing we arrive at a coprime triple that satisfies this equation. But by the above that cannot exist. □

2.7.3 The Ramanujan-Nagell theorem

An integer is called a *triangular number* if it is of the form $1+2+3+\cdots+m = \frac{m(m+1)}{2}$. An integer is called a *Mersenne number* if it is of the form $2^s - 1$. Now one can ask which integers are both triangular and Mersenne. In other words, one wants to solve the equation $2^s - 1 = \frac{m(m+1)}{2}$ for m, s. After a bit of rewriting this is seen to be equivalent to $2^{s+3} = (2m+1)^2 + 7$. So we are led to consider the equation $2^n = x^2 + 7$ for $n \geq 3$ and positive integers x. The famous indian mathematician Srinivasa Ramanujan found some solutions to this equation, and asked whether there were any more (question 464 in the *Journal of the Indian Mathematical Society*, 1913). In 1948 Trygve Nagell showed that Ramanujan's solutions were the only ones.

Theorem 2.7.9 *Consider the equation $2^n = x^2 + 7$ for positive integers x, n. The only solutions to this equation are $(x, n) = (1, 3), (3, 4), (5, 5), (11, 7), (181, 15)$.*

Before showing a proof of this theorem we first look at the ring $\mathbb{Z}[\alpha]$, where $\alpha = \frac{1+\sqrt{-7}}{2}$ is a root of $f = x^2 - x + 2$. Its other root is $\bar{\alpha} = \frac{1-\sqrt{-7}}{2}$. For the norm we have $N(a + b\alpha) = |a^2 + ab + 2b^2|$. So $N(\alpha) = N(\bar{\alpha}) = 2$, implying that $\alpha, \bar{\alpha}$ are irreducible.

We show that $\mathbb{Z}[\alpha]$ is Euclidean. We use a similar approach as for $\mathbb{Z}[i]$, see Example 2.4.22. Let $\eta, \xi \in \mathbb{Z}[\alpha]$. We have to show that there are $q, r \in \mathbb{Z}[\alpha]$ with $\eta = q\xi + r$ and $N(r) < N(\xi)$. Write $\eta\xi^{-1} = u + v\alpha$ with $u, v \in \mathbb{Q}$ (here for a short wile we work in $\mathbb{Q}(\alpha)$). Let $u = u_0 + x_0$, $v = v_0 + y_0$, where $u_0, v_0 \in \mathbb{Z}$ and $|x_0|, |y_0| \leq \frac{1}{2}$. Furthermore, if $|x_0| = |y_0| = \frac{1}{2}$ then we choose u_0, v_0 in such a way that x_0, y_0 have opposite sign (for example, such that $x_0 = \frac{1}{2}$, $y_0 = -\frac{1}{2}$). Set $q = u_0 + v_0\alpha$ and $r = \xi(x_0 + y_0\alpha)$. Then $N(r) = N(\xi)|x_0^2 + x_0y_0 + 2y_0^2|$. Now $|x_0^2 + x_0y_0 + 2y_0^2| \leq 4 \cdot \frac{1}{4} = 1$. But equality can only happen if $|x_0| = |y_0| = \frac{1}{2}$. However, because of our choice of signs, in that case $|x_0^2 + x_0y_0 + 2y_0^2| = \frac{1}{2}$. We see that in all cases $N(r) < N(\xi)$.

We also need a little technical lemma.

Lemma 2.7.10 *Let a, l, e be positive integers and $m = a + 7^l e$. Let $0 \leq t \leq a$ and $s \geq 1$. Then $\binom{m}{t} \cdot 7^s \equiv \binom{a}{t} \cdot 7^s \bmod 7^{l+s}$ if $t < 7$ and $\binom{m}{t} \cdot 7^s \equiv \binom{a}{t} \cdot 7^s \bmod 7^{l+s-1}$ if $7 \leq t < 14$.*

Proof. We have that

$$m(m-1)\cdots(m-t+1) = (a+7^l e)(a-1+7^l e)\cdots(a-t+1+7^l e) = a(a-1)\cdots(a-t+1) + u \cdot 7^l.$$

And therefore

$$\binom{m}{t} \cdot 7^s = \binom{a}{t} \cdot 7^s + \frac{u \cdot 7^{l+s}}{t!}.$$

Now $u \cdot 7^l$ is divisible by $t!$. So if $t < 7$ then already u is divisible by $t!$. If $7 \leq t < 14$ then $7u$ is divisible by $t!$. This implies both cases of the lemma. □

Proof.(Of Theorem 2.7.9.) First suppose that n is even. Then $2^n - x^2 = (2^{\frac{n}{2}} - x)(2^{\frac{n}{2}} + x)$. Since 7 is prime (in $\mathbb{Z}$) and $2^{\frac{n}{2}} + x > 1$ we have that $2^{\frac{n}{2}} + x = 7$ and $2^{\frac{n}{2}} - x = 1$. By adding these equations we obtain $2^{\frac{n}{2}+1} = 8$. Therefore we have $n = 4$ and $x = 3$ as the only solution.

Now suppose that n is odd. Since $n = 1$ clearly yields no solution, and $n = 3$ gives $x = 1$, we may assume that $n > 3$.

Note that $\alpha\bar{\alpha} = 2$ and $x^2 + 7 = 2^n$ is equivalent to $\frac{x^2+7}{4} = 2^{n-2}$. The latter is the same as

$$\frac{x + \sqrt{-7}}{2}\frac{x - \sqrt{-7}}{2} = \alpha^{n-2}\bar{\alpha}^{n-2}.$$

If α would divide both factors on the left, it would divide their difference, $\sqrt{-7}$. In that case we would have $\sqrt{-7} = \eta\alpha$ for a certain $\eta \in \mathbb{Z}[\alpha]$. By taking norms that leads to $7 = 2N(\eta)$ which clearly is impossible. Similarly, $\bar{\alpha}$ can only divide one of the two factors on the left. Because α, $\bar{\alpha}$ are irreducible, we have that $\frac{x+\sqrt{-7}}{2}$ is equal to $\pm\alpha^{n-2}$ or equal to $\pm\bar{\alpha}^{n-2}$ (note that the only invertible elements of $\mathbb{Z}[\alpha]$ are ± 1).

If $\frac{x+\sqrt{-7}}{2} = \pm\alpha^{n-2}$ then $\frac{x-\sqrt{-7}}{2} = \pm\bar{\alpha}^{n-2}$ and

$$\sqrt{-7} = \frac{x + \sqrt{-7}}{2} - \frac{x - \sqrt{-7}}{2} = \pm(\alpha^{n-2} - \bar{\alpha}^{n-2}).$$

If $\frac{x+\sqrt{-7}}{2} = \pm\bar{\alpha}^{n-2}$ then we get the same conclusion.

We claim that in the above equation we must have the minus sign. So suppose that $\sqrt{-7} = \alpha^{n-2} - \bar{\alpha}^{n-2}$. This is the same as $\alpha^{n-2} - \bar{\alpha}^{n-2} = \alpha - \bar{\alpha}$. Using $\alpha\bar{\alpha} = 2$ and $\alpha = 1 - \bar{\alpha}$ we see that

$$\alpha^2 = (1 - \bar{\alpha})^2 = 1 - 2\bar{\alpha} + \bar{\alpha}^2 = 1 - \alpha\bar{\alpha}^2 + \bar{\alpha}^2.$$

Let $I = \langle \bar{\alpha}^2 \rangle$ be the ideal of $\mathbb{Z}[\alpha]$ consisting of all elements that are divisible by $\bar{\alpha}^2$. Then we have just seen that $\alpha^2 = 1 + u$ where $u \in I$. Write $m = n - 2 = 2k + 1$. By hypothesis we have $n \geq 5$ so that $k \geq 1$. Then $\alpha^m = \alpha(\alpha^2)^k = \alpha(1+u)^k = \alpha + v$ for some $v \in I$. As also $\bar{\alpha}^m \in I$ we get from $\alpha^m - \bar{\alpha}^m = \alpha - \bar{\alpha}$ that $\bar{\alpha} \in I$, which is a contradiction as $\bar{\alpha}$ is not invertible.

It follows that we have $-\sqrt{-7} = \alpha^m - \bar{\alpha}^m$, where $m = n-2$ as before. Using the binomial theorem we have

$$\alpha^m = (\tfrac{1}{2})^m (1 + \sqrt{-7})^m = \sum_{i=0}^{m} \binom{m}{i} \sqrt{-7}^i$$
$$\bar{\alpha}^m = (\tfrac{1}{2})^m (1 - \sqrt{-7})^m = \sum_{i=0}^{m} \binom{m}{i} (-\sqrt{-7})^i.$$

Subtracting these we see that the terms with i even cancel, whereas the terms with i odd should sum to $-\sqrt{-7}$. This means that

$$-1 = (\tfrac{1}{2})^m \left(2\binom{m}{1} - 2\binom{m}{3} \cdot 7 + 2\binom{m}{5} \cdot 7^2 - \cdots \pm 2\binom{m}{m} \cdot 7^{\frac{m-1}{2}} \right).$$

So after multiplying with 2^{m-1} this yields

$$-2^{m-1} = \binom{m}{1} - \binom{m}{3} \cdot 7 + \binom{m}{5} \cdot 7^2 - \cdots \pm \binom{m}{m} \cdot 7^{\frac{m-1}{2}}. \tag{2.7.1}$$

One immediate consequence of (2.7.1) is that $-2^{m-1} \equiv m \bmod 7$. Observe that $2^6 = 64 \equiv 1 \bmod 7$. Now we can make a little table with three columns: one for $m = 3, 5, 7, \ldots$, the second one has $-2^{m-1} \bmod 7$ and the third contains $m \bmod 7$ (we leave this as an exercise). Let $m_1, m_2 \geq 3$ both be odd. Then $-2^{m_1} \equiv -2^{m_2} \bmod 7$ is equivalent to $m_1 \equiv m_2 \bmod 6$. So m_1, m_2 have the same entries in the second and third columns if and only if $m_1 \equiv m_2 \bmod 6$ and $m_1 \equiv m_2 \bmod 7$. By the Chinese remainder theorem (Theorem 2.5.16) this is equivalent to $m_1 \equiv m_2 \bmod 42$. So the table starts repeating itself at the line of $m = 45$. From the table it readily follows that the only odd m with $3 \leq m \leq 43$ and $-2^{m-1} \equiv m \bmod 7$ are $m = 3, 5, 13$. However, also $m = 3 + 42k$, $m = 5 + 42k$, $m = 13 + 42k$ satisfy this. We now show that when $k \neq 0$ these will not satisfy (2.7.1).

For this we first note the following identity (for integers $a \geq 1$, $0 \leq k \leq a$, $0 \leq s \leq k$)

$$\binom{a}{k} = \frac{a(a-1)\cdots(a-s+1)}{k(k-1)\cdots(k-s+1)} \binom{a-s}{k-s}.$$

Now suppose that $m \equiv 3 \bmod 42$ and write $m - 3 = 7^l \cdot 6 \cdot h$, where $l \geq 1$, $h \neq 0$ and 7 does not divide h. A typical term in (2.7.1) is $\binom{m}{2k+1} \cdot 7^k$ and for $k \geq 2$ we have

$$\binom{m}{2k+1} = \frac{m(m-1)(m-2)(m-3)}{(2k+1)(2k)(2k-1)(2k-2)} \binom{m-4}{2k-3}.$$

Now 7 can only divide one of the four factors in the denominator of the first factor on the right. As $7^{k-1} > 2k + 1$, this denominator can be divisible by 7^{k-2} but not by 7^{k-1}. In the numerator we see that m is divisible by 7^l so the whole term $\binom{m}{2k+1} \cdot 7^k$ is divisible by 7^{l+1} (in fact, we showed that it is divisible by 7^{l+2}, but we will not need that). So from (2.7.1) it follows that

$$-2^{m-1} \equiv m - \binom{m}{3} \cdot 7 \bmod 7^{l+1}.$$

From $m = 3 + 7^6 \cdot 6 \cdot h$ it follows that

$$2^{m-1} = 4(2^6)^{7^l h} = 4(1 + 9 \cdot 7)^{7^l h} = 4 \sum_{i=0}^{7^l h} \binom{7^l h}{i} (9 \cdot 7)^i = 4(1 + 7^l h \cdot 9 \cdot 7 + \cdots) \equiv 4 \bmod 7^{l+1}.$$

By Lemma 2.7.10 we have that $\binom{m}{3} \cdot 7 \equiv \binom{3}{3} \cdot 7 \bmod 7^{l+1} \equiv 7 \bmod 7^{l+1}$. Therefore from (2.7.1) it follows that $-4 \equiv m - 7 \bmod 7^{l+1}$, that is $m \equiv 3 \bmod 7^{l+1}$ contrary to the choice of h (such that $7 \not| h$).

The other cases are dealt with in a similar manner. If $m \equiv 5 \bmod 42$ then we write $m-5 = 7^l \cdot 6 \cdot h$ where $7 \not| h$. For $k \geq 3$ we have

$$\binom{m}{2k+1} = \tfrac{m(m-1)\cdots(m-5)}{(2k+1)(2k)\cdots(2k-4)} \binom{m-6}{2k-5}.$$

By a similar reasoning as above this implies that $\binom{m}{2k+1} \cdot 7^k$ is divisible by 7^{l+1}. So using Lemma 2.7.10, the right hand side of (2.7.1) becomes, modulo 7^{l+1}

$$m - \binom{m}{3} \cdot 7 + \binom{m}{5} \cdot 7^2 \equiv m - \binom{5}{3} \cdot 7 + \binom{5}{5} \cdot 7^2 \bmod 7^{l+1} \equiv m - 21 \bmod 7^{l+1}.$$

Furthermore, $2^{m-1} \equiv 2^4 \bmod 7^{l+1}$. Hence from (2.7.1) we get $m \equiv 5 \bmod 7^{l+1}$, contrary to the choice of h.

If $m \equiv 13 \bmod 42$ then we write $m - 13 = 7^l \cdot 6 \cdot h$ where $7 \not| h$. For $k \geq 7$ we have

$$\binom{m}{2k+1} = \tfrac{m(m-1)\cdots(m-13)}{(2k+1)(2k)\cdots(2k-12)} \binom{m-14}{2k-13}.$$

By a similar reasoning as above this implies that $\binom{m}{2k+1} \cdot 7^k$ is divisible by 7^{l+1}. So using Lemma 2.7.10, the right hand side of (2.7.1) becomes, modulo 7^{l+1}

$$m - \sum_{k=1}^{6} (-7)^k \binom{m}{2k+1} \equiv m - \sum_{k=1}^{6} (-7)^k \binom{13}{2k+1} \bmod 7^{l+1} \equiv m - 4109 \bmod 7^{l+1}.$$

Also $2^{m-1} \equiv 2^{12} \bmod 7^{l+1}$. Therefore, since $2^{12} = 4096$, from (2.7.1) we get $m \equiv 13 \bmod 7^{l+1}$, contrary to the choice of h.

The conclusion is that the only possibilities are $m = 3, 5, 13$ leading to $n = 5, 7, 15$. For these n there is indeed a solution for x. □

Chapter 3

Groups

Groups first appeared in the work of Évariste Galois who studied the problem of finding solutions to polynomial equations (like $x^3 + 9x - 2 = 0$) in terms of *radicals*, which are complex numbers that are given by expressions involving the arithmetic operations and taking n-th roots ($\sqrt[n]{\ }$). For example, the above equation has solution $\sqrt[3]{1+2\sqrt{7}} + \sqrt[3]{1-2\sqrt{7}}$. His main idea was to take the roots of the polynomial f in question and look at a certain *group* G_f of permutations of them. He was then able to relate the solvability of the equation $f = 0$ by radicals to a certain property of G_f.

After Galois, the idea of not directly looking at the object of interest but instead at sets of maps of the object to itself became more and more important. Nowadays this idea plays a key role in many areas of science like mathematics, physics, chemistry. When studying an object (this can be anything: the solutions of a polynomial equation, or of a differential equation, an atom, a molecule, a crystal, a black hole, a graph, a ring,...) one looks at its *symmetries*. What a symmetry is depends on the object in question. Very roughly speaking, a symmetry of an object O is a bijective map $f : O \to O$ preserving something important. To make this a bit clearer we have a look at an example.

Example 3.0.1 A *graph* is a pair $\Gamma = (V, E)$, where V is a set (of *vertices*) and $E \subset \{\{v, w\} \mid v, w \in V\}$. If $\{v, w\} \in E$ then we say that there is an *edge* between v and w. We can make a graphical representation of a graph by drawing a dot for each element of V, and drawing an edge between v, w if $\{v, w\} \in E$. Figure 3.1 has two examples with $V = \{1, 2, 3, 4\}$.

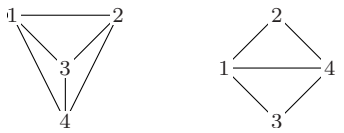


Figure 3.1: Two graphs.

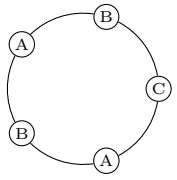
A symmetry (or automorphism) of a graph $\Gamma = (V, E)$ is a bijective map $\sigma : V \to V$ such that $\{v, w\} \in E$ if and only $\{\sigma(v), \sigma(w)\} \in E$. In the first graph of Figure 3.1 every vertex is connected to every other vertex. So any σ will be a symmetry. Therefore this graph has 24 symmetries. For the second graph it is easily seen that a symmetry can interchange $1, 4$ and/or $2, 3$. Therefore here we have just 4 symmetries.

In we take the set of all symmetries of an object, then usually we have the following properties:

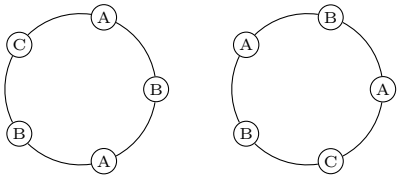
1. The identity map is a (rather trivial) symmetry.
2. If f is a symmetry then so is its inverse map f^{-1}.
3. If f, g are symmetries then so is their composition $f \circ g$ (defined by $f \circ g(o) = f(g(o))$).

Furthermore we note that composition of functions is always associative: $f \circ (g \circ h) = (f \circ g) \circ h$ (as both sides send an element o to $f(g(h(o)))$). These properties are taken to form the definition of the algebraic structure called group.

One more word about the use of symmetry groups. Although it has turned out to be an extremely fruitful idea in many contexts to look at (symmetry-) groups, it is not the case that attaching a group to a certain problem already solves it. The way a group is used depends strongly on the context, and usually quite some group theory has to be developed to make good use of the groups that have been defined. In Section 3.6 we will see an application of groups to the combinatorial problem of counting colourings. One interesting instance of this is the problem of counting necklaces. We briefly describe what this is about. A *necklace* of length n consists of n beads which are arranged equidistantly on a circle. All beads have the same shape, but possibly different colours. The colour of each bead is chosen from a fixed set of q colours. For example, suppose that we have 3 colours, A, B, C, then a necklace of length 5 is



Two necklaces (both of length n and beads of q colours) are said to be *equivalent* if one can be obtained from the other by a *rotation* or by *reflection* (this corresponds to picking it up from the table, turning it upside down, and putting it back). For example, the following two necklaces are equivalent to the one above (the first one is obtained by a rotation, the second one by a reflection from the one above):



This notion of equivalence comes from considering necklaces in the real world, i.e., beads connected by pieces of string of equal length. Two necklaces are equivalent if and only if one can be moved in such a way that the two become equal. The question is, given n and q, how many inequivalent necklaces there are. Here we will see how groups can be used to show that the number of inequivalent necklaces of length 5 with q colours is

$$\frac{1}{10}(q^5 + 5q^3 + 4q)$$

(so for $q = 3$ we have 39 inequivalent necklaces).

3.1 Definition and examples

In this section we define what a group is and study two classes of examples, the symmetric and the dihedral groups.

Definition 3.1.1 *A* group *is a set G with an operation $\cdot : G \times G \to G$ with*

1. *$\cdot$ is associative, i.e., $g_1 \cdot (g_2 \cdot g_3) = (g_1 \cdot g_2) \cdot g_3$ for all $g_1, g_2, g_3 \in G$,*
2. *there is an element $1 \in G$ (called the* neutral element*) with $1 \cdot g = g \cdot 1 = g$ for all $g \in G$,*
3. *for each $g \in G$ there is a $g^{-1} \in G$ (called the* inverse of g*) such that $g^{-1} \cdot g = g \cdot g^{-1} = 1$.*

Remark 3.1.2

- Let G be a group. If $g_1 \cdot g_2 = g_2 \cdot g_1$ for all $g_1, g_2 \in G$ then we say that G is *commutative* or *abelian*.
- Two elements g_1, g_2 of a group are said to *commute* if $g_1 g_2 = g_2 g_1$.

- Sometimes we use $+$ instead of $\cdot$ to denote the operation of a group. In those cases the convention is that the group is commutative, that we write 0 instead of 1 and $-g$ instead of g^{-1}.
- Usually, when the group operation is written $\cdot$ we omit it and write g_1g_2 instead of $g_1 \cdot g_2$.
- Let G be a group. Then its neutral element 1 is uniquely determined. (Indeed, suppose that $g_0 \in G$ has the property $g_0 g = g$ for all $g \in G$. Multiplying on the right by g^{-1} we obtain $g_0 = 1$.) Furthermore, for each $g \in G$, its inverse is uniquely determined. (We leave it as an exercise to verify that.)
- Let G be a group. Then its cardinality $|G|$ is also called the *order* of G.

Example 3.1.3
- Let R be a ring. Then with respect to $+$ we have that R is a commutative group.
- Let R be a commutative ring with unity. Let R^* be the set of invertible elements of R. Then R^* is a commutative group with respect to the multiplication $\cdot$.
- Let $\Gamma = (V, E)$ be a graph, and G the set of all symmetries of Γ (see Example 3.0.1). Then with respect to the composition of functions, G is a group. As already noted, it is immediate that composition is associative. The neutral element is the identity map. The only non-obvious thing is to show that for a symmetry σ its inverse σ^{-1} is also a symmetry, that is, it maps edges to edges. We leave the verification of that as an exercise.

3.1.1 Symmetric groups

Let X be a finite set. A *permutation* of X is a bijective map $\pi : X \to X$. Let S_X denote the set of all permutations of X. With respect to the composition of functions this is a group (indeed, composition of functions is associative, the neutral element is the identity map and the inverse of a permutation is simply its inverse map), called the *symmetric group* on X. If $X = \{1, 2, \ldots, n\}$ then we also write S_n instead of S_X.

Although the group operation is the composition of functions, we use multiplication to denote the group operation. So for $\pi, \sigma \in S_X$ we have $\pi\sigma = \pi \circ \sigma$. In other words, $\pi\sigma$ is the element of S_X defined by $(\pi\sigma)(x) = \pi(\sigma(x))$. It should be noted that here we *first* apply σ, and *then* we apply π.

Lemma 3.1.4 *Let $|X| = n$ then $|S_X| = n!$.*

Proof. The proof is by induction, the case $n = 1$ being obvious. Write $X = \{x_1, \ldots, x_n\}$. A $\pi \in S_X$ is uniquely determined by writing the x_i in some ordered sequence. Then $\pi(x_i)$ is the i-th element of the sequence. So we need to count how many sequences we can make with the elements of X. Fix $x_{i_0} \in X$. Then the sequences starting with x_{i_0} are formed by taking x_{i_0} followed by an arbitrary sequence of the elements of $X \setminus \{x_{i_0}\}$. By induction there are $(n-1)!$ of those. Since we have n possibilities for our i_0, and different choices for i_0 lead to different sequences, the total number is $n(n-1)! = n!$. □

Example 3.1.5 Define $\pi, \sigma \in S_5$ by

$$\pi : \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 4 & 5 & 2 \end{array} \qquad \sigma : \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 5 & 4 \end{array}$$

This means that $\pi(1) = 3$, $\pi(2) = 1$ and so on. We compute $\pi\sigma$ and $\sigma\pi$. This is done by simply computing all images of these permutations. For example $\pi\sigma(1) = \pi(\sigma(1)) = \pi(2) = 1$, $\pi\sigma(2) = \pi(\sigma(2)) = \pi(1) = 3$ and so on. We obtain

$$\pi\sigma : \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 4 & 2 & 5 \end{array} \qquad \sigma\pi : \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \\ 3 & 2 & 5 & 4 & 1 \end{array}$$

so in particular we see that S_5 is not commutative.

From this example we see that, if we are going to do a lot of work with permutations, we need a better notation for them. It turns out that a very useful way of describing a permutation is by a product of *disjoint cycles*.

Let $1 \leq k \leq |X|$. A *k-cycle* in S_X is a permutation $\pi \in S_X$ such that there are $x_{i_1}, \ldots, x_{i_k} \in X$ with

$$\pi(x_{i_1}) = x_{i_2}, \pi(x_{i_2}) = x_{i_3}, \ldots, \pi(x_{i_{k-1}}) = x_{i_k}, \pi(x_{i_k}) = x_{i_1}$$

and $\pi(x) = x$ if x is not one of the x_{i_j}. (The term "cycle" is obvious if one writes the x_{i_j} on a circle; then in a sense, π goes round this circle.) We write $\pi = (x_{i_1}, \ldots, x_{i_k})$. Two cycles $(x_{i_1}, \ldots, x_{i_k})$, $(x_{j_1}, \ldots, x_{j_l})$ are said to be *disjoint* if the sets $\{x_{i_1}, \ldots, x_{i_k}\}$, $\{x_{j_1}, \ldots, x_{j_l}\}$ are disjoint. *We note that disjoint cycles commute.*

Also note that a 1-cycle is the identity map. For this reason we avoid mentioning them. We say that a permutation $\pi \in S_X$ has a *factorization in disjoint cycles* if π can be written as a product of pairwise disjoint k-cycles (with $k \geq 2$) such that every $x \in X$ with $\pi(x) \neq x$ lies in (necessarily exactly) one of these cycles. For example, it is easily verified that in Example 3.1.5 we have $\pi = (1, 3, 4, 5, 2)$, $\sigma = (1, 2)(4, 5)$. We see that the factorization of σ does not "mention" 3, but that just means that $\sigma(3) = 3$.

Also we note that the order in which the cycles appear in a factorization is arbitrary because disjoint cycles commute.

Proposition 3.1.6 *Let X be a finite set and $\pi \in S_X$. Suppose that π is not the identity. Then π has a factorization in disjoint cycles. Moreover, apart from the order in which they appear, these cycles are uniquely determined by π.*

Proof. Define $N(\pi) = |\{x \in X \mid \pi(x) \neq x\}|$. The proof is by induction on $N(\pi)$. The induction hypothesis is that the theorem holds for all $\pi' \in S_X$ with $N(\pi') < N(\pi)$.

Write $X = \{x_1, \ldots, x_n\}$. We now define a sequence $i_1, \ldots, i_k$. First of all, i_1 is minimal such that $\pi(x_{i_1}) \neq x_{i_1}$ (as π is not the identity, i_1 exists). If, for $j \geq 1$, i_j is defined then we do the following. If $\pi(x_{i_j}) \in \{x_{i_1}, \ldots, x_{i_j}\}$ then we set $k = j$ and the sequence terminates. Otherwise we define i_{j+1} by $\pi(x_{i_j}) = x_{i_{j+1}}$. Note that the sequence always terminates as X is finite. Furthermore, if $\pi(x_{i_k}) = x_{i_j}$ with $2 \leq j \leq k$ then $\pi(x_{i_k}) = \pi(x_{i_{j-1}})$ which as π is injective implies $x_{i_k} = x_{i_{j-1}}$; but in that case the sequence would have terminated earlier (as then $\pi(x_{i_{k-1}}) \in \{x_{i_1}, \ldots, x_{i_{k-1}}\}$). It follows that $\pi(x_{i_k}) = x_{i_1}$. Set $\sigma = (x_{i_1}, \ldots, x_{i_k})$, $\hat{\pi} = \sigma^{-1}\pi$ and $Y = X \setminus \{x_{i_1}, \ldots, x_{i_k}\}$. Then $\hat{\pi}(x_{i_j}) = x_{i_j}$ for $1 \leq j \leq k$ and $\hat{\pi}(y) = \pi(y)$ for all $y \in Y$. Hence $N(\hat{\pi}) < N(\pi)$. So $\hat{\pi}$ is a product of uniquely determined disjoint cycles. Because $\hat{\pi}(x) = x$ if $x \notin Y$, these cycles all involve elements from Y only. We see that $\pi = \sigma\hat{\pi}$ is the product of the k-cycle σ and some cycles involving the elements of Y.

Now let $\pi = \sigma_1 \cdots \sigma_m = \pi_1 \cdots \pi_r$ be two factorizations of π into pairwise disjoint k-cycles ($k \geq 2$). The element x_{i_1} appears in exactly one σ_i, say σ_1. Similarly, we may assume that it appears in π_1. But then necessarily $\sigma_1 = (x_{i_1}, \ldots, x_{i_k}) = \pi_1$. After cancelling σ_1 and π_1, the proof is finished by induction. □

Example 3.1.7 The proof of the previous proposition also yields a method to find the factorization of a given $\pi \in S_X$. Indeed, write $X = \{x_1, \ldots, x_n\}$. The first cycle starts with x_1. If $\pi(x_1) = x_{i_2}$, $\pi(x_{i_2}) = x_{i_3}$, …, $\pi(x_{i_k}) = x_1$ then the first cycle is $(x_1, x_{i_2}, \ldots, x_{i_k})$. Let j_1 be minimal such that x_{j_1} does not appear in the first cycle. Then the second cycle starts with x_{j_1}. If $\pi(x_{j_1}) = x_{j_2}$, …, $\pi(x_{j_l}) = x_{j_1}$ then the second cycle is $(x_{j_1}, \ldots, x_{j_l})$. Then let k_1 be minimal such that x_{k_1} does not appear in the first two cycles. Again we complete the cycle containing x_{k_1}, and continue like that.

Example 3.1.8 (Multiplication of permutations) Here we consider the multiplication of two permutations given as products of disjoint cycles. We want to write the result immediately as a product of disjoint cycles. Let us consider the same example as in Example 3.1.5, $\pi = (1, 3, 4, 5, 2)$, $\sigma = (1, 2)(4, 5)$. As seen in the mentioned example: $\pi\sigma(1) = \pi(\sigma(1)) = \pi(2) = 1$, $\pi\sigma(2) = \pi(\sigma(2)) = \pi(1) = 3$ and so on. This way we can compute all images of $\pi\sigma$ and write it as a product of disjoint

cycles using the method of Example 3.1.7. However, if we write the cycles of σ after those of π on a single line

$$(1,3,4,5,2)(1,2)(4,5)$$

we see more clearly what is happening. In order to determine the image of i we *read from right to left* and we stop at the first occurrence of i. Then we see that i is mapped to j and we keep j in our mind and continue reading to the left. We stop at the first occurrence of j. Then we see that j is mapped to k. The image of i is then k. For example take $i = 4$. It first occurs in the first cycle from the right. There it is mapped to $j = 5$. Then we look where 5 appears in the cycles further to the left. It is contained in the left-most cycle, where 5 is mapped to 2. So the image of 4 is 2. Using this method we can also directly compute the product of more than two permutations. The only difference is that k may be mapped to l further down the left. For example, let $\tau = (1,2,4)(3,5)$ then $\pi\sigma\tau$ is

$$(1,3,4,5,2)(1,2)(4,5)(1,2,4)(3,5).$$

So reading from right to left, we see $1 \mapsto 2 \mapsto 1 \mapsto 3$, $3 \mapsto 5 \mapsto 4 \mapsto 5$, $5 \mapsto 3 \mapsto 4$, $4 \mapsto 1 \mapsto 2 \mapsto 1$, $2 \mapsto 4 \mapsto 5 \mapsto 2$. So that $\pi\sigma\tau = (1,3,5,4)$.

Remark 3.1.9 It is also possible to do the opposite and multiply permutations by reading from left to right. Of course the two ways of doing it should not be mixed. Furthermore, it is a choice with quite a few consequences. Firstly, if permutations are multiplied reading from left to right then also we must write $x\sigma$ (or x^σ) instead of $\sigma(x)$ (for $x \in X$, $\sigma \in S_X$). Secondly, when we study group actions they should be written on the right (so x^g instead of $g \cdot x$ as we will do). For a third example consider the proof of Theorem 3.5.12. In that proof left cosets are used. But when the group action is written on the right then one should use right cosets in order to make the same proof work.

Remark 3.1.10 It turns out that writing a permutation as a product of disjoint cycles is useful in many more ways than just for compactly describing a permutation. As an example we mention the following. Let G be a group. Then $g_1, g_2 \in G$ are said to be *conjugate* if there is a $h \in G$ with $g_1 = hg_2h^{-1}$. It can be shown that two elements of S_X are conjugate if and only if they have the same *cycle structure*, that is, if their factorizations as products of disjoint cycles have the same number of k-cycles for each k. For example, the elements $(1,3)(2,4,6)$, $(1,3,5)(2,4)$ are conjugate in S_6.

3.1.2 Dihedral groups

Let $n \geq 3$ and define the graph $C_n = (V, E)$ with $V = \mathbb{Z}/n\mathbb{Z} = \{[0],[1],\ldots,[n-1]\}$ and $E = \{\{[i],[i+1]\} \mid [i] \in \mathbb{Z}/n\mathbb{Z}\}$. So when $n = 6$ we have

$$E = \{\{[0],[1]\},\{[1],[2]\},\{[2],[3]\},\{[3],[4]\},\{[4],[5]\},\{[5],[0]\}\}$$

so that the graphical representation of C_6 is displayed in Figure 3.2.

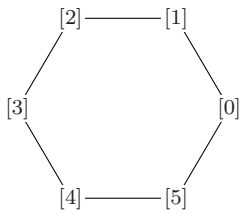


Figure 3.2: The graph C_6.

By D_n we denote the symmetry group of C_n (as defined in Example 3.0.1), which is called the *dihedral group* of order $2n$. (The order will be explained by the next proposition.)

For $[i] \in \mathbb{Z}/n\mathbb{Z}$ define the maps $\sigma_{[i]}, \rho_{[i]} : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}/n\mathbb{Z}$ by

$$\sigma_{[i]}([k]) = [k] + [i]$$
$$\rho_{[i]}([k]) = -[k] + [i].$$

Proposition 3.1.11 $D_n = \{\sigma_{[0]}, \ldots, \sigma_{[n-1]}, \rho_{[0]}, \ldots, \rho_{[n-1]}\}$.

Proof. First we show that the given maps lie in D_n. They are clearly bijective as $\sigma_{[i]}^{-1} = \sigma_{[-i]}$, $\rho_{[i]}^{-1} = \rho_{[i]}$. Let $\{[k], [k+1]\}$ be an edge of C_n. Then $\{[k]+[i], [k+1]+[i]\} = \{[k+i], [k+i+1]\}$, $\{-[k]+[i], -[k+1]+[i]\} = \{[-k+i-1], [-k+i]\}$ are also edges of C_n. So indeed the $\sigma_{[i]}, \rho_{[i]}$ map edges to edges and hence lie in D_n.

Now let $\tau \in D_n$ and let $\tau([0]) = [i]$. Because $\{\tau([0]), \tau([1])\}$ is an edge of C_n we see that $\tau([1])$ equals $[i+1]$ or $[i-1]$.

Suppose that $\tau([1]) = [i+1]$. Then by induction on k we show that $\tau([k]) = [k+i]$. For $k = 0, 1$ this holds. So suppose that $1 \leq k < n-1$ and $\tau([l]) = [l+i]$ for $0 \leq l \leq k$. As $\{\tau([k]), \tau([k+1])\}$ is an edge of C_n, we see that $\tau([k+1])$ is equal to $[k+i+1]$ or to $[k+i-1]$. But by induction $\tau([k-1]) = [k-1+i]$. As $n \geq 3$ we have that $[k-1] \neq [k+1]$ and therefore $\tau([k+1])$ cannot be equal to $[k+i-1]$. It follows that $\tau([k+1]) = [k+1+i]$ and $\tau = \sigma_{[i]}$.

Suppose that $\tau([1]) = [i-1]$. Then by induction on k we show that $\tau([k]) = [-k+i]$. For $k = 0, 1$ this holds. So suppose that $1 \leq k < n-1$ and $\tau([l]) = [-l+i]$ for $0 \leq l \leq k$. As $\{\tau([k]), \tau([k+1])\}$ is an edge of C_n, we see that $\tau([k+1])$ is equal to $[-k+i+1]$ or to $[-k+i-1]$. But by induction $\tau([k-1]) = [-k+1+i]$. As $n \geq 3$ we have that $[k-1] \neq [k+1]$ and therefore $\tau([k+1])$ cannot be equal to $[-k+i+1]$. It follows that $\tau([k+1]) = [-k-1+i]$ and $\tau = \rho_{[i]}$. □

It is straightforward to write down a multiplication table of D_n. For example $\sigma_{[i]}\sigma_{[j]}([k]) = \sigma_{[i]}(\sigma_{[j]}([k])) = \sigma_{[i]}([k]+[j]) = [k]+[j+i]$. Hence $\sigma_{[i]}\sigma_{[j]} = \sigma_{[i+j]}$. The other cases are left as exercises and we get

$$\begin{aligned}
\sigma_{[i]}\sigma_{[j]} &= \sigma_{[i+j]} \\
\sigma_{[i]}\rho_{[j]} &= \rho_{[i+j]} \\
\rho_{[i]}\sigma_{[j]} &= \rho_{[i-j]} \\
\rho_{[i]}\rho_{[j]} &= \sigma_{[i-j]}.
\end{aligned}$$

Remark 3.1.12 We can also give a geometric interpretation to the elements of D_n. For this put the nodes of C_n equidistantly on a circle. Then C_n becomes a regular n-gon. It is clear that $\sigma_{[i]}$ corresponds to a rotation of C_n by i units. Furthermore $\rho_{[i]}$ corresponds to a reflection, the axis of reflection being the line through the origin and perpendicular to the line connecting the points [0] and $[i]$.

3.2 Subgroups

A subgroup of a group G is a subset which is closed under the group operation of G, so that with the group operation of G it also becomes a group. More formally we have the following definition.

Definition 3.2.1 *Let G be a group. Then a subset $H \subset G$ is called a* subgroup *if $H \neq \emptyset$ and for all $g, h \in H$ we have that gh and g^{-1} lie in H.*

Remark 3.2.2 Let G be a group and $H \subset G$. Then H is a subgroup if and only if $1 \in H$ and $gh^{-1} \in H$ for all $g, h \in H$.

Example 3.2.3

- Consider the group $\mathbb{Z}$ with operation $+$. Let $n \in \mathbb{Z}$ then $n\mathbb{Z} = \{nk \mid k \in \mathbb{Z}\}$ is a subgroup of $\mathbb{Z}$.
- It is straightforward to check that $\{1, (1,2)(3,4), (1,3)(2,4), (1,4)(2,3)\}$ is a subgroup of S_4 (where 1 denotes the identity permutation).
- $\{\sigma_{[i]} \mid [i] \in \mathbb{Z}/n\mathbb{Z}\}$ is a subgroup of D_n.

Let G be a group with subgroup H. Let $g \in G$ then the set

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

is called a (left-) *coset* of H in G. The element g is said to be a *representative* of the coset gH.

Let $g_1, g_2 \in G$. Note that $g_2 \in g_1H$ if and only if $g_2 = g_1h$ for a certain $h \in H$ if and only if $g_1^{-1}g_2 \in H$. Now define the relation R_H on G by $g_1R_Hg_2$ if and only if $g_1^{-1}g_2 \in H$. We leave it as an exercise to verify that R_H is an equivalence relation. The equivalence class of g_1 is

$$\{g_2 \in G \mid g_1R_Hg_2\} = \{g_2 \in G \mid g_2 \in g_1H\} = g_1H,$$

hence it is preciesly the coset g_1H. So by Proposition 1.4.2 we see that two cosets are either disjoint or equal, and that $g_1H = g_2H$ if and only if $g_2 \in g_1H$, which happens if and only if $g_2 = g_1h$ for a certain $h \in H$.

Definition 3.2.4 *Let G be a group with subgroup H. Then the number of cosets of H in G is called the* index *of H in G. It is denoted $[G : H]$.*

Example 3.2.5 Here we look at the groups and subgroups given in Example 3.2.3. Let $G = \mathbb{Z}$, $n > 1$ and $H = n\mathbb{Z}$. The cosets of $n\mathbb{Z}$ in $\mathbb{Z}$ are precisely

$$0 + n\mathbb{Z} = \mathbb{Z}, 1 + n\mathbb{Z}, \ldots, (n-1) + n\mathbb{Z}$$

(note that here the group operation is +!). Indeed, the given cosets are all different as $i+n\mathbb{Z} = j+n\mathbb{Z}$ is equivalent to $j - i \in n\mathbb{Z}$ which is the same as $n|(j-i)$. Secondly, every $m \in \mathbb{Z}$ lies in one of the listed cosets, so there are not more of them. In fact, these cosets are exactly the congruence classes studied in Section 2.5.1. It follows that the index of $n\mathbb{Z}$ in $\mathbb{Z}$ is n.

Let $n \geq 3$, $G = D_n$, $H = \{\sigma_{[0]}, \ldots, \sigma_{[n-1]}\}$. Then H has two cosets in G, namely H and $\rho_{[0]}H$. These are different (as $\rho_{[0]} \notin H$) and as $\rho_{[0]}\sigma_{[i]} = \rho_{[-i]}$ we have $\rho_{[0]}H = \{\rho_{[0]}, \ldots, \rho_{[n-1]}\}$ so that every element of G lies in a coset. So in this case the index is 2.

Finally let $G = S_4$, $H = \{1, (1,2)(3,4), (1,3)(2,4), (1,4)(2,3)\}$. We list a few cosets:

$$\begin{aligned}
1H &= \{1, (1,2)(3,4), (1,3)(2,4), (1,4)(2,3)\} \\
(1,2)H &= \{(1,2), (3,4), (1,3,2,4), (1,4,2,3)\} \\
(1,3)H &= \{(1,3), (1,2,3,4), (2,4), (1,4,3,2)\} \\
(1,4)H &= \{(1,4), (1,2,4,3), (1,3,4,2), (2,3)\} \\
(1,2,3)H &= \{(1,2,3), (1,3,4), (2,4,3), (1,4,2)\} \\
(1,3,2)H &= \{(1,3,2), (2,3,4), (1,2,4), (1,4,3)\}.
\end{aligned}$$

Again we see that these are different and that each element of S_4 is contained in exactly one of them. So these are all cosets, and therefore the index is 6 in this case.

Proposition 3.2.6 *Let G be a finite group and $H \subset G$ a subgroup. Then $|G| = [G : H]|H|$.*

Proof. Let $g \in G$. First we show that $|H| = |gH|$. Let $f_g : G \to G$ be defined by $f_g(h) = gh$. Then f_g has an inverse map, namely $f_{g^{-1}}$. So f_g is bijective. Moreover, $f_g(H) = gH$, and therefore both sets have the same cardinality.

Let $g_1H, \ldots, g_mH$ be the distinct cosets of H in G. These form a partition of G (see Section 1.4). Hence

$$|G| = \sum_{i=1}^{m} |g_iH| = \sum_{i=1}^{m} |H| = m|H| = [G : H]|H|.$$

□

Corollary 3.2.7 (Lagrange's theorem) [1] *Let G be a finite group and $H \subset G$ a subgroup. Then $|H|$ divides $|G|$.*

[1] It is funny to realize that in Lagrange's time groups had not yet been introduced. However, in his work a statement can be found that is a special case of this theorem. Hence the naming is not unjustified.

3.3 Normal subgroups

Definition 3.3.1 *Let G be a group and $H \subset G$ a subgroup. Then H is said to be* normal *if $ghg^{-1} \in H$ for all $g \in G$ and $h \in H$.*

Example 3.3.2 Let $n \geq 3$, $G = D_n$ and $H = \{\sigma_{[0]}, \ldots, \sigma_{[n-1]}\}$. Then H is a subgroup of G. We have $\sigma_{[j]}\sigma_{[i]}\sigma_{[j]}^{-1} \in H$ as H is a subgroup. Furthermore $\rho_{[j]}\sigma_{[i]}\rho_{[j]}^{-1} = \rho_{[j]}\sigma_{[i]}\rho_{[j]} = \rho_{[j]}\rho_{[i+j]} = \sigma_{[-i]}$. It follows that H is a normal subgroup.

Let G be a group and $H \subset G$ a normal subgroup. By G/H we denote the set of all cosets of H in G. On G/H we define an operation by

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

However, since this operation is defined using representatives of the cosets, rather than the cosets themselves, we have to show that it is well-defined. (This is very similar to the operations defined in Section 2.5.1 and in Section 2.6.1.) So let $g_i, \hat{g}_i \in G$ be such that $g_iH = \hat{g}_iH$ for $i = 1, 2$. Then there are $h_i \in H$ with $\hat{g}_i = g_ih_i$ for $i = 1, 2$. Hence

$$(\hat{g}_1\hat{g}_2)H = (g_1h_1g_2h_2)H = (g_1g_2)(g_2^{-1}h_1g_2h_2)H = (g_1g_2)H$$

(note that $g_2^{-1}h_1g_2 \in H$ as H is normal, and therefore $g_2^{-1}h_1g_2h_2 \in H$ and in general $hH = H$ for all $h \in H$). It follows that the above definition defines an operation on cosets.

Proposition 3.3.3 *Let G be a group and $H \subset G$ a normal subgroup. With the operation defined above the set G/H is a group.*

Proof. Everything follows directly from the definition of the operation on G/H and the fact that the original operation of G makes G into a group. Indeed, the neutral element is $1H = H$. The inverse of gH is $g^{-1}H$ (as $(gH)(g^{-1}H) = (gg^{-1})H = 1H = H$). Finally the associativity of the operation follows immediately from the associativity of the operation of G. □

The group G/H given by this proposition is called the *quotient group* of G by H. As with rings, it is not immediately obvious what a quotient group looks like. However, if the quotient is finite then we can just make a list of its elements, and compute a multiplication table.

Example 3.3.4 Let $n \geq 3$, $G = D_n$ and $H = \{\sigma_{[0]}, \ldots, \sigma_{[n-1]}\}$. In Example 3.3.2 we have seen that H is normal. Example 3.2.5 has the cosets of H in G: H and $\rho_{[0]}H$. The only product of interest is the one of $\rho_{[0]}H$ with itself. We have

$$(\rho_{[0]}H)(\rho_{[0]}H) = (\rho_{[0]}\rho_{[0]})H = \sigma_0H = H.$$

Example 3.3.5 Now let $G = D_4 = \{\sigma_{[0]}, \ldots, \sigma_{[3]}, \rho_{[0]}, \ldots, \rho_{[3]}\}$. Let $H = \{\sigma_{[0]}, \sigma_{[2]}\}$. As $\sigma_{[0]}$ is the neutral element, $g\sigma_{[0]}g^{-1} = \sigma_{[0]}$ for all $g \in G$. Since we know the multiplication table of G (see Section 3.1.2) it is straightforward to see that also $g\sigma_{[2]}g^{-1} = \sigma_{[2]}$ for all $g \in G$. It follows that H is a normal subgroup of G. Since $|G| = 8$ and $|H| = 2$, by Proposition 3.2.6 we have that $|G/H| = \frac{8}{2} = 4$. Four cosets of H are: H, $\sigma_{[1]}H = \{\sigma_{[1]}, \sigma_{[3]}\}$, $\rho_{[0]}H = \{\rho_{[0]}, \rho_{[2]}\}$, $\rho_{[1]}H = \{\rho_{[1]}, \rho_{[3]}\}$. Now we can compute the multiplication table of G/H. For example, $(\sigma_{[1]}H)(\rho_{[1]}H) = (\sigma_{[1]}\rho_{[1]})H = \rho_{[2]}H = \rho_{[0]}H$; where the last equality follows from the fact that $\rho_{[2]} \in \rho_{[0]}H$.

For brevity we write $\bar{g}$ instead of gH, so that G/H consists of $\bar{1}$, $\bar{\sigma}_{[1]}$, $\bar{\rho}_{[0]}$, $\bar{\rho}_{[1]}$. After some computations we arrive at the following multiplication table:

	$\bar{1}$	$\bar{\sigma}_{[1]}$	$\bar{\rho}_{[0]}$	$\bar{\rho}_{[1]}$
$\bar{1}$	$\bar{1}$	$\bar{\sigma}_{[1]}$	$\bar{\rho}_{[0]}$	$\bar{\rho}_{[1]}$
$\bar{\sigma}_{[1]}$	$\bar{\sigma}_{[1]}$	$\bar{1}$	$\bar{\rho}_{[1]}$	$\bar{\rho}_{[0]}$
$\bar{\rho}_{[0]}$	$\bar{\rho}_{[0]}$	$\bar{\rho}_{[1]}$	$\bar{1}$	$\bar{\sigma}_{[1]}$
$\bar{\rho}_{[1]}$	$\bar{\rho}_{[1]}$	$\bar{\rho}_{[0]}$	$\bar{\sigma}_{[1]}$	$\bar{1}$

3.4 Homomorphisms of groups

Definition 3.4.1 *Let G_1, G_2 be groups. A map $f : G_1 \to G_2$ is called a* group homomorphism *if $f(g_1g_2) = f(g_1)f(g_2)$ for all $g_1, g_2 \in G_1$. (Note that in this equality we use the operation of G_1 on the left, whereas on the right we use the operation of G_2.)*

Lemma 3.4.2 *Let G_1, G_2 be groups with respective neutral elements 1_{G_1}, 1_{G_2}. Let $f : G_1 \to G_2$ be a homomorphism. Then $f(1_{G_1}) = 1_{G_2}$ and $f(g^{-1}) = f(g)^{-1}$ for all $g \in G_1$.*

Proof. Note that $f(1_{G_1}) = f(1_{G_1} \cdot 1_{G_1}) = f(1_{G_1}) \cdot f(1_{G_1})$. We multiply this equation on both sides with $f(1_{G_1})^{-1}$ (which is an element of G_2!), and obtain the desired result.

Let $g \in G_1$. Then using the first part we have

$$1_{G_2} = f(1_{G_1}) = f(g \cdot g^{-1}) = f(g) \cdot f(g^{-1}).$$

So multiplying by $f(g)^{-1}$ we obtain $f(g)^{-1} \cdot 1_{G_2} = f(g)^{-1}f(g)f(g^{-1})$ which simplyfies to $f(g^{-1}) = f(g)^{-1}$. □

Example 3.4.3 Let $G_1 = \mathbb{Z}$ which is a group with respect to addition. Let

$$G_2 = \{\left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right) \mid m \in \mathbb{Z}\},$$

with the operation of matrix multiplication. It is straightforward to see that also G_2 is a group. Define $f : G_1 \to G_2$ by $f(m) = \left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)$. Then f is a homomorphism of groups, because

$$f(m+n) = \left(\begin{smallmatrix} 1 & m+n \\ 0 & 1 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & m \\ 0 & 1 \end{smallmatrix}\right)\left(\begin{smallmatrix} 1 & n \\ 0 & 1 \end{smallmatrix}\right) = f(m)f(n).$$

Note that the operation of G_1 is addition, whereas the operation of G_2 is multiplication. The neutral element of G_1 is 0 which is indeed mapped to the neutral element of G_2, which is $\left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$.

A bijective group homomorphism is called a *group isomorphism*. It is immediate that the homomorphism of Example 3.4.3 is an isomorphism. Two groups G, H are said to be *isomorphic* if there is a group isomorphism $f : G \to H$. In that case we write $G \cong H$.

Example 3.4.4 Define the following elements of S_4: $\pi_1 = 1$ (the identity permutation), $\pi_2 = (1,2)(3,4)$, $\pi_3 = (1,3)(2,4)$, $\pi_4 = (1,4)(2,3)$. As already observed in Example 3.2.3, $M = \{\pi_1, \pi_2, \pi_3, \pi_4\}$ is a subgroup of S_4. Some calculations show that the multiplication table of M is

	π_1	π_2	π_3	π_3
π_1	π_1	π_2	π_3	π_3
π_2	π_2	π_1	π_4	π_3
π_3	π_3	π_4	π_1	π_2
π_4	π_4	π_3	π_2	π_1

Now consider the group G/H from Example 3.3.5. Set $g_1 = \bar{1}$, $g_2 = \bar{\sigma}_{[1]}$, $g_3 = \bar{\rho}_{[0]}$, $g_4 = \bar{\rho}_{[1]}$. If we write the g_i into the multiplication table of G/H in the mentioned example, then we see that we get exactly the same table as for M, the only difference being that in the table for M we used the letter π and in the table for G/H the letter g. From this it follows that the map sending π_i to g_i is a group homomorphism. It is obviously bijective, so in fact it is an isomorphism.

Definition 3.4.5 *Let $f : G_1 \to G_2$ be a group homomorphism. The set $\{g \in G_1 \mid f(g) = 1_{G_2}\}$ (where 1_{G_2} is the neutral element of G_2) is called the* kernel *of f and denoted* $\ker(f)$.

Remark 3.4.6 With f as in the previous definition we have that f is injective if and only if $\ker(f) = \{1_{G_1}\}$. Indeed, suppose that f is injective. Then by Lemma 3.4.2 we see that $1_{G_1} \in \ker(f)$. Let $g \in G_1$ lie in $\ker(f)$. Then $f(g) = 1_{G_2}$ and by the injectivity of f it follows that $g = 1_{G_1}$. Hence $\ker(f) = \{1_{G_1}\}$. Conversely, suppose that $\ker(f) = \{1_{G_1}\}$. Let $g_1, g_2 \in G_1$ be such that $f(g_1) = f(g_2)$. Then $f(g_2^{-1}g_1) = 1_{G_2}$, from which it follows that $g_2^{-1}g_1 = 1_{G_1}$ and $g_1 = g_2$. We conclude that f is injective.

The next theorem can often be used to show that a given quotient group is isomorphic to another given group.

Theorem 3.4.7 *Let $f : G_1 \to G_2$ be a group homomorphism and set $H_1 = \ker(f)$, $H_2 = f(G_1) = \{f(g) \mid g \in G_1\}$. Then H_1 is a normal subgroup of G_1 and H_2 is a subgroup of G_2. Moreover, the map $\bar{f} : G_1/H_1 \to H_2$, $\bar{f}(gH_1) = f(g)$ is well-defined and an isomorphism of groups, so that $G_1/H_1 \cong H_2$.*

Proof. We denote the neutral element of G_i by 1_{G_i} for $i = 1, 2$.

By Lemma 3.4.2, $1_{G_1} \in \ker(f)$. Using the same lemma we see that for $g_1, g_2 \in H_1$ we have $f(g_1 g_2^{-1}) = f(g_1)f(g_2)^{-1} = 1_{G_2} \cdot 1_{G_2}^{-1} = 1_{G_2}$. So $g_1 g_2^{-1} \in H_1$. As observed in Remark 3.2.2, this shows that H_1 is a subgroup of G_1. Let $h \in H_1$, $g \in G_1$. Again using Lemma 3.4.2 we obtain $f(ghg^{-1}) = f(g)f(h)f(g)^{-1} = f(g)1_{G_2}f(g)^{-1} = 1_{G_2}$. Hence $ghg^{-1} \in H_1$ and we see that H_1 is normal.

Lemma 3.4.2 also shows that $1_{G_2} \in H_2$. Let $g_1, g_2 \in H_2$, then there are $h_1, h_2 \in G_1$ with $g_i = f(h_i)$. Now $g_1 g_2^{-1} = f(h_1)f(h_2)^{-1} = f(h_1 h_2^{-1})$ and we have $g_1 g_2^{-1} \in H_2$. It follows that H_2 is a subgroup of G_2.

Now let $g_1, g_2 \in G_1$ and suppose that $g_1 H_1 = g_2 H_1$. Then $g_2 = g_1 h$ for a certain $h \in H_1$. Hence $f(g_2) = f(g_1 h) = f(g_1)f(h) = f(g_1) \cdot 1_{G_1} = f(g_1)$. It follows that the map $\bar{f}$ is well-defined. The fact that $\bar{f}$ is a homomorphism follows directly from the analogous property of f: $\bar{f}(g_1 H_1 \cdot g_2 H_2) = \bar{f}(g_1 g_2 H_1) = f(g_1 g_2) = f(g_1)f(g_2) = \bar{f}(g_1 H_1)\bar{f}(g_2 H_1)$. Furthermore, $\bar{f}(g_1 H_1) = \bar{f}(g_2 H_1)$ if and only if $f(g_1) = f(g_2)$ if and only if $g_1^{-1} g_2 \in H_1$ if and only if $g_2 = g_1 h$ for a certain $h \in H_1$ if and only if $g_1 H_1 = g_2 H_1$. So $\bar{f}$ is injective. Finally $\bar{f}$ is surjective because $H_2 = f(G_1) = \bar{f}(G_1/H_1)$. □

Remark 3.4.8 The previous theorem is sometimes called the First Isomorphism Theorem. This suggests that there are more theorems of this kind, and indeed we have the following:

- (Second Isomorphism Theorem) Let G be a group with subgroups H, N and suppose that N is a normal subgroup of G. Then $H \cap N$ is a normal subgroup of H and $HN = \{hn \mid h \in H, n \in N\}$ is a subgroup of G and $H/(H \cap N) \cong HN/N$.
- (Third Isomorphism Theorem) Let G be a group with normal subgroups H, N with $N \subset H$. Then H/N is a normal subgroup of G/N and $(G/N)/(H/N) \cong G/H$.

(Compare Remark 2.6.6.)

3.5 Actions of groups

When considering the symmetries of an object, the way these operate is often encoded by a so-called group action.

Definition 3.5.1 *Let G be a group and X a set. We say that G* acts on *X if there is a function $\alpha : G \times X \to X$ with*

1. *$\alpha(1, x) = x$ for all $x \in X$ (where $1 \in G$ is its neutral element),*
2. *$\alpha(g, \alpha(h, x)) = \alpha(gh, x)$ for all $g, h \in G$ and $x \in X$.*

If it is clear which α we are using then we usually write $g \cdot x$ instead of $\alpha(g, x)$. With that notation the requirements of the previous definition become

1. $1 \cdot x = x$ for all $x \in X$,
2. $g \cdot (h \cdot x) = (gh) \cdot x$ for all $g, h \in G$ and $x \in X$.

Example 3.5.2 Let's start with a rather trivial example. Let $G = \mathbb{R}_{>0} = \{a \in \mathbb{R} \mid a > 0\}$, which is a group with respect to the multiplication. Let $X = \mathbb{R}$. For $a \in G$ and $x \in X$ we set $a \cdot x = ax$. This is a group action: $1 \cdot x = 1x = x$ and $a \cdot (b \cdot x) = a \cdot (bx) = a(bx) = (ab)x = (ab) \cdot x$.

Remark 3.5.3 Let the group G act on the set X. Let $g \in G$ and define $\psi_g : X \to X$ by $\psi_g(x) = g \cdot x$. Then this map is a bijection as its inverse map is $\psi_{g^{-1}}$. (Indeed: $\psi_{g^{-1}}(\psi_g(x)) = \psi_{g^{-1}}(g \cdot x) = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = 1 \cdot x = x$.)

The following lemma says that having a group action is the same as having a group homomorphism. So it would be possible to dispense with group actions, and only talk about homomorphisms. However, it turns out that in many situations the language and notation of group actions makes it easier to formulate theorems and proofs.

Lemma 3.5.4 *Let G be a group and X a set.*

1. *Suppose that G acts on X and define $f : G \to S_X$ by $f(g) = \psi_g$ (notation as in Remark 3.5.3). Then f is a group homomorphism.*
2. *Let $f : G \to S_X$ be a group homomorphism, and write $\pi_g = f(g)$ for $g \in G$. Then $g \cdot x = \pi_g(x)$ defines an action of G on X.*

Proof. First observe that Remark 3.5.3 says that $\psi_g \in S_X$. So the range of f in fact is S_X. Now let $g, h \in G$ and $x \in X$, then $\psi_g \circ \psi_h(x) = \psi_g(\psi_h(x)) = \psi_g(h \cdot x) = g \cdot (h \cdot x) = (gh) \cdot x = \psi_{gh}(x)$. Hence $\psi_g \circ \psi_h = \psi_{gh}$ which is the same as saying that f is a group homomorphism.

For the second part, note that by Lemma 3.4.2, $\pi_1 = f(1)$ is the neutral element of S_X which is the identity mapping $X \to X$. So $1 \cdot x = \pi_1(x) = x$. Secondly, for $g, h \in G$ we have that $f(gh) = f(g)f(h)$ is the same as $\pi_g \circ \pi_h = \pi_{gh}$. So for $x \in X$ we have $g \cdot (h \cdot x) = \pi_g(\pi_h(x)) = \pi_g \circ \pi_h(x) = \pi_{gh}(x) = (gh) \cdot x$. □

Example 3.5.5 Let $G = D_4 = \{\sigma_{[0]}, \ldots, \sigma_{[3]}, \rho_{[0]}, \ldots, \rho_{[3]}\}$. Set $X = \{x_1, x_2\}$ where $x_1 = \{[0], [2]\}$, $x_2 = \{[1], [3]\}$. For such a subset $x = \{[a], [b]\}$ and $\tau \in G$ we set $\tau \cdot x = \{\tau([a]), \tau([b])\}$. We claim that this defines an action of G on X.

The first thing that we have to check for this is that $\tau \cdot x$ does lie in X for all $\tau \in G$, $x \in X$. This is not explicitly stated in Definition 3.5.1, but it is implicit in the assumption that we have a map $\alpha : G \times X \to X$. In other words, we have to check that indeed we have such a map. This is just done by checking all possibilities. For example, $\sigma_{[1]} \cdot x_1 = \{[1], [3]\} = x_2$, $\sigma_{[1]} \cdot x_2 = \{[2], [0]\} = x_1$. Continuing like this we see that $\tau \cdot x \in X$ for all $\tau \in G$, $x \in X$.

Next we check the two conditions of Definition 3.5.1. Let $x = \{[a], [b]\} \in X$. Then $\sigma_{[0]} \cdot x = \{\sigma_{[0]}([a]), \sigma_{[0]}([b])\} = \{[a], [b]\} = x$. Secondly, for $\tau, \pi \in G$ we have $\tau \cdot (\pi \cdot x) = \tau \cdot \{\pi([a]), \pi([b])\} = \{\tau(\pi([a])), \tau(\pi([b]))\} = \{(\tau\pi)([a]), (\tau\pi)([b])\} = (\tau\pi) \cdot x$. We conclude that our definition gives an action of G on X.

Now let f be the homomorphism provided by Lemma 3.5.4. We write elements of S_X as products of disjoint cycles. Then $f(\sigma_{[0]}) = 1$ (by which we denote the identity mapping on X), $f(\sigma_{[1]}) = (x_1, x_2)$, as seen above. By direct computation (which, by the way, we already perfomed when we checked that $\tau \cdot x$ lies in X for all $\tau \in G$ and $x \in X$) we see that

$$f(\sigma_{[2]}) = 1,\ f(\sigma_{[3]}) = (x_1, x_2),\ f(\rho_{[0]}) = 1,\ f(\rho_{[1]}) = (x_1, x_2),\ f(\rho_{[2]}) = 1,\ f(\rho_{[3]}) = (x_1, x_2).$$

We see that $\ker(f) = \{\sigma_{[0]}, \sigma_{[2]}, \rho_{[0]}, \rho_{[2]}\}$. This is a normal subgroup of G and by Theorem 3.4.7 we have $G/\ker(f) \cong S_X$.

Definition 3.5.6 *Let G be a group acting on the set X. For $x \in X$ the set $\{g \cdot x \mid g \in G\}$ is called the* orbit *of x, and denoted $G \cdot x$.*

Example 3.5.7 Let $G = D_4$ as in Example 3.5.5. But this time we let X be the set that consists of all subsets of $\mathbb{Z}/4\mathbb{Z}$ having two elements, i.e.,

$$X = \{\{[0], [1]\}, \{[0], [2]\}, \{[0], [3]\}, \{[1], [2]\}, \{[1], [3]\}, \{[2], [3]\}\}.$$

The action of G is defined in the same way as in Example 3.5.5, that is $\tau \cdot \{[a], [b]\} = \{\tau([a]), \tau([b])\}$. Again we have that this defines an action of G on X. (Note that in this case we do not need to check

that $\tau \cdot x$ really lies in X as $\tau \cdot x$ is a set of two distinct elements (τ being injective) and X contains all such sets.) We compute a few orbits of G on X:

$$G \cdot \{[0],[1]\} = \{\{[0],[1]\},\{[1],[2]\},\{[2],[3]\},\{[0],[3]\}\}$$
$$G \cdot \{[0],[2]\} = \{\{[0],[2]\},\{[1],[3]\}\}.$$

Remark 3.5.8 Let a group G act on a set X. Then we define a relation R_G on X by xR_Gy if and only if $y \in G \cdot x$. This an equivalence relation. Indeed:

- $x = 1 \cdot x$, so that xR_Gx for all $x \in X$,
- if for $x, y \in X$ we have xR_Gy then $y \in G \cdot x$ so that there is a $g \in G$ with $y = g \cdot x$; then $g^{-1} \cdot y = g^{-1} \cdot (g \cdot x) = (g^{-1}g) \cdot x = 1 \cdot x = x$, and we see that $x \in G \cdot y$ and therefore yR_Gx,
- if for $x, y, z \in X$ we have xR_Gy and yR_Gz then there are $g, h \in G$ with $y = g \cdot x$ and $z = h \cdot y$, whence $(hg) \cdot x = h \cdot (g \cdot x) = z$ and we conclude xR_Gz.

The equivalence classes of this relation are exactly the orbits of G. Therefore, as seen in Section 1.4, two orbits are either equal or disjoint, in other words, they the form a partition of X. The two orbits found in Example 3.5.7 cover all elements of X. So there can be no other orbits, and we have that the set X of the mentioned example has exactly two orbits under the action of G.

Definition 3.5.9 *Let G be a group acting on the set X. For $x \in X$ the set $\{g \in G \mid g \cdot x = x\}$ is called the* stabilizer *of x in G, and denoted G_x.*

Lemma 3.5.10 *Let G be a group acting on the set X. Let $x \in X$; then G_x is a subgroup of G.*

Proof. Note that $1 \in G_x$ by the first item of Definition 3.5.1. Let $g, h \in G_x$, so that $g \cdot x = h \cdot x = x$. Then $h^{-1} \cdot x = h^{-1} \cdot (h \cdot x) = (h^{-1}h) \cdot x = 1 \cdot x = x$. Hence $h^{-1} \in G_x$. Furthermore, $(gh) \cdot x = g \cdot (h \cdot x) = g \cdot x = x$ and $gh \in G_x$. We conclude that G_x is a subgroup of G. □

Example 3.5.11 Let G, X be as in Example 3.5.7. Let $x = \{[0],[1]\}$ then by checking for all $\tau \in G$ whether $\tau \cdot x = x$ we see that $G_x = \{\sigma_{[0]}, \rho_{[1]}\}$. Let $y = \{[0],[2]\}$ then again by checking the elements of G we obtain $G_y = \{\sigma_{[0]}, \sigma_{[2]}, \rho_{[0]}, \rho_{[2]}\}$.

In this example we see the larger orbit corresponds to the smaller stabilizer. This is intuitively clear because the more elements of G stabilize x the fewer elements of G are available to move x to some other element. The next theorem makes this precise.

Theorem 3.5.12 *Let G be a* finite *group acting on the set X. Let $x \in X$. Then $|G \cdot x| = [G : G_x]$.*

Proof. By G/G_x we denote the set of cosets gG_x of G_x in G. However, note that in general this is not a group because G_x need not be a normal subgroup. By definition, $|G/G_x| = [G : G_x]$.

Define $f : G \cdot x \to G/G_x$ by $f(g \cdot x) = gG_x$. There is a potential ambiguity in this definition as it may happen that $g \cdot x = h \cdot x$ for $g, h \in G$ and $g \neq h$. But in that case, $(h^{-1}g) \cdot x = h^{-1} \cdot (g \cdot x) = h^{-1} \cdot (h \cdot x) = (h^{-1}h) \cdot x = 1 \cdot x = x$ and $h^{-1}g \in G_x$, implying $gG_x = hG_x$. So the image under f of $y \in G \cdot x$ is uniquely defined.

It is obvious that f is surjective as we can let g run through G, and so $f(g \cdot x) = gG_x$ runs through G/G_x. Furthermore, $f(g \cdot x) = f(h \cdot x)$ is the same as $gG_x = hG_x$, which is equivalent to $h^{-1}g \in G_x$, or $(h^{-1}g) \cdot x = x$. By acting with h on both sides we see that the latter is equivalent to $g \cdot x = h \cdot x$ and we conclude that f is injective.

We have a bijection between $G \cdot x$ and G/G_x so it follows that both sets have the same cardinality. □

Combining this with Proposition 3.2.6 we immediately obtain the following corollary.

Corollary 3.5.13 (orbit-stabilizer theorem) *Let G be a finite group acting on the set X. Let $x \in X$. Then $|G \cdot x||G_x| = |G|$.*

3.6 Counting colourings

In this section we look at the problem mentioned at the beginning of the chapter: counting the number of non-equivalent necklaces of length n with beads of q colours. The trick is to translate the notion of equivalence to an action of a group, in such a way that two necklaces are equivalent if and only if they lie in the same orbit. Then we will see how a theorem, called Burnside's lemma, can be applied to solve the problem.

The same method works in many situations where one wants to count the number of inequivalent ways in which a certain object can be coloured. Therefore we will describe it in more general terms, and talk about colourings of a set with a group action. But first we will show Burnside's lemma.

Lemma 3.6.1 *Let G be a finite group acting on a set X. Let $x, y, z \in X$ be such that $y, z \in G \cdot x$. Then $|G_y| = |G_z|$.*

Proof. We have $G \cdot y = G \cdot z = G \cdot x$ (indeed, the orbits are the equivalence classes of the equivalence relation R_G of Remark 3.5.8 and x, y, z all lie in the same class). Therefore the lemma follows from Corollary 3.5.13. □

The next theorem was actually first proved by Frobenius in 1887 (although some form of it appears also in work of Cauchy of 1845). In his 1897 book on group theory, Burnside correctly attibuted it to Frobenius. But in the 1911 second edition of the book this attribution was no longer there. This may have led later authors to think that it is due to Burnside.

Theorem 3.6.2 (Burnside's lemma) *Let G be a finite group acting on the finite set X. Let N be the number of orbits of G on X. For $g \in G$ define $F(g) = \{x \in X \mid g \cdot x = x\}$. Then*

$$N = \frac{1}{|G|} \sum_{g \in G} |F(g)|.$$

Proof. Define $U = \{(g, x) \in G \times X \mid g \cdot x = x\}$. For $g \in G$ define $A_g = \{(g, x) \mid x \in F(g)\}$ which is a subset of U. Note that trivially $A_{g_1} \cap A_{g_2} = \emptyset$ if $g_1 \neq g_2$ and that U is the union of the sets A_g as g runs through G. Furthermore, $|A_g| = |F(g)|$. Therefore, $|U| = \sum_{g \in G} |A_g| = \sum_{g \in G} |F(g)|$.

For $x \in X$ define $B_x = \{(g, x) \mid g \in G_x\}$ which is a subset of U. Note that trivially $B_{x_1} \cap B_{x_2} = \emptyset$ if $x_1 \neq x_2$ and that U is the union of the sets B_x as x runs through X. Furthermore, $|B_x| = |G_x|$. Therefore, $|U| = \sum_{x \in X} |B_x| = \sum_{x \in X} |G_x|$.

Let $G \cdot x_1, \ldots, G \cdot x_N$ be the orbits of G in X. By Lemma 3.6.1 we have $|G_y| = |G_z|$ for y, z lying in the same $G \cdot x_i$. Hence, using Corollary 3.5.13,

$$\sum_{x \in X} |G_x| = \sum_{i=1}^{N} \sum_{y \in G \cdot x_i} |G_y| = \sum_{i=1}^{N} |G \cdot x_i||G_{x_i}| = \sum_{i=1}^{N} |G| = N|G|.$$

Putting things together we obtain $N|G| = |U| = \sum_{g \in G} |F(g)|$. □

Example 3.6.3 Let $G = D_4$ and X be as in Example 3.5.7. Then we have $N = 2$, $|G| = 8$. So by the previous theorem we must have $\sum_{g \in G} |F(g)| = 16$. Let us check that. As $\sigma_{[0]}$ is the identity we have $F(\sigma_{[0]}) = X$, having 6 elements. By direct checking we see that

$$\begin{aligned}
&F(\sigma_{[1]}) = F(\sigma_{[3]}) = \emptyset,\\
&F(\sigma_{[2]}) = F(\rho_{[0]}) = F(\rho_{[2]}) = \{\{[0], [2]\}, \{[1], [3]\}\},\\
&F(\rho_{[1]}) = \{\{[0], [1]\}, \{[2], [3]\}\},\\
&F(\rho_{[3]}) = \{\{[0], [3]\}, \{[1], [2]\}\}.
\end{aligned}$$

We see that the sum of the sizes of the various $F(g)$ is indeed 16.

Now we turn our attention to colourings. Let X be a finite set. Let C be another finite set, whose elements are called *colours*. Write $|C| = q$. A map $\gamma : X \to C$ is said to be a *q-colouring* of X. (This reasonable: γ associates a colour to every element of X.)

Let G be a group acting on X. Two q-colourings γ, δ are called *G-equivalent* if there is a $g \in G$ such that

$$\delta(x) = \gamma(g^{-1} \cdot x) \text{ for all } x \in X.$$

(It obviously comes down to the same thing to require that there is a $g \in G$ with $\delta(x) = \gamma(g \cdot x)$, but for technical reasons, that become clear later, we prefer the g^{-1}.)

Example 3.6.4 Let $X = \mathbb{Z}/5\mathbb{Z}$ and $G = D_5$. Since G consists of bijections $X \to X$, it has a natural action on X by $\tau \cdot x = \tau(x)$. Of course, G is also the symmetry group of the graph C_5 (see Section 3.1.2). We let $C = \{\mathrm{A}, \mathrm{B}, \mathrm{C}\}$. We consider two colourings γ, δ of X, which we give by writing the graph C_5 and next to each node we put its colour.

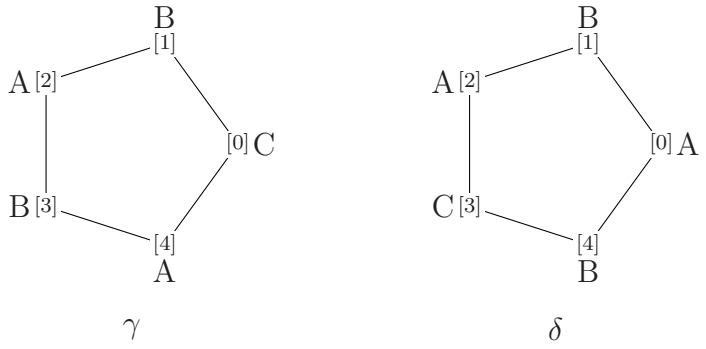


Figure 3.3: Two 3-colourings of $\mathbb{Z}/5\mathbb{Z}$.

We see that the second colouring is obtained from the first one by rotating the graph by three units. The $\sigma_{[i]}$ are the rotations in D_5, whereas the $\rho_{[i]}$ are the reflections. So we try to see whether γ, δ are equivalent via σ_3. In fact,

$$\begin{aligned} 0 &\xrightarrow{\sigma_3^{-1}} 2 \xrightarrow{\gamma} \mathrm{A} \\ 1 &\longrightarrow 3 \longrightarrow \mathrm{B} \\ 2 &\longrightarrow 4 \longrightarrow \mathrm{A} \\ 3 &\longrightarrow 0 \longrightarrow \mathrm{C} \\ 4 &\longrightarrow 1 \longrightarrow \mathrm{B} \end{aligned}$$

from which we see that $\delta(x) = \gamma(\sigma_3^{-1} \cdot x)$ for all $x \in X$.

Now let

$$\Gamma = \{q\text{-colourings } \gamma : X \to C\}$$

be the set of all q-colourings of X. We translate the notion of G-equivalence to an action of G, that is, we define an action of G on Γ by

$$(g \cdot \gamma)(x) = \gamma(g^{-1} \cdot x)$$

(so $g \cdot \gamma$ is the colouring that maps x to $\gamma(g^{-1} \cdot x)$). We have to check that this indeed defines an action. Firstly,

$$(1 \cdot \gamma)(x) = \gamma(1^{-1} \cdot x) = \gamma(1 \cdot x) = \gamma(x),$$

so that $1 \cdot \gamma = \gamma$ for all $\gamma \in \Gamma$. Secondly, for $g, h \in G$, $\gamma \in \Gamma$ and $x \in X$ we have

$$(g \cdot (h \cdot \gamma))(x) = (h \cdot \gamma)(g^{-1} \cdot x) = \gamma(h^{-1} \cdot (g^{-1} \cdot x)) = \gamma((h^{-1}g^{-1}) \cdot x) = \gamma((gh)^{-1} \cdot x) = ((gh) \cdot \gamma)(x),$$

from which it follows that $g \cdot (h \cdot \gamma) = (gh) \cdot \gamma$. (Here we used the general fact that $h^{-1}g^{-1} = (gh)^{-1}$; this computation is also the reason why we used g^{-1} in the definition of G-equivalence instead of g.)

Concluding, we have that two q-colourings of X are G-equivalent if and only if they lie in the same G-orbit in Γ. Therefore the number of non-equivalent G-colourings is equal to the number of orbits of G in Γ. In order to compute this number we can use Burnside's lemma. For that we need to compute the various $|F(g)|$. The following lemma shows how to do that. In order to formulate it we need one more piece of terminology. Let $\pi \in S_n$ and consider its factorization as a product of disjoint k-cycles for $k \geq 2$ (Section 3.1.1). Then for each i with $\pi(i) = i$ we add the 1-cycle (i) to this factorization. The result is called the *complete factorization* of π.

Lemma 3.6.5 *Let G be a group acting on the finite set X. Let C be a set of colours of size q, and $\Gamma = \{\gamma : X \to C\}$ the set of q-colourings of X. Let $f : G \to S_X$ the homomorphism corresponding to the action of X (Lemma 3.5.4). For $g \in G$ let $c(g)$ be the number of cycles in the* complete *factorization of $f(g)$. For $g \in G$ let $F(g) = \{\gamma \in \Gamma \mid g \cdot \gamma = \gamma\}$. Then $|F(g)| = q^{c(g)}$.*

Proof. Write $X = \{x_1, \ldots, x_n\}$. Let $g \in G$. Then we say that a $\gamma \in \Gamma$ is *g-flat* if for every cycle $(x_{i_1}, \ldots, x_{i_k})$ appearing in the complete factorization of $f(g)$ we have $\gamma(x_{i_1}) = \gamma(x_{i_2}) = \cdots = \gamma(x_{i_k})$. We claim that $\gamma \in F(g)$ if and only if γ is g-flat. First suppose that $\gamma \in F(g)$ and consider a cycle $(x_{i_1}, \ldots, x_{i_k})$ appearing in the complete factorization of $f(g)$. Then

$$g \cdot x_{i_1} = x_{i_2}, g \cdot x_{i_2} = x_{i_3}, \cdots, g(x_{i_k}) = x_{i_1}.$$

Hence $g^{-1} \cdot x_{i_1} = x_{i_k}, g^{-1} \cdot x_{i_k} = x_{i_{k-1}}, \cdots, g^{-1} \cdot x_{i_2} = x_{i_1}$. So for $2 \leq j \leq k$ it follows that $\gamma(x_{i_j}) = (g \cdot \gamma)(x_{i_j}) = \gamma(g^{-1} \cdot x_{i_j}) = \gamma(x_{i_{j-1}})$. Therefore $\gamma(x_{i_j}) = \gamma(x_{i_1})$ for $2 \leq j \leq k$ and we conclude that γ is g-flat.

Secondly, suppose that γ is g-flat. Then again we consider a cycle as above. This time using the assumption of g-flatness, we infer that $(g \cdot \gamma)(x_{i_j}) = \gamma(g^{-1} \cdot x_{i_j}) = \gamma(x_{i_{j-1}}) = \gamma(x_{i_j})$ for $2 \leq j \leq k$. But this means that $\gamma \in F(g)$. The claim is proved.

It follows that the choice of an element of C for each cycle appearing in the complete factorization of $f(g)$ uniquely determines an element of $F(g)$. Since C has q elements, and there are $c(g)$ cycles, the number of such choices is eaxctly $q^{c(g)}$. □

Combining this with Theorem 3.6.2 we immediately obtain the following corollary.

Corollary 3.6.6 *Let the notation be as in the previous lemma. Then the number of G-nonequivalent q-colourings of X is*

$$\frac{1}{|G|} \sum_{g \in G} q^{c(g)}.$$

Example 3.6.7 Now we look at the problem of counting non-equivalent necklaces that was mentioned at the beginning if this chapter. We consider necklaces with n beads and q colours. First we identify such a necklace in the obvious way with a q-colouring of the graph C_n. Secondly we observe that two necklaces are equivalent if and only if the two corresponding q-colourings of C_n are G-equivalent where $G = D_n$ is the dihedral group of order $2n$. Indeed, two necklaces are equivalent if one can be obtained from the other by a rotation or a reflection. In D_n the $\sigma_{[i]}$ are rotations, whereas the $\rho_{[i]}$ are relfections. So we can use Corollary 3.6.6 to compute the number of non-equivalent necklaces. Here we execute this for $n = 5$.

In order to ease notation a bit we write i instead of $[i]$. So G consists of the elements σ_i and ρ_i for $0 \leq i \leq 4$, and $X = \{0, 1, 2, 3, 4\}$. Now for each element of G we have to compute the number of cycles appearing in its complete factorization. For example we have $\rho_2(k) = -k + 2 \bmod 5$. (Since we do not write the brackets any more, we have to stress that the sum is modulo 5.) So $\rho_2(0) = 2$, $\rho_2(1) = 1$, $\rho_2(2) = 0$, $\rho_2(3) = 4$, $\rho_2(4) = 3$. Hence the complete factorization of $f(\rho_2)$ is $(0, 2)(1)(3, 4)$. In Table 3.1 we list the complete factorization of each element of G.

element	factorization	number of cycles
σ_0	$(0)(1)(2)(3)(4)$	5
σ_1	$(0,1,2,3,4)$	1
σ_2	$(0,2,4,1,3)$	1
σ_3	$(0,3,1,4,2)$	1
σ_4	$(0,4,3,2,1)$	1
ρ_0	$(0)(1,4)(2,3)$	3
ρ_1	$(1,0)(2,4)(3)$	3
ρ_2	$(0,2)(1)(3,4)$	3
ρ_3	$(0,3)(1,2)(4)$	3
ρ_4	$(0,4)(1,3)(2)$	3

Table 3.1: The factorizations of the elements of D_5.

We see that we have one element with five cycles, four elements with one cycle, and five elements with three cycles. So the number of non-equivalent necklaces with five beads and q colours is

$$\frac{1}{10}(q^5 + 5q^3 + 4q).$$

Example 3.6.8 We can use the same ideas to count other coloured objects. In this example we look at flags. A *flag* of length n and q colours consists of n stripes, arranged next to each other horizontally, an each stripe having one of q colours. For example, if we have $n = 7$ and $q = 3$ and the colours are denoted A,B,C, then we can have

B	C	A	A	B	A	C

Two flags are equivalent if either they are the same (obviously), or one can be obtained from the other by reversing it. So the above flag is equivalent to

C	A	B	A	A	C	B

Now the question is to count the non-equivalent flags of length n and with q colours.

First we translate the notion of equivalence into the language of groups. Let $X = \{s_1, \ldots, s_n\}$, where each s_i is just a symbol, corresponding to the i-th stripe. Let $\sigma_0, \sigma_1 \in S_X$ be defined by $\sigma_0(s_i) = s_i$ and $\sigma_1(s_i) = s_{n+1-i}$ for $1 \leq i \leq n$. That is, σ_0 leaves the flag inviariant, wereas σ_1 reverses it. Let $G = \{\sigma_0, \sigma_1\}$ then G is a subgroup of S_X because $\sigma_1\sigma_1 = \sigma_0$. Let C be a set of q colours. Then a flag of length n corresponds to a q-colouring $\gamma : X \to C$.

Now we look at the complete factorizations of the elements of G as products of disjoint cycles. First of all, the factorization of σ_0 consists of n cycles of length 1. Secondly the factorization of σ_1 is

$$(s_1, s_n)(s_2, s_{n-1}) \cdots (s_{\frac{n}{2}}, s_{\frac{n}{2}+1})$$

if n is even and

$$(s_1, s_n)(s_2, s_{n-1}) \cdots (s_{\frac{n-1}{2}}, s_{\frac{n-1}{2}+2})(s_{\frac{n+1}{2}})$$

if n is odd. So it has $\frac{n}{2}$ cycles if n is even, and $\frac{n+1}{2}$ cycles if n is odd. So by Corollary 3.6.6 the number of non-equivalent flags of length n having q colours is $\frac{1}{2}(q^n + q^{\frac{n}{2}})$ is n is even, and $\frac{1}{2}(q^n + q^{\frac{n+1}{2}})$ if n is odd.

Example 3.6.9 Let $p \in \mathbb{Z}$ be a prime, and $X = \mathbb{Z}/p\mathbb{Z}$. This time we do not act on X with the group D_p but with its subgroup $H = \{\sigma_{[0]}, \ldots, \sigma_{[p-1]}\}$. Let C be a set of a colours, where $a \geq 1$, and consider colourings $\gamma : X \to C$. We look at the cycles in the complete factorization of the image of $\sigma_{[i]}$ in S_X. First of all, $\sigma_{[0]}$, being the identity, has p cycles of length 1. Let $i \geq 1$ and consider the cycle containing $[0]$ of $\sigma_{[i]}$. This cycle starts $([0], [i], [2i], [3i], \ldots)$. Its last element is $[ki]$ and then $[(k+1)i] = [0]$. However, that means that $p|(k+1)i$ and as $1 \leq i \leq p-1$, this implies that $p|(k+1)$. But also, $[pi] = [0]$. We see that $k+1 = p$ and the cycle has p elements. The conclusion is that $\sigma_{[i]}$ has exactly one cycle. Therefore by Corollary 3.6.6 the number of non-equivalent colourings is

$$\frac{1}{p}(a^p + (p-1)a).$$

In particular, this number is an integer. Hence $a^p + (p-1)a \equiv 0 \bmod p$. But that is the same as $a^p \equiv a \bmod p$ (which is Fermat's little theorem, Theorem 2.5.22).

Example 3.6.10 Many other things can be counted this way. Ed Russell and Frazer Jarvis have applied Burnside's lemma to the problem of counting non-equivalent Sudoku's:

`http://www.afjarvis.staff.shef.ac.uk/sudoku/russell_jarvis_spec2.pdf.`

3.7 Cyclic groups and orders of group elements

Definition 3.7.1 *Let G be a group and $g \in G$. If there is an $n \in \mathbb{Z}_{>0}$ with $g^n = 1$ then we say that g has finite* order *and the order of g is the minimal positive n with $g^n = 1$. We write $|g| = n$.*

Lemma 3.7.2 *Let G be a group, and $g \in G$ of finite order n. Set $H = \{1, g, g^2, \ldots, g^{n-1}\}$ then H is a subgroup of G and $|H| = n$.*

Proof. It is immediate that H contains 1. Let $g_1, g_2 \in H$, $g_1 = g^i$, $g_2 = g^j$ with $0 \leq i, j < n$. Let $q, r \in \mathbb{Z}$ be such that $i + j = qn + r$ and $0 \leq r < n$. Then $g_1 g_2 = g^{i+j} = (g^n)^q g^r = g^r$ (see Section 1.3 for the justification of these steps) which lies in H. Furthermore, suppose that $i > 0$ and set $k = -i + n$. Then $g_1 g^k = g^i g^{n-i} = g^n = 1$. It follows that H also contains the inverse of g_1, and therefore H is a subgroup of G.

If $g^i = g^j$ and $0 \leq i < j \leq n-1$, then $g^{j-i} = 1$ and $j - i < n$ contrary to the definition of n. It follows that $|H| = n$. □

Definition 3.7.3 *The group H of the previous lemma is called the* cyclic group *generated by g. A group G is said to be* cyclic *if there is a $g \in G$ such that G is equal to the cyclic group generated by g.*

Proposition 3.7.4 *Let G be a finite group. Then every element of G has finite order, and the order of each element divides $|G|$.*

Proof. Let $g \in G$. Consider $g, g^2, g^3, \ldots$. As G is finite there are i, j with $1 \leq i < j$ and $g^i = g^j$, so that $g^{j-i} = 1$. It follows that g has finite order.

Write $n = |g|$. By Lemma 3.7.2 we see that G has a subgroup of order n. So by Lagrange's theorem (Corollary 3.2.7) n divides $|G|$. □

Corollary 3.7.5 *Let G be a finite group, $|G| = n$. Then $g^n = 1$ for all $g \in G$.*

Lemma 3.7.6 *Let G be a group and let $g \in G$. Suppose that $g^k = 1$ for a certain $k > 0$. Then $|g|$ divides k.*

Proof. Write $n = |g|$. There are $q, r \in \mathbb{Z}$ with $k = qn + r$ and $0 \leq r < n$. Then $1 = g^k = (g^n)^q g^r = g^r$. So by definition of n it follows that $r = 0$ and $n|k$. □

Proposition 3.7.7 *Let G be a group. Let $g, h \in G$ be commuting elements with finite orders, and assume that* $\gcd(|g|, |h|) = 1$. *Then gh has finite order and $|gh| = |g||h|$.*

Proof. Write $m = |g|$, $n = |h|$.

Using the fact that g, h commute we see that

$$(gh)^{mn} = (g^m)^n (h^n)^m = 1 \cdot 1 = 1.$$

So gh has finite order and writing $k = |gh|$, by Lemma 3.7.6 we have that $k|mn$.

Furthermore,

$$1 = ((gh)^k)^m = (g^m)^k h^{km} = h^{km}.$$

So from Lemma 3.7.6 it follows that n divides km. But as $\gcd(m, n) = 1$ this implies that $n|k$. In an analogous manner we see that $m|k$. Because $\gcd(m, n) = 1$ that means that mn divides k. It follows that $k = mn$. □

Example 3.7.8 Consider the additive group $\mathbb{Z}/10\mathbb{Z}$. Then [2] has order 5 ([2]+[2]+[2]+[2]+[2] = [0]) and [5] has order 2 ([5] + [5] = [0]). So from the proposition it follows that [7] has order 10 (which can of course also be verified directly).

In order to see that we really need the orders to be coprime, consider the additive group $\mathbb{Z}/8\mathbb{Z}$. Here [2] has order 4 and [4] has order 2. But [6] has order 4 and not 8.

Proposition 3.7.9 *Let F be a field and consider the multiplicative group $F^* = F \setminus \{0\}$. Let G be a* finite *subgroup of F^*. Then G is cyclic.*

Proof. Write $n = |G|$ and write $n = p_1^{e_1} \cdots p_s^{e_s}$ where the p_i are distinct primes and $e_i > 0$ for all i. The polynomial

$$x^{\frac{n}{p_i}} - 1$$

has at most $\frac{n}{p_i}$ roots in F. Hence there exist $a_i \in G$ with $a_i^{\frac{n}{p_i}} \neq 1$. Define

$$b_i = a_i^{\frac{n}{p_i^{e_i}}}.$$

Then with Corollary 3.7.5 we have

$$b_i^{p_i^{e_i}} = a_i^n = 1 \text{ but } b_i^{p_i^{e_i - 1}} = a_i^{\frac{n}{p_i}} \neq 1.$$

Using Lemma 3.7.6 we now infer that the order of b_i divides $p_i^{e_i}$, but does not divide $p_i^{e_i - 1}$. Hence the order of b_i is $p_i^{e_i}$. So by Proposition 3.7.7 we see that the order of $b = b_1 \cdots b_s$ is n and G is equal to the cyclic group generated by b. □

Chapter 4

Fields

In this chapter we study fields, with a special emphasis on the finite fields. The latter are of paramount importance in coding theory, which is an area at the basis of modern digital communication. In the second half of the chapter we will give an introduction to this theory.

4.1 The characteristic of a field

In Section 2.1 a *field* has been defined as a commutative ring with unity 1, where $1 \neq 0$, such that every nonzero element has a multiplicative inverse. Using the language of group theory of Chapter 3, it is also possible to define a field as a set F with two operations $+$ and $\cdot$, such that

1. with $+$, F is an abelian group with neutral element 0,
2. set $F^* = F \setminus \{0\}$; with $\cdot$, F^* is an abelian group with neutral element 1,
3. $a \cdot (b + c) = a \cdot b + a \cdot c$ for all $a, b, c \in F$.

In Chapter 2 we have already encountered some examples of fields: $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$, $\mathbb{Q}(\alpha)$ (Remark 2.4.3), $\mathbb{Z}/p\mathbb{Z}$ (where p is a prime, Theorem 2.5.9).

Since fields are a special class of rings we can talk of ring homomorphisms between fields, or between a field and a ring (see Section 2.5.2). In this context we record the next useful observation.

Remark 4.1.1 Let F be a field and R a ring, and let $f : F \to R$ be a ring homomorphism. Then $f(a) = 0$ for all $a \in F$, or f is injective. Indeed, suppose that f is not injective. Then there are $a, b \in F$ with $a \neq b$ and $f(a) = f(b)$. Set $c = a - b$, then this means that $c \neq 0$ and $f(c) = 0$. But then $f(1) = f(c^{-1}c) = f(c^{-1})f(c) = 0$ and this yields $f(u) = f(1 \cdot u) = f(1)f(u) = 0$ for all $u \in F$.

The next lemma allows us to construct a homomorphism between $\mathbb{Z}$ and a field.

Lemma 4.1.2 *Let F be a field with zero 0_F and unity 1_F. For $n \in \mathbb{Z}$, $n > 0$, we define*

$$\bar{n} = \underbrace{1_F + 1_F + \cdots + 1_F}_{n \text{ summands}},$$

and $\bar{0} = 0_F$, $\overline{-n} = -\bar{n}$. Then for $m, n \in \mathbb{Z}$ we have $\overline{m+n} = \bar{m} + \bar{n}$, $\overline{mn} = \bar{m} \cdot \bar{n}$.

Proof. The first assertion is a special case of the rule for exponentiation (1.3.1). We first prove the second assertion for $m \geq 0$ and all $n \in \mathbb{Z}$, using induction on m. Firstly, $\overline{0 \cdot n} = \bar{0} = 0_F$ and $\bar{0} \cdot \bar{n} = 0_F \cdot \bar{n} = 0_F$. Secondly, suppose that $\overline{mn} = \bar{m} \cdot \bar{n}$ for a certain $m \geq 0$ and all $n \in \mathbb{Z}$. Then $\overline{m+1} \cdot \bar{n} = (\bar{m} + 1_F) \cdot \bar{n} = \bar{m} \cdot \bar{n} + \bar{n} = \overline{mn} + \bar{n} = \overline{mn + n} = \overline{(m+1)n}$.

Now let $m \in \mathbb{Z}$, $m > 0$. Then $\overline{-m} \cdot \bar{n} = (-\bar{m}) \cdot (\bar{n}) = -(\bar{m} \cdot \bar{n}) = -\overline{mn} = \overline{-mn}$. □

The lemma says that the map $\psi : \mathbb{Z} \to F$ defined by $\psi(m) = \bar{m}$, is a ring homomorphism. We can consider its kernel, $\ker \psi = \{m \in \mathbb{Z} \mid \psi(m) = 0_F\}$. As seen in Proposition 2.6.5 this is an ideal of $\mathbb{Z}$. So by Example 2.6.3 $\ker \psi = n\mathbb{Z}$ for a certain $n \in \mathbb{Z}$ and we may assume that $n \geq 0$. We consider two cases:

1. $n = 0$, that is, $\ker \psi = \{0\}$. In this case F is said to be of *characteristic 0*. Furthermore, ψ is injective (Section 2.5.2) and induces an injective homomorphism $\psi : \mathbb{Q} \to F$ by $\psi(\frac{r}{s}) = \bar{r} \cdot (\bar{s})^{-1}$ (we leave the verification that this indeed defines a homomorphism to the reader; note that first it needs to be shown that it is well-defined, that is, if $\frac{r}{s} = \frac{u}{v}$ then $\psi(\frac{r}{s}) = \psi(\frac{u}{v})$; note also that injectivity is immediate by Remark 4.1.1). Let $M = \psi(\mathbb{Q})$, then M is a field because $\mathbb{Q}$ is. So in this case we see that F contains a subfield isomorphic to $\mathbb{Q}$.

2. $n > 0$; in this case n is prime. Indeed, suppose that $n = kl$, where $2 \leq k, l < n$. Then $0_F = \psi(kl) = \psi(k)\psi(l)$, so that at least one of k, l lies in $\ker \psi$. Suppose $k \in \ker \psi$. Then k is divisible by n, but that is clearly impossible and we conclude that n is prime. In this case we say that F is of characteristic n. Using Proposition 2.6.5 we have that $\bar{\psi} : \mathbb{Z}/n\mathbb{Z} \to F$, $\bar{\psi}([k]_n) = \bar{k}$, is an injective ring homomorphism. Let $M = \bar{\psi}(\mathbb{Z}/n\mathbb{Z})$. Then M is a subring of F isomorphic to $\mathbb{Z}/n\mathbb{Z}$. As the latter is a field, so is M. It follows that F contains a subfield isomorphic to $\mathbb{Z}/n\mathbb{Z}$.

Example 4.1.3 The fields $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ are all of characteristic 0. The field $\mathbb{Z}/p\mathbb{Z}$ is of characteristic p.

4.2 Vector spaces

Let F be a field. A *vector space* over F is a set V with two operations $+ : V \times V \to V$ (mapping (v, w) to $v + w$), and $\cdot : F \times V \to V$ (mapping (α, v) to αv) with

1. $(V, +)$ is an abelian group,
2. $\alpha \cdot (v + w) = \alpha \cdot v + \alpha \cdot w$,
3. $(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$,
4. $\alpha \cdot (\beta \cdot v) = (\alpha\beta) \cdot v$,
5. $1 \cdot v = v$

for all $\alpha, \beta \in F$, $v, w \in V$. Usually the neutral element of a vector space V (as it is an abelian group) is denoted 0.

Example 4.2.1 Let $V = F \times F \times \cdots \times F$ (n factors F). The elements of V are written $(\alpha_1, \ldots, \alpha_n)$, where $\alpha_i \in F$. The addition is defined component wise:

$$(\alpha_1, \ldots, \alpha_n) + (\beta_1, \ldots, \beta_n) = (\alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n).$$

Secondly,

$$\alpha \cdot (\alpha_1, \ldots, \alpha_n) = (\alpha\alpha_1, \ldots, \alpha\alpha_n).$$

It is straightforward to check that this defines a vector space over F. We write $V = F^n$.

Let V be a vector space over the field F.

Let $v_1, \ldots, v_s \in V$ and $\alpha_1, \ldots, \alpha_s \in F$; then $\alpha_1 v_1 + \cdots + \alpha_s v_s$ is called a *linear combination* of the v_i. The set of elements of V that are linear combinations of the v_i is called the *linear span* of $v_1, \ldots, v_s$.

Elements $v_1, \ldots, v_r \in V$ are said to be *linearly independent* if $\alpha_1 v_1 + \cdots + \alpha_r v_r = 0$ is only possible when $\alpha_1 = \ldots = \alpha_r = 0$.

Suppose that there are $v_1, \ldots, v_n \in V$ whose linear span equals V. Let $w_1, \ldots, w_m \in V$ with $m > n$. Then we claim that the w_j are linearly dependent. Indeed, we can write $w_j = \alpha_{j1} v_1 + \cdots + \alpha_{jn} v_n$, where $1 \leq j \leq m$ and $\alpha_{ji} \in F$. Consider the following system of n linear equations in the m unknowns λ_j:

$$\alpha_{1i}\lambda_1 + \cdots + \alpha_{mi}\lambda_m = 0$$

($1 \leq i \leq n$). Because $m > n$ this has a solution with not all λ_j equal to 0. But then $\lambda_1 w_1 + \cdots + \lambda_m w_m = 0$ and we see that the w_j are linearly dependent.

Let n be minimal such that there are $v_1, \ldots, v_n \in V$ whose linear span equals V. This n is called the *dimension of* V. We write $\dim V = n$. Then $v_1, \ldots, v_n$ is linearly independent. Indeed, otherwise there exist $\alpha_i \in F$, not all zero, such that $\alpha_1 v_1 + \cdots + \alpha_n v_n = 0$. We may suppose that $\alpha_n \neq 0$. Then by dividing by $-\alpha_n$ we find β_i such that $v_n = \beta_1 v_1 + \cdots + \beta_{n-1} v_{n-1}$. But then the linear span of $v_1, \ldots, v_{n-1}$ is equal to the linear span of $v_1, \ldots, v_n$. The set $\{v_1, \ldots, v_n\}$ is called a *basis* of V.

Remark 4.2.2 It is not immediate, using the above definition, to find a basis of a given vector space. However, we have the following simple criterion. Let V be a vector space. Let $v_1, \ldots, v_n \in V$ be *linearly independent* and such that V *is their span*, then they form a basis of V. Indeed, suppose that V is the span of m elements with $m < n$, then by what we have seen above it follows that $v_1, \ldots, v_n$ is linearly dependent, contradicting the first hypothesis. Hence the minimal number of elements spanning V is n.

Let $B = \{v_1, \ldots, v_n\}$ be a basis of V. Let $v \in V$. Then $B \cup \{v\}$ is linearly dependent. So there exist α_i, $1 \leq i \leq n$, β in F with $\alpha_1 v_1 + \cdots + \alpha_n v_n + \beta v = 0$ and not all coefficients are zero. Here β cannot be zero, as otherwise B would be linearly dependent. By dividing by $-\beta$ we obtain $\beta_i \in F$ with

$$v = \beta_1 v_1 + \cdots + \beta_n v_n.$$

Furthermore, the linear independence of B immediately implies that these coefficients are unique. This is expressed by saying that every element of V is a unique linear combination of the elements of a basis of V.

Of course, this whole argument depends on the existence of a finite number of vectors whose linear span is V. If this does not exist then V is said to be *infinite dimensional*.

Example 4.2.3 Let F be a field and $V = F^n$ (Example 4.2.1). Then $\dim V = n$ and the vectors

$$(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \cdots, (0, \ldots, 0, 1)$$

form a basis of V.

Let $F[x]$ be the polynomial ring over F in the indeterminate x (Section 2.3.1). Then $F[x]$ is an infinite dimensional vector space over F.

4.3 Field extensions

Definition 4.3.1 *Let E, F be two fields with $F \subset E$. If the operations $+$ and $\cdot$ are the restrictions of the same operations of E, then we say that E is an* extension *of F.*

For example $\mathbb{R}$ is an extension of $\mathbb{Q}$ and $\mathbb{C}$ is an extension of $\mathbb{R}$.

Often when we want to indicate that E is an extension of F we write E/F, and also speak of the extension E/F. It is important to keep in mind that no quotient is intended here; "the extension E/F" is just short for "E is a field extension of F".

Remark 4.3.2 Let E/F be a field extension. Let $0_F, 0_E$ be the zeros of F and E respectively; then $0_F = 0_E$. Indeed, let $\delta \in E$ be the opposite of 0_F (that is, we have $\delta + 0_F = 0_E$). Then $0_E = \delta + 0_F = \delta + (0_F + 0_F) = (\delta + 0_F) + 0_F = 0_E + 0_F = 0_F$. In an analogous manner one shows that the unities of E and F are the same element.

Let E be an extension of F. Then E is a vector space over F because E has an addition and a multiplication by elements of F. It is immediate that the conditions of Section 4.2 are satisfied. By $[E : F]$ we denote the dimension of this vector space. It is called the *degree* of E over F. For example, we have $[\mathbb{R} : \mathbb{Q}] = \infty$ (this is not immediate), $[\mathbb{C} : \mathbb{R}] = 2$.

Proposition 4.3.3 (Degree formula) *Let K/E and E/F be extensions of finite degree. Then also K/F is an extension of finite degree and $[K : F] = [K : E][E : F]$.*

Proof. Let $\alpha_1, \ldots, \alpha_m$ be a basis of K over E and $\beta_1, \ldots, \beta_n$ a basis of E over F. We claim that $\alpha_i\beta_j$ for $1 \leq i \leq m$ and $1 \leq j \leq n$ form a basis of K over F. For this we use Remark 4.2.2.

First we show that these elements are linearly independent. So suppose that there are $a_{ij} \in F$ with

$$0 = \sum_{i,j} a_{ij}\alpha_i\beta_j = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij}\beta_j \right) \alpha_i.$$

The term in brackets lies in E. So because the α_i are linearly independent over E it follows that

$$\sum_{j=1}^{n} a_{ij}\beta_j = 0$$

for $1 \leq i \leq m$. But the β_j are linearly independent over F, whence $a_{ij} = 0$ for all i, j.

Now we show that K is the linear span of the $\alpha_i\beta_j$. Let $a \in K$. Then there are $b_j \in E$ such that $a = \sum_{j=1}^{m} b_j\alpha_j$. Furthermore, there are $c_{ij} \in F$ with $b_j = \sum_{i=1}^{n} c_{ij}\beta_i$. Hence

$$a = \sum_{j=1}^{m} \sum_{i=1}^{n} c_{ij}\alpha_j\beta_i.$$

We conclude that the $\alpha_i\beta_j$ form a basis of K over F. □

Let E/F be a field extension. An $\alpha \in E$ is called *algebraic* over F if there are polynomials $f \in F[x]$, $f \neq 0$, such that $f(\alpha) = 0$. (If there are no such polynomials we say that α is *transcendental* over F.) For example, the following elements of $\mathbb{C}$ are algebraic over $\mathbb{Q}$: i, $\sqrt{3}$, $\frac{-1+\sqrt{-3}}{2}$.

If $\alpha \in E$ is algebraic over F then there is a unique monic polynomial $f \in F[x]$ of minimal degree with $f(\alpha) = 0$. (Indeed, if there were two such polynomials then their difference would be a polynomial of smaller degree vanishing on α.) This polynomial is called the *minimal polynomial* of α over F.

Proposition 4.3.4 *Let E/F be a field extension. Let $\alpha \in E$ be algebraic over F and let $f \in F[x]$ be its minimal polynomial. Then f is irreducible and for $g \in F[x]$ with $g(\alpha) = 0$ we have that f divides g. Conversely, if $h \in F[x]$ is monic and irreducible with $h(\alpha) = 0$ then $h = f$.*

Proof. For the first assertion, let $g \in F[x]$ be such that $g(\alpha) = 0$. Write $g = qf + r$, where $\deg(r) < \deg(f)$ (see Proposition 2.3.8). Then $r(\alpha) = g(\alpha) - q(\alpha)f(\alpha) = 0$. From the minimality of $\deg f$ it follows that $r = 0$ and hence g is divisible by f.

If $f = gh$ with $g, h \in F[x]$ then $0 = f(\alpha) = g(\alpha)h(\alpha)$, implying that $g(\alpha) = 0$ or $h(\alpha) = 0$. We may assume that $g(\alpha) = 0$. Then from the first part of the proof it follows that f divides g. But that is only possible if $g = \gamma f$, $\gamma \in F$. But then $h \in F$. It follows that f is irreducible.

For the second part let h be monic and irreducible with $h(\alpha) = 0$. Then by what we have seen above f divides h. Since h is irreducible and both f and h are monic, it follows that $f = h$. □

A problem that immediately comes to mind is how to compute the minimal polynomial of a given $\alpha \in E$. It is usually not difficult to find a polynomial that vanishes on α: we list the elements $1, \alpha, \alpha^2, \ldots$ until we see a linear dependence among them. This linear dependence then defines the polynomial. For example, let $\alpha = \sqrt{2} + \sqrt{3} \in \mathbb{R}$, and we want to find its minimal polynomial over $\mathbb{Q}$. We find the elements

$$1, \alpha = \sqrt{2} + \sqrt{3}, \alpha^2 = 5 + 2\sqrt{6}, \alpha^3 = 11\sqrt{2} + 9\sqrt{3}, \alpha^4 = 49 + 20\sqrt{6},$$

and we see that $\alpha^4 - 10\alpha^2 + 1 = 0$ giving the polynomial $f = x^4 - 10x^2 + 1$. But how can we show that this really is the minimal polynomial over $\mathbb{Q}$? By the previous proposition it suffices to show that it is irreducible. So let us try to do that. First we investigate whether it has linear factors. If β is a root of f then $\beta^2 = \frac{10 \pm \sqrt{96}}{2}$ so that $\beta \notin \mathbb{Q}$ (if $\beta \in \mathbb{Q}$ then $\beta^2 \in \mathbb{Q}$ which would imply $\sqrt{6} \in \mathbb{Q}$). It follows that f has no factors of degree 1. If f has factors of degree 2 then

$$f = (x^2 + ax + b)(x^2 + cx + d) = x^4 + (c + a)x^3 + (d + b + ac)x^2 + (ad + bc)x + bd.$$

Therefore

$$\begin{aligned} bd &= 1 \\ ad + bc &= 0 \\ d + b + ac &= -10 \\ c + a &= 0. \end{aligned}$$

From the first and fourth equation we get $d = b^{-1}$, $c = -a$. Hence the second equation becomes $a(b^{-1} - b) = 0$. If $a = 0$ it follows from the third equation that $b^2 + 10b + 1 = 0$ and $b \notin \mathbb{Q}$. If $b^{-1} - b = 0$ we get $b = \pm 1$. From $b = 1$ it follows that $a^2 = 12$, and $b = -1$ implies $a^2 = 8$; in both cases $a \notin \mathbb{Q}$. So f does not have factors of degree 2, and therefore it is irreducible and the minimal polynomial of α.

4.4 Construction of extensions I

In this section we will see a formal way to construct extensions of a given field F. The method is to take a quotient of $F[x]$ by an ideal. We have seen the construction of the quotient of a ring by an ideal in Section 2.6 and here we review it for polynomial rings. Then we show that starting from an irreducible polynomial in $F[x]$ this gives us a field extension of F. We know a basis for this extension and hence we know its degree.

Let F be a field and $f \in F[x]$ a polynomial. Then we consider the ideal $\langle f \rangle$, that is

$$\langle f \rangle = \{gf \mid g \in F[x]\}.$$

For $h \in F[x]$ we write

$$[h]_f = h + I = \{h + gf \mid g \in F[x]\}$$

which is called the class of h modulo f. (It is the equivalence class of h with respect to the equivalence relation $\sim_f$, where $h_1 \sim_f h_2$ if and only if $f|(h_2 - h_1)$.) If it is clear which f we mean then we also write $[h]$ instead of $[h]_f$. We set

$$F[x]/\langle f \rangle = \{[h]_f \mid h \in F[x]\}.$$

We note that $[h_1]_f = [h_2]_f$ if and only if $h_1 \in [h_2]_f$ if and only if f divides $h_2 - h_1$.

We define an addition and multiplication on $F[x]/\langle f \rangle$ by

$$\begin{aligned} [h_1]_f + [h_2]_f &= [h_1 + h_2]_f \\ [h_1]_f \cdot [h_2]_f &= [h_1 h_2]_f. \end{aligned}$$

As usual we first need to check that these operations are well defined. For that let $h_i, \hat{h}_i \in F[x]$, $i = 1, 2$, be such that $[h_i]_f = [\hat{h}_i]_f$. This means that there are $g_i \in F[x]$ with $h_i = \hat{h}_i + g_i f$. Then $h_1 h_2 = \hat{h}_1 \hat{h}_2 + (\hat{h}_1 g_2 + \hat{h}_2 g_1 + g_1 g_2 f) f$ and therefore $[h_1 h_2]_f = [\hat{h}_1 \hat{h}_2]_f$. Analogously we see that $[h_1 + h_2]_f = [\hat{h}_1 + \hat{h}_2]_f$. We conclude that the operations are well defined, and with them $F[x]/\langle f \rangle$ becomes a commutative ring with unity (see also Proposition 2.6.4).

Proposition 4.4.1 *Let $f \in F[x]$ be irreducible. Then $F[x]/\langle f \rangle$ is a field.*

Proof. Since we already know that $F[x]/\langle f \rangle$ is a commutative ring with unity (the unity is $[1]_f$) the only thing that we need to verify is that every nonzero element has a multiplicative inverse. So let $[g]_f \in F[x]/\langle f \rangle$, $[g]_f \neq [0]_f$. Since $[0]_f = \{hf \mid h \in F[x]\}$ we see that $[g]_f \neq [0]_f$ is equivalent to saying that f does not divide g. Because f is irreducible that implies that $\gcd(f, g) = 1$. Hence there are $u, v \in F[x]$ with $uf + vg = 1$ (Theorem 2.3.11). But then $[v]_f [g]_f = [vg]_f = [1 - uf]_f = [1]_f$ and it follows that $[v]_f$ is the inverse of $[g]_f$. □

Now let $f \in F[x]$ be irreducible and set $K = F[x]/\langle f \rangle$, which is a field by the previous proposition. The map $\psi : F \to K$, $\psi(a) = [a]$, is an injective ring homomorphism. So F is isomorphic to the subfield $\psi(F)$ of K. For this reason we identify F and $\psi(F)$, and write a instead of $[a]$ for $a \in F$. In this sense K is an extension of F.

Proposition 4.4.2 *We use the notation from above. We write $\alpha = [x]_f$. Let $n = \deg(f)$. Then $[K : F] = n$ and $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$ is a basis of K over F. (In other words, every element of K can uniquely be written as $a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}$ with $a_i \in F$.)*

Proof. Let $g \in F[x]$ and let $q, r \in F[x]$ be such that $g = qf + r$ and $\deg(r) < n$ (Proposition 2.3.8). Write $r = a_0 + a_1x + \cdots + a_{n-1}x^{n-1}$. Then

$$[g] = [r + qf] = [r] = [a_0] + [a_1][x] + \cdots + [a_{n-1}][x]^{n-1} = a_0 + a_1\alpha + \cdots + a_{n-1}\alpha^{n-1}.$$

So the elements $1, \alpha, \ldots, \alpha^{n-1}$ span K as vector space over F. Suppose that $b_0+b_1\alpha+\cdots+b_{n-1}\alpha^{n-1} = 0$ for certain $b_i \in F$. This is the same as saying that f divides $b_0 + b_1x + \cdots b_{n-1}x^{n-1}$. Because of the degrees of the two polynomials that is only possible when all b_i are zero. It follows that $1, \alpha, \ldots, \alpha^{n-1}$ are linearly independent, and hence form a basis of K over F. □

Example 4.4.3 Let $f = x^2 - 2 \in \mathbb{Q}[x]$. Then f is irreducible (because $\deg(f) = 2$ this follows from the fact that f has no roots in $\mathbb{Q}$; alternatively it follows from Eisenstein's criterion, Theorem 2.5.14). Let $K = \mathbb{Q}[x]/\langle f \rangle$ and write $\alpha = [x]$ as in Proposition 4.4.2. From the same proposition it follows that $K = \{a + b\alpha \mid a, b \in \mathbb{Q}\}$. Furthermore, $\alpha^2 = [x]^2 = [x^2] = [2] = 2$, from which it follows that α is a zero of f.

Now consider the field $\mathbb{Q}(\sqrt{2})$ (see Remark 2.4.3). We have $\mathbb{Q}(\sqrt{2}) = \{a+b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ and also $\sqrt{2}^2 = 2$. In fact, it is straightforward to see that the map $\varphi : \mathbb{Q}(\sqrt{2}) \to K$, $\varphi(a + b\sqrt{2}) = a + b\alpha$ is an isomorphism.

However, one has to appreciate the difference between the two fields. The field K was constructed artificially, whereas the field $\mathbb{Q}(\sqrt{2})$ arises "in nature" as a subfield of $\mathbb{C}$.

4.5 Construction of extensions II

In this section we look at the fields that arise "in nature" like $\mathbb{Q}(\sqrt{2})$ in Example 4.4.3. We will show that in many cases we have an isomorphism with an artificially constructed field.

Lemma 4.5.1 *Let E be a field. Let A be a subset of E. Then there exists a unique subfield of E with*

- *$A \subset K$,*
- *if $K' \subset E$ is a subfield with $A \subset K'$ then $K \subset K'$.*

Proof. Observe that if $K_1, K_2 \subset E$ are subfields containing A, then $K_1 \cap K_2$ is a subfield containing A. So if we let K be the intersection of all subfields that contain A, then clearly K satisfies the conditions of the lemma. □

We remark that it is by no means clear what the field K of the previous lemma looks like. We know that K exists, but we have no general method to answer questions about it (one such question could be to ask what $[E : K]$ is). Here we will see that for a special class of such field we do know a lot of things.

Definition 4.5.2 *Let E/F be a field extension and $\alpha_1, \ldots, \alpha_n \in E$. Then $F(\alpha_1, \ldots, \alpha_n)$ is defined to be the smallest subfield of E containing F and $\alpha_1, \ldots, \alpha_n$. (Note that this exists by Lemma 4.5.1.)*

Proposition 4.5.3 *Let E/F be a field extension and $\alpha \in E$ algebraic over F with minimal polynomial $f \in F[x]$. Then the map $\varphi : F[x]/\langle f \rangle \to F(\alpha)$, defined by $\varphi([g]) = g(\alpha)$, is a field isomorphism.*

Proof. First we have to show that φ is well defined. So let $g_1, g_2 \in F[x]$ be such that $[g_1] = [g_2]$. Then $g_1 = g_2 + hf$ for a certain $h \in F[x]$. Hence $g_1(\alpha) = g_2(\alpha) + h(\alpha)f(\alpha) = g_2(\alpha)$. We see that φ is well defined.

We have $\varphi([g_1]+[g_2]) = \varphi([g_1+g_2]) = (g_1+g_2)(\alpha) = g_1(\alpha)+g_2(\alpha) = \varphi([g_1])+\varphi([g_2])$. Analogously one sees that $\varphi([g_1][g_2]) = \varphi([g_1])\varphi([g_2])$. So φ is a homomorphism.

If $\varphi([g_1]) = \varphi([g_2])$ then $(g_1 - g_2)(\alpha) = 0$. Hence f divides $g_1 - g_2$ (Proposition 4.3.4) and therefore $[g_1] = [g_2]$. We conclude that φ is injective. Hence the image of φ is a subfield of $F(\alpha)$ and it contains F and α. So this image is equal to $F(\alpha)$. □

Now the following corollary immediately follows by putting together the previous proposition and Proposition 4.4.2.

Corollary 4.5.4 *Let E/F be a field extension. Let $\alpha \in E$ be algebraic over F with minimal polynomial $f \in F[x]$. Write* $\deg(f) = n$. *Then $[F(\alpha) : F] = n$ and $1, \alpha, \alpha^2, \ldots, \alpha^{n-1}$ form a basis of $F(\alpha)$ over F.*

Consider, for example, $\alpha = \sqrt{2}+\sqrt{3} \in \mathbb{C}$. In Section 4.3 we have seen that its minimal polynomial over $\mathbb{Q}$ has degree 4. Hence $[\mathbb{Q}(\alpha) : \mathbb{Q}] = 4$ with basis $1, \alpha, \alpha^2, \alpha^3$.

Another application of Corollary 4.5.4 is the following. If we can compute $[\mathbb{Q}(\alpha) : \mathbb{Q}]$ in an independent manner (that is, without first computing the minimal polynomial of α), then we know the degree of the minimal polynomial of α and there is no longer the necessity to show that a given polynomial that vanishes on α is irreducible.

Again, let $\alpha = \sqrt{2} + \sqrt{3}$. By computing some powers of α we see that $\sqrt{2}, \sqrt{3} \in \mathbb{Q}(\alpha)$. Hence $\mathbb{Q}(\sqrt{2}, \sqrt{3}) \subset \mathbb{Q}(\alpha)$. Conversely, obviously $\alpha \in \mathbb{Q}(\sqrt{2}, \sqrt{3})$, so that $\mathbb{Q}(\alpha) \subset \mathbb{Q}(\sqrt{2}, \sqrt{3})$. We conclude that $\mathbb{Q}(\alpha) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. We have that $x^2 - 2$ is irreducible in $\mathbb{Q}[x]$ (it is of degree 2 and has no roots) so that $[\mathbb{Q}(\sqrt{2}) : \mathbb{Q}] = 2$. It is straightforward to see that there are no $a, b \in \mathbb{Q}$ with $(a + b\sqrt{2})^2 = 3$. Hence $x^2 - 3$ is irreducible in $\mathbb{Q}(\sqrt{2})[x]$. Hence $[\mathbb{Q}(\sqrt{2})(\sqrt{3}) : \mathbb{Q}(\sqrt{2})] = 2$. Furthermore, we note that $\mathbb{Q}(\sqrt{2})(\sqrt{3}) = \mathbb{Q}(\sqrt{2}, \sqrt{3})$. So the degree formula (Theorem 4.3.3) says that

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

The conclusion is that the degree of the minimal polynomial of $\sqrt{2}+\sqrt{3}$ is 4. Because $f = x^4 - 10x^2 + 1$ vanishes on α, this has to be the minimal polynomial of α.

4.6 Splitting fields

In many situations it is necessary to work with the roots of a given polynomial $f \in F[x]$ (where F is a field). However these roots do not necessarily all lie in F. In Section 4.4 we have seen a construction of an extension $K \supset F$ containing at least one root of f: let g be an irreducible factor of f, set $K = F[x]/\langle g \rangle$ and write $\alpha = [x]$. Then $\alpha \in K$ is a root of f. If we need more (or indeed all) roots of f then we can iterate this construction: find $h \in K[x]$ with $f = (x - \alpha)h$, take an irreducible factor of h, and construct the corresponding extension of K, which will then contain at least two roots of f. We can continue this process, and the field that we find at the end is called a *splitting field* of f. Here we define the concept of spliting field in a more intrinsic way (that is, not depending on a particular construction). Then we show that every polynomial has a splitting field (this is essentially the argument just given). Finally we show that two splitting fields of the same polynomial have to be isomorphic (so a polynomial essentially has just one splitting field).

Definition 4.6.1 *Let F be a field and $f \in F[x]$. An extension E/F is called a* splitting field *of f over F if there are $\alpha_1, \ldots, \alpha_n \in E$ with*

1. *$f = \gamma(x - \alpha_1) \cdots (x - \alpha_n)$ for some $\gamma \in F$,*
2. *$E = F(\alpha_1, \ldots, \alpha_n)$.*

So E is a splitting field of f if it contains all roots of f and any intermediate field K (that is, $F \subset K \subsetneq E$) does not contain all roots of f.

Example 4.6.2 We claim that $\mathbb{Q}(\sqrt{2}) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\}$ is a splitting field of the polynomial $x^2 - 2 \in \mathbb{Q}[x]$. Indeed, $\mathbb{Q}(\sqrt{2})$ contains $\pm\sqrt{2}$ and $x^2 - 2 = (x - \sqrt{2})(x + \sqrt{2})$. Secondly, let K be a field with $\mathbb{Q} \subset K \subset \mathbb{Q}(\sqrt{2})$ and $\pm\sqrt{2} \in K$. The elements of $\mathbb{Q}(\sqrt{2})$ are of the form $a + b\sqrt{2}$ where $a, b \in \mathbb{Q}$. Hence K contains all of $\mathbb{Q}(\sqrt{2})$ and thus must be equal to it. We conclude that $\mathbb{Q}(\sqrt{2})$ is a splitting field of $x^2 - 2$.

Example 4.6.3 Let $f = x^4 - 2 \in \mathbb{Q}[x]$. Let $E = \mathbb{Q}(\sqrt[4]{2}, i) \subset \mathbb{C}$. Then E contains four roots of f, which are $\pm\sqrt[4]{2}$, $\pm i\sqrt[4]{2}$. Let $K \subset E$ be a field containing $\mathbb{Q}$ and the roots of f. Then K also contains $\frac{1}{2}(i\sqrt[4]{2})(\sqrt[4]{2})^3$, which is equal to i. Hence $E \subset K$ and thus $E = K$. We conclude that E is a splitting field of f.

Remark 4.6.4 Let $E = F(\alpha_1, \ldots, \alpha_n)$ be a splitting field of $f \in F[x]$. Then every α_i is algebraic over F. By repeatedly applying Corollary 4.5.4 and the degree formula (Theorem 4.3.3) we see that $[E : F]$ is finite.

Lemma 4.6.5 *Let F be a field and $f \in F[x]$. Then there exists a spliting field of f over F.*

Proof. Let $g \in F[x]$ be an irreducible factor of f and $K = F[x]/\langle g \rangle$, which is a field by Proposition 4.4.1. As in Proposition 4.4.2 we write $\alpha = [x]$. Write also $f = a_0 + a_1 x + \cdots + a_n x^n$ then

$$f(\alpha) = a_0 + a_1[x] + \cdots + a_n[x]^n = [a_0 + a_1 x + \cdots + a_n x^n] = [f].$$

But $[f] = [0]$ because $g | f$. So K is an extension of F having a zero of f.

Now we prove the lemma by induction on $\deg f$. The induction hypothesis is that for all fields E and all polynomials $h \in E[x]$ of degree $< \deg(f)$ there exists a splitting field of h over E. By what we have seen above, there exists an extension K/F containing a root α of f. In $K[x]$ we can write $f = (x - \alpha)h$. Since $\deg(h) = \deg(f) - 1$, by the induction hypothesis there exists a splitting field L of h over K. Then there are $\alpha_2, \ldots, \alpha_n \in L$ with $h = \gamma(x - \alpha_2) \cdots (x - \alpha_n)$, for a certain $\gamma \in K$. Note that in fact $\gamma \in F$ as it is the coefficient of the highest degree term of f. So by seting $\alpha_1 = \alpha$ we obtain that $F(\alpha_1, \ldots, \alpha_n) \subset L$ is a splitting field of f over F. □

Now we want to show that two splitting fields of the same polynomial are isomorphic. We do this by induction. However, in the induction step we will see that all of a sudden there appear two fields that are isomorphic. So the induction hypothesis also has to be concerned with two isomorphic fields. For this reason we start with two fields, F, $\overline{F}$, and an isomorphism $a \mapsto \bar{a}$ (so this map is bijective and we have $\overline{a + b} = \bar{a} + \bar{b}$, $\overline{ab} = \bar{a}\bar{b}$ for all $a, b \in F$). This isomorphism extends to a ring isomorphism $F[x] \to \overline{F}[x]$, $h \mapsto \bar{h}$, where

$$\bar{h} = \bar{b}_0 + \bar{b}_1 + \cdots + \bar{b}_m x^m \text{ if } h = b_0 + b_1 x + \cdots + b_m x^m.$$

We start with a lemma that will also have other applications.

Lemma 4.6.6 *Let $g \in F[x]$ be irreducible and let $K \supset F$ be an extension containing a root α of g. Let $\overline{K} \supset \overline{F}$ be an extension containing a root $\bar{\alpha}$ of $\bar{g}$. Then there exists a unique isomorphism $\psi : F(\alpha) \to \overline{F}(\bar{\alpha})$ with $\psi(a) = \bar{a}$ for $a \in F$ and $\psi(\alpha) = \bar{\alpha}$.*

Proof. We start by proving uniqueness. Suppose that $\psi' : F(\alpha) \to \overline{F}(\bar{\alpha})$ is an isomorphism with the same properties as ψ. Write $n = \deg g$. As g is irreducible it is the minimal polynomial of α. Hence every $\beta \in F(\alpha)$ can be written as $\beta = b_0 + b_1\alpha + \cdots + b_{n-1}\alpha^{n-1}$ with $b_i \in F$ (Corollary 4.5.4). But then

$$\begin{aligned}\psi'(\beta) &= \psi'(b_0) + \psi'(b_1)\psi'(\alpha) + \cdots + \psi'(b_{n-1})\psi'(\alpha)^{n-1} \\ &= \bar{b}_0 + \bar{b}_1\bar{\alpha} + \cdots + \bar{b}_{n-1}\bar{\alpha}^{n-1} = \psi(\beta)\end{aligned}$$

so that $\psi' = \psi$.

The fact that $h \mapsto \bar{h}$ is a ring isomorphism $F[x] \to \overline{F}[x]$ implies that $\bar{g}$ is irreducible as well. So $F[x]/\langle g\rangle$ and $\overline{F}[x]/\langle \bar{g}\rangle$ are fields (Proposition 4.4.1). Now we observe that $\theta : F[x]/\langle g\rangle \to \overline{F}[x]/\langle \bar{g}\rangle$, $\theta([h]_g) = [\bar{h}]_{\bar{g}}$, is an isomorphism. (One has to show that it is well defined, bijective, and that it respects addition and multiplication; we leave that as an exercise.) From Proposition 4.5.3 we obtain isomorphisms $\varphi_1 : F[x]/\langle g\rangle \to F(\alpha)$, $\varphi_2 : \overline{F}[x]/\langle \bar{g}\rangle \to \overline{F}(\bar{\alpha})$ with $\varphi_1([x]_g) = \alpha$, $\varphi_2([x]_{\bar{g}}) = \bar{\alpha}$. Hence $\psi = \varphi_2 \circ \theta \circ \varphi_1^{-1}$ is an isomorphism $\psi : F(\alpha) \to \overline{F}(\bar{\alpha})$ with $\psi(a) = \bar{a}$ for $a \in F$. Moreover, $\psi(\alpha) = \varphi_2(\theta(\varphi_1^{-1}(\alpha))) = \varphi_2(\theta([x]_g)) = \varphi_2([x]_{\bar{g}}) = \bar{\alpha}$. This proves the existence part. □

Theorem 4.6.7 *Let $f \in F[x]$ with splitting field $E \supset F$. Let $\overline{E}$ be a splitting field of $\bar{f}$ over $\overline{F}$. Then there exists an isomorphism $\sigma : E \to \overline{E}$ such that $\sigma(a) = \bar{a}$ for all $a \in F$.*

Proof. The proof is by induction on $[E : F]$, which is finite by Remark 4.6.4. We consider quintuples $(K, \overline{K}, L, \overline{L}, \tau)$, where K, $\overline{K}$ are fields, $\tau : K \to \overline{K}$ is an isomorphism, $L \supset K$ is a splitting field of $h = a_0 + a_1x + \cdots + a_mx^m \in K[x]$ and $\overline{L} \supset \overline{K}$ is a splitting field of $\tau(h) = \tau(a_0) + \tau(a_1)x + \cdots + \tau(a_m)x^m \in \overline{K}[x]$. The induction hypothesis is that for each quintuple $(K, \overline{K}, L, \overline{L}, \tau)$ with $[L : K] < [E : F]$ there exists an isomorphism $\varphi : L \to \overline{L}$ with $\varphi(a) = \tau(a)$ for all $a \in K$.

If $E = F$ then F contains all roots of f. In that case $\overline{F}$ contains all roots of $\bar{f}$ (they are $\bar{\alpha}$ for α a root of f) and $\overline{E} = \overline{F}$. In this case we can define σ by $\sigma(a) = \bar{a}$ for all $a \in F$.

If $E \neq F$ then f has an irreducible factor g of degree > 1 (as otherwise all roots of f are in F and $E = F$). Then $\bar{g}$ is an irreducible factor of $\bar{f}$. Since E is a splitting field of f, it contains a root α of g (in fact, it contains all roots of g, but that does not matter to us now). For the same reason, $\overline{E}$ contains a root $\bar{\alpha}$ of $\bar{g}$.

Let $\psi : F(\alpha) \to \overline{F}(\bar{\alpha})$ be the isomorphism from Lemma 4.6.6. Then we have the following picture.

$$\begin{array}{ccc} F & \xrightarrow{\bar{\ }} & \bar{F} \\ \cap & & \cap \\ F(\alpha) & \xrightarrow{\psi} & \overline{F}(\bar{\alpha}) \\ \cap & & \cap \\ E & & \overline{E} \end{array}$$

By the degree formula we have $[E : F] = [E : F(\alpha)][F(\alpha) : F]$. But $[F(\alpha) : F] = \deg(g) > 1$ so that $[E : F] > [E : F(\alpha)]$. Also note that E is a splitting field of f over $F(\alpha)$ and that $\overline{E}$ is a splitting field of $\bar{f}$ over $\overline{F}(\bar{\alpha})$. Therefore we can apply the induction hypothesis and conclude that there is an isomorphism $\sigma : E \to \overline{E}$ with $\sigma(a) = \psi(a)$ for all $a \in F(\alpha)$. In particular, $\sigma(a) = \bar{a}$ for all $a \in F$. □

By applying this theorem with $F = \overline{F}$ and $\bar{a} = a$ for all $a \in F$ we immediately obtain the following corollary.

Corollary 4.6.8 *Let F be a field and $f \in F[x]$. Two splitting fields of f over F are isomorphic.*

4.7 Finite fields

In Section 2.5.2 we already encountered the finite fields $\mathbb{Z}/p\mathbb{Z}$. But there are many more examples of finite fields. Here we will show that a finite field has cardinality p^n for a prime p and an $n \geq 1$, and that for each p and $n \geq 1$ there exists precisely one finite field of cardinality p^n. Secondly, we will give a construction of the finite field of cardinality p^n starting from a primitive polynomial. Throughout we write $\mathbb{F}_p$ for $\mathbb{Z}/p\mathbb{Z}$.

4.7.1 Some preliminary observations

Lemma 4.7.1 *Let p be a prime and A a domain in which $\underbrace{1 + 1 + \cdots + 1}_{p \text{ summands}} = 0$. (For example, A can be a field of characteristic p, or A can be $F[x]$, where F is of characteristic p.) Then for $a, b \in A$ we*

have $(a+b)^p = a^p + b^p$.

Proof. Let $1 \leq i \leq p-1$. Note that $p! = \binom{p}{i} i!(p-i)!$. Because p divides $p!$, it has to divide one of the factors on the right. But p clearly does not divide $i!$ nor $(p-i)!$. Hence p divides $\binom{p}{i}$. It follows that for $c \in A$ we have $\binom{p}{i}c = 0$. Hence

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

□

Let F be a field. Let $f = a_0 + a_1 x + \cdots + a_n x^n \in F[x]$. Then we write $f' = a_1 + 2a_2 x + \cdots + n a_n x^{n-1}$, which is called the *derivative* of f. This defines a map $' : F[x] \to F[x]$ with the following properties

$$\begin{aligned} (af)' &= af' \\ (f+g)' &= f' + g' \\ (fg)' &= f'g + fg' \end{aligned}$$

for $a \in F$, $f, g \in F[x]$. The first two equations are clear. For the third we note that if we have $(f_i g)' = f_i' g + f_i g'$, for $i = 1, 2$, then with $f = f_1 + f_2$ we also have $(fg)' = f'g + fg'$. So it suffices to show the last equation for $f = ax^k$. For the same reason it is enough to consider $g = bx^l$. But for those the equation follows readily.

Lemma 4.7.2 *Let $f \in F[x]$ and suppose that f has a root α of multiplicity at least 2 in an extension of F. Then α is also a root of f'.*

Proof. Let E/F be a field extension with $\alpha \in E$. We have $f = (x-\alpha)^2 g$ with $g \in E[x]$. Then $f' = 2(x-\alpha)g + (x-\alpha^2)g'$ and the lemma follows. □

4.7.2 Existence and uniqueness of finite fields

Lemma 4.7.3 *Let F be a finite field. Then there is a prime p and an integer $n > 0$ such that F has p^n elements, and F is of characteristic p.*

Proof. We use the notation $\bar{m}$ from Lemma 4.1.2. Since the set $\{\bar{m} \mid m \in \mathbb{Z}, m > 0\}$ (being a subset of F) is finite, there are $k, l > 0$, $l > k$, with $\bar{k} = \bar{l}$. But that means that $\overline{l-k} = 0$. As seen in Section 4.1 this implies that the charactistic of F is a prime p.

In Section 4.1 it is also observed that F contains a subfield isomorphic to $\mathbb{F}_p$. We identify this subfield and $\mathbb{F}_p$. Hence F is an extension of $\mathbb{F}_p$ and $[F : \mathbb{F}_p] = n$ is finite because F is finite. Let $e_1, \ldots, e_n$ be a basis of F over $\mathbb{F}_p$. Then the elements of F can uniquely be written as $a_1 e_1 + \cdots + a_n e_n$, with $a_i \in \mathbb{F}_p$. We see that there are exactly p^n such elements. □

Theorem 4.7.4 *Let p be a prime and n a positive integer. Then there exists a field of p^n elements. Moreover, two fields of cardinality p^n are isomorphic.*

Proof. Set $q = p^n$ and consider the polynomial $f_q = x^q - x \in \mathbb{F}_p[x]$. Let E be the splitting field of f_q over $\mathbb{F}_p$. Because $f' = qx^{q-1} - 1 = -1$, it follows from Lemma 4.7.2 that f only has roots of multiplicity 1. Let $K = \{\alpha_1, \ldots, \alpha_q\}$ be the set of the roots of f_q in E.

Note that for $\alpha \in E$ we have $\alpha \in K$ if and only if $\alpha^q = \alpha$. This implies that $0, 1 \in K$ and that K is closed under addition and multiplication (for the addition use Lemma 4.7.1). It follows that K is subfield of E. From $1 \in K$ we immediately see that $\mathbb{F}_p \subset K$. Furthermore K contains all roots of f_q and therefore $K = E$. In particular we see that E has cardinality q.

Now let L be a second field of cardinality q. Then L has characteristic p (as otherwise $|L|$ would be a power of a different prime, see Lemma 4.7.3). Hence L has a subfield M isomorphic to $\mathbb{F}_p$. Let $\phi : \mathbb{F}_p \to M$ be an isomorphism. Furthermore, $L^* = \{\alpha \in L \mid \alpha \neq 0\}$ is a multiplicative group with $q-1$ elements. Hence $\alpha^{q-1} = 1$ for all $\alpha \in L^*$ (Corollary 3.7.5). It follows that every element of L is a root of f_q. Because f_q has degree q, it splits into linear factors over L. We conclude that L is a splitting field of f_q over M. Hence by Theorem 4.6.7 we see that there is an isomorphism $\sigma : E \to L$ (such that $\sigma(a) = \phi(a)$ for $a \in \mathbb{F}_p$). □

We denote the field of $q = p^n$ elements by $\mathbb{F}_q$.

Remark 4.7.5 Let p be a prime and $q = p^n$. From the proof of Theorem 4.7.4 it follows that $\mathbb{F}_q$ is the splitting field of $x^q - x$ over $\mathbb{F}_p$, and that $\alpha^q = \alpha$ for all $\alpha \in \mathbb{F}_q$. Furthermore, from the proof of Lemma 4.7.3 it follows that $[\mathbb{F}_q : \mathbb{F}_p] = n$.

4.7.3 Constructing finite fields

For given p and n we want to construct the finite field $\mathbb{F}_q$, $q = p^n$. Furthermore, we want this construction to be explicit, that is, we want a set S of size q whose elements we easily can write down, together with efficient methods to compute the sum and product of two elements of S. The first idea that comes to mind is to construct $\mathbb{F}_q$ as the splitting field of $x^q - x$ (Remark 4.7.5). This, however, does not easily give us a set S as above: we would have to factorize $x^q - x$ over $\mathbb{F}_q$, take an irreducible factor g, construct the extension $\mathbb{F}_p[x]/\langle g \rangle$, and so on. These steps are all rather cumbersome. Instead, we would like to have an irreducible polynomial f of degree n over $\mathbb{F}_p$. Then we can construct the field $K = \mathbb{F}_p[x]/\langle f \rangle$. As seen in Proposition 4.4.2 we have $[K : \mathbb{F}_p] = n$, so that $|K| = p^n$. Moreover, the proposition gives us a basis of K and using that we can easily write down all elements of K. This leads us to consider the problem of finding irreducible polynomials of degree n in $\mathbb{F}_p[x]$. (Note that it is not even immediately obvious that such polynomials exist.) For this we look at a special class of elements of $\mathbb{F}_q$.

Let $\mathbb{F}_q^* = \{\alpha \in \mathbb{F}_q \mid \alpha \neq 0\}$. Then $\mathbb{F}_q^*$ is a finite multiplicative group, so from Proposition 3.7.9 it follows that $\mathbb{F}_q^*$ is cyclic. That means that there is an $\alpha_0 \in \mathbb{F}_q^*$ that generates $\mathbb{F}_q^*$, or in other words, such that $\mathbb{F}_q^* = \{1, \alpha_0, \alpha_0^2, \ldots, \alpha_0^{q-2}\}$ and $\alpha_0^{q-1} = 1$. Such an α_0 is called a *primitive element* of $\mathbb{F}_q$.

Example 4.7.6 Consider the field $\mathbb{F}_7 = \mathbb{Z}/7\mathbb{Z}$, whose elements we write $0, 1, \ldots, 6$. We have $2^2 = 4$, $2^3 = 1$, so 2 is not primitive. Let's try 3: $3^2 = 2$, $3^3 = 6$, $3^4 = 4$, $3^5 = 5$, $3^6 = 1$, so 3 is a primitive element.

Now let $\alpha_0 \in \mathbb{F}_q$ be primitive. Then the field $\mathbb{F}_p(\alpha_0) \subset \mathbb{F}_q$ contains all elements of $\mathbb{F}_q$ and is therefore equal to $\mathbb{F}_q$. So the minimal polynomial of α_0 over $\mathbb{F}_p$ has degree n (Corollary 4.5.4) and being a minimal polynomial it has to be irreducible (Proposition 4.3.4).

Definition 4.7.7 *A polynomial $f \in \mathbb{F}_p[x]$ is called* primitive *if it is the minimal polynomial of a primitive element in $\mathbb{F}_q$, where $q = p^n$, $n = \deg(f)$.*

So, in our quest for an irreducible polynomial of degree n in $\mathbb{F}_p[x]$ we try to find a primitive polynomial. At first this does not seem to be much easier than to find an irreducible polynomial. However, the next proposition gives a very useful criterion. For polynomials $g, h \in F[x]$ we write $g \equiv h \bmod f$ if f divides $g - h$.

Proposition 4.7.8 *Let $f \in \mathbb{F}_p[x]$ be monic of degree n. Then f is primitive if and only if*

$$x^{p^n-1} \equiv 1 \bmod f \text{ and} \tag{4.7.1}$$

$$x^{\frac{p^n-1}{r}} \not\equiv 1 \bmod f \text{ for every prime } r \text{ that divides } p^n - 1. \tag{4.7.2}$$

Proof. Suppose that f is primitive. Then f is irreducible, and hence $K = \mathbb{F}_p[x]/\langle f \rangle$ is a field of p^n elements. Let $\mathbb{F}_q$ be a field of q elements where $q = p^n$. Then $\mathbb{F}_q = \mathbb{F}_p(\alpha_0)$ where α_0 is a primitive

α	$(0,1,0,0)$	α^6	$(0,0,1,1)$	α^{11}	$(0,1,1,1)$
α^2	$(0,0,1,0)$	α^7	$(1,1,0,1)$	α^{12}	$(1,1,1,1)$
α^3	$(0,0,0,1)$	α^8	$(1,0,1,0)$	α^{13}	$(1,0,1,1)$
α^4	$(1,1,0,0)$	α^9	$(0,1,0,1)$	α^{14}	$(1,0,0,1)$
α^5	$(0,1,1,0)$	α^{10}	$(1,1,1,0)$	α^{15}	$(1,0,0,0)$

Table 4.1: Logarithm table for $\mathbb{F}_{16}$.

element of $\mathbb{F}_q$ such that $f(\alpha_0) = 0$. Then there is an isomorphism $\varphi : \mathbb{F}_q \to K$ with $\varphi(a) = a$ for $a \in \mathbb{F}_p$ and $\varphi(\alpha_0) = [x]_f$ (Proposition 4.5.3). This immediately implies (4.7.1) and (4.7.2).

Now we assume that (4.7.1) and (4.7.2) hold, and again consider $K = \mathbb{F}_p[x]/\langle f\rangle$. In this case we just know that K is a commutative ring with unity. We now prove that it is a field.

Write $\alpha = [x]_f$ and $q = p^n$. Let $A = \{1, \alpha, \alpha^2, \ldots, \alpha^{q-2}\}$. Then A is a multiplicative group. Indeed, from (4.7.1) it follows that $\alpha^s\alpha^t = \alpha^{s+t \bmod (q-1)}$. So A is closed under multiplication. Furthermore $\alpha^{q-1} = 1$, and thus the inverse of α^s is α^{q-1-s}. It follows that A is an abelian group.

Let m be the order of α in A. Then m divides $q-1$ by Lemma 3.7.6. If $m < q-1$ then there exists a prime r dividing $q-1$ such that m divides $\frac{q-1}{r}$. But then $\alpha^{\frac{q-1}{r}} = 1$ which is excluded by (4.7.2). Hence A consists of $q-1$ elements, and therefore A consists exactly of all non-zero elements of K. It follows that every non-zero element of K has a multiplicative inverse. We conclude that K is a field.

We have also just seen that α is a primitive element of K. But its minimal polynomial is f. It follows that f is primitive. □

Remark 4.7.9 The conditions (4.7.1), (4.7.2) do not seem to have anything to do with the irreducibility of f. But if f satisfies them then f is necessarily irreducible because then it is the minimal polynomial of a primitive element of a field.

Remark 4.7.10 Let $f \in \mathbb{F}_p[x]$ be primitive of degree n and set $K = \mathbb{F}_p[x]/\langle f\rangle$, $\alpha = [x]$. Then α is a primitive element of K, that is, the non-zero elements of K are exactly $1, \alpha, \alpha^2, \ldots, \alpha^{q-2}$ ($q = p^n$). Because of this property the primitive polynomials are very useful for constructing a finite field. Indeed, by using a primitive polynomial we have two ways of listing the elements of the finite field: as powers of α or as linear combinations of a basis. Giving elements in the first way makes it easy to perform multiplication, whereas in the second form they are easy to add. Since we want to perform both operations efficiently, it is a good idea to make a table containing both representations of each element. We call such a table a *logarithm table* for $\mathbb{F}_q$.

Example 4.7.11 We write $\mathbb{F}_2 = \{0, 1\}$ and our objective is to construct the field $\mathbb{F}_{16}$. For that we have to find a primitive polynomial of degree 4 in $\mathbb{F}_2$. We have no systematic method for that: we just try a few until we hit one. Let's try $f = x^4 + x + 1 \in \mathbb{F}_2[x]$. We check (4.7.1) and (4.7.2) for f. We have $x^4 \equiv x + 1 \bmod f$ so that $x^8 \equiv x^2 + 1 \bmod f$. Continuing: $x^{12} \equiv (x^2 + 1)(x + 1) \bmod f \equiv x^3 + x^2 + x + 1 \bmod f$, $x^{15} \equiv x^6 + x^5 + x^4 + x^3 \bmod f \equiv 1 \bmod f$. The primes dividing 15 are 3 and 5. But $x^5 \not\equiv 1 \bmod f$ and $x^3 \not\equiv 1 \bmod f$. So by Proposition 4.7.8, f is primitive. It follows that $\mathbb{F}_{16} \cong \mathbb{F}_2[x]/\langle f\rangle$.

As usual we write $\alpha = [x]$. Then $\mathbb{F}_{16} = \{0, \alpha, \alpha^2, \ldots, \alpha^{15}\}$. From Proposition 4.4.2 it follows that every element of $\mathbb{F}_{16}$ can uniquely be written as $a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3$ with $a_i \in \mathbb{F}_2$. Table 4.1 is the logarithm table for $\mathbb{F}_{16}$. For each α^i it contains the vector (a_0, a_1, a_2, a_3) such that $\alpha^i = a_0 + a_1\alpha + a_2\alpha^2 + a_3\alpha^3$.

As noted in Remark 4.7.10 this table makes it easy to perform the arithmetic operations in $\mathbb{F}_{16}$. For example $\alpha^3 + \alpha^{13} = (0,0,0,1) + (1,0,1,1) = (1,0,1,0) = \alpha^8$.

4.8 Coding theory

Coding theory has arisen as an attempt to tackle the problem of inteference that occurs when transmitting messages. Because of this interference the received message is not the same as the one that was sent, and the problem is to reconstruct the original message when given only the corrupted one. One example of this occurs when using everyday language. Suppose that we read "raite the frag". In this sentence there are two words that do not belong to the english language: "raite" and "frag". We see that whoever wrote this must have made an error. Also we note that by making only two changes we can obtain something that makes perfect sense: "raise the flag". Of course, if we assume more errors have been made, the original massage could have been anything. In order to get a handle on this, we assume that the number of errors that have been made is as low as possible, something that we also do in everyday life.

This example illustrates two of the basic principles of error correcting codes:

- A message is being represented by something longer than absolutely necessary, so that two "words" always differ in a few positions.
- When reconstructing the original messsage, from the message with errors, we assume that as few errors have been made as necessary to obtain a message which makes sense.

A famous application of error correcting codes has been the development of the compact disc. If such a disc is scratched not too severely, then one does not hear a difference in the music being played. This is due to the fact that the information on the disc has been encoded using error correcting codes.

4.8.1 Hamming distance

We formalize the idea of when two "words" differ in few positions using a metric on the set of all words.

Let A be a set, and $V = A^n$ which consists of all n-tuples $a = (a_1, \ldots, a_n)$ where $a_i \in A$. Let $a, b \in A^n$, then the *Hamming distance* between them is defined as

$$d(a, b) = |\{i \mid a_i \neq b_i\}|.$$

For example let $A = \{0, 1\}$ then $d((0, 1, 0, 1), (1, 1, 0, 0)) = 2$.

Definition 4.8.1 *Let X be a set. A* metric *on X is a function $m : X \times X \to \mathbb{R}$ such that for all $x, y, z \in X$ we have the following*

1. $m(x, y) \geq 0$,
2. $m(x, y) = 0$ *if and only if* $x = y$,
3. $m(x, y) = m(y, x)$,
4. $m(x, y) \leq m(x, z) + m(y, z)$ *(triangle inequality).*

Proposition 4.8.2 *The Hamming distance is a metric on A^n.*

Proof. The only property that needs some proof is the triangle inequality. For $a, b \in A^n$ define $D(a, b) = \{i \mid a_i \neq b_i\}$; then $d(a, b) = |D(a, b)|$. Let $a, b, c \in A^n$. If $a_i \neq b_i$ then we must have $a_i \neq c_i$ or $b_i \neq c_i$ (or both). Hence $D(a, b) \subset D(a, c) \cup D(b, c)$. Hence

$$d(a, b) = |D(a, b)| = |D(a, c) \cup D(b, c)| \leq |D(a, c)| + |D(b, c)| = d(a, c) + d(b, c).$$

□

In our applications we will always have $A = \mathbb{F}_q$, where $q = p^m$, p a prime. It turns out to be a good idea to consider the distance of an element to the origin; for $x \in \mathbb{F}_q^n$ we define its *weight* as

$$w(x) = |\{i \mid x_i \neq 0\}| = d(x, (0, 0, \ldots, 0)).$$

4.8.2 Linear codes

The idea of coding theory is to code pieces of information as words, which are elements of some A^n. More precisely, we start with a set A and an idea of what elementary pieces of information are. (The latter can be, for instance, all strings of 0 and 1 of a given length k.) The encoding is a map from the set of elementary pieces of information to A^n. Ideally, decoding would boil down to applying the inverse of the encoding function. However, because of the presence of errors, the element a of A^n that we want to decode may not lie in the image of the encoding map. Then we try to find an element x of the image of the encoding map with the smallest Hamming distance to a, and apply the inverse of the encoding map to x. This search is called *decoding*.

The set of elements of A^n that lie in the image of the encoding map is called a *code*. How well we are able to deal with errors in received words depends very much on the code that we use. In order to study these codes, and to determine good ones, we need codes with some structure. It has turned out that a very useful structure in this respect is that of a vector space.

In what follows, $\mathbb{F}_q$ will always be a finite field of $q = p^m$ elements, where p is a prime.

Definition 4.8.3 *A* linear code *is a subspace of an* $\mathbb{F}_q^n$.

Let $C \subset \mathbb{F}_q^n$ be a linear code. Then we say that n is the *length* of C, $k = \dim C$ is called the *dimension* of C, and

$$d = \min_{x,y \in C, x \neq y} d(x, y)$$

is called the *minimum distance* of C. We say that C is an $[n, k, d]$-code, and the integers n, k, d are the *parameters* of the code.

Example 4.8.4 Let $C = \{(0, 0, \ldots, 0), (1, 1, \ldots, 1)\} \subset \mathbb{F}_2^n$. Then C is an $[n, 1, n]$-code.

Definition 4.8.5 *Let $C \subset \mathbb{F}_q^n$ be an $[n, k, d]$-code. Then we say that C corrects t errors if $d \geq 2t + 1$.*

The point of this definition is the following. Suppose that a codeword $c \in C$ was sent, and $x \in \mathbb{F}_q^n$ received. Suppose further that t errors have occurred, that is, $d(c, x) = t$. If C corrects t errors, then there is no other element of C at distance $\leq t$ from x. Indeed, suppose that there is a $c' \in C$ with $d(c', x) \leq t$ and $c' \neq c$. Then by the triangle inequality: $d(c, c') \leq d(c, x) + d(x, c') \leq 2t$, which is not possible because the minimum distance of C is at least $2t + 1$. So from x we can determine c as the unique element of C that is closest to x, and decode correctly. How to find this c quickly is quite another matter.

From Example 4.8.4 we see that it is possible to conctruct codes that correct any number of errors. However, the only encoding procedure with the code from that example is to send 0 to $(0, \ldots, 0)$ and 1 to $(1, \ldots, 1)$ (or the other way round). But that makes the sent message n times longer than the original one. One of the main challenges of coding theory is to find codes that limit the factor by which a message gets longer while correcting possibly many errors.

Now we address two problems. The first is how to describe a linear code. A linear code C is a subspace of some $\mathbb{F}_q^n$. So two ways of describing C come to mind: by giving a basis of C, or a set of linear equations whose solution space is exactly C. The second problem is how to determine the minimum distance of a code.

Definition 4.8.6 *Let $C \subset \mathbb{F}_q^n$ be a linear code. Let $x^{(1)}, \ldots, x^{(k)}$ be a basis of C. Then the matrix*

$$G = \begin{pmatrix} x^{(1)} \\ x^{(2)} \\ \vdots \\ x^{(k)} \end{pmatrix}$$

with rows $x^{(1)}, \ldots, x^{(k)}$ is called a generator matrix *of C.*

Example 4.8.7 Consider the code C from Example 4.8.4. Then $G = \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix}$ is a generator matrix of C.

Example 4.8.8 Let

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 1 & 0 \end{pmatrix}.$$

Then G is a generator matrix of a $[6, 3, d]$-code in $\mathbb{F}_2^6$.

Definition 4.8.9 *Let G be the generator matrix of a linear $[n, k, d]$-code. Then G is said to be in* standard form *if*

$$G = \left(\begin{array}{c|c} I_k & P \end{array} \right),$$

where I_k is a $k \times k$-identity matrix and P is a $k \times (n-k)$-matrix

Remark 4.8.10 Not every code has a generator matrix in standard form. For example consider the code in $\mathbb{F}_2^5$ with generator matrix $(0, 1, 1, 1, 1)$.

Two linear codes $C, C' \subset \mathbb{F}_q^n$ are said to be *equivalent* if there is a $\pi \in S_n$ (the symmetric group of degree n, see Section 3.1.1) with $C' = \{(c_{\pi(1)}, \dots, c_{\pi(n)}) \mid (c_1, \dots, c_n) \in C\}$. This means that by permuting the coordinates in a fixed way we send C to C'. It is straightforward to see that every linear code is equivalent to a code having a generator matrix in standard form.

Here we do not let us be bothered by this problem, and we always assume that our codes have a generator matrix in standard form.

The generator matrix can be used for the encoding procedure. Let C be an $[n, k, d]$-code over $\mathbb{F}_q$ with generator matrix G. Then our basic pieces of information are vectors in $\mathbb{F}_q^k$. Such a vector $v \in \mathbb{F}_q^k$ is simply encoded as vG, which is an element of C. If G is in standard form then this is particularly suggestive, because in that case

$$vG = \left(\begin{array}{c|c} v & vP \end{array} \right).$$

We see that the encoding of v consists of v followed by some extra information. For example, if we take C and G from Example 4.8.8 then $(1, 1, 0)G = (1, 1, 0, 1, 1, 0)$.

Now we see how to describe a linear code by a set of linear equations. Later we will see that this is useful for decoding.

Definition 4.8.11 *Let $C \subset \mathbb{F}_q^n$ be an $[n, k, d]$-code. Let H be an $(n-k) \times n$-matrix over $\mathbb{F}_q$. Then H is called a* parity check matrix *for C if $C = \{c \in \mathbb{F}_q^n \mid cH^T = (0, \dots, 0)\}$.*

Lemma 4.8.12 *Let $G = \left(\begin{array}{c|c} I_k & P \end{array} \right)$ be an $k \times n$-matrix over $\mathbb{F}_q$ and set $H = \left(\begin{array}{c|c} -P^T & I_{n-k} \end{array} \right)$, which is an $(n-k) \times n$-matrix over $\mathbb{F}_q$. Then G is a generator matrix of the linear code C if and only if H is a parity check matrix of C.*

Proof. Let $C' = \{c \in \mathbb{F}_q^n \mid cH^T = (0, \dots, 0)\}$. As H has $n-k$ linearly independent rows it follows that $\dim C' = k$. Secondly,

$$GH^T = \left(\begin{array}{c|c} I_k & P \end{array} \right) \begin{pmatrix} -P \\ I_{n-k} \end{pmatrix} = -P + P = 0.$$

Hence the rows of G lie in C', and thus $C \subset C'$. But their dimensions are equal and therefore $C = C'$. □

Example 4.8.13 Let C be the code from Example 4.8.8. Then

$$H = \begin{pmatrix} 0 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$$

is a parity check matrix for C.

Now we come to the second of our problems, that is, to determine the minimum distance of a given linear code. The basic ingredient is the following lemma.

Lemma 4.8.14 *Let $C \subset \mathbb{F}_q^n$ be a linear code. Then its minimum distance d is equal to the minimal weight of a nonzero element of C. In other words,*

$$d = \min_{\substack{c \in C \\ c \neq (0,\ldots,0)}} w(c).$$

Proof. Let $\bar{0} = (0, \ldots, 0)$, so that $w(c) = d(c, \bar{0})$. Let ω denote the minimal weight of a nonzero element of C. As $\bar{0} \in C$ we have that $w(c) = d(c, \bar{0}) \geq d$ for every $c \in C$, $c \neq \bar{0}$. Hence $\omega \geq d$.

On the other hand, let $x, y \in C$ be such that $d(x, y) = d$. Observe that $d(x, y) = d(x - y, \bar{0}) = w(x - y)$. As $x - y \in C$ we have that $w(x - y) \geq \omega$, and we are done. □

If we want to determine the minimum distance of a linear code C by checking the distance between all pairs of words, then we have to check approximately $\frac{N^2}{2}$ pairs, where $N = |C|$. By using the lemma it is enough to compute the weight of each element of C, which boils down to checking N elements.

Example 4.8.15 Let C be the code of Example 4.8.8. Since this is a code over $\mathbb{F}_2$ there are the following nonzero elements: the rows of G, sums of two rows of G, the sum of all rows of G. Every row of G has weight 3. Each sum of two rows of G has weight 4. The sum of all rows of G has weight 3. We see that the minimum weight is 3, and hence the minimum distance of C is 3 as well.

Lemma 4.8.16 *Let C be a linear code with parity check matrix H. Then the minimum distance of C is equal to the minimum number of columns of H that are linearly dependent.*

Proof. Write $H = (h^{(1)}, \ldots, h^{(n)})$, where $h^{(i)} \in \mathbb{F}_q^{n-k}$ is a column vector. Let $c \in C$ then $cH^T = (0, \ldots, 0)$. But also $cH^T = c_1 h^{(1)} + \cdots + c_n h^{(n)}$. So c gives a linear dependence between some of the columns of H and the number of columns involved is exactly the weight of c. The conclusion follows by Lemma 4.8.14. □

Example 4.8.17 Again consider the code C of Example 4.8.8. Its parity check matrix H is given in Example 4.8.13. There are no zero columns in H, so there is no single linearly dependent column. Every column has at least one 0, and no two columns have their zeros in the same places. This implies that no two columns are linearly dependent. Finally, there are sets of three columns that are linearly dependent (for example, the first column along with the last two columns). It follows that the minimum distance of C is 3.

4.8.3 Syndrome decoding

Now we look at the decoding problem. Let C be a linear code, and suppose the code word c is sent over some channel, and the element $x \in \mathbb{F}_q^n$ is received. The problem is to find an element $c' \in C$ of minimal Hamming distance to x, i.e., such that

$$d(c', x) = \min_{c'' \in C} d(c'', x).$$

Ideally this element is c itself. However, if many errors have been made, then it may happen that $c' \neq c$. It can also happen that more than one element of C is at minimum distance to x. Then we have to make a choice, and even if c is among those elements, we cannot guarantee that we choose c. But if the minimum distance of C is $\geq 2t + 1$ and up to t errors have been made, then we recover c.

In the previous section we have seen that the generator matrix of a code can be used for the encoding procedure. Here we show how the parity check matrix can be used for decoding. So let H denote the parity check matrix of C. For a vector $e \in \mathbb{F}_q^n$ we say that eH^T is the *syndrome* of e. Note that $eH^T \in \mathbb{F}_q^{n-k}$. The first thing we do is to assemble a *syndrome table*. That is a table listing for every $s \in \mathbb{F}_q^{n-k}$ a vector $e \in \mathbb{F}_q^n$, *of minimal weight*, such that $eH^T = s$.

Example 4.8.18 Let C be the code of Example 4.8.8. Its parity check matrix H is given in Example 4.8.13. We consider all elements of $\mathbb{F}_2^{n-k} = \mathbb{F}_2^3$. We have to write each element as a sum of a minimal number of columns of H; putting a 1 in the positions of those columns gives us the element e. For example, $(1,0,1)$ is the second column of H, so $(1,0,1)$ corresponds to $(0,1,0,0,0,0)$. For another example, $(1,1,1)$ is not equal to a column of H, but it can be written as the sum of two columns in more than one way. In this case we just pick one, for instance, it is the sum of the first and fourth cloumns, so it corresponds to $(1,0,0,1,0,0)$. (Alternatively, we could choose $(0,1,0,0,1,0)$ or $(0,0,1,0,0,1)$.) We see that we get the following table

syndrome	$e \in \mathbb{F}_2^6$
(0,0,0)	(0,0,0,0,0,0)
(1,0,0)	(0,0,0,1,0,0)
(0,1,0)	(0,0,0,0,1,0)
(0,0,1)	(0,0,0,0,0,1)
(1,1,0)	(0,0,1,0,0,0)
(1,0,1)	(0,1,0,0,0,0)
(0,1,1)	(1,0,0,0,0,0)
(1,1,1)	(1,0,0,1,0,0)

Write $x = c' + e$, where $c' \in C$ and $e \in \mathbb{F}_q^n$ is the error vector. Then $xH^T = c'H^T + eH^T = eH^T$. We see that x and e have the same syndrome. So we compute the syndrome $s = xH^T$, look it up in the syndeome table, and obtain a vector $e \in \mathbb{F}_q^n$. Then we set $c' = x - e$ and c' is the result of our decoding procedure.

Example 4.8.19 Let the set up be the same as in the previous example. Let $v = (1,0,1)$, which is the elementary piece of information that we want to send. This is encoded as $vG = (1,0,1,1,0,1)$. Suppose that the received word is $x = (1,1,1,1,0,1)$. Then $xH^T = (1,0,1)$. From the syndrome table we get $e = (0,1,0,0,0,0)$. So $c = x - e = (1,0,1,1,0,1)$ is indeed the sent element, and we recover $v = (1,0,1)$.

Remark 4.8.20 Observe that the requirement that the elements e of the syndrome table be of minimal weight reflects the general assumption that as few errors have been made as possible, given the received word x.

4.8.4 Hamming codes

In this section we look at an interesting family of linear codes, the so-called Hamming codes.

Definition 4.8.21 *Let $\ell \geq 2$ and set $n = 2^\ell - 1$. Let H be a matrix having all* nonzero *elements of $\mathbb{F}_2^\ell$ as columns. Let $C \subset \mathbb{F}_2^n$ be the linear code with parity check matrix H. Then C is called the binary* Hamming code *of length n.*

Example 4.8.22 Let $\ell = 2$. Then $H = \left(\begin{smallmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{smallmatrix}\right)$. By Lemma 4.8.12 a generator matrix of C is $G = (1,1,1)$. Hence C is a $[3,1,3]$-code.

Example 4.8.23 Let $\ell = 3$. Then

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}$$

and therefore by Lemma 4.8.12 a generator matrix is

$$G = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{pmatrix}.$$

Let d denote the minimum distance of this code. No two columns of H have their zeros in the same places, so $d > 2$. On the other hand, there are sets of three columns that are linearly dependent (for example, columns 2, 5, 6). Hence $d = 3$. It follows that we have a $[7,4,3]$-code.

The argument in this example works in general to show that a Hamming code has minimum distance 3. So a Hamming code is a $[n, n-\ell, 3]$-code, where $n = 2^\ell - 1$. These are examples of so-called perfect codes.

Definition 4.8.24 *Let $C \subset \mathbb{F}_q^n$ be an $[n,k,d]$-linear code. Then C is said to be* perfect *if $d = 2t+1$ and for every $x \in \mathbb{F}_q^n$ there is a* unique *$c \in C$ with $d(x,c) \leq t$.*

For $c \in \mathbb{F}_q^n$ define $B_r(c) = \{x \in \mathbb{F}_q^n \mid d(c,x) \leq r\}$ (this is the sphere with centre c and radius r). If the minimum distance of a code C is $2t+1$ then two spheres $B_t(c)$, $B_t(c')$, for $c, c' \in C$, do not intersect. The definition says that the code is perfect if the union of the disjoint spheres $B_t(c)$ covers the entire space $\mathbb{F}_q^n$. For example the code of Example 4.8.8 is not perfect, as $(1,0,0,1,0,0)$ is not at distance ≤ 1 from a code word. This is reflected in the fact that its syndrome corresponds to more than one error vector.

Proposition 4.8.25 *Let $\ell \geq 2$ and $n = 2^\ell - 1$. The $[n, n-\ell, 3]$-Hamming code is perfect.*

Proof. Let $x \in \mathbb{F}_2^n$, then $B_1(x)$ consists of x and all vectors in $\mathbb{F}_2^n$ at distance 1 from x. A vector at distance 1 from x is obtained from x by changing exactly one coordinate. Since the field is $\mathbb{F}_2$ that can happen only in one way. It follows that $|B_1(x)| = n+1$. Furthermore, $|C| = 2^k = 2^{n-\ell}$. So

$$\sum_{c \in C} |B_1(c)| = (n+1)2^{n-\ell} = 2^\ell \cdot 2^{n-\ell} = 2^n = |\mathbb{F}_2^n|.$$

It follows that the non-overlapping spheres $B_1(c)$ exhaust all of $\mathbb{F}_2^n$, and therefore the code is perfect. □

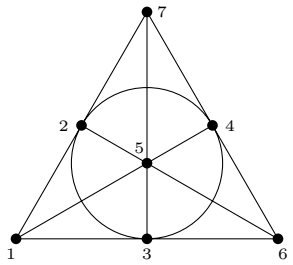
Now we briefly describe the connection between the $[7,4,3]$-Hamming code and an interesting geometrical object. Let $P = \mathbb{F}_2^7 \setminus \{(0,0,0)\}$. We say that three elements $u, v, w \in P$ are on a line if $u + v + w = (0,0,0)$. We set

$$u_1 = (1,1,1),\ u_2 = (1,1,0),\ u_3 = (1,0,1),\ u_4 = (0,1,1)$$
$$u_5 = (1,0,0),\ u_6 = (0,1,0),\ u_7 = (0,0,1).$$

Then the lines are

$$\{\{1,2,7\},\{1,3,6\},\{1,4,5\},\{2,3,4\},\{2,5,6\},\{3,5,7\},\{4,6,7\}$$

(where $\{i,j,k\}$ denotes the line $\{u_i, u_j, u_k\}$). We can display this in a picture:



Here the circle in the middle is also a line. This is a configuration with 7 points and 7 lines, such that any two points are on a unique line, and any two lines meet in a unique point. It is called the *Fano plane*, and it is an example of a finite projective plane.

We can use the Fano plane to give a description of all elements of the $[7,4,3]$-Hamming code.

Proposition 4.8.26 *Let C be the $[7,4,3]$-Hamming code with parity check matrix as in Example 4.8.23. Then the elements of C, ordered according to weight, are characterized as follows:*

weight 0: only $(0,0,0,0,0,0,0)$,

weight 3: let $x \in \mathbb{F}_2^7$ be of weight 3; then $x \in C$ if and only if x has a 1 on the coordinates of a line in the Fano plane,

weight 4: let $x \in \mathbb{F}_2^7$ be of weight 4; then $x \in C$ if and only if x has a 0 on the coordinates of a line in the Fano plane,

weight 7: only $(1,1,1,1,1,1,1)$.

Proof. Since $d = 3$ there are no code words of weights $1, 2$. Let $x \in \mathbb{F}_2^7$ be of weight 3, and i, j, k be such that $x_i = x_j = x_k = 1$. Then $xH^T = (0,0,0)$ is the same as $u_i + u_j + u_k = (0,0,0)$, which is equivalent to $\{u_i, u_j, u_k\}$ being a line of the Fano plane.

Write $\bar{1} = (1,1,1,1,1,1,1)$ and for $x \in \mathbb{F}_2^7$ write $\bar{x} = x + \bar{1}$. Since $\bar{1} \in C$ we have $x \in C$ if and only if $\bar{x} \in C$. Now let x be of weight 4. Then $\bar{x}$ is of weight 3, and we use the previous point.

If $x \in C$ would be of weight 5,6 then $\bar{x}$ would be of weight 1,2. Hence there are no such x in C. □

4.8.5 Turning Turtles

Here we look at a game, called Turning Turtles, and show how linear codes (and especially the $[7,4,3]$-Hamming code) yield a strategy for winning the game.

The game is played by two people, let's call them Alice and Bob (fed up with exchanging encrypted messages, they decide to spend the evening together at Alice's place, and to relax a bit they play this game).

At the start of the game they choose an anteger $n > 0$ and put n coins heads up in a row. After that they take turns making moves. A move has two parts, the first is obligatory and the second is optional (that is, it can be executed or not, as the player wishes). The obligatory part is to turn a coin from heads up to tails up. The optional part is to take a coin *to the left of the one turned previously* and turn it (from heads to tails, or from tails to heads). The player who makes the last move wins (that is, the player after whose move all coins have tails up).

For example, consider the following situation (we number the coins so as to be able to refer to them):

1	2	3	4	5	6	7
T	H	H	T	H	T	T

Some possible moves are:

- Turn coin 3 to T.
- Turn coin 3 to T and then coin 1 to H.
- Turn coin 5 to T.
- Turn coin 5 to T and then coin 2 to T.

We show that this game always terminates. For that let the coins be numbered from 1 to n (as above). To the row of coins we associate a vector $c = (c_1, \ldots, c_n)$ with $c_i = 0, 1$. We stipulate that if coin i is heads up then $c_i = 1$, otherwise $c_i = 0$. We call this a *position vector* of the game. Define $N(c) = \sum_{i=1}^{n} c_i 2^i$. Then $N(c) \geq 0$ and $N(c) = 0$ if and only if the game has terminated. Obviously, if a player moves from c to c' then $N(c') < N(c)$. Hence the game has to terminate.

The next question is how to play this game well. Are there any winning strategies in a given position? In order to analyze this we first look at a rather particular group, called the group of nimbers.

The nimber group

Let $\mathcal{N} = \{m \in \mathbb{Z} \mid m \geq 0\}$ which is the set of *nimbers*. For a subset $S \subset \mathcal{N}$ we define

$$\mathrm{mex}(S) = \min\{m \in \mathcal{N} \mid m \notin S\}$$

(the *m*inimal *ex*cluded value of S). Now for $a, b \in \mathcal{N}$ we set

$$a \oplus b = \mathrm{mex}(\{a' \oplus b \mid 0 \leq a' < a\} \cup \{a \oplus b' \mid 0 \leq b' < b\}).$$

For example, $0 \oplus 0 = \mathrm{mex}(\emptyset) = 0$, $1 \oplus 0 = 0 \oplus 1 = \mathrm{mex}(\{0\}) = 1$, $1 \oplus 1 = \mathrm{mex}(\{1\}) = 0$, $0 \oplus 2 = \mathrm{mex}(\{0, 1\}) = 2$, $1 \oplus 2 = \mathrm{mex}(\{0, 1, 2\}) = 3$.

Lemma 4.8.27 *For $a, b, c \in \mathcal{N}$ we have*

1. $a \oplus 0 = 0 \oplus a = a$,
2. $a \oplus b = a \oplus c$ *if and only if* $b = c$,
3. $a \oplus b = b \oplus a$,
4. $(a \oplus b) \oplus c = a \oplus (b \oplus c)$,
5. $a \oplus a = 0$.

Proof.

1. We use induction on a, the induction hypothesis being that $a' \oplus 0 = a'$ for all $a' < a$. But $a \oplus 0 = \mathrm{mex}\{a' \oplus 0 \mid 0 \leq a' < a\}$. By induction the latter set is equal to $\{0, 1, \ldots, a-1\}$. Hence $a \oplus 0 = a$. In the same way we see that $0 \oplus a = a$.

2. Suppose that $a \oplus b = a \oplus c$ and $b \neq c$. Then we may assume that $b > c$. But then $a \oplus b = \mathrm{mex}(S)$, where S is a set containing $a \oplus c$. Hence we cannot have $a \oplus b = a \oplus c$. This contradiction shows point 2.

3. This is proved by induction on the pair (a, b), where these pairs are ordered lexicographically: $(c, d) < (a, b)$ if and only if $c < a$ or $c = a$ and $d < b$. The induction hypothesis is that $c \oplus d = d \oplus c$ for all (c, d) with $(c, d) < (a, b)$. Now $a \oplus b = \mathrm{mex}(S)$, where $S = \{a' \oplus b \mid 0 \leq a' < a\} \cup \{a \oplus b' \mid 0 \leq b' < b\}$. By induction $S = \{b \oplus a' \mid 0 \leq a' < a\} \cup \{b' \oplus a \mid 0 \leq b' < b\}$. But the mex of that set is precisely $b \oplus a$.

4. We first note the following general property of mex:

$$A \subset A' \subset \mathcal{N}, B \subset \mathcal{N}, \mathrm{mex}(A \cup B) \notin A' \Rightarrow \mathrm{mex}(A \cup B) = \mathrm{mex}(A' \cup B). \tag{4.8.1}$$

   The proof is by induction on the triple (a, b, c), where again these triples are ordered lexicographically (and we leave the precise formulation of this order and the induction hypothesis to the reader). We have $(a \oplus b) \oplus c = \mathrm{mex}(S)$, where $S = A \cup B$ with $A = \{u \oplus c \mid 0 \leq u < a \oplus b\}$, $B = \{(a \oplus b) \oplus c' \mid 0 \leq c' < c\}$. Set $A' = \{(a' \oplus b) \oplus c \mid 0 \leq a' < a\} \cup \{(a \oplus b') \oplus c \mid 0 \leq b' < b\}$. By definition $a \oplus b$ is the minimal element of $\mathcal{N}$ not contained in $\{a' \oplus b \mid 0 \leq a' < a\} \cup \{a \oplus b' \mid 0 \leq b' < b\}$. So if $0 \leq u < a \oplus b$ then u is contained in the latter set, and hence $u \oplus c \in A'$. It follows that $A \subset A'$. Furthermore, by 2. we see that $(a \oplus b) \oplus c \notin A'$. So by (4.8.1) we obtain $(a \oplus b) \oplus c = \mathrm{mex}(A' \cup B)$. By induction $A' \cup B$ is equal to

$$\{a' \oplus (b \oplus c) \mid 0 \leq a' < a\} \cup \{a \oplus (b' \oplus c) \mid 0 \leq b' < b\} \cup \{a \oplus (b \oplus c') \mid 0 \leq c' < c\}.$$

   By the same reasoning as before we have that the mex of this set is equal to $a \oplus (b \oplus c)$.

5. This is shown by induction on a. We have $a \oplus a = \mathrm{mex}(S)$ with $S = \{a' \oplus a \mid 0 \leq a' < a\} \cup \{a \oplus a' \mid 0 \leq a' < a\}$. If $0 \leq a' < a$ then by induction $a' \oplus a' = 0$. So by 2. both $a' \oplus a$ and $a \oplus a'$ are nonzero. Hence 0 is not contained in S and therefore $a \oplus a = 0$.

□

Corollary 4.8.28 *With the operation $\oplus$, $\mathcal{N}$ becomes an abelian group with neutral element 0 and such that every element of $\mathcal{N}$ has order 2.*

We call $\mathcal{N}$ the *nimber group.*

For $\delta \in \mathcal{N}$, $\delta \geq 1$, we define $H_\delta = \{0, 1, \ldots, \delta - 1\}$. The next thing we want to prove is that if $\delta = 2^k$, then H_δ is a subgroup of $\mathcal{N}$.

Lemma 4.8.29 *Let $\delta \geq 1$ and suppose that H_δ is a subgroup of $\mathcal{N}$. Then for $a \in H_\delta$ we have $a \oplus \delta = a + \delta$ (where the latter operation is the normal addition of $\mathbb{Z}$).*

Proof. We use induction on a, the induction hypothesis being that $a' \oplus \delta = a' + \delta$ for all a' with $a' < a$.

By definition $a \oplus \delta = \mathrm{mex}(S)$ with $S = \{a' \oplus \delta \mid 0 \leq a' < a\} \cup \{a \oplus \delta' \mid 0 \leq \delta' < \delta\}$. The δ' in the second set runs over H_δ, so because this is a subgroup we have that $\{a \oplus \delta' \mid 0 \leq \delta' < \delta\} = H_\delta = \{0, 1, \ldots, \delta - 1\}$. By induction $\{a' \oplus \delta \mid 0 \leq a' < a\} = \{a' + \delta \mid 0 \leq a' < a\} = \{\delta, \delta + 1, \ldots, \delta + a - 1\}$. Hence $S = \{0, 1, \ldots, \delta + a - 1\}$ and $a \oplus \delta = \mathrm{mex}(S) = a + \delta$. □

Proposition 4.8.30 *Let $\delta \in \mathcal{N}$ be such that H_δ is a subgroup of $\mathcal{N}$. Then also $H_{2\delta}$ is a subgroup of $\mathcal{N}$.*

Proof. Note that if A is any subset of $\mathcal{N}$ then A contains the inverses of its elements. So it is enough to show that $H_{2\delta}$ is closed under $\oplus$. So let $a, b \in H_{2\delta}$ if $a, b \leq \delta - 1$ then $a, b \in H_\delta$ and $a \oplus b \in H_\delta \subset H_{2\delta}$ by hypothesis. If $a \leq \delta - 1$ and $\delta \leq b \leq 2\delta - 1$ then $b = b' + \delta$ for a $b' \in H_\delta$. By Lemma 4.8.29, $a \oplus b = a \oplus (b' + \delta) = a \oplus (b' \oplus \delta) = (a \oplus b') \oplus \delta = (a \oplus b') + \delta$, which lies in $H_{2\delta}$. The other cases are dealt with by similar arguments. □

Since H_1 is a subgroup of $\mathcal{N}$ this immediately implies that H_{2^k} are subgroups of $\mathcal{N}$ for $k \geq 0$. Now consider two powers of 2, 2^k, 2^l with $k < l$. Then $2^k \in H_{2^l}$ and therefore by Lemma 4.8.29, $2^k \oplus 2^l = 2^k + 2^l$. We can use this to quickly compute $a \oplus b$ for given $a, b \in \mathcal{N}$. Indeed, we write a, b as sums of powers of 2, and use the above rule. For example let's compute $7 \oplus 11$. We have $7 = 1 + 2 + 4$ and $11 = 1 + 2 + 8$. Hence

$$7 \oplus 11 = (1+2+4) \oplus (1+2+8) = (1 \oplus 2 \oplus 4) \oplus (1 \oplus 2 \oplus 8) = 1 \oplus 1 \oplus 2 \oplus 2 \oplus 4 \oplus 8 = 4 \oplus 8 = 4 + 8 = 12.$$

We can also see this a bit differently. Let k be such that $a, b < 2^k - 1$ and write $a = \sum_{i=0}^{k-1} a_i 2^i$, $b = \sum_{i=0}^{k-1} b_i 2^i$ where $a_i, b_i = 0, 1$ (that is, we compute the binary expansions of a, b, see also Section 2.5.3). We consider the vectors $v_a = (a_0, \ldots, a_{k-1})$, $v_b = (b_0, \ldots, b_{k-1})$, which we view as elements of $\mathbb{F}_2^k$. In this vector space we compute $v_c = v_a + v_b = (c_0, \ldots, c_{k-1})$. Then $a \oplus b = \sum_{i=0}^{k-1} c_i 2^i$. For example, if we have $a = 7$, $b = 11$ again, then $v_a = (1, 1, 1, 0)$, $v_b = (1, 1, 0, 1)$ and $v_c = (0, 0, 1, 1)$ so that $a \oplus b = 2^2 + 2^3 = 12$.

The Sprague-Grundy function

Now we look at the Turning Turtles game again. We identify a position in the game with its position vector $c = (c_1, \ldots, c_n)$ with $c_i = 0, 1$ for all i. We define a function $\mathcal{G} : \{0, 1\}^n \to \mathcal{N}$ by

$$\mathcal{G}(c) = \mathrm{mex}\{\mathcal{G}(c') \mid c' \text{ is obtainable from } c \text{ in one move}\}$$

which is called the *Sprague-Grundy function.* For example $\mathcal{G}(0, \ldots, 0) = \mathrm{mex}(\emptyset) = 0$. Suppose that c has exactly one coordinate equal to 1: $c_i = 1$, $c_j = 0$ for $j \neq i$. If $i = 1$ then only $(0, \ldots, 0)$ is reachable in one move, so $\mathcal{G}(c) = \mathrm{mex}\{0\} = 1$. If $i = 2$ then $(0, \ldots, 0)$ and $(1, 0, \ldots, 0)$ are reachable in one move, so that $\mathcal{G}(c) = \mathrm{mex}\{0, 1\} = 2$. For general i we have $\mathcal{G}(c) = i$, as is straightforward to show by induction.

The function $\mathcal{G}$ is very useful for studying winning strategies in the Turning Turtles game. Indeed, let $\mathcal{G}(c) = u$ for some $u \in \mathcal{N}$. Then every c' which is reachable from c in one move has $\mathcal{G}(c') < u$. Conversely, for all $v \in \mathcal{N}$ with $0 \leq v < u$ there is such a c' with $\mathcal{G}(c') = v$. This implies that the player *who can move to a position c with $\mathcal{G}(c) = 0$ has a winning strategy.* Indeed, suppose that this player is Alice. Then Alice moves to c, after which Bob moves to c'. But necessarily $\mathcal{G}(c') > 0$, as otherwise $\mathcal{G}(c)$ could not have been equal to 0. But then Alice can move to c'' with $\mathcal{G}(c'') = 0$ again. It continues like this: Alice moves to a position where the Sprague-Grundy function takes the value 0, and after that Bob is forced to move to a position where the function is > 0. The last position is $(0, \ldots, 0)$, where the Sprague-Grundy function takes the value 0, so it has to be Alice who moves to that position, and therefore Alice wins.

Conversely, if Alice moves to a position c with $\mathcal{G}(c) > 0$ then Bob has a winning strategy because he can move to a position c' with $\mathcal{G}(c') = 0$.

We see that in order to play this game well we have to be able to determine the positions c with $\mathcal{G}(c) = 0$. For that we have the following proposition.

Proposition 4.8.31 *$\mathcal{G}(c) = c_1 \cdot 1 \oplus c_2 \cdot 2 \oplus \cdots \oplus c_n \cdot n$, where $c_k \cdot k$ is k if $c_i = 1$ and it is 0 if $c_i = 0$.*

Proof. As before we define $N(c) = \sum_{i=1}^{n} c_i 2^i$. Furthermore we say that n is the *length* of our game. The proof is by induction on the pair $(n, N(c))$, where these pairs are ordered lexicographically.

We consider the game of length n and the game of length $n-1$. We denote their Sprague-Grundy functions by $\mathcal{G}_{n-1}$ and $\mathcal{G}_n$ respectively.

Let $d = (d_1, \ldots, d_n) \in \{0,1\}^n$ be such that $d_n = 0$. Write $\hat{d} = (d_1, \ldots, d_{n-1})$. Then by induction $\mathcal{G}_{n-1}(\hat{d}) = d_1 \cdot 1 \oplus \cdots \oplus d_{n-1} \cdot (n-1)$.

Furtheremore, we claim that $\mathcal{G}_n(d) = \mathcal{G}_{n-1}(\hat{d})$. This is proved by induction on $N(d)$. Note that $N(d) = N(\hat{d})$. Secondly, every position reachable from d in one move is of the form $(\hat{d}', 0)$, where $\hat{d}'$ is reachable from $\hat{d}$ in one move. Using induction we have $\mathcal{G}_n(d) = \mathrm{mex}\{\mathcal{G}_n(\hat{d}', 0)\} = \mathrm{mex}\{\mathcal{G}_{n-1}(\hat{d}')\} = \mathcal{G}_{n-1}(\hat{d})$ (where in the sets $\hat{d}'$ varies over all positions reachable from $\hat{d}$ in one move).

Write $\hat{c} = (c_1, \ldots, c_{n-1})$. Suppose that $c_n = 0$. By the above claim, $\mathcal{G}_n(c) = \mathcal{G}_{n-1}(\hat{c})$. As noted above $\mathcal{G}_{n-1}(\hat{c}) = c_1 \cdot 1 \oplus \cdots \oplus c_{n-1} \cdot (n-1)$. But that is equal to $c_1 \cdot 1 \oplus \cdots \oplus c_n \cdot n$ as $c_n = 0$.

Now suppose that $c_n = 1$. Then from c we can reach two types of positions.

The first type is of the form $(\hat{c}', 1)$, where $\hat{c}' = (c'_1, \ldots, c'_{n-1})$ is reachable from $\hat{c}$ in one move. All such positions have smaller N-value, so by induction $\mathcal{G}_n(\hat{c}', 1) = c'_1 \cdot 1 \oplus \cdots \oplus c'_{n-1} \cdot (n-1) \oplus n = \mathcal{G}_{n-1}(\hat{c}') \oplus n$. Furthermore, by varying $\hat{c}'$, we have that $\mathcal{G}_{n-1}(\hat{c}')$ takes all values between 0 and $\mathcal{G}_{n-1}(\hat{c}) - 1$, it does *not* take the value $\mathcal{G}_{n-1}(\hat{c})$, and possibly also takes some values $\geq \mathcal{G}_{n-1}(\hat{c}) + 1$.

The second type is of the form $(\hat{c}'', 0)$, where $\hat{c}'' = \hat{c}$, or $\hat{c}''$ is obtained from $\hat{c}$ by changing the i-th coordinate (from 0 to 1, or from 1 to 0), where $1 \leq i \leq n-1$. As seen above we have $\mathcal{G}_n(\hat{c}, 0) = \mathcal{G}_{n-1}(\hat{c})$. If $\hat{c}''$ is obtained from $\hat{c}$ by changing the i-th coordinate then $\mathcal{G}_{n-1}(\hat{c}'') = \mathcal{G}_{n-1}(\hat{c}) \oplus i$ (this trivially holds when the change on position i is $0 \to 1$, but as $i \oplus i = 0$, also when it is $1 \to 0$). It follows that $\mathcal{G}_n(\hat{c}'', 0)$ takes the values $\mathcal{G}_{n-1}(\hat{c})$, $\mathcal{G}_{n-1}(\hat{c}) \oplus i$.

Let $S = \{\mathcal{G}_n(c') \mid c' \text{ can be obtained from } c \text{ in one move}\}$. Then $S = A' \cup B$. Here $A' = \{u \oplus n \mid u \in Q\}$, where Q consists of $0, \ldots, \mathcal{G}_{n-1}(\hat{c}) - 1$ and possibly some values $\mathcal{G}_{n-1}(\hat{c}) + \epsilon$ for some $\epsilon \geq 1$. Secondly, $B = \{\mathcal{G}_{n-1}(\hat{c}) \oplus i \mid 0 \leq i < n\}$. Let $A = \{u \oplus n \mid 0 \leq u < \mathcal{G}_{n-1}(\hat{c})\}$. Then by definition, $\mathcal{G}_{n-1}(\hat{c}) \oplus n = \mathrm{mex}(A \cup B)$. Also $\mathcal{G}_{n-1}(\hat{c}) \oplus n \notin A'$ as that would require $u = \mathcal{G}_{n-1}(\hat{c}) \in Q$ (see Lemma 4.8.27.2). Hence, using (4.8.1), we obtain $\mathcal{G}_n(c) = \mathrm{mex}(S) = \mathrm{mex}(A' \cup B) = \mathrm{mex}(A \cup B) = \mathcal{G}_{n-1}(\hat{c}) \oplus n$. This proves the proposition. □

With the methods that we have seen for computing $a \oplus b$, this makes it straightforward to compute the value of $\mathcal{G}$. However, using the method for computing $a \oplus b$ by letting a, b correspond to vectors in $\mathbb{F}_2^k$, we can also do this a bit differently. Let k be such that $2^k - 1 \geq n$. Write $1, \ldots, n$ as vectors in $\mathbb{F}_2^k$, and let H be the matrix with those vectors as columns. Let $c = (c_1, \ldots, c_n)$ be a position vector of the Turning Turtles game, and view c as a vector in $\mathbb{F}_2^n$. Then by Proposition 4.8.31 we have that $\mathcal{G}(c) = 0$ if and only if

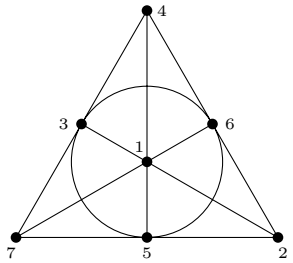
$$cH^T = (0, 0, \ldots, 0),$$

that is, if and only if c lies in the linear code with parity check matrix H. Let C be this linear code. Then the winning strategy of the Turning Turtles game is to *move to an element of* C. In other words, if Alice finds herself in a position that does not correspond to a code word, then she can make a move after which the position of the game does correspond to a code word. On the other hand, if her position corresponds to a code word, then necessarily she has to move to a position outside C, meaning that Bob has a winning strategy.

This is particularly interesting if $n = 7$, because in that case

$$H = \begin{pmatrix} 1&0&1&0&1&0&1 \\ 0&1&1&0&0&1&1 \\ 0&0&0&1&1&1&1 \end{pmatrix}.$$

But this is the parity check matrix of the $[7, 4, 3]$-Hamming code! With respect to Example 4.8.23 the columns are permuted, but that just leads to an equivalent code. As seen above, the best way to play the game is to move to an element of the code. Furthermore, we can use Proposition 4.8.26 and the Fano plane to quickly decide which vectors are code words. For this the numbering of the points of the Fano plane has to be changed a bit, as the columns of H have been permuted. The numbering that results is the following



For example, suppose that Alice finds herself in the following position 1110110 (for brevity we write vectors in $\mathbb{F}_2^7$ this way). This is not an element of the Hamming code because it has weight 5. Alice can try to reach an element of the code of weight 4. For that she has to turn a single 1 into a 0, in such a way that the three resulting 0's lie in a line of the Fano plane. This is achieved by turning the 1 on position 3. So Bob is served with 1100110. Now Bob changes the 1 on position 6 and the 0 on position 3, and reaches 1110100. But then Alice changes the 1 on position 5 and again reaches an element of the code, 1110000. Now it is easily seen that whatever Bob does, on the next move Alice wins.

Note also that, if Alice starts the game, she is in a losing position, as she starts from a code word (1111111). Alternatively, if they play the game with $n = 8$, then at the start Alice is in a winning position because in that case the starting position (11111111) is not a code word. In fact, Alice can immediately move to 11111110 and present Bob with what is essentially the starting position of the game with $n = 7$.

4.8.6 Reed-Solomon codes

Reed-Solomon codes form a class of linear codes over $\mathbb{F}_q$ with some very good properties. For this reason they find a lot of applications. Some examples of their practical use are:

- The *compact disc* (CD) uses a rather complicated scheme for error correction involving several Reed-Solomon codes. This scheme is called CIRC (cross interleaved Reed-Solomon codes).
- The *digital versatile disc* (DVD) uses a different scheme (Reed-Solomon product code, or RSPC), also involving several Reed-Solomon codes.
- The *Voyager* deep space missions to Jupiter and Saturn used Reed-Solomon codes to transmit the data obtained by them.

The starting point of the theory of the Reed-Solomon codes is the following lemma.

Lemma 4.8.32 (Singleton bound) *Let $C \subset \mathbb{F}_q^n$ be a linear $[n,k,d]$-code. Then $d \leq n-k+1$.*

Proof. Let H be a parity check matrix of C. By Lemma 4.8.16 we have that d is the minimum number of columns that are linearly dependent. But the columns of H lie in $\mathbb{F}_q^{n-k}$. Hence any $n-k+1$ columns are linearly dependent. It follows that $d \leq n-k+1$. □

Definition 4.8.33 *Let $C \subset \mathbb{F}_q^n$ be a linear $[n,k,d]$-code. If $d = n-k+1$ then C is said to be* MDS *(maximum distance separable).*

Now we construct the Reed-Solomon codes. Let, as usual, $q = p^m$ and set $n = q$ (that is, our codes will be of length q). Write $\mathbb{F}_q = \{\beta_1, \ldots, \beta_q\}$ and for $2 \leq k \leq q$ define the map $\psi_k : \mathbb{F}_q^k \to \mathbb{F}_q^q$ by

$$\psi_k : (a_0, \ldots, a_{k-1}) \mapsto f = a_0 + a_1 z + \cdots + a_{k-1} z^{k-1} \mapsto (f(\beta_1), \ldots, f(\beta_q)).$$

Then ψ_k is obviously a linear map, and it is injective as $k \leq q$ (indeed, if $\psi_k(a_0, \ldots, a_{k-1}) = \psi_k(b_0, \ldots, b_{k-1})$ then the polynomial $(a_0 - b_0) + (a_1 - b_1)z + \cdots + (a_{k-1} - b_{k-1})z^{k-1}$ is of degree $\leq q-1$, but has q zeros, and therefore is the zero polynomial). Set $C = \psi_k(\mathbb{F}_q^k)$. Then $C \subset \mathbb{F}_q^q$ is a linear $[n,k,d]$-code where $n = q$. It is called a *Reed-Solomon code*.

Lemma 4.8.34 *We have $d = n-k+1$, in other words, C is MDS.*

Proof. Let $x = (x_1, \ldots, x_q) \in C$, with not all x_i equal to 0. Then there are $a_0, \ldots, a_{k-1} \in \mathbb{F}_q$ with $x_i = f(\beta_i)$, where $f = a_0 + a_1 z + \cdots + a_{k-1} z^{k-1}$. Then $x_i = 0$ if and only if $f(\beta_i) = 0$. But f has degree $\leq k-1$ and therefore at most $k-1$ zeros. It follows that for the weight of x we have $w(x) \geq q-k+1 = n-k+1$. So by Lemma 4.8.14 $d \geq n-k+1$. But by the Singleton bound (Lemma 4.8.32) it follows that $d \leq n-k+1$. We conclude that $d = n-k+1$. □

Example 4.8.35 Let $n = q = 4$. We construct the field $\mathbb{F}_4$ using the primitive polynomial x^2+x+1. The logarithm table of $\mathbb{F}_4$ is as follows

α	$(0,1)$
α^2	$(1,1)$
α^3	$(1,0)$

(as usual, α is a primitive element of $\mathbb{F}_4$, see Section 4.7.3). Take $k = 2$. Let $(\alpha, 1) \in \mathbb{F}_4^2$. Then $\psi_2(\alpha, 1) = (f(0), f(\alpha), f(\alpha^2), f(\alpha^3)) = (\alpha, 0, \alpha^3, \alpha^2)$ with $f = \alpha + z$. So we have found an element of C. Continuing like this it is straightforward to list all 16 elements of C.

Let $q = 2^n$ and let C be a Reed-Solomon code over $\mathbb{F}_q$. The logarithm table of $\mathbb{F}_q$ associates a vector in $\mathbb{F}_2^n$ to each element of $\mathbb{F}_q$. So by writing the elements of $\mathbb{F}_q$ as these vectors, instead of in the form α^i, each element of C is a word in the symbols 0,1. This is important for applications, where data are stored as sequences of 0's and 1's. For example, the element $(\alpha, 0, \alpha^3, \alpha^2)$ of Example 4.8.35 is written as 01001011.

There is another advantage to this approach. In applications, errors often occur in "bursts" (that is, a lot of errors close together - think of inteference because of lightning, or a scratch on a CD). Because each coordinate of a codeword is written as a vector of length n, this is unlikely to affect many coordinates. Therefore it may very well be that the code is capable of correcting these errors. This of course goes for any code in $\mathbb{F}_{2^n}$, but because Reed-Solomon codes are MDS they are particularly well-suited for error correction.

Now we turn to the problem of encoding and decoding with Reed-Solomon codes. Encoding is straightforward. Let C be a Reed-Solomon code with paramaters $n = q$, k, $d = n-k+1$. That is, $C = \psi_k(\mathbb{F}_q^k)$. Then a vector $(a_0, \ldots, a_{k-1})$ is simply encoded as $\psi_k(a_0, \ldots, a_{k-1})$.

Decoding is a lot trickier, and a field of active research. Because $n-k = d-1$ is often a rather large number it is not possible to work with a syndrome table. (Take, for example, $n = 256$, $k = 224$,

$d = 33$; then a syndrome table would have to list all elements of $\mathbb{F}_{256}^{32}$.) So we have to resort to other methods. One way of doing it is as follows.

Now let $x \in \mathbb{F}_q^q$ be a received word which we want to decode. Write $x = c + e$ with $c \in C$ and $e \in \mathbb{F}_q^q$ is the error vector. Define the polynomial

$$E = \prod_{e_j \neq 0} (z - \beta_j)$$

which is called the *error locator polynomial*. Let $a_0, \ldots, a_{k-1} \in \mathbb{F}_q$ be such that $c = \psi_k(a_0, \ldots, a_{k-1})$ (of course we do not know these a_i; the problem is to find them). Let $f = a_0 + a_1 z + \cdots + a_{k-1} z^{k-1}$ and

$$Q = E(y - f)$$

which is a polynomial in $\mathbb{F}_q[z, y]$. If $e_j = 0$ then $x_j = c_j = f(\beta_j)$ so that $Q(\beta_j, x_j) = 0$ (here we substitute β_j for z and x_j for y). On the other hand, if $e_j \neq 0$ then $E(\beta_j) = 0$ and again $Q(\beta_j, x_j) = 0$.

Now write $g = Ef$ which is a polynomial in z only. Then $Q = yE - g$. We assume that $d = 2t + 1$ and that there have been up to t errors. Then $\deg E \leq t$ and $\deg g \leq k + t - 1$. We treat the coefficients of E and g as unknowns (because of the degrees we see that thus we have $k + 2t + 1$ unknowns). The condition $Q(\beta_j, x_j) = 0$ translates to

$$x_j E(\beta_j) - g(\beta_j) = 0 \text{ for } 1 \leq j \leq q.$$

This gives us a system of linear equations for the coefficients of E and g. We solve it, and in the solution space we look for a solution (E, g) such that E divides g (we don't go into the problem as to how exactly to do that). Then from E and g we arrive at $f = g/E$ and f corresponds to the decoding of x.

Index

Bibliography

[Cla94] David A. Clark. A quadratic field which is Euclidean but not norm-Euclidean. *Manuscripta Math.*, 83(3-4):327–330, 1994.

[DH76] Whitfield Diffie and Martin E. Hellman. New directions in cryptography. *IEEE Trans. Information Theory*, IT-22(6):644–654, 1976.

[Edw77] Harold M. Edwards. *Fermat's last theorem*, volume 50 of *Graduate Texts in Mathematics*. Springer-Verlag, New York-Berlin, 1977. A genetic introduction to algebraic number theory.

[Gal78] Steven Galovich. Unique factorization rings with zero divisors. *Math. Mag.*, 51(5):276–283, 1978.

[Mig83] Maurice Mignotte. How to share a secret. In *Cryptography (Burg Feuerstein, 1982)*, volume 149 of *Lecture Notes in Comput. Sci.*, pages 371–375. Springer, Berlin, 1983.

[RSA78] R. L. Rivest, A. Shamir, and L. Adleman. A method for obtaining digital signatures and public-key cryptosystems. *Comm. ACM*, 21(2):120–126, 1978.

[Sie13] Aaron N. Siegel. *Combinatorial game theory*, volume 146 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2013.

[Ste07] Ian Stewart. *Why beauty is truth.* Basic Books, New York, 2007. A history of symmetry.