Chapter 1

Permutations

1.1 Permutations

Keywords: function, injective, surjective, bijective, bijection, permutation, cardinality, set of permutations, composition, permutation group.

1.1.1 Prelude: A little bit on functions in general

Before starting out with the theory of permutations, let us first recall some terminology and for functions. If we have given two sets A and B, a function f from A to B assigns to any $a \in A$ an element of B. There is a compact notation to capture all this information, namely $f : A \to B$. In words we say that f is a function f from A to B. The value of a function f in a specific element a in A will be denoted by $f[a]$. In words $f[a]$ is often called the image of a under f or sometimes also the evaluation of f at a. Instead of saying that the evaluation of f at a equals $f[a]$ one can also briefly write $a \mapsto f[a]$. You may see a different notation in other books, since it is also common to write $f(a)$ instead of $f[a]$. However, to avoid "overloading" the use of the usual parentheses (and), we will stick to $f[a]$. All the notation sofar can be compactly given as follows:

$$\begin{aligned} f &: A \to B \\ a &\mapsto f[a] \end{aligned}$$

If two functions $f : A \to B$ and $g : B \to C$ are given, it makes sense to consider the function

$$\begin{aligned} h &: A \to C \\ a &\mapsto g[f[a]] \end{aligned}$$

This function is usually denoted by $g \circ f$ (pronounce g after f) and called the composition of g and f. Hence we have $(g \circ f)[a] = g[f[a]]$ by definition.

Given a function $f : A \to B$, we say that the function f is injective, if any two distinct elements from A are mapped to distinct elements of B. Writing this in terms of logical symbols, this means that:

$$f : A \to B \text{ is injective if } \forall\, a_1, a_2 \in A : \ a_1 \neq a_2 \Rightarrow f[a_1] \neq f[a_2].$$

A result from propositional logic says that for statements P and Q the statement $\neg Q \Rightarrow \neg P$ is logically equivalent to the statement $P \Rightarrow Q$. Therefore we can also say that:

$$f : A \to B \text{ is injective if } \forall\, a_1, a_2 \in A : \ f[a_1] = f[a_2] \Rightarrow a_1 = a_2.$$

This reformulation is sometimes more convenient in practice.

A function $f : A \to B$ is called surjective if any element from B is in the image of f, that is:

$$f : A \to B \text{ is surjective if } \forall\, b \in B\, \exists\, a \in A :\ b = f(a).$$

An example of a function that is injective, but not surjective, is $f : \mathbb{R} \to \mathbb{R}$ given by $f[x] = e^x$. An example of a function that is surjective, but not injective is $g : \mathbb{R} \to [-1, 1]$ given by $g[x] = \sin[x]$.

A function $f : A \to B$ is called bijective if it is both injective and surjective. A bijective function is also called a bijection. An example of a bijection is the function $h : \{0, 1, 2\} \to \{0, 1, 2\}$ given by $h[x] = x^3 + 1 \mod 3$. Note that $h[0] = 1$, $h[1] = 2$ and $h[2] = 9 \mod 3 = 0$, so indeed we can see that h is a bijection from $\{0, 1, 2\}$ to $\{0, 1, 2\}$. Combining the definitions of injective and surjective, we see that function $f : A \to B$ is bijective precisely if for each $b \in B$ there exists a unique $a \in A$ such that $f[a] = b$. In logical notation:

$$f : A \to B \text{ is bijective if } \forall\, b \in B\, \exists!\, a \in A :\ b = f[a].$$

There is a very practical reformulation of this using inverse functions. Let us for completeness first define what the inverse of a function is.

Definition 1 *Let $f : A \to B$ be a function. A function $g : B \to A$ is called the inverse function of f if $f \circ g = \mathrm{id}_B$ (the identity function on B) and $g \circ f = \mathrm{id}_A$ (the identity function on A). The inverse of f will be denoted by f^{-1}.*

As we have seen, a function $f : A \to B$ is bijective precisely if for any $b \in B$ there exists a unique $a \in A$ such that $f[a] = b$. The uniqueness of a implies that we can define a function $g : B \to A$ as $b \mapsto a$. We will show that this is nothing but the inverse function of f. In fact we have the even stronger following result.

Lemma 2 *Suppose that A and B are sets and let $f : A \to B$ be a function. Then f is a bijection if and only if f has an inverse function.*

Proof. Suppose that $f : A \to B$ is a bijection. We have already describe that in that case one can define the function $g : B \to A$ defined by $g[b] \mapsto a$, where a is the unique element of A such that $f[a] = b$. It is not hard to see that $f \circ g = \mathrm{id}_B$ and $g \circ f = \mathrm{id}_A$, which by Definition 1 means that $g = f^{-1}$. Indeed if $b = f[a]$, we have

$$(f \circ g)[b] = f[g[b]] = f[a] = b \text{ and} (g \circ f)[a] = g[f[a]] = g[b] = a.$$

Conversely, if f has an inverse function, then the equation $f[a] = b$ implies that $f^{-1}[f[a]] = f^{-1}[b]$. Since $a = (f^{-1} \circ f)[a] = f^{-1}[f[a]]$, we see that $a = f^{-1}[b]$. Hence for any $b \in B$, there exists a unique element $a \in A$ such that $f[a] = b$ (namely $a = f^{-1}[b]$). Hence f is bijective. ■

This lemma can be very convenient when checking is a function is bijective or not. In this chapter we are particularly interested in the case where the set A is finite, that is to say, where the set A only contains finitely many elements. In this case there is an even easier way to determine whether or not a function is a bijection. For this we need a notation for the number of elements in a set A. The number of elements in A, called the cardinality of A, will be denoted by $|A|$. Hence for a set A containing n elements, we have $|A| = n$. Another notation that is used quite often is $\#A$ instead of $|A|$. We will stick to $|A|$ for the cardinality of a set in these notes. The empty set $\emptyset$ contains no elements and hence we have $|\emptyset| = 0$. With this notation in place, we can formulate the following lemma.

Lemma 3 *Suppose that A and B are finite sets and let $f : A \to B$ be a function. If f is injective, then $|A| \leq |B|$. If f is surjective, then $|A| \geq |B|$.*

Proof. If f is injective, then $f(A) = \{f[a] \mid a \in A\}$ contains exactly $\#A$ elements. Since $f(A)$ is a subset of B, we see that B contains at least as many elements as A.

Similarly, if f is surjective, then $f(A) = B$ and for each element of B there exists at least one element $a \in A$ such that $f[a] = b$. Therefore A contains at least as many elements as B. ■

If f is bijective, this lemma implies that $|A| = |B|$. This has a nice consequence that is only true for finite sets of equal cardinality.

Lemma 4 *Suppose that A and B are finite sets and let $f : A \to B$ be a function. Further suppose that $|A| = |B|$. Then if f is injective, it is in fact bijective. Similarly, if f is surjective, it is bijective.*

Proof. Suppose that f is injective, but not surjective. Since f is injective, we have $|f(A)| = |A| = |B|$, where the last equality follows by the assumption that $|A| = |B|$. On the other hand, since f is not surjective, we have then $f(A) \subsetneq B$ and hence $|f(A)| < |B|$. This gives a contradiction. Apparently if f is injective, it needs to be surjective as well.

Now suppose that f is surjective, but not injective. Since f is surjective, we have $f(A) = B$ and hence $|A| = |f(A)|$, since we assumed that $|A| = |B|$. On the other hand, if f is not injective, we have $|A| > |f(A)|$. Again we arrive at a contradiction. Apparently, if f is surjective, it needs to be injective as well. ■

1.1.2 Definition of permutations

In this chapter we are interested in a particular kind of bijections called permutations. A bijection $f : A \to A$ is called a permutation of the set A. The point is that a permutation is a bijective function from a set A to itself. The function $h : \{0, 1, 2\} \to \{0, 1, 2\}$ given by $h[x] = x^3 + 1 \bmod 3$. defined above is for example a permutation. Since a permutation by definition is a bijection, it is always both injective and surjective. Moreover, it always has an inverse by Lemma 2.

Let us from now on assume that the set A is finite, say $|A| = n$ for some natural number n.

Definition 5 *Let A be a set. The set of all permutations $f : A \to A$ is denoted by S_A. In case $A = \{1, 2, \ldots, n\}$, one writes S_n, rather than $S_{\{1,2,\ldots,n\}}$.*

To write a permutation $f \in S_A$ down explicitly, we need to find a way to explicitly write down $f[a]$ for all $a \in A$. We could do this using a table with two rows: The first row lists the elements a of A, the second row the corresponding values of $f[a]$. This really boils down to describing a permutation using a 2 by n matrix.

$$f = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ f[a_1] & f[a_2] & \cdots & f[a_n] \end{pmatrix}.$$

Let us looks at an example.

Example 6 Suppose $A = \{1, 2, 3\}$. Then the function $f \in S_3$ defined by $f[1] = 2$, $f[2] = 3$ and $f[3] = 1$ is a permutation. In matrix notation, we obtain:

$$f = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}.$$

There are $n!$ many permutations of the set $\{1, 2, \ldots, n\}$ (recall that $n! = 1 \cdot 2 \cdots n$ is the factorial function of n, which outputs the product of all integers between 1 and n). A way to see this is the following: to create a permutation f we can first specify $f[1]$ in n ways. We can namely choose $f[1]$ to be any element of $\{1, 2, \ldots, n\}$ we want. Next we specify $f[2]$. Since f has to be a permutation, we cannot choose $f[2]$ equal to $f[1]$, but any other element of $\{1, 2, \ldots, n\}$ is possible. Therefore we have $n - 1$ possibilities left for $f[2]$. Continuing like this, we see that there are $n-2$ possibilities left for $f[3]$, and so on... All in all we have $n \cdot (n-1) \cdot (n-2) \cdots 1 = n!$ possibilities for f.

Example 7 This is a continuation of Example 6. The set S_3 contains $3! = 6$ permutations. More precisely, we have:

$$S_3 = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \right\}.$$

If we have two permutations f and g of the same set A, then we can compose these permutations to a new one $f \circ g$ (pronounce "f after g" or "f composed with g"). More precisely, the permutation $f \circ g$ is defined as:

$$\text{For } a \in A \ (f \circ g)[a] := f[g[a]]. \tag{1.1}$$

It is an exercise to show that indeed the function $f \circ g$ is a bijection.

Example 8 This is a continuations of Example 7. If

$$f = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \text{ and } g = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$

then

$$f \circ g = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

Also, we have

$$g \circ f = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}.$$

From this example it is clear that $f \circ g$ is not the same permutation as $g \circ f$ in general (though it may happen for specific permutations). If $f \circ g = g \circ f$ one says that f and g commute.

Since $f \in S_A$ is a bijection, it has an inverse function. This inverse, usually denoted by f^{-1} is also a permutation. Here the matrix notation comes in handy: to find the inverse of a permutation f, we simply read the matrix describing f from bottom row to top row. If we compose a permutation with its inverse, one obtains the permutation fixing each element of A (that is to say $f[a] = a$ for all $a \in A$). This permutation is called the identity permutation and is denoted by id. In other words, we have

$$\text{id} = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ a_1 & a_2 & \cdots & a_n \end{pmatrix}.$$

Then we have $f \circ f^{-1} = \text{id}$ and $f^{-1} \circ f = \text{id}$.

Example 9 This is a continuations of Example 8. The identity element in S_3 is give by

$$\text{id} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}.$$

If

$$f = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} \text{ and } g = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$

then

$$f^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} \text{ and } g^{-1} = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}.$$

Note that $g^{-1} = g$.

A common notation is to write f^2 instead of $f \circ f$, f^3 instead of $f \circ f \circ f$, etc. Also negative exponents are possible by defining $f^{-m} = (f^{-1})^m$. With these notations in place, we collect the central properties for the composition of permutations in the following theorem:

Theorem 10 *Let A be a set and S_A the set of all permutations from A to itself. Further let us denote by $\circ$ the composition of permutations from S_A, then*

- *$\forall f, g, h \in S_n$ we have $f \circ (g \circ h) = (f \circ g) \circ h$ (the composition map is associative),*
- *the identity permutation $\mathrm{id} \in S_A$ satisfies $\mathrm{id} \circ f = f$ and $f \circ \mathrm{id} = f$ for any $f \in S_A$,*
- *for any $f \in S_A$ there exists an inverse permutation $g \in S_A$ such that $f \circ g = \mathrm{id}$ and $g \circ f = \mathrm{id}$ (this inverse is denoted by f^{-1}).*

Proof. We prove the first item of the theorem: For any $a \in A$ we have

$$(f \circ (g \circ h))[a] = f[(g \circ h)[a]] = f[g[h[a]]],$$

while

$$((f \circ g) \circ h)[a] = (f \circ g)[h[a]] = f[g[h[a]]].$$

We conclude that for any $a \in A$ it holds that $(f \circ (g \circ h))[a] = ((f \circ g) \circ h)[a]$. This means that the permutations $f \circ (g \circ h)$ and $(f \circ g) \circ h$ are one and the same. ■

Definition 11 *The pair $(S_A, \circ)$ is called the permutation group on the set A. In case $A = \{1, 2, \ldots, n\}$ one says that $(S_n, \circ)$ is the permutation group on n letters.*

The S in the notation S_A is historic and stands for "symmetric". In fact some people prefer to say "symmetric group on n letters" rather than "permutation group on n letters".

1.2 Cycle notation

Keywords: cycle, disjoint cycles, disjoint cycle decomposition.

The matrix description of a permutation used in the previous section is not a very practical notation for a permutation. A much more useful description of permutation is what is known as the disjoint cycle description. Before we can explain that, we first need to know what a cycle is:

Definition 12 *Let $m \geq 1$ be an integer and $a_1, a_2, \ldots, a_m$ distinct elements of A. The permutation $(a_1 a_2 \ldots a_{m-1} a_m) \in S_n$ is the permutation sending a_1 to a_2, a_2 to a_3, $\ldots$, a_{m-1} to a_m and a_m to a_1. It holds all other elements of A fixed. Such a permutation is called a cycle, or more precisely an m-cycle.*

Two cycles $(a_1 a_2 \dots a_{m-1} a_m)$ and $(b_1 b_2 \dots b_\ell)$ are called disjoint, if the sets $\{a_1, a_2, \dots, a_m\}$ and $\{b_1, b_2, \dots, b_\ell\}$ have no elements in common. Even though we have seen in Example 8 that in general it does not holds that $f \circ g = g \circ f$, this does hold if f and g are disjoint cycles. To show this is is an exercise.

Example 13 The permutation $(1\,2\,3) \in S_5$ is the permutation given in matrix notation by

$$(1\,2\,3) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 3 & 1 & 4 & 5 \end{pmatrix}.$$

The permutation $(1\,2\,3)$ is a 3-cycle. Note that $(1\,2\,3) = (2\,3\,1) = (3\,1\,2)$, so there are more than one way to write the same cycle down.

The 1-cycle $(4) \in S_5$ is in matrix notation given by

$$(4) = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{pmatrix}.$$

Therefore (4) is just the identity element id. In general any 1-cycle is just the identity permutation.

Cycles can be seen as elementary building blocks of permutations. We will namely show in a moment that any permutation can be written as the composition of cycles that are mutually disjoint. Let us first look at an example.

Example 14 We consider the permutation $f \in S_{11}$ defined by

$$f = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ 2 & 6 & 7 & 8 & 9 & 1 & 3 & 11 & 5 & 10 & 4 \end{pmatrix}.$$

To write f as a composition of disjoint cycles, we start by determining the cycle permutation c containing the element 1. We may assume that this cycle starts with 1, that is to say $c = (1 \dots)$. Since 1 is sent to 2 (in other words, since $f[1] = 2$), the element after 1 in c has to be 2. Now we know that $c = (1\,2 \dots)$. Similarly, since $f[2] = 6$, we see that $c = (1\,2\,6 \dots)$. Finally, since $f[6] = 1$, we can conclude that $c = (1\,2\,6)$. We now have dealt with the elements 1, 2 and 6. The next element we consider is 3. Reasoning as before we see that the cycle containing 3 is $(3\,7)$, since $f[3] = 7$ and $f[7] = 3$. Continuing like this we find

$$f = (1\,2\,6) \circ (3\,7) \circ (4\,8\,11) \circ (5\,9) \circ (10).$$

Note that the cycle (10) is just the identity element id. just means that $f[10] = 10$, in other words that 10 is fixed by f. It is customary not to write such 1-cycles down. If an element does not occur in the disjoint cycle decomposition of a permutation f, it is simply fixed by f. Also it is customary not to write the composition symbol between two cycles. With these conventions we obtain

$$f = (1\,2\,6)(3\,7)(4\,8\,11)(5\,9).$$

Usually the order in which one writes the cycles is important, but for disjoint cycles the order does not matter. For example we have

$$(1\,2\,6)(3\,7)(4\,8\,11)(5\,9) = (3\,7)(1\,2\,6)(5\,9)(4\,8\,11).$$

Now we show that with a similar procedure one can always write a permutation as a composition of mutually disjoint cycles. We already mentioned that two cycles are called disjoint if the elements from A occurring in one cycle do not occur in the other cycle. More generally, we say that a collection of cycles is mutually disjoint, if any two of them are disjoint.

Theorem 15 *Let n be a natural number and let A be a set of cardinality n. Any permutation $f \in S_A$ can be written as the composite of mutually disjoint cycles.*

Proof. First note that the identity permutation $\mathrm{id} \in S_A$ is a composition of n disjoint 1-cycles, so the theorem holds for the identity permutation.

Now we prove the theorem by induction on n. First we look at the case $n = 1$. The only permutation in S_1 is the identity permutation id. Therefore the theorem is true for $n = 1$.

Suppose that $n \geq 2$ and that the theorem is true for S_{n-1}. Let $f \in S_n$ be a permutation. We claim that the sequence $n, f[n], f^2[n], f^3[n], \ldots$ contains the element n more than once. Indeed, since the sequence takes values in a finite set with n elements, at least one element has to occur twice, say $f^i[n] = f^j[n]$ for some i and j with $i < j$. Composing with f^{-i}, we see that $n = f^{j-i}[n]$. This shows that the element n occurs more than once in the sequence, as was claimed.

Now let m be the smallest element such that $f^m[n] = n$. Then we can define the m-cycle $c = (n\, f[n] \ldots f^{m-1}[n])$. The permutation $c^{-1} \circ f$ fixes the element n, since $(c^{-1} \circ f)[n] = c^{-1}[f[n]] = n$. In fact by a similar reasoning, we see that $c^{-1} \circ f$ fixes all the elements $n, f[n], f^2[n], \ldots, f^{m-1}[n]$. This means that we can interpret $c^{-1} \circ f$ as a permutation on $n-1$ elements, fixing the elements $n, f[n], f^2[n], \ldots, f^{m-1}[n]$. Using the induction hypothesis, we can write $c^{-1} \circ f$ as a product of disjoint cycles, say

$$c^{-1} \circ f = c_1 \circ \cdots \circ c_\ell,$$

where none of the cycles $c_1, \ldots, c_\ell$ contains any of the elements $n, f[n], \ldots, f^{m-1}[n]$. Therefore

$$f = \mathrm{id} \circ f = (c \circ c^{-1}) \circ f = c \circ (c^{-1} \circ f) = c \circ (c_1 \circ \cdots c_\ell).$$

This concludes the induction step.

By the induction principle, we conclude that the theorem is true. ■

This theorem assures us that any permutation is the composite of mutually disjoint cycles. This is called the disjoint cycle decomposition of a permutation. We have already remarked that it is customary to refrain from writing 1-cycles and that any two disjoint cycles commute. It is not so hard to see that there is essentially only one way to write a permutation as a composite of mutually disjoint cycles. The "essentially" in this statement, just means that the only freedom one has is to change order of the m-cycles in the decomposition.

Given two permutations written as a product of mutually disjoint cycles, we can readily write their composition as a product of disjoint cycles as well. We should remember to read the composition from right to left and determine the disjoint cycles one at a time. Let us consider an example:

Example 16 For example let us consider the composition $f \circ g$ with $f = (1\,2\,6)(3\,7)(4\,8\,11)(5\,9)$ and $g = (1\,7\,10\,2)(3\,9\,5)$. Then

$$(f \circ g)[1] = f[g[1]] = f[7] = 3,$$

$$(f \circ g)^2[1] = (f \circ g)[(f \circ g)[1]] = (f \circ g)[3] = f[g[3]] = f[9] = 5,$$

$$(f \circ g)^3[1] = (f \circ g)[(f \circ g)^2[1]] = (f \circ g)[5] = f[g[5]] = f[3] = 7,$$

$$(f \circ g)^4[1] = (f \circ g)[(f \circ g)^3[1]] = (f \circ g)[7] = f[g[7]] = f[10] = 10,$$
$$(f \circ g)^5[1] = (f \circ g)[(f \circ g)^4[1]] = (f \circ g)[10] = f[g[10]] = f[2] = 6,$$
$$(f \circ g)^6[1] = (f \circ g)[(f \circ g)^5[1]] = (f \circ g)[6] = f[g[6]] = f[6] = 1.$$

Therefore the first cycle in the disjoint cycle description of $f \circ g$ will be $(1\,3\,5\,7\,10\,6)$. Continuing like this one finds

$$(1\,2\,6)(3\,7)(4\,8\,11)(5\,9)(1\,7\,10\,2)(3\,9\,5) = (1\,3\,5\,7\,10\,6)(4\,8\,11).$$

1.3 Symmetries of a square

Keywords: permutation description of symmetries.

Permutations come up in a variety of settings. One which we look into now is to describe symmetries of geometric objects. We will as an example consider symmetries of a square. A symmetry of the square is rotation or reflection of the plane that sends the square to itself. For example a counterclockwise rotation of the plane over $\pi/2$ radians (90 degree) with center of rotation in the center of the square is such a symmetry. Also a reflection in the line passing through to opposite vertices of the square is a symmetry of the square. Such symmetries can be described using permutations by bookkeeping what happens with the four vertices. First we enumerate the vertices of the square as in Figure 1.1.

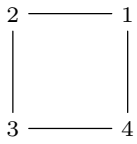


Figure 1.1: A square

Then for each symmetry, we just keep track of how the vertices are permuted. Hence a symmetry of the square can be described by an element of S_4. For example the counterclockwise rotation over $\pi/2$ radians with the center of rotation in the center of the square has the following effect:

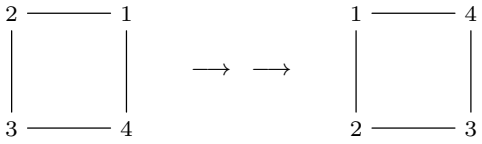


Figure 1.2: Counterclockwise rotation of a square

In other words, the rotation gives rise to the permutation $(1\,2\,3\,4) \in S_4$.

What about the reflection in the line connecting vertices 1 and 3? Graphically its effect is as given in Figure 1.3.

The reflection in the line connecting vertices 2 and 4 gives rise to the permutation $(2\,4) \in S_4$.

If we compose the above two symmetries, say first rotate then take the reflection, we obtain another symmetry. We can obtain the corresponding permutation by composing: $(2\,4)(1\,2\,3\,4) =$

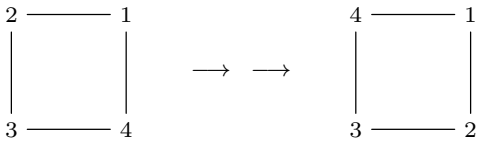


Figure 1.3: Mirror image in a diagonal of a square

$(1\,4)(2\,3)$. This permutation corresponds to another symmetry of the square, namely the reflection in the line passing through the center of the square and dividing the edges connecting 1 and 4, respectively 2 and 3 in equal parts.

There are more symmetries to be considered. We can rotate not only by $\pi/2$ radians counterclockwise, but also by 0, π or $3\pi/2$ radians. The rotation by 0 radians, is simply the identity id. Also there are four lines in which we can reflect (the two diagonals and the two lines dividing the square in to equal rectangles). In this way we obtain 8 distinct symmetries. It is left as an exercise to work out the exact permutations these 8 symmetries give rise to, but the total set of 8 permutations turns out to be

$$\{\mathrm{id}, (1\,2\,3\,4), (1\,3)(2\,4), (1\,4\,3\,2), (1\,3), (1\,4)(2\,3), (2\,4), (1\,2)(3\,4)\}.$$

These 8 permutations form a subset of S_4, but this subset has an additional property. Since the composition of two symmetries of the square is a symmetry of the square itself, these eight permutations form a subset of S_4 that is closed under composition. In the next chapters we will come back to such subsets in a more general setting.

1.4 The sign of a permutation

Keywords: permutation matrices, sign of a permutation, even and odd permutations.

In this section we will define what is known as the sign-map $\mathrm{sign} : S_A \to \{-1, 1\}$ from the permutation group S_A to the set of numbers ± 1. For convenience we will work with the set $A = \{1, 2, \ldots, n\}$, so that $S_A = S_n$. We define the sign using permutation matrices:

Definition 17 *For any permutation $f \in S_n$, we define an $n \times n$ matrix M_f*

$$(M_f)_{ij} := \begin{cases} 1 & \text{if } f(i) = j \\ 0 & \text{otherwise.} \end{cases}$$

Matrices of this form are called permutation matrices. One nice property they have is the following:

$$M_{g \circ f} = M_f \cdot M_g, \text{ where } \cdot \text{ denotes matrix multiplication.} \tag{1.2}$$

To show this is an exercise. We then define the sign of a permutation $f \in S_n$ as follows.

$$\mathrm{sign}(f) := \det(M_f). \tag{1.3}$$

Equation (1.2) implies that

$$\mathrm{sign}(f)\mathrm{sign}(g) = \det(M_f)\det(M_g) = \det(M_f M_g) = \det(M_{g \circ f}) = \mathrm{sign}(g \circ f). \tag{1.4}$$

This is a very useful property to investigate the sign of a permutation. Let us for example show that $\mathrm{sign}(f)$ only can take the values ± 1.

Lemma 18 *Let n be a natural number and $f \in S_n$. Then* $\operatorname{sign}(f) \in \{-1, 1\}$.

Proof. Since $\operatorname{sign}(f)$ is the determinant of a matrix only containing 0's and 1's, it is clear that $\operatorname{sign}(f) \in \mathbb{Z}$. Since this statement is valid for any permutation f, we also have $\operatorname{sign}(f^{-1}) \in \mathbb{Z}$. Moreover, by Equation (1.4)we have

$$\operatorname{sign}(f)\operatorname{sign}(f^{-1}) = \operatorname{sign}(f^{-1} \circ f) = \operatorname{sign}(\mathrm{id}) = 1.$$

This implies that $\operatorname{sign}(f^{-1}) = 1/\operatorname{sign}(f)$. Since, as we already saw, both $\operatorname{sign}(f)$ and $\operatorname{sign}(f^{-1})$ are integers, we may conclude that $\operatorname{sign}(f) \in \{-1, 1\}$. ■

The permutations in S_n with sign equal to 1 are called even permutations. Those with sign equal to -1 are called odd permutations.

Example 19 Let $n = 3$ and $f = (1\,2)$. Then

$$M_f = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ and hence } \operatorname{sign}(f) = \det(M_f) = -1.$$

Hence $(1\,2)$ is an odd permutation.

Remark 20 *Just as an aside: It is possible to give a compact formula for the determinant of a matrix using the sign function. It turns out that for a matrix*

$$A = \begin{pmatrix} a_{1\,1} & a_{1\,2} & \cdots & a_{1\,n} \\ a_{2\,1} & a_{2\,2} & \cdots & a_{2\,n} \\ \vdots & \vdots & & \vdots \\ a_{n\,1} & a_{n\,2} & \cdots & a_{n\,n} \end{pmatrix},$$

one has

$$\det(A) = \sum_{f \in S_n} \operatorname{sign}(f) \prod_{i=1}^{n} a_{i\,f(i)}.$$

Equation (1.4) can be used to compute the sign of permutations without using Equation (1.3) directly. For example, we have

Lemma 21 *The sign of an m-cycle is* $(-1)^{m-1}$.

Proof. We will show this using induction with induction basis $m = 2$. The basis case is an exercise. Assuming the result for $m - 1$, note that an m-cycle $(a_1\, a_2\, \ldots\, a_m)$ can be written as

$$(a_1\, a_2\, \ldots\, a_m) = (a_1\, a_2) \circ (a_2\, \ldots\, a_m).$$

To show this equality, it is enough to check that both permutations assign to any $a \in A$ the same value. We distinguish two cases:

1. $a \notin \{a_1, \ldots, a_m\}$. In this case both $(a_1\, a_2\, \ldots\, a_m)$ and $(a_1\, a_2) \circ (a_2\, \ldots\, a_m)$ keep a fixed.
2. $a = a_1$. In this case we have $(a_1\, a_2\, \ldots\, a_m)[a_1] = a_2$ and $((a_1\, a_2) \circ (a_2\, \ldots\, a_m))[a_1] = (a_1\, a_2)[a_1] = a_2$.
3. $a \in \{a_2, \ldots, a_{m-1}\}$, say $a = a_i$ for $1 < i < m$. In this case we have $(a_1\, a_2\, \ldots\, a_m)[a_i] = a_{i+1}$ and $((a_1\, a_2) \circ (a_2\, \ldots\, a_m))[a_i] = (a_1\, a_2)[a_{i+1}] = a_{i+1}$.

4. $a = a_m$. In this case we have $(a_1\, a_2\, \ldots\, a_m)[a_m] = a_1$ and $((a_1\, a_2) \circ (a_2\, \ldots\, a_m))[a_m] = (a_1\, a_2)[a_2] = a_1$.

All in all, we have now shown that $(a_1\, a_2\, \ldots\, a_m)$ and $(a_1\, a_2) \circ (a_2\, \ldots\, a_m)$ are the same permutations. Using this, Equation (1.4) and the induction hypothesis we then obtain that

$$\begin{aligned} \text{sign}(\,(a_1\, a_2\, \ldots\, a_m)\,) &= \text{sign}(\,(a_1\, a_2) \circ (a_2\, a_2\, \ldots\, a_m)\,) \\ &= \text{sign}(\,(a_1\, a_2)\,)\text{sign}(\,(a_2\, a_2\, \ldots\, a_m)\,) \\ &= (-1) \cdot (-1)^{m-2} = (-1)^{m-1}. \end{aligned}$$

By the induction principle, we have now shown that the lemma is true. ■

By Theorem 15 any permutation can be written as the composite of cycles. Using the above lemma and Equation (1.4), it is now easy to compute the sign of a permutation.

Example 22 This example is a continuation of Example 16. Let $f = (1\,2\,6)(3\,7)(4\,8\,11)(5\,9)$ and $g = (1\,7\,10\,2)(3\,9\,5)$. Then

$$\text{sign}(f) = \text{sign}(\,(1\,2\,6)\,)\text{sign}(\,(3\,7)\,)\text{sign}(\,(4\,8\,11)\,)\text{sign}(\,(5\,9)\,) = (-1)^2(-1)(-1)^2(-1) = 1,$$

and

$$\text{sign}(g) = \text{sign}(\,(1\,7\,10\,2)\,)\text{sign}(\,(3\,9\,5)\,) = (-1)^3(-1)^2 = -1.$$

To compute the sign of $f \circ g$ we actually do not have to compute its disjoint cycle decomposition. We can simply use Equation (1.4) an obtain that: $\text{sign}(f \circ g) = \text{sign}(f)\text{sign}(g) = -1$. Just to check, let us use the computation from Example 16 as well. There we saw that $f \circ g = (1\,3\,5\,7\,10\,6)(4\,8\,11)$. Therefore we obtain that

$$\text{sign}(f \circ g) = \text{sign}(\,(1\,3\,5\,7\,10\,6)\,)\text{sign}(\,(4\,8\,11)\,) = (-1)^5(-1)^2 = -1,$$

just as we expected.

1.5 Exercises

1. Compute the disjoint cycle description of the permutation

$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 7 & 5 & 4 & 1 & 6 \end{pmatrix}.$$

2. Given the permutations $f = (14)(236)$ and $g = (132)(456)$, compute the disjoint cycle description of the permutations $f \circ g$ and $g \circ f$.

3. In this exercise you are asked to prove some of the statements from the text.

   (a) Show that if $f, g \in S_A$, then $f \circ g$ as defined in Equation (1.1) is a permutation from A to itself.

   (b) Check the second item in Theorem 10: The identity permutation $\text{id} \in S_A$ satisfies $\text{id} \circ f = f$ and $f \circ \text{id} = f$ for any $f \in S_A$.

4. In this exercise you are asked to find a permutation description of the rotational symmetries of a tetrahedron. First use a model of a tetrahedron to find and describe all rotational symmetries. Now numerate the vertices of a tetrahedron from 1 up to 4 and consider the permutations from S_4 one can obtain from the rotational symmetries of the tetrahedron. Are all elements of S_4 be obtained?

5. (a) Show that any 3-cycle can be written as the composite of two 2-cycles.
   (b) Show more generally that an m-cycle can be written as the composite of $m-1$ many 2-cycles.
   (c) Conclude that any permutation can be written as the composite of 2-cycles.

6. Let f and g in S_A be two disjoint cycles. Show that $f \circ g = g \circ f$.

7. Show Equation (1.2): $M_{g\circ f} = M_f \cdot M_g$, where $\cdot$ denotes matrix multiplication. Hint: recall that the (i,j)-th entry for a product of two $n \times n$ matrices A and B is given by the formula $(A \cdot B)_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}$.

8. Show that the sign of a 2-cycle is -1 using Equation (1.3) directly.

9. Describe all rotation symmetries of a cube (there are 24 of them). Now identify them with the elements in S_4. Hint: Enumerate the four diagonals of the cube and describe the permutations of these diagonals for the distinct rotation symmetries of the cube.

10. Show that the permutation (1 2 3) cannot be written as the composite of an odd number of 2-cycles. Hint: use the sign function. This explains the terminology "even" and "odd" permutations. An even (respectively odd) permutation can be written as the composite of an even (respectively odd) number of 2-cycles.

Extra exercise for the energetic: the 15-puzzle.

The 15-puzzle is a game where the goal is to rearrange 15 squares from any given constellation into the following target constellation

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	

There is one empty position and the only way to move squares is by pushing a square neighbouring the empty position into this position. We will call such changes "legal moves". You can read more on the puzzle on wikipedia. As in the lectures, we will denote the empty square as square 16. Also as in the lectures, we will use permutations in S_{16}, to describe all constellations. More precisely, a permutation $f \in S_{16}$ corresponds to the constellation:

f[1]	f[2]	f[3]	f[4]
f[5]	f[6]	f[7]	f[8]
f[9]	f[10]	f[11]	f[12]
f[13]	f[14]	f[15]	f[16]

In particular, the target constellation corresponds to the identity permutation id $\in S_{16}$. The main goal of this exercise is to investigate whether or not the initial constellation given by $f = (14\,15)$, that is so say

1	2	3	4
5	6	7	8
9	10	11	12
13	15	14	16

can be rearranged into the target constellation using legal moves. In other words, the question that needs answering is:

"Can one, using a sequence of legal moves, interchange squares 14 **and** 15 **without changing the position of any of the other squares?"**

To arrive at the answer, please address the following questions:

1. Suppose that a constellation C is described by $f \in S_{16}$ and define $a := f^{-1}[16]$. Show that the constellation obtained from c by interchanging 16 and one of its neighbours. is described by a permutation of the form $f \circ (a\, b)$, where b is a neighbouring square of square a.

2. Suppose that several legal moves are performed changing a constellation C, but that 16 ends up in its own position again. Show that then the number of moves has to be even.

3. Let a constellation C, described by $f \in S_{16}$ be changed by legal moves into a constellation D described by $g \in S_{16}$. Suppose that $f^{-1}[16] = g^{-1}[16]$, that is to say that the empty square 16 is at the same position in C and D. Show that in this case $\text{sign}(f) = \text{sign}(g)$.

4. Can the constellation described by $f = (14\,15)$ be changed to the target constellation using legal moves?

5. Bonus question: Which constellations can be rearranged, using legal moves only, into the target constellation?

Chapter 2

Groups

2.1 Abstract groups

Keywords: abstract group, examples of groups, order of an element

It is possible to capture the essence of permutation groups and define a more abstract structure called a group . It involves elements from a set G and a group operation $\cdot$. More precisely:

Definition 23 *A pair $(G, \cdot)$ consisting of a set G and a group operation $\cdot : G \times G \to G$ is called a group if the following three properties (usually called groups axioms) are satisfied:*

- *for any elements $f, g, h \in G$ we have $f \cdot (g \cdot h) = (f \cdot g) \cdot h$ (one says that the group operation is associative),*
- *there exists an element $e \in G$, called the identity element of G, such that $e \cdot f = f$ and $f \cdot e = f$ for any $f \in G$,*
- *for any $f \in G$ there exists an element $g \in G$ such that $f \cdot g = e$ and $g \cdot f = e$ (the element g is called the inverse of f and will be denoted by f^{-1}).*

Example 24 As a first example of a group $(G, \cdot)$, we can take $(S_n, \circ)$. The group axioms in this case are exactly the properties mentioned in Theorem 10.

Example 25 Let us consider the integers $\mathbb{Z}$ with the usual addition as operation. Then $(\mathbb{Z}, +)$ is a group. The identity element is given by 0, since for any integer f it holds that $0 + h = h$ and $h + 0 = h$. The inverse of an integer f is given by $-f$. The notation for the inverse of an element f in Definition 23 was f^{-1}, but when the group operation is an addition, it is much more common to write $-f$ for the inverse of f, just as we did for integers. It takes quite a bit of work to prove that the associative law $f + (g + h) = (f + g) + h$ holds for integers. A possible proof involves induction on f, g and h, but we will not go through such a proof here.

The group $(\mathbb{Z}, +)$ is an example of an group with infinitely many elements. The group operation $+$ satisfies the group axioms, but also has the additional property that $f + g = g + f$ for all $f, g \in \mathbb{Z}$. In general, if a group $(G, \cdot)$ satisfies the additional axiom $f \cdot g = g \cdot f$ for all $f, g \in G$, it is called an abelian group, after the mathematician Niels Abel. Other examples of abelian groups are $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$. Here $\mathbb{Q}$ denote the set of rational numbers, $\mathbb{R}$ the set of real numbers, and $\mathbb{C}$ the set of complex numbers.

From the three group axioms, one can build up a whole theory that occurs in many areas of discrete mathematics (for example: graph theory, coding theory, cryptography), but also outside discrete mathematics (for example: geometry, organic chemistry, the theory of relativity, and quantum mechanics). The proofs in group theory can always be brought back to a clever use of the three group axioms. One advantage of this way of proving things is that once one has shown a property of an abstract group, one later does not have to show this again for a particular example of a group. Let us for example consider the following lemma.

Lemma 26 *Let $(G, \cdot)$ be a group. Then it has exactly one identity element.*

Proof. By the second group axiom, we know that there exists at least one identity element. We need to show that there exists only one. We will prove this using the proof-by-contradiction method. Let us therefore assume that there are two distinct elements e_1 and e_2 satisfying:

$$e_1 \cdot f = f \text{ and } f \cdot e_1 = f \text{ for any } f \in G \tag{2.1}$$

and

$$e_2 \cdot f = f \text{ and } f \cdot e_2 = f \text{ for any } f \in G. \tag{2.2}$$

Choosing f equal to e_2 in equation (2.1), we find $e_1 \cdot e_2 = e_2$, but choosing f equal to e_1 in equation (2.2), we can deduce $e_1 \cdot e_2 = e_1$. Combining these two, we conclude $e_1 = e_2$. This is a contradiction to the assumption that e_1 and e_2 were distinct. Apparently, this assumption was wrong and there exists only one identity element. ■

Let us look at some further examples of groups.

Example 27 Let $G = \mathbb{R}^3$ and denote by $+$ the vector addition. Then $(\mathbb{R}^3, +)$ is an abelian group. The identity element is the vector $(0, 0, 0)$, while the inverse of a vector (a, b, c) is given by $-(a, b, c) = (-a, -b, -c)$.

Example 28 Let $\mathrm{GL}(2, \mathbb{R})$ denote the set of all invertible 2×2 matrices and $\cdot$ the usual matrix multiplication. Then $(\mathrm{GL}(2, \mathbb{R}), \cdot)$ is a group. The identity elements is the identity matrix, while inverses are defined in the usual way for matrices. In other words

$$M_F = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

The group axioms, especially the associativity of the multiplication, then follow from the usual properties of matrices. One can show that this group is not abelian.

Example 29 Let n be a positive integer. If a is any integer, then there is exactly is exactly one pair of integers (q, r) such that $a = qn + r$ and $0 \leq r \leq n - 1$. For example if $n = 10$ and $a = 198$, then $r = 8$ and $q = 19$. The integer r is called the remainder of a modulo n and often denoted by $a \mod n$. Hence we have $198 \mod 10 = 8$. We will denote by $\mathbb{Z} \mod n$ the set of all possible remainders modulo n, that is to say $\mathbb{Z} \mod n = \{0, 1, 2, \ldots, n - 1\}$. We can define and addition modulo n, denoted by $+_n$, as follows $a +_n b := (a + b) \mod n$. For example $7 +_{10} 9 = 6$. Then $(\mathbb{Z} \mod n, +_n)$ is a abelian group. The neutral element is given by 0. Under this group operation, the inverse of $a \in \mathbb{Z} \mod n$ is given by $n - a$ if a is nonzero, while 0 it its own inverse.

Example 30 Let n be a positive integer and a, b two arbitrary integers. Then the multiplication modulo n, denoted by $\cdot_n$, is defined as $a \cdot_n b := (a \cdot b) \mod n$. Here $\cdot$ denotes the usual multiplication of integers. For example $8 \cdot_{10} 9 = 72 \mod 10 = 2$. Now let $(\mathbb{Z} \mod n)^* := \{a \in \mathbb{Z}$

mod $n \mid \gcd(a,n) = 1\}$.. Then $((\mathbb{Z} \mod n)^*, \cdot_n)$ is a group. This group occurs in the RSA crypto system (in that case n is the product of two suitably chosen prime numbers). The identity element of the group is 1. To compute inverse elements in this group, one can use the extended Euclidean algorithm. More precisely, let $a \in (\mathbb{Z} \mod n)^*$. Since $\gcd(a,n) = 1$, one can use the extended Euclidean algorithm to find natural numbers r and s such that $r \cdot a + s \cdot n = \gcd(a,n) = 1$. Then r is the inverse of a under the operation $\cdot_n$, since

$$r \cdot_n a = (r \cdot a) \mod n = (1 - s \cdot n) \mod n = 1.$$

If the group operation is clear, one often does not write the operation at all. In other words, one often writes fg instead of $f \cdot g$. As for permutations, a common notation in group theory is to write $f^2 = f \cdot f$, $f^3 = f \cdot f \cdot f$, etc. Negative exponents are also possible by defining $f^{-m} = (f^{-1})^m$.

Definition 31 *Let $(G, \cdot)$ be a group and $g \in G$. The smallest positive natural number n (if it exists) such that $g^n = e$ is called the order of g . If for all positive natural numbers n it holds that $g^n \neq e$, the order of g is said to be infinite. We will use the notation* $\mathrm{ord}(g)$ *for the order of g. The order of a group $(G, \cdot)$ is defined as the number of elements in G, that is to say as $|G|$.*

Example 32 We claim that the order $\mathrm{ord}(f)$ of the permutation $f = (123)(47)$ is 6. First we determine this by calculating the first six powers of f:

$$\begin{aligned} f &= (123)(47) \\ f^2 &= (123)(47)(123)(47) = (132) \\ f^3 &= f^2 \circ f = (132)(123)(47) = (47), \\ f^4 &= f^3 \circ f = (47)(123)(47) = (123) \\ f^5 &= f^4 \circ f = (123)(123)(47) = (132)(47) \\ f^6 &= f^5 \circ f = (132)(47)(123)(47) = \mathrm{id} \end{aligned}$$

However, there is a faster way: since the 2-cycle (47) and the 3-cycle (123) involve different elements from $\{1, \ldots, n\}$, they commute with each other. That is to say, we have $(123)(47) = (47)(123)$. Therefore we have $f^n = (123)^n(47)^n$. Then it is clear that $f^n = \mathrm{id}$ is only possible if n is a multiple of 6, since $(47)^n = \mathrm{id}$ if and only if n is a multiple of two and $(123)^n = \mathrm{id}$ if and only if n is a multiple of three.

2.2 Symmetries of a regular n-gon

Keywords: cyclic group, dihedral group.

In this section we consider the symmetries of a regular n-gon. For $n \geq 2$, we consider the regular n-gon having n vertices $(\cos(2k\pi/n), \sin(2k\pi/n)) \in \mathbb{R}^2$, with $k = 0, 1, \ldots, n-1$ and n edges, connecting consecutive vertices. For $n = 3$ one obtains a regular triangle, for $n = 4$ a square, etc.

Denote by r the rotation over the angle $2\pi/n$ and center $(0,0)$. The rotation r is a (rotational) symmetry of the regular n-gon. The n-gon has more rotational symmetries, since we could also consider rotations over the angle $2k\pi/n$ and center $(0,0)$ for $k = 0, 1, \ldots, n-1$. However, all these rotational symmetries can be described as the form r^k for $k = 0, 1, \ldots, n-1$. Using composition $\circ$ as group operation, we obtain in this way a group $(C_n, \circ)$ with

$$C_n := \{e, r, \ldots, r^{n-1}\}.$$

Note that for $0 \leq k \leq n-1$ we have

$$r^{-k} = (r^k)^{-1} = r^{n-k}.$$

and for $a, b \in \{0, 1, \ldots, n-1\}$ we have

$$r^a \circ r^b = r^{a+_n b},$$

since if $a + b = c + n$ for $c \in \{0, 1, \ldots n-1\}$, then $r^{a+b} = r^{a+b}e = r^{a+b}r^{-n} = r^{a+b-n} = r^c$.

The group $(C_n, \circ)$ is an example of what is called a cyclic group:

Definition 33 *A finite group $(G, \cdot)$ of order $|G| = n$ is called cyclic if there exists a $g \in G$ such that* $\mathrm{ord}(g) = n$.

What this really means is that all elements in G can be described as powers of g, so in other words a group $(G, \cdot)$ is cyclic precisely if there exists $g \in G$ such that $G = \{e, g, g^2, \ldots, g^{n-1}\}$. Such a g is called a generator of the cyclic group. In particular the group $(C_n, \circ)$ is cyclic, since we may choose r as generator.

Coming back to symmetries of the regular n-gon, we can observe that the reflection in the x-axis also is one of its symmetries, which we will denote by s. There are more symmetries, since we for example can compose r with s (obtaining $r \circ s$). If we take the composition of several of these elements, we can get a complicated expression, like $srsrr$ (suppressing the group operation $\circ$'s in the notation as is customary), but we can use the following lemma to simplify such expressions:

Lemma 34 *Let r and s be as above. Then we have*

1. $r^{-1} = r^{n-1}$ *and* $s^{-1} = s$.
2. $sr = r^{-1}s$.

Proof.

1. Since r is a rotation of the n-gon over the angle $2\pi/n$ and center $(0, 0)$, the rotation r^n keeps all vertices of the n-gon fixed. That is to say, we have that $r^n = e$. This means that $r^{-1} = r^{n-1}$. Similarly, $s^2 = e$, since taking the mirror image in the x-axis twice fixed all vertices. Therefore we have $s^{-1} = s$.

2. To show that $sr = r^{-1}s$, we compute the effect of both sr and $r^{-1}s$ on a vertex

   $$(\cos(2k\pi/n), \sin(2k\pi/n)).$$

   In the first place we have that:

   $$r \text{ sends } (\cos(2k\pi/n), \sin(2k\pi/n)) \text{ to } (\cos(2(k+1)\pi/n), \sin(2(k+1)\pi/n)),$$

   while

   $$s \text{ sends } (\cos(2k\pi/n), \sin(2k\pi/n)) \text{ to } (\cos(2k\pi/n), -\sin(2k\pi/n)),$$

   which equals $(\cos(-2k\pi/n), \sin(-2k\pi/n))$. Therefore:

   $$sr \text{ sends } (\cos(2k\pi/n), \sin(2k\pi/n)) \text{ to } (\cos(-2(k+1)\pi/n), \sin(-2(k+1)\pi/n)),$$

   while

   $$r^{-1}s \text{ sends } (\cos(2k\pi/n), \sin(2k\pi/n)) \text{ to } (\cos(2(-k-1)\pi/n), \sin(2(-k-1)\pi/n)).$$

   We see that sr and $r^{-1}s$ have the same effect on all vertices of the regular n-gon. Therefore they are the same.

■

The above lemma can be used to simplify the expression $srsrr$. Strictly speaking we should have put parentheses, for example $(((sr)s)r)r$, although we know from the associative law that the choice of parentheses does not matter. With this choice of parentheses and the above lemma we obtain using Lemma 34:

$$(((sr)s)r)r = (r^{-1}s)srr = ((r^{-1}(ss))r)r = ((r^{-1}e)r)r = (r^{-1}r)r = er = r.$$

With this lemma we can also the show the following:

Theorem 35 *Let $n \geq 2$ be an integer and define $D_n := \{e, r, \ldots, r^{n-1}, s, rs, \ldots, r^{n-1}s\}$. The pair $(D_n, \circ)$ forms a group.*

Proof. We first need to show that any composition of two elements is again in the same set D. This means that we need to show that for any pair of natural numbers i and j (both between 0 and $n-1$) that $r^i r^j \in S$, $r^i s r^j \in S$, $r^i s r^j s \in S$, and $r^i r^j s \in S$. Using that $r^n = e$, it is not hard to see that $r^i r^j = r^{i+j} \in D$: if $i+j < n$, this is clear, while if $i+j \geq n$, then $r^{i+j} = r^{i+j-n} \in D$. A similar argument shows that $r^i r^j s$ can be rewritten such that is in D.

Using induction and Lemma 34, one can show that for any natural number j it holds that $sr^j = r^{-j}s$. This observation can be used to see that $r^i s r^j = r^{-i+j}s$. If $-i+j \geq 0$, we already have $r^{-i+j}s \in D$, while if $-i+j < 0$, then $r^{-i+j}s = r^{n-i+j}s \in D$. Similarly one show that $r^i s r^j s = r^{-i+j}s^2 = r^{-i+j}$ can always be rewritten in a form occurring in the set D.

To check the remaining group axioms can be slightly tedious, but can all be done checking some cases. We give some examples.

Associativity:

$$(r^i s r^j)r^k s = (r^{i-j}s) \circ (r^k s) = r^{i-j-k},$$

while

$$r^i s \circ (r^j \circ r^k s) = (r^i s) \circ (r^{j+k}s) = r^{i-(j+k)},$$

which is the same element.

Inverses: $(r^i)^{-1} = r^{n-i}$ and $(r^i s)^{-1} = r^i s$. The second of these statements is true, since

$$(r^i s) \circ (r^i s) = r^{i-i}s^2 = e.$$

■

The group in the above theorem is called the dihedral group and is denoted by $(D_n, \circ)$. Warning: Some books write D_{2n} instead of D_n, which may be confusing when considering for example D_4. In these notes D_n will always denote the groups of symmetries of a regular n-gon. Therefore D_4 consists in these notes of 8 elements, not 4.

2.3 Subgroups

Keywords: subgroup, subgroup generated by an element, alternating group.

Definition 36 *Let $H \subset G$ be a subset of G. Then H is called a subgroup of $(G, \cdot)$ if the following conditions are satisfied:*

- *$e \in H$,*
- *for any $f \in H$ also $f^{-1} \in H$.*

- *for any $f, g \in H$ also $f \cdot g \in H$.*

A subgroup $H \subset G$ simply inherits the group operation of the larger group $(G, \cdot)$. In other words: the group operation in H is the restriction of the group operation of G. The third property in the definition of a subgroup makes sure that this operation send two elements of H to an element of H again.

Example 37 The set of permutations $\{\mathrm{id}, (1234), (13)(24), (1432), (13), (14)(23), (24), (12)(34)\}$ found in Section 1.3 is a subgroup of $(S_4, \circ)$.

Example 38 The set of even integers $2\mathbb{Z} = \{\dots, -4, -2, 0, 2, 4, \dots\}$ is a subgroup of $(\mathbb{Z}, +)$.

Example 39 Let $G = \mathbb{R}^3$ and denote by $+$ the vector addition. Further denote by $V \subset \mathbb{R}^3$ the linear subspace defined by

$$V := \{(v_1, v_2, v_3) \in \mathbb{R}^3 \,|\, v_3 = 0\}.$$

Then V is a subgroup of $(\mathbb{R}^3, +)$.

Example 40 The cyclic group C_n is a subgroup of the dihedral group $(D_n, \circ)$.

Definition 41 *Let $(G, \circ)$ be a group and let $g \in G$ be a group element. The set $\langle g \rangle := \{g^n \,|\, n \in \mathbb{Z}\}$ is a subgroup of G. This subgroup is said to be the subgroup generated by g .*

Note that indeed $\langle g \rangle$ is a subgroup of G, since $e = g^0 \in \langle g \rangle$; if $f = g^i \in \langle g \rangle$, then $f^{-1} = g^{-i} \in \langle g \rangle$; if $g^i, g^j \in \langle g \rangle$, then $g^i \cdot g^j = g^{i+j} \in G$.

Lemma 42 *Let $(G, \cdot)$ be a group and let $g \in G$ be a group element. Then the order of the group $\langle g \rangle$ is the same as the order of the element g.*

Proof. If g has finite order, say order m, we can express any power g^i of g in the form g^j, with $j \in \{0, 1, \dots, m-1\}$. Indeed, since we can write $i = mq + j$ with $0 \leq j \leq m-1$ and some integer q, we see that $g^i = (g^m)^q \circ g^j = e \circ g^j = g^j$. Moreover, we claim that any two group elements of the form g^k and g^ℓ, with $k < \ell$ and both k and ℓ between 0 and $m-1$, are distinct. Indeed if not, we would obtain that $e = g^k \circ g^{-k} = g^\ell \circ g^{-k} = g^{\ell-k}$, implying that the order of g would be less than m. This means that the group $\langle g \rangle$ has order m.

Now assume conversely that the group $\langle g \rangle$ has order m. We consider the sequence of group elements $g^0, g^1, g^2, g^3, \dots$. Since $\langle g \rangle$ has finite order, at least one element of $\langle g \rangle$ has to occur twice in the sequence, which means that there exist i, j such that $0 \leq i < j \leq m$ and $g^i = g^j$. This implies that $e = g^i \circ g^{-i} = g^j \circ g^{-i} = g^{j-i}$, so g has order at most m. In fact, the order of g has to be exactly m, since otherwise the first part of the proof would imply that the order of the group $\langle g \rangle$ has order less than m.

We have shown so far implies that the element g has finite order if and only if the group $\langle g \rangle$ has finite order. But then the element g has infinite order if and only if the group $\langle g \rangle$ has infinite order. This concludes the proof. ■

Another example of a subgroup, which we will study in more detail, is a subgroup of $(S_n, \circ)$. It is defined as follows:

Definition 43 *Let $A_n := \{f \in S_n \,|\, \mathrm{sign}(f) = 1\}$, the set of all even permutations. Then A_n is a subgroup of $(S_n, \circ)$ called the alternating group.*

Example 44 Let $n = 4$, then

$$A_4 = \{\mathrm{id}, (123), (132), (124), (142), (134), (143), (234), (243), (13)(24), (14)(23), (12)(34)\}. \quad (2.3)$$

We have seen that any permutation can be written as a composition of 2-cycles. A result somewhat like it holds for even permutations.

Theorem 45 *Let n be a natural number. Any permutation $f \in A_n$ can be written as the composition of 3-cycles.*

Proof. The identity permutation id can be seen as the composition of zero 3-cycles. Therefore the theorem is true for $n < 3$, since $A_n = \{\mathrm{id}\}$ in these cases.

We will show the theorem by induction on n, the induction basis being the case $n = 2$. From now suppose $n \geq 3$ and let $f \in A_n$ be an even permutation. Moreover assume as induction hypothesis that the theorem is true for $n - 1$. If f has a fixed point, that is to say if f sends i to i for some $i \in \{1, \ldots, n\}$, then we can interpret f as a permutation on $n - 1$ elements by ignoring i. Then by the induction hypothesis, f can be written as the composition of 3-cycles. If f does not have a fixed point, then $f(1) = a$ for some a different from 1. Since $n \geq 3$, we can choose $b \in \{1, \ldots, n\}$ different from both 1 and a. The permutation $g = (a\,1\,b) \circ f \in A_n$ will have 1 as a fixed point. This implies as we have seen before, that g can be written as $c_1 \circ \cdots \circ c_\ell$ for suitably chosen 3-cycles $c_1, \ldots, c_\ell$. But then we find that f can be written as a composition of 3-cycles, namely:

$$f = (a\,1\,b)^{-1} \circ c_1 \circ \cdots \circ c_\ell = (b\,1\,a) \circ c_1 \circ \cdots \circ c_\ell.$$

This concludes the induction step.

By the induction principle, the theorem is true for all values of n. ■

Example 46 Most of the elements in A_4 are 3-cycles, as can be seen from equation (2.3), with four exceptions. Let us write the permutation $(12)(34)$ as composition of 3-cycles, following the recipe given in the proof of Theorem 45. Since $f = (12)(34)$ has no fixed points and sends 1 to 2, we first compose it from the left with the a 3-cycle of the form $(2\,1\,b)$. We choose $b = 3$ and obtain $g = (213)(12)(34) = (234)$. Therefore $(12)(34) = (213)^{-1}(234) = (123)(234)$.

2.4 Rotational symmetries of a dodecahedron

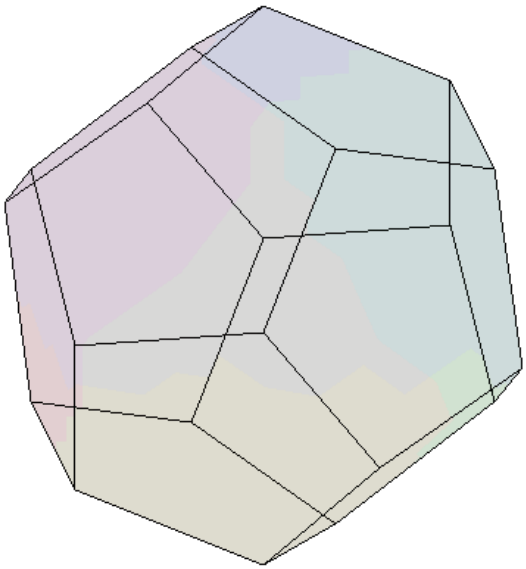
Keywords: generators of the alternating group, rotation symmetries of the dodecahedron.

In this section we will investigate the group of rotational symmetries of a regular dodecahedron, see Figure 2.1.

One type of rotation symmetry consists of a rotation over $2\pi/3$ or $4\pi/3$ radians with rotation axis connecting two opposite vertices of the dodecahedron. Since a regular dodecahedron has 20 vertices, this gives rise to 20 rotation symmetries.

Now consider five cubes inside the dodecahedron as in Figure 2.2. The vertices of each of these cubes are vertices of the dodecahedron as well. Therefore any rotation symmetry can be described by an element of S_5. Moreover, each vertex of the dodecahedron is exactly a vertex of two out of these five cubes. This means that each of the 20 rotation symmetries with rotation axis passing through opposite vertices will keep two cubes fixed, while permuting the other three in a 3-cycle. This means that these 20 rotation symmetries give rise to 20 3-cycles in A_5.

Figure 2.1: The regular dodecahedron.



The number of 3-cycles in A_5 can be computed as follows: first consider all strings of the form abc, with $a, b, c \in \{1, \ldots, 5\}$. There are $5 \cdot 4 \cdot 3 = 60$ of them and each string abc gives rise to a 3-cycle (abc). However, since $(abc) = (cab) = (bca)$, we do not obtain 60, but 20 3-cycles. It is clear that any 3-cycle in S_5 will be obtained in this way. We conclude that there are exactly 20 3-cycles in A_5.

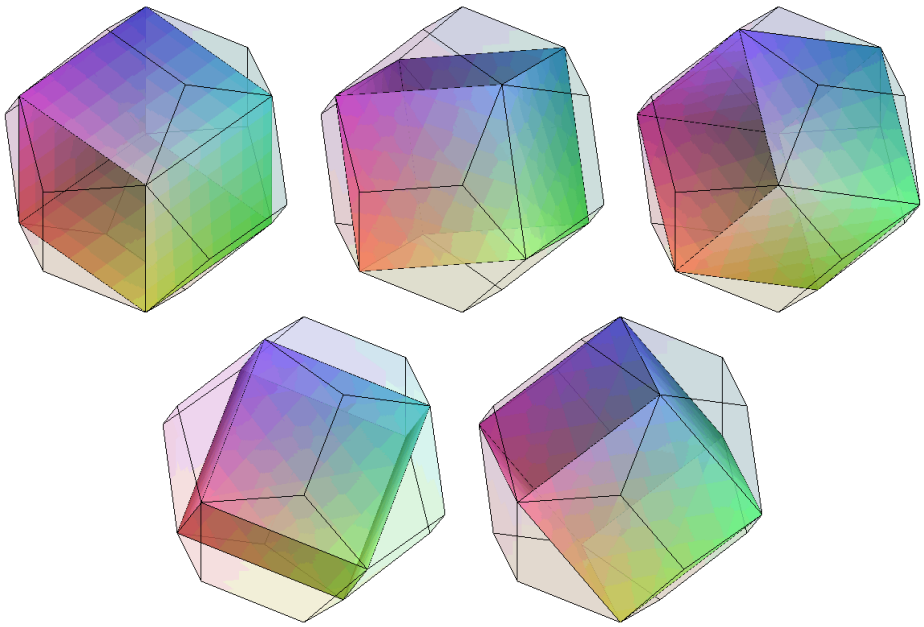
Combining the above, we see that the above 20 rotation symmetries of the dodecahedron give rise to all 20 3-cycles in A_5. By Theorem 45 we can conclude that considering all rotation symmetries of the dodecahedron, we will obtain at least all 60 permutations in A_5. However, we may obtain even more. To see that this is not the case, we count the number of rotation symmetries.

A regular dodecahedron has 20 vertices, 30 edges and 12 faces. All twenty vertices of a dodecahedron have the same distance from the center of the dodecahedron. This means that these vertices lie on a sphere with middle point is the center of the dodecahedron. Therefore, any rotation sending a dodecahedron to itself, is a symmetry of the sphere as well. This means that its rotational axis passes through this center. Now we assume this to be the case. Looking at the surface of the dodecahedron and the effect a rotation has on it, we see that the rotational axis of a rotation symmetry either passes through a vertex, the midpoint of a edge, or the midpoint of a face. This means that we find the following list of rotation symmetries:

- 1 identity rotation.
- 20 rotation symmetries with rotation axis passing though a vertex.
- 15 rotation symmetries with rotation axis passing though the midpoint of an edge.
- 24 rotation symmetries with rotation axis passing though the midpoint of a face.

The group of rotation symmetries of a regular dodecahedron apparently has order 60. All in

Figure 2.2: Five cubes inside the regular dodecahedron.



all, we have shown that this group can be identified with A_5 using the permutations of the five inscribed cubes the rotation symmetries give rise to.

2.5 Exercises

1. Is $(\mathbb{N}, +)$ a group?

2. Denote by $\mathbb{R}_+$ the set of positive real numbers. Is $(\mathbb{R}_+, \cdot)$ a group? You may assume that multiplication of real numbers is an associative operation.

3. Show that $(\mathrm{GL}(2, \mathbb{R}), \cdot)$ is not an abelian group.

4. We have simply introduced the notation f^{-1} for the element $g \in G$ such that $f \cdot g = e$ and $g \cdot f = e$. However, this notation only makes sense if there is exists only one inverse for f.

   (a) Show that inverses in a group are unique. In other words, show the following: if for a given $f \in G$ there exist elements $g_1, g_2 \in G$ such that $f \cdot g_2 = f \cdot g_1 = e$ and $g_2 \cdot f = g_1 \cdot f = e$, then $g_1 = g_2$.

   (b) Show that $(f \cdot g)^{-1} = g^{-1} \cdot f^{-1}$.

5. As before we denote by $(S_n, \circ)$ the permutation group on a set with n elements. Is $\{\mathrm{id}, (12), (123), (132)\}$ a subgroup of $(S_3, \circ)$?

6. Let $(G, \cdot)$ be a group and let H be a nonempty subset of G. Show that in this case $H \subset G$ is a subgroup of $(G, \cdot)$ if and only if for all $f, g \in H$ it holds that $f \cdot g^{-1} \in H$.

7. Show that any element in the group $((\mathbb{Z} \mod 8)^*, \cdot_8)$ has order 1 or 2 and use this to determine all possible subgroups (see the examples in the first section of Chapter 2 for a definition of this group).

8. Let n be a prime number and let $C_n \subset D_n$ be the subset of the dihedral group $(D_n, \circ)$ consisting of all rotations (so $C_n = \{e, r, \ldots, r^{n-1}\}$). Show that the only subgroups of C_n are $\{e\}$ and C_n. (Hint: you may use that if $\gcd(n, m) = 1$, then there exist $a, b \in \mathbb{Z}$ such that $an + bm = 1$.)

9. The aim of this exercise is to describe the subgroup $H \subset (\mathrm{GL}(2, \mathbb{R}), \cdot)$ consisting of all rotation and mirror symmetries of a unit circle.

   (a) First show that any matrix $R \in \mathrm{GL}(2, \mathbb{R})$ that gives rise to a rotation symmetry of the unit circle is of the form

   $$R = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix},$$

   for some $\theta \in \mathbb{R}$.

   (b) Show that any reflection symmetry M of the unit circle can be written in the form $M = R \cdot S$, with R a suitably chosen rotation symmetry and S the matrix

   $$S = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

10. Consider one of the 20 rotation symmetries of the dodecahedron with rotation axis passing through two opposite vertices. We consider the action of such a rotation on the 5 inscribed cubes depicted in Figure 2.2.

    (a) What is the order of such a rotation?

    (b) Show using theoretical arguments only that this implies that the action on the 5 inscribed cubes either is described by the identity permutation or by a 3-cycle.

    (c) Check, using the Maple file on campusnet, that the action of such a rotation on the 5 inscribed cubes is in fact described by a 3-cycle.

11. Write the permutation $(1\,2)(3\,4)$ as a composition of 3-cycles.

12. Show that if $H \subset G$ and $K \subset G$ are two subgroups of a group $(G, \cdot)$, then $H \cap K$ is also a subgroup of $(G, \cdot)$. Is the same necessarily true for $H \cup K$?

Chapter 3

Cosets

3.1 Equivalence relations

Keywords: equivalence relation, equivalence class.

In the next sections we will study properties of subgroups of a group. One very useful tool for this study is the notion of an equivalence relation. Let us briefly recall what a relation is. The word relation comes from the idea that sometimes one can give a relation between elements of a set. For example, if $A = \mathbb{R}$, we can relate elements by size using the symbol $\leq$. Then $1 \leq 2$ can be interpreted as: 1 is related to 2 under the relation $\leq$. Many more elements are related to 2, namely all elements in $\mathbb{R}$ less than or equal to 2. One can also relate elements in $\mathbb{R}$ using the symbol $=$. In this case only a itself is related to a. In general for a relation $_R$ on a set A, one has for any $a, b \in A$ that either a is related to b (in which case we write $a \,_R\, b$) or that a is not related to b (in which case we write $\neg(a \,_R\, b)$).

We will need a special kind of relation called an equivalence relation:

Definition 47 *Let A be a set. An* equivalence relation $\sim$ *on A, is a relation on A satisfying the following:*

- *For all $a \in A$ we have $a \sim a$ (reflexivity).*
- *For all $a, b \in A$ we have $a \sim b$ implies $b \sim a$ (symmetry).*
- *For all $a, b, c \in A$ we have $a \sim b$ and $b \sim c$ imply $a \sim c$ (transitivity).*

Note that the relation $\leq$ on real numbers is not an equivalence relation, since the symmetry does not hold in general. Indeed $1 \leq 2$, but it does not hold that $2 \leq 1$. On the other hand, the relation $=$ on real numbers is an equivalence relation. Given an equivalence relation $\sim$ on a set A and an element $a \in A$, we define the *equivalence class* of a to be the set:

$$[a]_\sim := \{b \in A \mid a \sim b\}.$$

For example, if $A = \mathbb{R}$ and the equivalence relation is the one defined by the usual $=$ sign, then $[a] = \{a\}$.

Another example of an equivalence relation is given in the following example:

Example 48 Let $A = \mathbb{Z}$, the set of integers. Let $n \geq 1$ be a natural number. For $a, b \in \mathbb{Z}$ we write $a \equiv b \pmod{n}$ if n divides $b - a$. It is an exercise to show that this in fact defines an equivalence relation. The equivalence classes are $a + n\mathbb{Z} := \{a + kn \mid k \in \mathbb{Z}\}$ for $a = 0, 1, \ldots, n-1$.

Equivalence classes have several nice properties that turn out to be useful when studying subgroups of a group. We collect some properties in the following theorem.

Theorem 49 *Let A be a set and $\sim$ an equivalence relation on A. Then we have:*

1. *For any $a \in A$ we have $a \in [a]_\sim$.*
2. *The set A is covered by equivalence classes: $\cup_{a\in A}[a]_\sim = A$.*
3. *For any $a, b \in A$ we have $[a]_\sim \cap [b]_\sim = \emptyset$ or $[a]_\sim = [b]_\sim$.*
4. *For any $a, b \in A$ we have $[a]_\sim = [b]_\sim$ if and only if $a \sim b$.*

Proof. By reflexivity we have that $a \sim a$. Therefore $a \in [a]_\sim$. This proves the first part. The second part follows immediately from the first part, since any element $a \in A$ is in some equivalence class, for example in $[a]_\sim$. The third part is the most laborious to show. The given statement is equivalent to showing that

$$[a]_\sim \cap [b]_\sim \neq \emptyset \text{ implies } [a]_\sim = [b]_\sim.$$

So let us assume that there exists $x \in A$ such that $x \in [a]_\sim \cap [b]_\sim$. In this case we know by definition of equivalence classes that $a \sim x$ and $b \sim x$. We will show that $[a]_\sim = [b]_\sim$ by showing that $[a]_\sim \subset [b]_\sim$ and $[b]_\sim \subset [a]_\sim$.

$[a]_\sim \subset [b]_\sim$: Assume that $c \in [a]_\sim$. Then by definition $a \sim c$. We wish to show that $b \sim c$, since then $c \in [b]_\sim$. First of all we know that $a \sim x$ and $a \sim c$. Using symmetry on the first equivalence and transitivity afterwards, we may conclude that $x \sim c$. Since we know that $b \sim x$ and (as we have just seen) $x \sim c$, we can use transitivity again to conclude that $b \sim c$. This is exactly what we wanted to show.

$[b]_\sim \subset [a]_\sim$: The proof of this is very similar to the proof that we have just given for the reverse inclusion. We only need to reverse the roles of a and b.

Finally to see that the fourth statement is correct, we need to show the implication $[a]_\sim = [b]_\sim$ implies $a \sim b$ and the reverse implication $a \sim b$ implies $[a]_\sim = [b]_\sim$. If $[a]_\sim = [b]_\sim$, then using part one of the theorem, we see that $b \in [b]_\sim$. However, since we assume that $[a]_\sim = [b]_\sim$, this implies that $b \in [a]_\sim$. By definition of an equivalence class, we see that $a \sim b$. Conversely, if we assume that $a \sim b$, then we see that $b \in [a]_\sim$. Since, again by the first part of the theorem, we also know that $b \in [b]_\sim$, we see that $b \in [a]_\sim \cap [b]_\sim$. Apparently $[a]_\sim \cap [b]_\sim \neq \emptyset$, so by the third part of the theorem we obtain that $[a]_\sim = [b]_\sim$. ■

One says that the set A is partitioned into equivalence classes. The word partitioned is appropriate, since the set A divided into mutually disjoint pieces, namely the various equivalence classes.

3.2 Cosets of a subgroup

Keywords: multiplication of subsets, cosets of a subgroup, cosets are equivalence classes.

In this section we establish a fundamental concept from group theory: cosets. We will use this concept in various ways later on. We will start indicating how one can define a multiplication of subsets of a group.

Definition 50 *Let $(G, \cdot)$ be a group and let $M \subset G$ and $N \subset G$ be two subsets of G. Then we define*

$$M \cdot N := \{f \cdot g \,|\, f \in M\, g \in N\}.$$

Example 51 If we consider the group $(\mathbb{Z}, +)$ and choose

$$M = \{50, 100, 200, 500, 1000\} \text{ and } N = \{0, 50, 100, 200, 500, 1000\},$$

then we would get

$$M + N = \{50, 100, 150, 200, 250, 300, 400, 500, 550, 600, 700, 1000, 1050, 1100, 1200, 1500, 2000\}.$$

The numbers in $M + N$ are in fact exactly the amounts one could pay for using either one or two Danish banknotes.

Definition 52 *Let H be a subgroup of a group $(G, \cdot)$ and let $f \in G$. Then we define the left coset of H in G by f to be:*

$$f \cdot H := \{f\} \cdot H = \{f \cdot h \,|\, h \in H\}.$$

Similarly we define the right coset of H in G by f as

$$H \cdot f := H \cdot \{f\} = \{h \cdot f \,|\, h \in H\}.$$

In an abelian group, there is no difference between left and right cosets, since in that case $f \cdot H = H \cdot f$. Therefore in an abelian group G, we can simply talk about cosets of H in G by f. For non-abelian groups $f \cdot H \neq H \cdot f$ in general, though there are subgroups H for which this does hold. If for a subgroup H it does hold that $f \cdot H = H \cdot f$ for all $f \in G$, then we call the subgroup *normal*.

Example 53 Let $G = S_4$ and

$$H = \{e, (1\,2\,3\,4), (1\,3)(2\,4), (1\,4\,3\,2), (1\,3), (1\,4)(2\,3), (2\,4), (1\,2)(3\,4)\}$$

as in Example 37. Then

$$(1\,2) \circ H = \{(1\,2), (2\,3\,4), (1\,3\,2\,4), (1\,4\,3), (1\,3\,2), (1\,4\,2\,3), (1\,2\,4), (3\,4)\}$$

and

$$H \circ (1\,2) = \{(1\,2), (1\,3\,4), (1\,4\,2\,3), (2\,4\,3), (1\,2\,3), (1\,3\,2\,4), (1\,4\,2), (3\,4)\}.$$

This shows that left and right cosets of a subgroup by the same element are not necessarily the same.

Example 54 Let $G = \mathbb{R}^3$ and $H = V$ be as in Example 39. The coset of V in $\mathbb{R}^3$ by $(5, 6, -4) \in \mathbb{R}^3$ is given by

$$\begin{aligned}
(5,6,-4) + V &= \{(5,6,-4) + (v_1, v_2, v_3) \,|\, (v_1, v_2, v_3) \in V\} \\
&= \{(5,6,-4) + (v_1, v_2, 0) \,|\, v_1, v_2 \in \mathbb{R}\} \\
&= \{(5 + v_1, 6 + v_2, -4) \,|\, v_1, v_2 \in \mathbb{R}\} \\
&= \{(w_1, w_2, -4) \,|\, w_1, w_2 \in \mathbb{R}\} \qquad (3.1)
\end{aligned}$$

We can see the coset $(5,6,-4)+V$ as a translation of the linear subspace V by the vector $(5,6,-4)$. Different translations could produce the same result. For example if we translate V by the vector $(1,0,0)$, we simply get V back again, since the vector $(1,0,0)$ lies within V itself. In other words:

$$(0,0,0)+V=(1,0,0)+V.$$

Similarly, from equation (3.1) we see for example that the coset $(0,0,-4)+V$ is the same as the coset $(5,6,-4)+V$.

We will now describe cosets as equivalence classes of suitably chosen equivalence relations. It turns out we need two relations, one for left cosets and one for right cosets.

Definition 55 *Let $(G,\cdot)$ be a group and $H\subset G$ a subgroup. For $f,g\in G$ we write $f\sim_H g$ if $f^{-1}\cdot g\in H$. We write $f\ {}_H\sim g$ if $g\cdot f^{-1}\in H$.*

The beautiful thing is that these relations in fact are equivalence relations. Let us show this for the relation $\sim_H$ only, since the proof for the relation ${}_H\sim$ is very similar.

Symmetry: for any $f\in G$ we have $f^{-1}\cdot f=e$. Since H is a subgroup, we have $e\in H$. Hence $f\sim_H f$.

Reflexivity: if $f\sim_H g$, then by definition $f^{-1}\cdot g\in H$. Since H is a subgroup we also have $(f^{-1}\cdot g)^{-1}\in H$. However, we have

$$(f^{-1}\cdot g)^{-1}=g^{-1}\cdot(f^{-1})^{-1}=g^{-1}\cdot f$$

and hence $g\sim_H f$.

Transitivity: given that $f\sim_H g$ and $g\sim_H h$ is the same as stating that $f^{-1}\cdot g\in H$ and $g^{-1}\cdot h\in H$. Since H is a subgroup, this implies that $(f^{-1}\cdot g)\cdot(g^{-1}\cdot h)\in H$. However, we have

$$(f^{-1}\cdot g)\cdot(g^{-1}\cdot h)=f^{-1}\cdot((g\cdot g^{-1})\cdot h)=f^{-1}\cdot(e\cdot h)=f^{-1}\cdot h$$

and hence $f\sim_H h$ as desired.

The point of the equivalence relations $\sim_H$ and ${}_H\sim$ is that their equivalence classes actually are cosets. We will show this now:

Lemma 56 *Let $(G,\cdot)$ be a group and $H\subset G$ a subgroup. For $f\in G$ we have*

$$[f]_{\sim_H}=f\cdot H \text{ and } [f]_{{}_H\sim}=H\cdot f.$$

Proof. We will only show that $[f]_{\sim_H}=f\cdot H$, since the statement $[f]_{{}_H\sim}=H\cdot f$ can be shown similarly.

First we show that $[f]_{\sim_H}\subset f\cdot H$: If $g\in[f]_{\sim_H}$, then by definition we have $f\sim_H g$, that is to say $f^{-1}\cdot g\in H$. But then we can write $f^{-1}\cdot g=h$ for some $h\in H$ and hence we have $g=f\cdot h$ for a certain $h\in H$. This means that $g\in f\cdot H$. Since g was chosen arbitrarily from $[f]_{\sim_H}$, we conclude that $[f]_{\sim_H}\subset f\cdot H$.

Now we show the converse inclusion, namely that $f\cdot H\subset[f]_{\sim_H}$: Assume that $g\in f\cdot H$. This is equivalent to saying that there exists $h\in H$ such that $g=f\cdot h$. This implies that $f^{-1}\cdot g=h\in H$.

Hence $f \sim_H g$, meaning that $g \in [f]_{\sim_H}$. We have shown that for any $g \in f \cdot H$ it holds that $g \in [f]_{\sim_H}$, which implies that $f \cdot H \subset [f]_{\sim_H}$.

Combining that we now know that $[f]_{\sim_H} \subset f \cdot H$ and $f \cdot H \subset [f]_{\sim_H}$, we may conclude that $f \cdot H = [f]_{\sim_H}$, which was what we wanted to show. ■

Now that we have identified left and right cosets of H in G as equivalence classes under $\sim_H$ and ${}_H\sim$, we can apply Theorem 49 to these equivalence relations. The result is the following:

Theorem 57 *Let $(G, \cdot)$ be a group and $H \subset G$ a subgroup. Then the following holds:*

1. *For all $f \in G$ we have $f \in f \cdot H$ and $f \in H \cdot f$.*
2. *We have $G = \cup_{f \in G} f \cdot H$ and $G = \cup_{f \in G} H \cdot f$.*
3. *For any $f, g \in G$ we have $f \cdot H \cap g \cdot H = \emptyset$ or $f \cdot H = g \cdot H$. Similarly we have $H \cdot f \cap H \cdot g = \emptyset$ or $H \cdot f = H \cdot g$.*
4. *For any $f, g \in G$ we have $f \cdot H = g \cdot H$ if and only if $f^{-1}g \in H$. Similarly we have $H \cdot f = H \cdot g$ if and only if $gf^{-1} \in H$.*

Proof. This follows from Lemma 56 and Theorem 49. ■

Since $e \cdot H = H$, we see from part four of the above theorem that in particular $H = f \cdot H$ if and only if $f \in H$. Similarly $H = H \cdot f$ if and only if $f \in H$.

Example 58 One can get a good intuitive understanding of Theorem 57 by considering Examples 39 and 54 again. First of all, part one of Theorem 57 just says that $v \in v + V$, which is quite clear: if we translate a plane with translation vector v, then v will be in the translated plane. We can interpret the coset $v + V$ of V as a translations of V by the vector v. It is clear that two translated planes $v + V$ and $w + V$ can be the same, in which case the difference vector $-v + w$ is in V (this is just part four of Theorem 57). In case the translated planes $v + V$ and $w + V$ are distinct, they are parallel to each other. Therefore if $v + V \neq w + V$, we have $(v+V) \cap (w+V) = \emptyset$, just as part three of Theorem 57 predicts. Finally, the union of all possible translated planes $v + V$ fills out the entire space $\mathbb{R}^3$. This is exactly what part two of Theorem 57 boils down to in this example.

3.3 The order of a subgroup and of an element

Keywords: order of a subgroup, index of a subgroup, Euler's theorem, Fermat's little theorem.

A first useful application of left (or right) cosets is the following:

Theorem 59 *Let $(G, \cdot)$ be a finite group and let $H \subset G$ be a subgroup. Then the number of elements in H, that is to say $|H|$, divides $|G|$.*

Proof. By Theorem 57, one can write G as the disjoint union of left cosets of H (or of right cosets of H). Since G is finite, we see that we can write G as the disjoint union of finitely many left cosets of H, say

$$G = \cup_{i=1}^{n} f_i \cdot H. \tag{3.2}$$

We claim that any left coset $f \cdot H$ of H in G contains the same number of elements as H itself. Consider the map ℓ from H to $f \cdot H$ defined by $\ell(h) = f \cdot h$. By definition of a coset, this map is surjective. The map is also injective. For if $f \cdot h_1 = f \cdot h_2$, then

$$f^{-1} \cdot (f \cdot h_1) = f^{-1} \cdot (f \cdot h_2),$$

which implies that

$$h_1 = (f^{-1} \cdot f) \cdot h_1 = f^{-1} \cdot (f \cdot h_1) = f^{-1} \cdot (f \cdot h_2) = (f^{-1} \cdot f) \cdot h_2 = h_2.$$

But then $|(f \cdot H)|$ (the number of elements in the coset $f \cdot H$) is equal to $|H|$ (the number of elements in H) for any $f \in G$. Equation (3.2) now implies that $|G| = n|H|$. ■

The following notation concerning the number of cosets is often used:

Definition 60 *Let $(G, \cdot)$ be a group and H a subgroup. The number of left (or right) cosets H in G is denoted by $[G : H]$ and called the index of H in G.*

Example 61 Let $G = D_8$ and let $H = \{e, r^2, r^4, r^6\}$. Then $[G : H] = |D_8|/|H| = 16/4 = 4$. Indeed, H has four left cosets in G, namely H, $rH = \{r, r^3, r^5, r^7\}$, $sH = \{s, r^6s, r^4s, r^2s\}$ and $rsH = \{rs, r^7s, r^5s, r^3s\}$.

If G is a finite group, we see from the proof of Theorem 59 that $|G|/|H| = [G : H]$. A consequence of Theorem 59 is the following:

Proposition 62 *Let $(G, \cdot)$ be a finite group and let $g \in G$ be a group element. Then $\mathrm{ord}(g)$ divides $|G|$. In particular we have $g^{|G|} = e$ for any group element $g \in G$.*

Proof. The first part of the proposition follows from Lemma 42 and Theorem 59. To show that $g^{|G|} = e$ for any $g \in G$ note that we can write $|G| = \mathrm{ord}(g)n$ for some natural number n. But then we have

$$g^{|G|} = g^{\mathrm{ord}(g)n} = (g^{\mathrm{ord}(g)})^n = e^n = e,$$

for any $g \in G$. ■

This proposition has a number of interesting consequences that also turn up in applications, for example in the RSA crypto system.

Corollary 63 *Let $d, n \in \mathbb{Z}$ be two integers and assume that $\gcd(d, n) = 1$. Then $d^{\phi(n)} \equiv 1 \bmod n$. Here ϕ denotes the Euler totient function defined by $\phi(n) = |(\mathbb{Z} \mod n)^*|$.*

Proof. The corollary follows by applying Proposition 62 to the group $((\mathbb{Z} \mod n)^*, \cdot_n)$. ■

This theorem is sometimes known as Euler's theorem. In case n is a prime it is known as Fermat's little theorem.

Corollary 64 *Let p be a prime number and d a natural number such that p does not divide d. Then $d^{p-1} \equiv 1 \bmod p$.*

Proof. This corollary follows from Corollary 63 once we show that $\phi(p) = p - 1$. However,

$$\phi(n) = \#(\mathbb{Z} \mod n)^* = |\{d \in 0, \ldots, n-1 \mid \gcd(d, n) = 1\}|,$$

so if n is a prime p we find that

$$\phi(p) = |\{d \in 0, \ldots, p-1 \mid \gcd(d, p) = 1\}| = |\{1, \ldots, p-1\}| = p - 1.$$

■

3.4 Exercises

1. Show that the relation on $\mathbb{Z}$ defined in Example 48 is an equivalence relation. Describe the equivalence classes and convince yourself that Theorem 49 indeed holds for this example.

2. Also in linear algebra, equivalence relations occur regularly:

   (a) Let $\mathrm{Mat}_{n\times n}(\mathbb{R})$ be the set of $n \times n$ matrices with coefficients in $\mathbb{R}$. For $M, N \in \mathrm{Mat}_{n\times n}(\mathbb{R})$, we say that $M \sim N$, if there exists an invertible matrix A such that $N = A \cdot M$. Show that $\sim$ is an equivalence relation. Remark: It turns out any equivalence class contains exactly one matrix in reduced row echelon form. When you are using Gaussian elimination to solve a system of linear equations, you are actually replacing the coefficient matrix M with an easier matrix N such that $M \sim N$.

   (b) Two matrices $M, N \in \mathrm{Mat}_{n\times n}(\mathbb{R})$ are called similar, notation $M \,\mathrm{sim}\, N$ if there exists an invertible matrix Q such that $N = Q^{-1}MQ$. Show that sim is an equivalence relations. Remark: M and N represent the same linear operator under two different bases, with Q being the change of basis matrix. Hence similar matrices are just ways to represent the same linear operator in different ways.

3. Let $G = S_4$ and H the subgroup of S_4 defined in Examples 37 and 53. Compute $(1\,2\,3)H$.

4. Let G and H be as in the previous exercise. Determine without computing cosets whether or not $(1\,2)H = (1\,3\,4)H$.

5. Let p and q be distinct prime numbers and denote by ϕ the Euler totient function (see Corollary 63). Show that $\phi(pq) = (p-1)(q-1)$. This value of the Euler totient function is used in the RSA cryptosystem.

6. A group $(G, \cdot)$ is called cyclic if there exists $g \in G$ such that $\langle g \rangle = G$ or in other words, if there exists $g \in G$ such that $\mathrm{ord}(g) = |G|$.

   (a) Show that any finite group of order a prime is cyclic (and hence abelian).

   (b) Is the following statement true? If p and q are any distinct primes, then a group of order pq is cyclic.

7. Let $H \subset G$ be a subgroup of a group $(G, \cdot)$ and suppose that $[G : H] = 2$. Show that in this case $fH = Hf$ for any $f \in G$ (in other words, there is no distinction between left and right cosets). Remark: A subgroup for which $fH = Hf$ for all $f \in G$ is called a *normal* subgroup.

8. Assume that $n \geq 3$ and let $f \in S_n$ an arbitrary element. Show that $f(1\,2\,3)f^{-1} = (f[1]\, f[2]\, f[3])$ is a 3-cycle. Also show that as f varies among the elements of S_n, all 3-cycles in S_n are obtained. Remark: Similarly one could show that the cycle types of g and fgf^{-1} are the same for any $f, g \in S_n$ and that all permutations with that cycle type are obtained as f varies among the elements of S_n.

Chapter 4

Group actions and Burnside's lemma

4.1 Group actions

Keyword: group action

We have seen that groups can describe symmetries of various objects: platonic solids (tetrahedron, cube, dodecahedron, etc.) or regular n-gons. For all these examples, we could see the effect of a symmetry by describing the permutation it gave rise to on the (parts of the) object. For example, the symmetries of a cube give rise to permutations of its four diagonals. In this example we can say that the group of symmetries of a cube gives rise to permutations of the elements of the set A of diagonals. Similarly, given a set A, elements in S_A give by their very nature rise to a permutation of the element of A. In this chapter we capture these examples into one concept: group actions.

Let us take another look at Section 1.3, where the symmetries of a square were considered. We enumerated the vertices from 1 to 4, so let us write $A = \{1, 2, 3, 4\}$ for this set of vertices. Any of the symmetries of the square considered in Section 1.3 then gives rise to a permutation of A. For example, the counter clockwise rotation over 90 degrees gives rise to the permutation $(1\,2\,3\,4) \in S_4$, while the reflection in the x-axis gives rise to the permutation $(2\,4)$. Therefore we can in this way construct a map from D_4, the symmetries of the square, to S_4, the permutations of $\{frm[o]--, 2, 3, 4\}$. This idea gives rise to following definition:

Definition 65 *Let a group $(G, \cdot)$ and a set A be given. A group action of G on A is a map $\varphi : G \to S_A$, such that*

1. *$\varphi(e) = \mathrm{id}$, where $e \in G$ is the identity element of G and $\mathrm{id} \in S_A$ is the identity permutation on A.*
2. *$\varphi(f \cdot g) = \varphi(f) \circ \varphi(g)$ for all $f, g \in G$.*

One also says that the group acts on the set A. It is a bit cumbersome to write $\varphi(f)$, especially if one wants to evaluate the permutation $\varphi(f)$ in an element of A so we will often write φ_f instead of $\varphi(f)$. Then a group action of G on A satisfies

$$\varphi_e = \mathrm{id} \quad \text{and} \quad \varphi_{f \cdot g} = \varphi_f \circ \varphi_g.$$

From the two properties of a group action, we can derive the following:

Lemma 66 *Let a group $(G, \cdot)$, A a set and $\varphi : G \to S_A$ a group action. Then for any $f \in G$ we have $\varphi_{f^{-1}} = \varphi_f^{-1}$.*

Proof. Since $\varphi_{f \cdot g} = \varphi_f \circ \varphi_g$ for any $f, g \in G$ and $\varphi_e = \mathrm{id}$, we have that

$$\mathrm{id} = \varphi_e = \varphi_{f \cdot f^{-1}} = \varphi_f \circ \varphi_{f^{-1}}$$

and similarly

$$\mathrm{id} = \varphi_e = \varphi_{f^{-1} \cdot f} = \varphi_{f^{-1}} \circ \varphi_f.$$

This shows that $\varphi_{f^{-1}} = \varphi_f^{-1}$, which is what we wanted to show. ■

Example 67 For any matrix $M \in \mathrm{GL}(2, \mathbb{R})$, the set of invertible 2 by 2 matrices with coefficients in $\mathbb{R}$, we can obtain a bijection φ_M from $\mathbb{R}^2$ to itself by defining $\varphi_M(x) := M \cdot x$. Here $M \cdot x$ denotes the usual product of a matrix M and a column vector x. As usual we denote by I the identity matrix. The resulting map $\varphi : \mathrm{GL}(2, \mathbb{R}) \to S_{\mathbb{R}^2}$ is in fact a group action. In the first place, φ_I fixes every element of $\mathbb{R}^2$ (and hence is the identity permutation on $\mathbb{R}^2$), since $\varphi_I[x] = I \cdot x = x$ for all $x \in \mathbb{R}^2$. In the second place it holds for any $x \in \mathbb{R}^2$ that

$$\varphi_{MN}[x] = (MN) \cdot x = M \cdot (N \cdot x) = M \cdot \varphi_N[x] = \varphi_M[\varphi_N[x]] = (\varphi_M \circ \varphi_N)[x],$$

which means that $\varphi_{MN} = \varphi_M \circ \varphi_N$.

Example 68 The symmetric group $(S_n, \circ)$ acts on the set $\{1, 2, \ldots, n\}$ by defining $\varphi_f := f$. In other words, we just define $\varphi : S_n \to S_n$ as $\varphi(f) = f$.

Example 69 The group of rotational symmetries of the dodecahedron acts on the set of five inscribed cubes mentioned in Section 2.4.

Example 70 The rotation symmetry group of a cube acts on the set of 4 diagonals of the cube. It also acts on the set of 12 edges, the set of 6 faces, or the set of 8 vertices.

4.2 Orbits and stabilizers

Keyword: orbit, stabilizer, orbit-stabilizer theorem The notion of a group action gives rise to a wealth of mathematical applications. Group actions connect the abstract notion of a group to the more concrete permutation groups we started out with. Also properties of regular geometric objects can be investigated using the action of its group of symmetries on (parts of) the geometric object itself. This gives for example rise to applications of group theory in the study of crystals.

To key notions concerning group actions in general are the following:

Definition 71 *Let $(G, \cdot)$ be a group and $\varphi : G \to S_A$ an action on a set A. Then we define*

$$O_a := \{\varphi_g[a] \mid g \in G\}$$

the orbit of $a \in A$ and

$$G_a := \{g \in G \mid \varphi_g[a] = a\},$$

the stabilizer of $a \in A$.

Example 72 Consider the action of the rotation symmetry group of the cube on the set of the eight vertices of the cube. Then there will be only one orbit and the stabilizer of a vertex is a group of order three generated by the rotation with axis through the vertex and its antipodal vertex.

Example 73 A more abstract example of a group action is given by the following: Let $(G, \cdot)$ be a group and $H \subset G$ a subgroup. Further let $A = G$. Then H acts on the set A by defining $\varphi_h[f] = f \cdot h$. In this case the orbit of an element f is the left coset $f \cdot H$, since

$$O_f = \{\varphi_h[f] \,|\, h \in H\} = \{f \cdot h \,|\, h \in H\} = f \cdot H.$$

The stabilizer of f consists of the identity element only, since $f \cdot h = f$ implies $h = e$.

From these examples, one may expect that two orbits, just like two left cosets of a subgroup, are either identical or disjoint. This is indeed the case.

Proposition 74 *Let a group $(G, \cdot)$, A a set and $\varphi : G \to S_A$ a group action. Then*

1. *For any $a \in A$ we have $a \in O_a$.*
2. *The set A is covered by equivalence classes: $\cup_{a \in A} O_a = A$.*
3. *For any $a, b \in A$ we have $O_a \cap O_b = \emptyset$ or $O_a = O_b$.*
4. *For any $a, b \in A$ we have $O_a = O_b$ if and only if there exists $f \in G$ such that $a = \varphi_f[b]$.*

Proof. We consider the relation $\sim$ on A defined by $a \sim b$ if and only if there exists $f \in G$ such that $a = \varphi_f[b]$. We claim that $\sim$ is an equivalence relation. Indeed $a \sim a$ is clear, since $\varphi_e[a] = a$. Further if $a \sim b$, then for some $f \in G$ we have $a = \varphi_f[b]$. Using Lemma 66, we obtain that $b = \varphi_{f^{-1}}[a]$ implying that $b \sim a$. Finally, if $a \sim b$ and $b \sim c$, that is to say if $a = \varphi_f[b]$ and $b = \varphi_g[c]$ for certain $f, g \in G$, then we have

$$a = \varphi_f[b] = \varphi_f[\varphi_g[c]] = (\varphi_f \circ \varphi_g)[c] = \varphi_{f \cdot g}[c],$$

where in the last equality we used the second defining property of a group action. Now the proposition follows from Theorem 49. ■

This establishes the main facts on orbits. Now we study stabilizers.

Lemma 75 *Let $\varphi : G \to S_A$ be a group action of a group $(G, \cdot)$ on a set A. Further let $a \in A$. Then G_a is a subgroup of G. If $b \in O_a$ and $f \in G$ satisfies $\varphi_f[a] = b$, then the left coset $f \cdot G_a$ consists precisely of those elements $g \in G$ such that $\varphi_g[a] = b$, that is to say*

$$f \cdot G_a = \{g \in G \,|\, \varphi_g[a] = b\}.$$

Proof. First we show that G_a is a subgroup. Since $\varphi_e = \mathrm{id}$, we have $e \in G_a$. Also, if $f \in G_a$ (that is to say if $\varphi_f[a] = a$), we see by Lemma 66 that $\varphi_{f^{-1}}[a] = a$. Hence we have shown that $f \in G_a$ implies that $f^{-1} \in G_a$. Finally, if $f, g \in G_a$, then $f \cdot g \in G_a$, since $\varphi_{f \cdot g}[a] = \varphi_f[\varphi_g[a]] = \varphi_f[a] = a$.

Now we show the statement $f \cdot G_a = \{g \in G \,|\, \varphi_g[a] = b\}$. First of all if $g \in f \cdot G_a$, then $g = f \cdot h$ for some $h \in G_a$. Then we see that $\varphi_g[a] = \varphi_{f \cdot h}[a] = \varphi_f[\varphi_h[a]] = \varphi_f[a] = b$.

Conversely, if $g \in G$ and $\varphi_g[a] = b$, we claim that $f^{-1} \cdot g \in G_a$. Indeed this follows, since

$$\varphi_{f^{-1} \cdot g}[a] = \varphi_{f^{-1}}[\varphi_g(a)] = \varphi_{f^{-1}}[b] = \varphi_f^{-1}[b] = a.$$

Now that we know that $f^{-1} \cdot g \in G_a$, we see that $g = e \cdot g = (f \cdot f^{-1}) \cdot g = f \cdot (f^{-1} \cdot g)$ and hence $g \in f \cdot G_a$. ■

Before continuing to develop the general theory for orbits and stabilizers, we consider one more example.

Example 76 We have remarked before that the group of symmetries of the cube acts on the set of eight vertices of the cube. To be able to distinguish the vertices from one another, we introduce a coordinate system such that the vertices of the cube are the eight points $(\pm 1, \pm 1, \pm 1)$. Instead of looking at the action of the whole group on the set of vertices of the cube, we will look at the action of a subgroup on the set of vertices. Let us choose the diagonal of the cube connecting $(1, 1, 1)$ and $(-1, -1, -1)$ and denote by r the rotation over $2\pi/3$ radians with this diagonal as rotation axis. Then r has order three and the subgroup G generated by r is equal to $G = \{e, r, r^2\}$. This groups acts on the set of 8 vertices of the cube. There will be four different orbits in this case, namely $\{(1, 1, 1)\}$ and $\{(-1, -1, -1)\}$ (which are kept fixed by all elements in G), $\{(1, 1, -1), (1, -1, 1), (-1, 1, 1)\}$ and $\{(-1, -1, 1), (-1, 1, -1), (1, -1, -1)\}$. A stabilizer of a vertex is $\{e\}$, except for the vertices $(1, 1, 1)$ and $(-1, -1, -1)$ in which cases the stabilizer is the entire group G.

In the above example we see that the larger an orbit is, the smaller the stabilizer of an element from that orbit is. This turns out to be true in general in the following sense:

Theorem 77 *Let $(G, \cdot)$ be a group and suppose that $\varphi : G \to S_A$ is a group action of G on a set A. Then for any $a \in A$ we have*

$$[G : G_a] = |O_a|.$$

In case G is a finite group, we have

$$|G| = |G_a| \cdot |O_a|.$$

Proof. We define a map $M : G \to O_a$ by $M(f) := \varphi_f[a]$. By definition of an orbit, this map is surjective. Given $b \in O_a$, denote by f an element of G such that $\varphi_f[a] = b$. Then by Lemma 75, we see that $f \cdot G_a$ consists exactly of those group elements g such that $\varphi_g[a] = b$. This means that $f \cdot G_a$ consists exactly of those group elements g such that $M(g) = b$. All in all, we see that $[G : G_a]$, the number of left cosets of G_a in G, is exactly the same as the number of elements in O_a.

Finally, if G is a finite group, then $[G : G_a] = |G|/|G_a|$, implying that $|G| = |G_a| \cdot |O_a|$. ■

The above result is known as the orbit-stabilizer theorem. It is very useful in the next section where we will use it to solve several combinatorial problems (problems involving counting the number of certain structures). Group actions do not only occur in applications of group theory, but also in the abstract setting. We consider one example:

Example 78 Let a group $(G, \cdot)$ and two group elements $f, g \in G$ be given. We say that the element fgf^{-1} is a conjugate of g. A group $(G, \cdot)$ can act on itself by conjugation. More precisely, we can define a group action $\varphi : G \to S_G$ as $\varphi_f[g] := fgf^{-1}$.

4.3 Burnside's lemma

Keyword: number of orbits: Burnside's lemma

Now we will use the theory of group actions to solve the following problem: We take a cube and can give each of the six sides of the cube a colour of our choice. We can choose between 2 colours. A priori there are 2^6 possible colourings, since we can choose one of the two colours for every side. However, now we say that two colourings are the same if one can be obtained from the other by a rotation symmetry of the cube. How many distinct colourings does the cube have?

To connect this problem to group actions, we choose A to be the set of all 2^6 possible colourings. This means that within A we consider colourings to be distinct even if a rotation symmetry transforms one colouring into the other. The group of rotation symmetries acts on the set A. The point with this is that the to-be-identified colourings then exactly lie in the same orbit! So to solve the question, what we really want to know is the number of orbits that A has under this action. In this section we will derive a formula that is quite useful to count the number of orbits, hereby solving this and similar problems. This formula is often called Burnside's lemma.

Theorem 79 (Burnside's lemma) *Let $(G,\cdot)$ be a finite group and $\varphi : G \to S_A$ a group action on a finite set A. Define*

$$\mathrm{Fix}(g) := \{a \in A \,|\, \varphi_g[a] = a\}.$$

Then the number of distinct orbits is equal to:

$$\frac{1}{|G|}\sum_{g\in G} |\mathrm{Fix}(g)|.$$

Proof. The proof is based on counting the number of elements in the set

$$M := \{(g,a) \in G \times A \,|\, \varphi_g[a] = a\}$$

in two different ways. If $b \in O_a$ (that is to say, if $b = \varphi_f[a]$ for some $f \in G$), then by Proposition 74 we obtain that $O_a = O_b$. By Theorem 77, we may conclude that $|G_a| = |G_b|$. Therefore the cardinality of G_b does not vary if b varies within an orbit O_a. Now let us enumerate all distinct orbits by $O_{a_1}, \ldots, O_{a_n}$ and denote by $a_1, \ldots, a_n$ elements from these distinct orbits. Then we find

$$|M| = \sum_{a\in A} |G_a| = \sum_{i=1}^{n} |O_{a_i}| \cdot |G_{a_i}| = n|G|.$$

In the last equality we used Theorem 77.

Now we count $|M|$ in another way. Given $g \in G$, there are $|\mathrm{Fix}(g)|$ possibilities for $a \in A$ satisfying $\varphi_g[a] = a$. Therefore

$$|M| = \sum_{g\in G} |\mathrm{Fix}(g)|.$$

The theorem now follows. ■

Example 80 We consider again the possible colourings of the six sides of a cube by two colours. The group of 24 rotation symmetries acts on the set of 2^6 colourings of the cube. We count the number of distinct colourings by computing the number of orbits. There are in total 24 rotational symmetries, namely:

1. The identity symmetry e. It fixes all 2^6 colourings.
2. Rotations with rotation axis through the midpoints of opposite sides and rotation angle $\pi/2$ or $-\pi/2$. There are 6 such rotations. Each of them fixes 2^3 colourings.
3. Rotations with rotation axis through the midpoints of opposite sides and rotation angle π. There are 3 such rotations. Each of them fixes 2^4 colourings.
4. Rotations with rotation axis through the midpoints of opposite edges and rotation angle π. There are 6 such rotations. Each of them fixes 2^3 colourings.
5. Rotations with rotation axis through opposite vertices and rotation angle $2\pi/3$ or $-2\pi/3$. There are 8 such rotations. Each of them fixes 2^2 colourings.

Using Theorem 79 we can compute the total number of distinct colourings as follows:

$$\frac{1}{24}\left(2^6 + 6 \cdot 2^3 + 3 \cdot 2^4 + 6 \cdot 2^3 + 8 \cdot 2^2\right) = 10.$$

4.4 Exercises

1. In Section 1.3 a description of the symmetries of a square was given by describing the corresponding permutations of the 4 vertices of the square. More formally this can be described by a map φ from D_4 to S_4. It was mentioned in the text that this map φ is a group action. Check that this indeed is true.
2. Show that the conjugation action from Example 78 indeed is a group action.
3. Consider the set of necklaces consisting of six beads. The beads all have the same shape, but can each have one out of three colours. To necklaces are considered to be the same, if one can be obtained from another using a rotation (giving rise to a cyclic shift of the six beads). How many different necklaces are there if all possible colourings are considered?
4. Determine the possible of distinct colourings of the vertices of the tetrahedron if there are 4 colours to choose from and if two colourings are counted as the same if a rotational symmetry maps one colouring to the other.
5. Determine how many different colourings of the vertices a pentagon has using n colours. Two colourings are identified with each other if an element from the dihedral group $(D_5, \cdot)$ maps one colouring to the other.

Chapter 5

Maps between groups

5.1 Group homomorphisms

Keyword: group homomorphism, group isomorphism, kernel The groups $(\{0,1,2\},+_3)$ and $(\{\mathrm{id},(1\,2\,3),(1\,3\,2)\},\circ)$ at first sight looks very different. The first one has addition modulo three as group operation, while the second one has composition as group operation. In the first group (let us write $G_1 := \{0,1,2\}$ for now), the group elements are the numbers 0, 1 and 2, while in the second group (let us write $G_2 = \{\mathrm{id},(1\,2\,3),(1\,3\,2)\}$), the group elements are the permutations $\mathrm{id},(1\,2\,3)$ and $(1\,3\,2)$. Seen from a different point of view, there are similarities. Both group have order three and both groups are cyclic, since 1 generates the first group and $(1\,2\,3)$ generates the second group. Even stronger similarities exist when we start looking at the tables describing the group operators $+_3$ and $\circ$:

$+_3$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

$\circ$	id	$(1\,2\,3)$	$(1\,3\,2)$
id	id	$(1\,2\,3)$	$(1\,3\,2)$
$(1\,2\,3)$	$(1\,2\,3)$	$(1\,3\,2)$	id
$(1\,3\,2)$	$(1\,3\,2)$	id	$(1\,2\,3)$

Based on the similarity of these tables, one can define a function $\psi : G_1 \to G_2$ defined by $\psi(0) = \mathrm{id}$, $\psi(1) = (1\,2\,3)$ and $\psi(2) = (1\,3\,2)$ expressing the similarity between $(G_1,+_3)$ and $(G_2,\circ)$. The function ψ has for example the property that $\psi(f +_3 g) = \psi(f) \circ \psi(g)$ for any $f,g \in G_1$. Moreover, $\psi(0) = \mathrm{id}$.

Based on this example, we introduce the notion of a group homomorphism and a group isomorphism:

Definition 81 *Let $(G_1,\cdot_1)$ and $(G_2,\cdot_2)$ be two groups. A function $\psi : G_1 \to G_2$ is called a group homomorphism, if it satisfies*

1. *$\psi(e_1) = e_2$, with e_1 the identity element of G_1 and e_2 the identity element of G_2,*
2. *$\psi(f \cdot_1 g) = \psi(f) \cdot_2 \psi(g)$.*

If $\psi : G_1 \to G_2$ is bijective as well (that is to say both injective and surjective), it is called a group isomorphism.

Example 82 Let $(G_1,\cdot_1) := (\{0,1,2\},+_3)$ and $(G_2,\cdot_2) := (\{\mathrm{id},(1\,2\,3),(1\,3\,2)\},\circ)$. Define as before $\psi : \{0,1,2\} \to \{\mathrm{id},(1\,2\,3),(1\,3\,2)\}$ as $\psi(0) = \mathrm{id}$, $\psi(1) = (1\,2\,3)$ and $\psi(2) = (1\,3\,2)$. Then ψ is a group homomorphism. In fact it is a group isomorphism, since ψ is a bijection.

Another example we already discussed in Chapter 1:

Example 83 The sign map $\text{sign} : S_n \to \{1, -1\}$ is a group homomorphism if we take as group operation on the set $\{1, -1\}$ the usual multiplication and identity element 1. Indeed, the fact that sign is a group homomorphism follows mainly from Equation (1.4), where we saw that $\text{sign}(f \circ g) = \text{sign}(f)\text{sign}(g)$. We also need to check that $\text{sign}(\text{id}) = 1$, but this follows directly from the definition of the sign-function. The sign is not a bijection, unless $n = 2$. Therefore for $n \neq 2$, the sign function is a group homomorphism, but not a group isomorphism.

Yet another example comes from linear algebra:

Example 84 We denote by $\text{GL}(n, \mathbb{R})$ the set of invertible $n \times n$ matrices with coefficients in $\mathbb{R}$. Let $\det : \text{GL}(n, \mathbb{R}) \to \mathbb{R}\backslash\{0\}$ be the determinant map. If we see $(\mathbb{R}\backslash\{0\}, \cdot)$ as a group (with $\cdot$ the usual multiplication as group operation and $1 \in \mathbb{R}$ as identity element), then det is a group homomorphism. The reason is that $\det(I) = 1$ (here I denotes the identity matrix) and $\det(A \cdot B) = \det(A)\det(B)$.

If $n = 1$ the group $\text{GL}(1, \mathbb{R})$ just equals $\mathbb{R}\backslash\{0\}$. In this case the determinant gives a bijection and hence is a group isomorphism. If $n > 1$, the determinant function is a group homomorphism, but not a group isomorphism.

Finally, note how similar the definition of a group homomorphism is to our definition of a group action in Definition 65. Indeed any group action is in fact a group homomorphism:

Example 85 Let $(G, \cdot)$ be a group, A a set and $\varphi : G \to S_A$ a group action. Since $s_e = \text{id}_A$ and $\varphi_{f \cdot g} = \varphi_f \circ \varphi_g$ (using the definition of a group action), we see that $\varphi : G \to S_A$ is a group homomorphism.

Lemma 66 can be generalized to any group homomorphism.

Lemma 86 *Let $\psi : G_1 \to G_2$ be a group homomorphism. Then $\psi(f^{-1}) = \psi(f)^{-1}$.*

Proof. The definition of a group homomorphism implies that

$$e_2 = \psi(e_1) = \psi(f \cdot_1 f^{-1}) = \psi(f) \cdot_2 \psi(f^{-1})$$

and

$$e_2 = \psi(e_1) = \psi(f^{-1} \cdot_1 f) = \psi(f^{-1}) \cdot_2 \psi(f).$$

This implies that $\psi(f^{-1})$ is the inverse of the group element $\psi(f)$. ■

An important concept for a group homomorphism is its so-called kernel:

Definition 87 *Let $\psi : G_1 \to G_2$ be a group homomorphism and denote by e_2 the identity element in G_2. Then we define $\ker \psi$ the kernel of ψ as follows:*

$$\ker \psi := \{g \in G_1 \mid \psi(g) = e_2\}.$$

Note that the kernel of a group homomorphism ψ always contains e_1, since $\psi(e_1) = e_2$. If ψ is a group isomorphism, it is assumed to be injective. Therefore if $\psi(g) = e_2$, then $g = e_1$, since it also holds that $\psi(e_1) = e_2$. This implies that the kernel of a group isomorphism equals $\{e_1\}$.

In Example 83, the kernel of the sign function is A_n, the set of all even permutations.

Theorem 88 *Let $\psi : G_1 \to G_2$ be a group homomorphism. Then $\ker \psi \subset G_1$ is a subgroup of $(G_1, \cdot_1)$. Moreover, for any $f \in G_1$ it holds that $f \cdot_1 (\ker \psi) = (\ker \psi) \cdot_1 f$.*

Proof. First we show that $\ker\psi$ is a subgroup of G_1:

First of all we need to check that $e_1 \in \ker\psi$, but it follows directly from the definition of a group homomorphism that $\psi(e_1) = e_2$.

Next, we need to show that if $f \in \ker\psi$, then also $f^{-1} \in \ker\psi$. However, we have seen that $\psi(f^{-1}) = \psi(f)^{-1}$. Therefore, if $f \in \ker\psi$ (that is to say, if $\psi(f) = e_2$), then

$$\psi(f^{-1}) = \psi(f)^{-1} = e_2^{-1} = e_2.$$

This implies that $f^{-1} \in \ker\psi$.

Finally if $f, g \in \ker\psi$, then we know that $\psi(f) = e_2$ and $\psi(g) = e_2$. Therefore we have in this case that

$$\psi(f \cdot_1 g) = \psi(f) \cdot_2 \psi(g) = e_2 \cdot_2 e_2 = e_2.$$

In other words, $f \cdot_1 g \in \ker\psi$. This finishes the proof of the claim that $\ker\psi \subset G_1$ is a subgroup of G_1.

Now we show that for any $f \in G_1$ we have $f \cdot_1 (\ker\psi) = (\ker\psi) \cdot_1 f$.

If $g \in f \cdot_1 (\ker\psi)$, then there exists $h \in \ker\psi$ such that $g = f \cdot_1 h = (f \cdot_1 h \cdot_1 f^{-1}) \cdot_1 f$. However, since $h \in \ker\psi$ and ψ is a homomorphism, we have

$$\psi(f \cdot_1 h \cdot_1 f^{-1}) = \psi(f) \cdot_2 \psi(h) \cdot_2 \psi(f^{-1}) = \psi(f) \cdot_2 e_2 \cdot_2 \psi(f)^{-1} = e_2.$$

Therefore, the element $f \cdot_1 h \cdot_1 f^{-1}$ is in the kernel of ψ. This implies that $g = (f \cdot_1 h \cdot_1 f^{-1}) \cdot_1 f \in (\ker\psi) \cdot_1 f$.
Conversely, if $g \in (\ker\psi) \cdot_1 f$, then there exists $h \in \ker\psi$ such that $g = h \cdot_1 f = f \cdot_1 (f^{-1} \cdot_1 h \cdot_1 f)$. A very similar reasoning as above, shows that the element $f^{-1} \cdot_1 h \cdot_1 f$ is in the kernel of ψ. Then $g = f \cdot_1 (f^{-1} \cdot_1 h \cdot_1 f) \in f \cdot_1 (\ker\psi)$.
Combining the above, we see that $f \cdot_1 (\ker\psi) = (\ker\psi) \cdot_1 f$. ■

A subgroup H of a group $(G, \cdot)$ is called *normal*, if $f \cdot H = H \cdot f$ for all $f \in G$. The above theorem states in words that the kernel of a group homomorphism always is a normal subgroup. We have already seen one example when we looked at the sign-function. Here is another example:

Example 89 Let A be a 2×3 matrix with coefficients in $\mathbb{R}$. Then the map $\psi_A : \mathbb{R}^3 \to \mathbb{R}^2$ defined by $\psi_A(x) = A \cdot x$ is a group homomorphism from the group $(\mathbb{R}^3, +)$ to the group $(\mathbb{R}^2, +)$. Indeed, since ψ_A is a linear map, we have $\psi_A(0) = A \cdot 0 = 0$ and $\psi_A(x + y) = A \cdot (x + y) = A \cdot x + A \cdot y = \psi_A(x) + \psi_A(y)$.

The kernel of ψ_A consists of those vectors $x \in \mathbb{R}^3$ such that $A \cdot x = 0$, that is to say that $\ker\psi_A$ is equal to the null space of A.

Another example of a group homomorphism can be obtained using modular arithmetic:

Example 90 Given a natural number n and an integer a, then using division with remainder, one can find uniquely determined integers q and r such that $a = qn + r$ and $0 \leq r < n$. The integer r is called the remainder of the division of a by n and is denoted by $a \mod n$ as we did before. The operation $+_n$ occurring in the group $(\mathbb{Z} \mod n, +_n)$ is as before explicitly defined by: $a +_n b := (a + b) \mod n$.

The map

$$\psi : \mathbb{Z} \to \mathbb{Z} \mod n \text{ defined by } \psi(a) = a \mod n$$

is a group homomorphism from $(\mathbb{Z}, +)$ to $(\mathbb{Z} \mod n, +_n)$. To see this we first need to check that $\psi(0) = 0$, which is true since $\psi(0) = 0 \mod n = 0$. Next we need to check that $\psi(a + b) = \psi(a) +_n \psi(b)$. This amounts to showing that

$$(a + b) \mod n = ((a \mod n) + (b \mod n)) \mod n.$$

We know that $a = q_1 n + a \mod n$ and $b = q_2 n + b \mod n$, for certain integers q_1 and q_2. Therefore we have $a + b = (q_1 + q_2)n + (a \mod n + b \mod n)$, while $a + b = qn + (a + b) \mod n$ for a certain integer q.

This does not imply that $a + b \mod n = (a \mod n + b \mod n)$, since it may happen that $(a \mod n + b \mod n) \geq n$. However, we do have that $(a \mod n + b \mod n) \mod n$ and $(a + b) \mod n$ are the same modulo n and both between 0 and $n - 1$. Since the remainder after division by n is unique, se see that $(a + b) \mod n = ((a \mod n) + (b \mod n)) \mod n$, which is what we wanted to show.

All in all we may conclude that ψ is a group homomorphism.

The kernel of the group homomorphism ψ is given by

$$\ker \psi = \{a \in \mathbb{Z} \mid \psi(a) = 0\},$$

which equals the set of all those integers that are a multiple of n. Therefore, $\ker \psi = n\mathbb{Z}$.

5.2 Quotient groups

Keywords: quotient groups

In this section we will show that a normal subgroup $H \subset G$ of a group $(G, \cdot)$ can be used to construct a new group denoted by G/H. The elements of this group will be the cosets of the subgroup H in G. Since for a normal subgroup left and right cosets are the same, we can indeed speak about the cosets of H without specifying whether we mean left cosets or right cosets.

We will start by defining precisely what G/H is as a set and afterwards define a group operation on G/H.

Definition 91 *Let $(G, \cdot)$ be a group and H a normal subgroup of G. Then we define G/H to be the set consisting of all cosets of H in G.*

Now that we have defined G/H as a set, we would like to introduce a group operation on it. Using Definition 50 we can define a multiplication of two cosets. We simply define

$$(fH) \cdot (gH) := \{k \cdot \ell \mid k \in fH, \ell \in gH\}.$$

However, it is not at all clear that the resulting set is a left coset of H again and indeed if H is not a normal subgroup of $(G\cdot)$ it turns out that it will not be in general. It turns out that for normal subgroups, the situation is different. First recall that $fH = \{f\} \cdot H$ by Definition 52. If H is a normal subgroup, we obtain

$$(fH) \cdot (gH) = (\{f\} \cdot H) \cdot (\{g\} \cdot H) = \{f\} \cdot (H \cdot \{g\}) \cdot H = \{f\} \cdot (Hg) \cdot H,$$

which using that H is a normal subgroup can be simplified further:

$$\{f\} \cdot (Hg) \cdot H = \{f\} \cdot (gH) \cdot H = \{f\} \cdot (\{g\} \cdot H) \cdot H = (\{f\} \cdot \{g\}) \cdot (H \cdot H) = \{fg\} \cdot H = (f \cdot g)H.$$

Here we used that $H \cdot H = H$, which follows from the fact that H is a subgroup of G. Using the above, we have motivated the following definition of multiplication of cosets:

Definition 92 *Let $(G, \cdot)$ be a group and H a normal subgroup of $(G, \cdot)$. Let $C, D \in G/H$ be two cosets of H in G. Suppose that $f \in C$ and $g \in D$. Then we define $C \cdot D := (f \cdot g)H$. Note that this coset is independent on the choice of f and g.*

If C is a left coset of a normal subgroup H and $f \in C$, then f is called a representative of the coset C. The definition can be restated as: If f is a representative of a coset C and g a representative of a coset D, then $C \cdot D$ is again a coset of H and $f \cdot g$ is a representative for it. Definition 92 gives us a way to define a group operation on the set G/H.

Definition 93 *Let $(G, \cdot)$ be a group and let H be a normal subgroup of G and denote the set of cosets of H in G by G/H. Then multiplication of cosets $(fH) \cdot (gH) := (f \cdot g)H$ gives $(G/H, \cdot)$ the structure of a group. This group is called the quotient group of G by H.*

For completeness, let us check if the group axioms are satisfied:

1. For any $f, g, h \in G$ be have

$$(fH \cdot gH) \cdot hH = (f \cdot g)H \cdot h \cdot H = ((f \cdot g) \cdot h)H,$$

   while

$$fH \cdot (gH \cdot hH) = fH \cdot (g \cdot h)H = (f \cdot (g \cdot h))H.$$

   Therefore, the associativity of the multiplication of cosets is a consequence of the associativity of the original group operation on elements of G.

2. The identity element of G/H is $eH = H$, since $eH \cdot fH = e \cdot fH = fH$ and $fH \cdot eH = f \cdot eH = fH$ for any coset fH of H.

3. Given a coset fH, we have $fH \cdot f^{-1}H = f \cdot f^{-1}H = eH = H$ and $f^{-1}H \cdot fH = f^{-1} \cdot fH = eH = H$. Therefore we have $(fH)^{-1} = f^{-1}H$.

We see that indeed $(G/H, \cdot)$ with the operation $\cdot$ defined in Definition 93 is a group.

Example 94 Define as before $(S_n, \circ)$ to be the group of permutations on n elements. We assume that $n \geq 2$. The subgroup A_n was defined as the subgroup consisting of all even permutations, that is to say all permutations with sign 1. We have seen the A_n is a normal subgroup of S_n of index two.

The quotient group $(S_n/A_n, \circ)$ is therefore a group with two elements: A_n and fA_n for a suitably chosen $f \in S_n$. In fact f can be any odd permutation, for example $f = (12)$.

A multiplication table for $S_n/A_n = \{A_n, (12)A_n\}$ is the following:

$\circ$	A_n	$(12)A_n$
A_n	A_n	$(12)A_n$
$(12)A_n$	$(12)A_n$	A_n

This is a similar multiplication table as the table for the group $(\{1, -1\}, \cdot)$:

$\cdot$	1	-1
1	1	-1
-1	-1	1

We see that the groups $(S_n/A_n, \circ)$ and $(\{1, -1\}, \cdot)$ are isomorphic and that an isomorphism ψ is given by defining $\psi(A_n) = 1$ and $\psi((12)A_n) = -1$.

Example 95 Define as before $n\mathbb{Z} := \{nk \mid k \in \mathbb{Z}\}$, that is to say $n\mathbb{Z}$ is the set of all multiples of n. Then $n\mathbb{Z}$ is a subgroup of $\mathbb{Z}$ with addition as group operation. Since $\mathbb{Z}$ is an abelian group, any subgroup (and in particular $n\mathbb{Z}$) is a normal subgroup. The cosets of $n\mathbb{Z}$ are the sets $a + n\mathbb{Z} := \{a + nk \mid k \in \mathbb{Z}\}$. In principle a can be any element of $\mathbb{Z}$, but since $a + n\mathbb{Z} = b + n\mathbb{Z}$ if $a - b \in n\mathbb{Z}$, we already can describe all possible cosets of $n\mathbb{Z}$ by choosing a between 0 and $n - 1$. Therefore the group $\mathbb{Z}/n\mathbb{Z}$ has n elements.

For example if $n = 3$, we have $3\mathbb{Z} = \{\dots, -6, -3, 0, 3, 6, \dots\}$, $1+3\mathbb{Z} = \{\dots, -5, -2, 1, 4, 7, \dots\}$ and $2 + 3\mathbb{Z} = \{\dots, -4, -1, 2, 5, 8, \dots\}$. The sum of for example the cosets $1 + 3\mathbb{Z}$ and $2 + 3\mathbb{Z}$ is by definition equal to the coset $(1 + 2) + 3\mathbb{Z} = \{\dots, -3, 0, 3, 6, 9, \dots\} = 3\mathbb{Z}$. A full table is given by

$+$	$3\mathbb{Z}$	$1 + 3\mathbb{Z}$	$2 + 3\mathbb{Z}$
$3\mathbb{Z}$	$3\mathbb{Z}$	$1 + 3\mathbb{Z}$	$2 + 3\mathbb{Z}$
$1 + 3\mathbb{Z}$	$1 + 3\mathbb{Z}$	$2 + 3\mathbb{Z}$	$3\mathbb{Z}$
$2 + 3\mathbb{Z}$	$2 + 3\mathbb{Z}$	$3\mathbb{Z}$	$1 + 3\mathbb{Z}$

which is similar to the table

$+_3$	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

More precisely we will say that the groups $(\mathbb{Z}/3\mathbb{Z}, +)$ and $(\mathbb{Z} \mod 3, +_3)$ are isomorphic under the isomorphism defined by $\phi(a + 3\mathbb{Z}) = a$ for $a \in \{0, 1, 2\}$. More generally $(\mathbb{Z}/n\mathbb{Z}, +)$ is isomorphic to the group $(\mathbb{Z} \mod n, +_n)$.

5.3 The isomorphism theorem

Keywords: the isomorphism theorem

We have seen that different looking groups may be essentially the same. For example the groups $(\{0, 1, 2\}, +_3)$ and $(\{e, (123), (132)\}, \circ)$ are isomorphic. Also we have seen the the quotient group $(S_n/A_n, \circ)$ is isomorphic with the group $(\{1, -1\}, \cdot)$. These kind of identifications using group homomorphisms can be very useful to gain insight in the structure of a given group. For example, when defining a group action $s : G \to S_A$, one essentially defines a group homomorphism between an abstract group $(G, \cdot)$ and the more down to earth permutation group $(S_A, \circ)$. In this section we will obtain a general result that can be used to find relations between groups.

Theorem 96 *Let $\psi : G_1 \to G_2$ be a homomorphism of groups. Then the map $\overline{\psi} : G_1/\ker\psi \to \psi(G_1)$ defined by $\overline{\psi}(g \ker\psi) = \psi(g)$ is a group isomorphism.*

Proof. First we show that $\overline{\psi}$ is well defined: if $f \ker\psi = g \ker\psi$, then we need to show that $\overline{\psi}(f \ker\psi) = \overline{\psi}(g \ker\psi)$, or in other words that $\psi(f) = \psi(g)$.

Assuming that $f \ker\psi = g \ker\psi$, then $f^{-1}g \in \ker\psi$. Therefore, we find that $\psi(f)^{-1}\psi(g) = \psi(f^{-1}g) = e_2$, or in other words that $\psi(f) = \psi(g)$. This is exactly what we needed to show.

In order to show that $\overline{\psi} : G_1/\ker\psi \to \psi(G_1)$ is a group isomorphism, we need to show three things: 1) it is a group homomorphism, 2) it is surjective, 3) it is injective. To show 1), note that $\overline{\psi}(\ker\psi) = e_2$ and

$$\overline{\psi}((f \ker\psi)(g \ker\psi)) = \overline{\psi}((fg) \ker\psi) = \psi(fg) = \psi(f)\psi(g) = \overline{\psi}(f \ker\psi)\overline{\psi}(g \ker\psi).$$

As for 2), since the image of $\overline{\psi}$ is the same as that of ψ, it is surjective onto $\psi(G_1)$. Now we only need to show 3), namely that $\overline{\psi}$ is injective. Suppose that $\overline{\psi}(f \ker \psi) = \overline{\psi}(g \ker \psi)$, then $\psi(f) = \psi(g)$. This implies that $\psi(f^{-1}g) = e_2$ and hence that $f^{-1}g \in \ker \psi$. This implies that $f \ker \psi = g \ker \psi$, which is what we wanted to show. ■

Example 97 As we have seen in Example 94 we can for $n \geq 2$ identify the groups $(S_n/A_n, \circ)$ and $(\{1, -1\}, \cdot)$. This can also be seen using Theorem 96 as follows. We start with the map

$$\text{sign} : S_n \to \{1, -1\}$$

from Section 1.4. First of all, we see that sign is a group homomorphism by Equation (1.4) and the fact that $\text{sign}(\text{id}) = 1$. The kernel of sign consists exactly of all even permutations and therefore $\ker \text{sign} = A_n$. Finally, $\text{sign}(S_n) = \{1, -1\}$ if $n \geq 2$, since $\text{sign}(\text{id}) = 1$ and $\text{sign}(\,(12)\,) = -1$. Theorem 96 now implies that the groups $(S_n/A_n, \circ)$ and $(\{1, -1\}, \cdot)$ are isomorphic.

Example 98 The map $\det : \text{GL}(2, \mathbb{R}) \to \mathbb{R}\backslash\{0\}$ is a group homomorphism between the groups $(\text{GL}(2, \mathbb{R}), \cdot)$ and $(\mathbb{R}\backslash\{0\}, \cdot)$, since $\det(I) = 1$ and $\det(AB) = \det(A)\det(B)$. The image of det is $\mathbb{R}\backslash\{0\}$ and its kernel is $\text{SL}(2, \mathbb{R})$ (all matrices in $\text{GL}(2, \mathbb{R})$ with determinant one). Using Theorem 96, we see that the groups $(\text{GL}(2, \mathbb{R})/\text{SL}(2, \mathbb{R}), \cdot)$ and $(\mathbb{R}\backslash\{0\}, \cdot)$ are isomorphic.

Example 99 In Example 90, we saw that the map $\psi : \mathbb{Z} \to \mathbb{Z} \mod n$ defined by $\psi(a) = a \mod n$ is a group homomorphism from $(\mathbb{Z}, +)$ to $(\mathbb{Z} \mod n, +_n)$. The image of ψ is all of the group $\mathbb{Z} \mod n$ and the kernel of ψ is equal to $n\mathbb{Z}$. Therefore Theorem 96 gives that the $(\mathbb{Z}/n\mathbb{Z}, +)$ is isomorphic to $(\mathbb{Z} \mod n, +_n)$.

5.4 Exercises

1. Show that the map $\exp : \mathbb{R} \to \mathbb{R}\backslash\{0\}$ defined by $\exp(r) = e^r$ is group homomorphism from the group $(\mathbb{R}, +)$ to the group $(\mathbb{R}\backslash\{0\}, \cdot)$. Compute the kernel of exp.
2. Show that the map $\psi : \mathbb{R} \to \mathbb{C}\backslash\{0\}$ defined by $\psi(r) = e^{ir}$ is a group homomorphism from the group $(\mathbb{R}, +)$ to the group $(\mathbb{C}\backslash\{0\}, \cdot)$. What is the kernel of ψ?
3. Consider the subgroup $H = \{e, (12)\}$ of S_3. Show that it is not true for all $f, g \in S_3$ that $(fH)(gH) = (fg)H$. Theorem 92 is therefore not true in general if H is not a normal subgroup of G.
4. Given are two group homomorphisms $\psi_1 : G_1 \to G_2$ and $\psi_2 : G_2 \to G_3$. Show that the composition $\psi_2 \circ \psi_1$ is a group homomorphism from G_1 to G_3.
5. Let $(G, \cdot)$ be a group. Show that the quotient group $(G/\{e\}, \cdot)$ is isomorphic to $(G, \cdot)$.
6. Given a group homomorphism $\psi : G_1 \to G_2$. Show that $\psi(G_1)$ is a subgroup of G_2.
7. Work out what the isomorphism theorem implies when considering the group homomorphism from the first two exercises.
8. Let $(G, \cdot)$ be a group. Use the isomorphism theorem to show that the quotient group $(G/\{e\}, \cdot)$ is isomorphic to $(G, \cdot)$.
9. Show that the quotient group $(S_n/A_n, \circ)$ is isomorphic with the group $(\{1, -1\}, \cdot)$.

Chapter 6

Rings

6.1 Definition of a ring

Keywords: rings, commutative rings, units, zero-divisors Up till now we have looked at groups, for example the group $(S_n, \circ)$ of permutations on n elements. Now we turn our attention to structures where there are two operations: rings.

Definition 100 *A ring $(R, +_R, \cdot_R)$ is a set R together with two operations*

$$+_R : R \times R \to R$$

and

$$\cdot_R : R \times R \to R$$

satisfying:

1. *$(R, +_R)$ is an abelian group. Its neutral element is denoted by 0_R.*
2. *There exists a neutral element for the operation $\cdot_R$ denoted by 1_R. It satisfies*

   $$1_R \cdot_R x = x \text{ and } x \cdot_R 1_R = x$$

   for all $x \in R$.
3. *The operation $\cdot_R$ is associative:*

   $$x \cdot_R (y \cdot_R z) = (x \cdot_R y) \cdot_R z$$

   for all $x, y, z \in R$.
4. *The operations $+_R$ and $\cdot_R$ satisfy the distributive laws:*

   $$x \cdot_R (y +_R z) = x \cdot_R y +_R x \cdot_R z$$

   and

   $$(y +_R z) \cdot_R x = y \cdot_R x +_R z \cdot_R x$$

   for all $x, y, z \in R$.

It is not necessary to assume that $(R, +_R)$ is an *abelian* group. However, even if we only assume that $(R, +_R)$ is a group, the other ring axioms imply that the group is abelian. (that is to say that $x +_R y = y +_R x$ for all $x, y \in R$). For let us assume that $(R, +_R)$ is a group, not necessary abelian, but that all other ring axioms are satisfied. Then by the distributive laws we obtain:

$$(1 +_R 1) \cdot_R (x +_R y) = 1 \cdot_R (x +_R y) +_R 1 \cdot_R (x +_R y) = x +_R y +_R x +_R y,$$

but also

$$(1 +_R 1) \cdot_R (x +_R y) = (1 +_R 1) \cdot_R x +_R (1 +_R 1) \cdot_R y = x +_R x +_R y +_R y.$$

Together, these two equations imply that $y +_R x = x +_R y$ as desired.

The ring operation $\cdot_R$ is not necessary commutative (that is to say $x \cdot_R y \neq y \cdot_R x$ in general). If all ring axioms are satisfied and additionally it holds that $x \cdot_R y = y \cdot_R x$ for all $x, y \in R$, then the ring $(R, +_R, \cdot_R)$ is called a *commutative* ring.

Although it is more precise to write $0_R, 1_R, +_R$ and $\cdot_R$, many textbooks drop the index R in the notation and simply write $0, 1, +, \cdot$. The disadvantage of this is that one may confuse the ring elements 0_R, 1_R and ring operations $+_{\mathbb{R}}, \cdot_R$ with the numbers $0, 1$ and usual addition and multiplication. However, the advantage is that it makes equations easier to read and write. Therefore we will often also drop the index R in the notation.

Example 101 The set of integers $\mathbb{Z}$ together with operations $+$ and $\cdot$ (with $+$ and $\cdot$ the usual addition and multiplication of integers) is a ring. We call $(\mathbb{Z}, +, \cdot)$ the ring of integers. Indeed the notation for the operations and the elements 0 and 1 in a general ring is taken from this example. Similarly $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$ and $(\mathbb{C}, +, \cdot)$ are examples of rings, called the ring of rational numbers, the ring of real numbers and the ring of complex numbers. All these rings are commutative.

Example 102 Let $\mathrm{Mat}(n, \mathbb{R})$ be the set of all $n \times n$ matrices with coefficients in $\mathbb{R}$. Denote by $+$ and $\cdot$ the usual matrix addition and multiplication. Then $(\mathrm{Mat}(n, \mathbb{R}), +, \cdot)$ is a ring. It is called the ring of n times n matrices with coefficients in $\mathbb{R}$. For $n > 1$, these rings are not commutative (for $n = 2$ this follows from Exercise 3 in Chapter 2).

Example 103 Let $\mathbb{Z} \mod n$ be the set of integers $\{0, 1, 2, \ldots, n-1\}$ and denote by $+_n$ the addition and $\cdot_n$ the multiplication modulo n. Then $(\mathbb{Z} \mod n, +_n, \cdot_n)$ is a ring. If n is the product of two large prime numbers, this ring is used in the RSA cryptosystem.

Definition 104 *Let $(R, +_R, \cdot_R)$ be a ring. An element $x \in R$ is called a* unit *if there exists $y \in R$ such that $x \cdot_R y = y \cdot_R x = 1_R$. The set of all units in R is denoted by R^*. In other words, we have*

$$R^* := \{x \in R \mid \exists y \in R \;\; x \cdot_R y = y \cdot_R x = 1_R\}.$$

In other words: the set R^* consists of all elements from R having a multiplicative inverse. Just as for groups multiplicative inverses (if they exist) are unique and one writes x^{-1} for this inverse. In fact $(R^*, \cdot_R)$ is a group as we show now:

Lemma 105 *Let $(R, +_R, \cdot_R)$ be a ring. Then $(R^*, \cdot_R)$ is a group with neutral element 1_R.*

Proof. First we need to check that $\cdot_R$ gives rise to an operation on R^*. For this the only thing we need to check is that if $x_1, x_2 \in R^*$, then $x_1 \cdot_R x_2 \in R^*$. Since both x_1 and x_2 are units, they have multiplicative inverses x_1^{-1} and x_2^{-1}. We know from Chapter 2, Exercise 4 b) that $x_2^{-1} \cdot_R x_1^{-1}$ is the multiplicative inverse of $x_1 \cdot_R x_2$. Indeed we have that

$$(x_2^{-1} \cdot_R x_1^{-1}) \cdot_R (x_1 \cdot_R x_2) = x_2^{-1} \cdot_R (x_1^{-1} \cdot_R x_1) \cdot_R x_2 = x_2^{-1} \cdot_R 1_R \cdot_R x_2 = x_2^{-1} \cdot_R x_2 = 1_R$$

and

$$(x_1 \cdot_R x_2) \cdot_R (x_2^{-1} \cdot_R x_1^{-1}) = x_1 \cdot_R (x_2 \cdot_R x_2^{-1}) \cdot_R x_1^{-1} = x_1 \cdot_R 1_R \cdot_R x_1^{-1} = x_1 \cdot_R x_1^{-1} = 1_R.$$

This means that $\cdot_R$ can be seen as an operation on R^*.

Now we need to check that $\cdot_R$ is associative on R^*. However, the operation is already associative on R by the third ring axiom. Since R^* is a subset of R, it is therefore certainly associative on R^*. Similarly, since by the second ring axiom 1_R is a neutral element for the operation $\cdot_R$ on R, it is certainly a neutral element for the operation $\cdot_R$ when it is restricted to R^*.

The last property we need to check is if any element from R^* has an inverse with respect to the operation $\cdot_R$. However, the elements in R^* were exactly chosen to be the elements from R that have such an inverse. ■

Let us look at some examples:

Example 106 Let $(\mathbb{Z}, +, \cdot)$ be the ring of integers. Then we have $\mathbb{Z}^* = \{-1, 1\}$. Indeed assume $x \cdot y = 1$, then by taking absolute values, we see that $|x| = 1/|y|$. This means that if $|x| > 1$, then $|y| < 1$, which would imply $y = 0$, since $y \in \mathbb{Z}$. This is impossible, since $x \cdot 0 = 0$. This means that $|x| = 1$ and hence $x = 1$ or $x = -1$. This shows that $\mathbb{Z}^* \subset \{-1, 1\}$. Conversely, $\mathbb{Z}^* \supset \{-1, 1\}$, since both -1 and 1 are units (they are their own inverses).

Example 107 Let $(\mathbb{Q}, +, \cdot)$ be the ring of rational numbers. Then we have $\mathbb{Q}^* = \mathbb{Q}\backslash\{0\}$. Indeed any rational number number different from 0 has a multiplicative inverse. Similarly $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$ and $\mathbb{C}^* = \mathbb{C}\backslash\{0\}$.

Example 108 Consider the ring $(\mathbb{Z} \mod 6, +_6, \cdot_6)$. The units in this ring are exactly the numbers 1 and 5, that is to say the set of units is what we previously denoted by $(\mathbb{Z} \mod 6)^*$. This notation was in fact inspired by the more general ring-theoretical notation introduced in Definition 104. In general we have

$$(\mathbb{Z} \mod n)^* = \{a \in \mathbb{Z} \mod n \mid \gcd(a, n) = 1\}.$$

Indeed if $\gcd(a, n) = 1$, then there exist integers b and m such that $ab + nm = 1$. This means that $b \cdot_n a = a \cdot_n b = 1$ and hence that a is a unit. Conversely if a is a unit, there exists $b \in \mathbb{Z}$ mod n such that $a \cdot_n b = 1$. But then $a \cdot b$ is 1 plus a multiple of n, so there exists an integer m such that $a \cdot b + n \cdot m = 1$. This implies that $\gcd(a, n) = 1$, since any common divisor of a and n then also needs to divide $a \cdot b + n \cdot m = 1$.

Example 109 Let $(\mathrm{Mat}(n, \mathbb{R}), +, \cdot)$ be the ring of $n \times n$ matrices with coefficients in $\mathbb{R}$. Then we have $\mathrm{Mat}(n, \mathbb{R})^* = \mathrm{GL}(n, \mathbb{R})$, where $\mathrm{GL}(n, \mathbb{R})$ by definition is the set consisting of all invertible n times n matrices.

Units capture a property of invertibility, just as we have seen that in groups. The difference is that in a ring not all elements need to have a multiplicative inverse. In rings a very different phenomenon may also happen: the product of two elements different from zero may be zero. We will see some examples in a moment.

Definition 110 *Let $(R, +_R, \cdot_R)$ be a ring. An element $x \in R$ is called a zero-divisor if $x \neq 0_R$ and if there exists $y \in R$ different from 0_R such that $x \cdot_R y = 0_R$ or $y \cdot_R x = 0_R$.*

Warning: The possible existence of zero-divisors in a ring makes the rule that $x \cdot_R y = 0_R \Rightarrow x = 0_R \vee y = 0_R$ invalid in general. This rule only holds in a ring R that does not contain any zero-divisors.

Example 111 The ring $(\mathbb{Z}, +, \cdot)$ has no zero-divisors. Indeed if $a \neq 0$ and $b \neq 0$, then $a \cdot b \neq 0$. Equivalent to this is the statement: If $a \cdot b = 0$, then either $a = 0$ or $b = 0$.

Example 112 Let $(\mathrm{Mat}(2, \mathbb{R}), +, \cdot)$ be the ring of 2×2 matrices with coefficients in $\mathbb{R}$. Define the matrix

$$A := \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } B := \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

then we have $A \cdot B = 0$. Therefore A (and also B) is a zero-divisor.

Example 113 Consider again the ring $(\mathbb{Z} \mod 6, +_6, \cdot_6)$. The elements $2, 3$ and 4 are all zero-divisors. We have for example that

$$2 \cdot_6 3 = 0, \; 3 \cdot_6 2 = 0 \;\text{ and }\; 4 \cdot_6 3 = 0.$$

In general for a positive integer n in $(\mathbb{Z} \mod n, +_n, \cdot_n)$ all non-zero elements $a \in \mathbb{Z} \mod n$ such that $\gcd(a, n) > 1$ are zero-divisors: If $\gcd(a, n) > 1$, then there exists a prime number p dividing both a and n. Then

$$a \cdot_n (n/p) = a \cdot n/p \mod n = a/p \cdot n \mod n = 0.$$

Conversely if there exist non-zero $a, b \in \mathbb{Z} \mod n$ such that $a \cdot_n b = 0$, then $a \cdot b$ is a multiple of n, while b is not a multiple of n. This means that there exists a prime number p dividing n, but not dividing b. Since $a \cdot b$ is a multiple of n, this implies that p divides a. Therefore $\gcd(a, n) \geq p > 1$.

6.2 Domains and fields

Keywords: (integral) domains, fields, finite fields with p elements (p a prime number)
A commutative ring without zero-divisors is called an *integral domain* (also just called a domain). The typical example is $\mathbb{Z}$, the ring of integers. The point of singling out this special class of rings is that they have the following property in common (note that we started writing $+$ and $\cdot$ and no longer use $+_R$ and $\cdot_R$):

Proposition 114 *Let $(R, +, \cdot)$ be a domain and suppose that $x \cdot y = 0$. Then $x = 0$ or $y = 0$. Consequently, if $x \cdot y = x \cdot z$ and $x \neq 0$, then $y = z$.*

Proof. If $x \cdot y = 0$ and both $x \neq 0$ and $y \neq 0$, then x and y are zero-divisors. However, since R is a domain, it does not have zero-divisors. Therefore $x = 0$ or $y = 0$.

If $x \cdot y = x \cdot z$, then we claim that one can conclude that $x \cdot (y - z) = 0$. Here we have used the notation $-z$ for the additive inverse of z and the notation $y - z$ is short hand for $y + (-z)$. Indeed the equation $x \cdot y = x \cdot z$ implies that

$$x \cdot y + x \cdot (-z) = x \cdot z + x \cdot (-z).$$

Using the distributive law, the left-hand side can be simplified to $x \cdot y + x \cdot (-z) = x \cdot (y + (-z)) = x \cdot (y - z)$, while the right-hand side can be simplified to $x \cdot z + x \cdot (-z) = x \cdot (z + (-z)) = x \cdot 0 = 0$. For the last equality see Exercise 2.

From the equation $x \cdot (y - z) = 0$ we can using the first part of the proposition conclude that $x = 0$ or $y - z = 0$. Since we know that $x \neq 0$, we see that $y - z = 0$ implying that $y = z$. ■

It is tempting to try to prove the second part of the proposition by dividing with x. What this really would mean is that one multiplies with x^{-1} on the left on both sides of the equality sign in the equation $x \cdot y = x \cdot z$. However, this makes sense only if x^{-1} exists (in other words if x is a unit), which does not have to be the case. Because of Proposition 114, domains are said to satisfy the *cancelation law* (one can cancel the x in the equation $x \cdot y = x \cdot z$ if $x \neq 0$).

Example 115 The rings $(\mathbb{R}, +, \cdot)$ of real numbers is a domain. Indeed, since any non-zero element has an inverse the equation $x \cdot y = 0$ implies that either $x = 0$ or

$$y = (x^{-1} \cdot x) \cdot y = x^{-1} \cdot (x \cdot y) = x^{-1} \cdot 0 = 0.$$

Example 116 The ring $(\mathbb{Z}[i], +, \cdot)$ of Gaussian integers is a domain. Indeed is $x \cdot y = 0$, then taking absolute values, we obtain $|x| \cdot |y| = 0$. By the previous example, this implies that $|x| = 0$ or $|y| = 0$. Since the only complex number with absolute value zero is 0 itself, we see that if $x \cdot y = 0$, then $x = 0$ or $y = 0$.

Another important class of rings is given by fields:

Definition 117 *Let $(R, +, \cdot)$ be a commutative ring such that $R^* = R \backslash \{0\}$. Then R is called a* field.

Example 118 The rings $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$ and $(\mathbb{C}, +, \cdot)$ are all fields.

Example 119 The ring $(\mathbb{Z} \mod p, +_p, \cdot_p)$ is a field with p elements. Indeed, from Example 108, we see that any nonzero element in $\mathbb{Z} \mod p$ is a unit. In order to make it clear from the notation that it is a field, one usually writes $\mathbb{F}_p$ instead of $\mathbb{Z} \mod p$. Also it is common to abbreviate the notation and write $+$ an $\cdot$ for the operations $+_p$ and $\cdot_p$. Another commonly used notation for $\mathbb{Z} \mod p$ is $GF(p)$. GF stands for Galois Field (after the discoverer of finite fields Évariste Galois).

Since $(\mathbb{F}_p, +, \cdot)$ has finitely many elements, it is an example of what is called a finite field. Finite fields are used very frequently in discrete mathematics and in applications in coding theory and cryptography. The finite field $(\mathbb{F}_2, +, \cdot)$ is used pervasively in computer science as well, since we can think of the two elements of $\mathbb{F}_2$ as bits.

Warning: It is *not* true that $\mathbb{Z} \mod n$ is a field if n is not a prime number. We will see later that there exist finite field with p^e elements (p a prime number, $e \geq 1$ an integer), but these

cannot be identified with $\mathbb{Z} \mod p^e$ if $e > 1$. Indeed by Example 108, we see that if $e > 1$, the nonzero element p in $\mathbb{Z} \mod p^e$ will not be a unit. Hence for $e > 1$, the ring $(\mathbb{Z} \mod p^e, +_{p^e}, \cdot_{p^e})$ is not a field.

The reason fields are an important class of rings is that the theory of linear algebra can be developed over any field, not just $\mathbb{R}$ and $\mathbb{C}$. For example, it turns out that one can solve systems of linear equations with coefficients from a field using Gaussian elimination, one can define the null space and rank of a matrix with coefficients in a field, etc. Also one can define over any field the notions of vector space, subspace, basis, etc... . Especially when the field is finite (as in Example 119), this has applications in several areas of discrete mathematics such as coding theory and cryptography. We will not pursue these applications further here, but in the coming chapters develop a way to construct fields.

6.3 Polynomials with coefficients in a commutative ring

Keywords: polynomials, degree and leading coefficient In this section we give a precise definition of polynomials with coefficients in a ring $(R, +, \cdot)$. We will in this section always assume that the ring is commutative, that is to say that for all $r, s \in R$ we have $r \cdot s = s \cdot r$.

Intuitively, a polynomial $p(X) = r_0 + r_1 X + \cdots + r_d X^d$ in an indeterminate X, is uniquely determined by its coefficients sequence $(r_0, r_1, \ldots, r_d)$ of elements in R. For the definition of polynomials with coefficients from a ring, one can go a step further and say that a polynomial is a sequence of elements from R of the form $(r_0, r_1, \ldots, r_d)$. In order not to specify d, it is even easier to consider infinite sequences of the form $(r_0, r_1, \ldots, r_d, 0, 0, 0, \ldots)$ (we simply add infinitely many zeroes). Putting these ideas together, we obtain the following formal definition of a polynomial with coefficients from R.

Definition 120 *Let $(R, +, \cdot)$ be a ring. A polynomial p is defined to be an infinite sequence $(r_0, r_1, r_2, \ldots)$ of elements in R for which there exists a natural number $d \in \mathbb{N}$ such that $r_n = 0$ for all $n > d$. If d is the smallest natural number with this property, then d is called the degree of the polynomial, while r_d is called the leading coefficient of the polynomial. The degree of a polynomial p is denoted by* $\deg p$. *The zero polynomial is said to have degree $-\infty$, minus infinity.*

To get closer to the usual way to write down a polynomial, we can define $X^0 := (1, 0, 0, \ldots)$, $X := (0, 1, 0, 0, \ldots)$ and in general $r_d X^d$ to be the sequence having zeroes in all positions, except the d-th, where it has the value r_d. Note that we count the positions starting with zero. With these notations in place, we obtain that

$$\sum_{i=0}^{d} r_i X^i = (r_0, r_1, \ldots, r_d, 0, 0, \ldots).$$

The set of all polynomials is denoted by $R[X]$. We can give it a ring structure by defining addition and multiplication of polynomials as follows:

$$(r_0, r_1, \ldots, r_d, \ldots) + (s_0, s_1, \ldots, s_d, \ldots) := (r_0 + s_0, r_1 + s_1, \ldots, r_d + s_d, \ldots) \tag{6.1}$$

and

$$(r_0, r_1, \ldots, r_d, \ldots) \cdot (s_0, s_1, \ldots, s_d, \ldots) := (r_0 \cdot s_0, r_0 \cdot s_1 + r_1 \cdot s_0, \ldots, \sum_{i=0}^{d} r_i \cdot s_{d-i}, \ldots) \tag{6.2}$$

The multiplication is defined exactly such that $(r_iX^i)\cdot(r_jX^j) = (r_i \cdot r_j)X^{i+j}$ for any natural numbers i and j and elements $r_i, r_j \in R$. We write $R[X]$ for the set of all polynomials with coefficients in R, can be equipped with a ring structure:

Definition 121 *Let $(R,+,\cdot)$ be a ring and let $R[X]$ be the set of all polynomials with coefficients from R. Then $(R[X],+,\cdot)$ equipped with the operations from equations* (6.1) *and* (6.2) *is a ring, called the polynomial ring in the indeterminate X and coefficients in R.*

We did not check that $(R[X],+,\cdot)$ actually satisfies the ring axioms from Definition 100. First of all, we use $(0,0,\dots)$ and $(1,0,0,\dots)$ as neutral elements for addition and multiplication. Then most of the ring axioms for $(R[X],+,\cdot)$ can be shown directly using that $(R,+,\cdot)$ is a ring. A more complicated one to show is the associativity of multiplication. If $r = (r_0, r_1, r_2, \cdots)$, $s = (s_0, s_1, s_2, \cdots)$ and $t = (t_0, t_1, t_2, \cdots)$, then starting with the e-th coefficient of $(r \cdot s) \cdot t$ and using the distributive and associative law, we find

$$\sum_{d=0}^{e}\left(\sum_{i=0}^{d} r_i \cdot s_{d-i}\right)\cdot t_{e-d} = \sum_{d=0}^{e}\sum_{i=0}^{d}(r_i \cdot s_{d-i})\cdot t_{e-d} = \sum_{d=0}^{e}\sum_{i=0}^{d} r_i \cdot (s_{d-i} \cdot t_{e-d}).$$

Interchanging the order of summation and using the new index $j = d - i$, we then find

$$\sum_{d=0}^{e}\left(\sum_{i=0}^{d} r_i \cdot s_{d-i}\right)\cdot t_{e-d} = \sum_{i=0}^{e} r_i \sum_{d=i}^{e} s_{d-i} \cdot t_{e-d} = \sum_{i=0}^{e} r_i \sum_{j=0}^{e-i} s_j \cdot t_{e-i-j},$$

which is exactly the e-th coefficient of $r \cdot (s \cdot t)$. Therefore we have $(r \cdot s) \cdot t = r \cdot (s \cdot t)$.

We will take a special interest in polynomial rings if the ring of coefficients $(R,+,\cdot)$ is a domain or a field. For example we have that:

Theorem 122 *Suppose that $(D,+,\cdot)$ is a domain. Then the polynomial ring $(D[X],+,\cdot)$ is a domain as well.*

Proof. Let $p, q \in D[X]$ be two polynomials both different from zero, say of degree n and m. Let us write $p = a_nX^n + \cdots + a_0X^0$ and $q = b_mX^m + \cdots + b_0X^0$ with a_n and b_m both different from zero. Then $p \cdot q = (a_n \cdot b_m)X^{n+m} + \cdots + a_0 \cdot b_0$. Since D does not contain zero-divisors, we have $a_n \cdot b_m \neq 0$. This implies that $p \cdot q \neq 0$. ■

A consequence of the above theorem is also that for two non-zero polynomials p, q with coefficients in a domain, we have

$$\deg(p \cdot q) = \deg p + \deg q. \tag{6.3}$$

Since any field is a domain (see Exercise 4), the conclusion of Theorem 122 that $(D[X],+,\cdot)$ is a domain is also true if $(D,+,\cdot)$ is a field. Also Equation (6.3) is true for polynomials with coefficients in a field.

6.4 The division algorithm and roots of polynomials

Keywords: division with remainder, roots, number of roots

In this section we show a generalization of the fact that a polynomial of degree m with real or complex coefficients has at most m distinct roots. Just as for polynomial with coefficients in $\mathbb{R}$ or $\mathbb{C}$, a key ingredient in the proof of that fact is division with remainder algorithm. We start by studying if this algorithm also works over a general commutative ring.

Theorem 123 *Let $(R, +, \cdot)$ be a commutative ring and let two polynomials p and d in $R[X]$, both different from zero, be given. Assume that the leading coefficient of d is a unit. Then there exist polynomials q and r such that:*

1. *$p = r + d \cdot q$, and*
2. *the polynomial r is the zero polynomial or $\deg r < \deg d$.*

Proof. If $\deg p < \deg d$, we can choose $r = p$ and $q = 0$ and we are done. Therefore we assume from now on that $\deg p \geq \deg d$. We will show the theorem using induction on n on the statement that for any polynomial p of degree at most n and any polynomial d, there exist polynomials q and r having the properties mentioned in the theorem. If $n = 0$, i.e., if $\deg p = 0$, then $p = a_0$ for some $a_0 \in R\backslash\{0\}$. Since we assumed that $\deg p \geq \deg d$ this implies that $\deg d = 0$. Since the leading coefficient of d (in this case the constant term) was assumed to be a unit, we see that $d = b_0$ for some $b_0 \in R^*$. Now we can choose $r = 0$ and $q = b_0^{-1} \cdot a_0$. This completes the induction basis.

Now we consider the induction step. As induction hypothesis we assume that the theorem is true for any polynomials p and d as long as $\deg p \leq n - 1$. We wish to prove the theorem for polynomials p of degree n. Let us write

$$p = a_n X^n + \cdots + a_0 \text{ and } d = b_m X^m + \cdots + b_0.$$

Then we can write

$$p = p_1 + d \cdot (b_m^{-1} \cdot a_n) X^{n-m} \text{ with } p_1 = p - d \cdot (b_m^{-1} \cdot a_n) X^{n-m}.$$

Note that $\deg p_1 < n$. Therefore we can conclude from the induction hypothesis that there exist polynomials q_1 and r such that $p_1 = r + d \cdot q_1$, with r either equal to zero, or $\deg r < \deg d$. Combining the above, we see that

$$\begin{aligned} p &= p_1 + d \cdot (b_m^{-1} \cdot an_) X^{n-m} \\ &= r + d \cdot q_1 + d \cdot (b_m^{-1} \cdot a_n) X^{n-m} \\ &= r + d \cdot (q_1 + (b_m^{-1} \cdot a_n) X^{n-m}) \end{aligned}$$

The polynomials r and $q := q_1 + (b_m^{-1} \cdot a_n) X^{n-m}$ satisfy the properties from the theorem. This concludes the induction step and by the induction principle the proof of the theorem. ■

Just as for polynomials with real or complex coefficients, one can speak of the root of a polynomial in general (though we still assume that the ring $(R, +, \cdot)$ is commutative). If $p \in R[X]$ is a polynomial, say $p = \sum_{i=0}^{d} r_i X^i$, then we define the *evaluation* of p in an element $a \in R$ by

$$p(a) := \sum_{i=0}^{d} r_i \cdot a^i.$$

Further we call an element $a \in R$ a *root* of p if $p(a) = 0$. The division algorithm on polynomials relates the notions of evaluation and root to a specific way to write a polynomial:

Proposition 124 *Let $(R, +, \cdot)$ be a commutative ring and $p \in R[X]$ a non-zero polynomial of degree $d \geq 1$ and let $a \in R$. Then there exists a polynomial q of degree $d - 1$ such that*

$$p = (X - a) \cdot q + p(a). \tag{6.4}$$

Moreover $a \in R$ is a root of p if and only if there exists a polynomial $q \in R[X]$ of degree $d - 1$ such that $p = (X - a) \cdot q$.

Proof. Using the division algorithm on $p \in R[X]$ and $d := X - a$, we can find polynomials q and a constant $r_0 \in R$ such that $p = (X-a)\cdot q + r_0$. Evaluating these expressions for $X = a$, we find that $r_0 = p(a)$. Moreover, since the leading coefficient of $X - a$ equals 1 and hence is not a zero-divisor, we see that $\deg(X-a)\cdot q = 1 + \deg q$. On the other hand, since $d = \deg p \geq 1$ and $p = (X-a)\cdot q + p(a)$, we see that $\deg p = 1 + \deg q$. This shows that $\deg q = d - 1$ and finishes the proof of the first part of the proposition.

Now we prove the second part of the proposition: If $a \in R$ is a root of $p(X)$, then $p(a) = 0$. Equation 6.4 then implies that $p(X) = q(X)\cdot(X-a)$. Conversely, if $p(X) = q(X)\cdot(X-a)$ for some polynomial $q(X)$, then $p(a) = q(a)\cdot 0 = 0$, implying that a is a root of $p(X)$. ■

A direct consequence of this is the following statement concerning the number of roots a polynomial can have.

Corollary 125 *Let $(\mathbb{F}, +, \cdot)$ be a field and $p(X) \in \mathbb{F}[X]$ a non-zero polynomial of degree d. Then p has at most d roots in $\mathbb{F}$.*

Proof. We prove the corollary with induction on d. If $d = 0$, there is nothing to prove, since then $p = a_0$ for $a_0 \in \mathbb{F}\backslash\{0\}$. Assume now that the theorem is true for polynomials of degree $d-1$. If $a \in \mathbb{F}$ is a root of p, we can write $p = (X-a)\cdot q$. Moreover, $\deg q = d - 1$. If $b \in F$ is another root of p distinct from a, then $q(b)\cdot(b-a) = p(b) = 0$. Since $a \neq b$ and $\mathbb{F}$ is a field, the element $b - a$ has a multiplicative inverse. Therefore we can conclude that $q(b) = 0$ or in other words that b is a root of q. By the induction hypothesis, the polynomial q has at most $d-1$ roots in $\mathbb{F}$. This implies that p has at most d roots (namely a and the roots of q). ■

6.5 Exercises

1. Determine for the following examples: 1) if is is a ring, 2) if yes whether or not the ring is commutative, 3) if yes, whether or not the ring is a domain, 4) if yes, whether or not the ring is a field.

   (a) $(\mathbb{N}, +, \cdot)$

   (b) $(D, +, \cdot)$, where D is the set of 2×2 diagonal matrices with coefficients in $\mathbb{R}$.

   (c) $(\{0\}, +, \cdot)$. This ring is called the ring with one element.

2. In this exercise, we investigate some fundamental properties of the elements 0_R and 1_R in a ring.

   (a) Let $(R, +_R, \cdot_R)$ be a ring. Show that $0_R \cdot x = 0_R$ for all $x \in R$. Hint: use that $0_R + 0_R = 0_R$ and the ring axioms.

   (b) Let $(R, +_R, \cdot_R)$ be a ring. Show that $0_R \neq 1_R$ unless $R = \{0_R\}$. Hint: use the previous part of the exercise and the ring axioms.

3. Let us define the set $\mathbb{Z}[\sqrt{2}] := \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$. Show that $\mathbb{Z}[\sqrt{2}]^*$ is an infinite set. Hint: show first that $1+\sqrt{2}$ is a unit in $\mathbb{Z}[\sqrt{2}]$, then try to construct new units from $1+\sqrt{2}$.

4. Show that a unit cannot be a zero-divisor. Using this, conclude that a field necessarily is a domain as well.

5. Let us define the set $\mathbb{Z}[i] := \{a + bi \mid a, b \in \mathbb{Z}\}$ and consider the ring $(\mathbb{Z}[i], +, \cdot)$. It is called the ring of Gaussian integers.

(a) Show that $\mathbb{Z}[i]$ is a domain. Hint: show first that $(a+bi)(c+di) = (e+fi)$ implies that $(a^2+b^2)(c^2+d^2) = (e^2+f^2)$.

(b) Show that $(\mathbb{Z}[i])^* = \{1, -1, i, -i\}$.

6. Consider the field with two elements $(\mathbb{F}_2, +, \cdot)$. Show that the polynomial $X^3+X^2+X+1 \in \mathbb{F}_2[X]$ can be written as the product of polynomials of degree one.

7. Let $R = \mathbb{Z} \mod 6$. Find a polynomial $p \in R[X]$ of degree 2 that has more than two roots in R.

Chapter 7

The theory of ideals

In this chapter we will develop the theory of ideals for commutative rings. In particular, we will from now on only consider commutative rings. Ideals in ring theory play a similar role as normal subgroups in group theory. They play for example an important role in the construction of quotient rings, the analog of quotient groups in ring theory.

7.1 Ideals in a ring

Keywords: ideal, principal ideal

Definition 126 *Let $(R, +_R, \cdot_R)$ be a commutative ring. A subset $I \subset R$ is called an ideal if*

1. *$I \subset R$ is a subgroup of the additive group $(R, +_R)$ of the ring.*
2. *For any $r \in R$ and any $x \in I$ we have $r \cdot_R x \in I$.*

The second condition implies that an ideal is closed under multiplication (that is to say, if $x, y \in I$, then $y \cdot_R x \in I$), but it is in general much stronger since for any r from R and $x \in I$ we should have that $r \cdot_R x \in I$. This property turns out to be important when we later define quotient rings.

Example 127 Let $(R, +_R, \cdot_R)$ be a commutative ring. Then the sets $\{0_R\}$ and R are ideals. They are often called the trivial ideals, because any commutative ring has these ideals.

Example 128 We consider the ring $(\mathbb{Z}, +, \cdot)$. The set $2\mathbb{Z}$ of even integers is an ideal. In the first place $2\mathbb{Z} \subset \mathbb{Z}$ is a subgroup of $(\mathbb{Z}, +)$, but moreover, any integer multiple of an even number is even again. More general, if $n \in \mathbb{Z}$, then the set $n\mathbb{Z}$ of multiples of n is an ideal.

Ideals like in the previous example can be defined for any ring.

Definition 129 *Let $(R, +_R, \cdot_R)$ be a commutative ring and $x \in R$. Define*

$$xR := \{x \cdot r \,|\, r \in R\}.$$

An ideal $I \subset R$ such that $I = xR$ for some $x \in R$, is called a principal *ideal. Another commonly used notation for the ideal xR is $\langle x \rangle$. One says that the ideal $\langle x \rangle$ is generated by x. The element x itself is called a generator of the ideal $\langle x \rangle$.*

A more general construction of ideals is investigated in the following lemma:

Lemma 130 *Let $(R, +_R, \cdot_R)$ be a commutative ring and $x_1, \ldots, x_n \in R$. The set*

$$\langle x_1, \ldots, x_n \rangle := \{r_1 \cdot x_1 + \cdots + r_n \cdot x_n \,|\, r_1, \ldots, r_n \in R\}$$

is an ideal.

Proof. For convenience we write $I = \langle x_1, \ldots, x_n \rangle$. Also we write $+$ for $+_R$ and $\cdot$ for $\cdot_R$. First we check that $I \subset R$ is a subgroup of the additive group $(R, +_R)$ of the ring.

First of all using that addition is commutative and the distribute laws, we see that

$$(r_1 \cdot x_1 + \cdots + r_n \cdot x_n) + (s_1 \cdot x_1 + \cdots + s_n \cdot x_n) = (r_1 + s_1) \cdot x_1 + \cdots + (r_n + s_n) \cdot x_n.$$

Therefore I is closed under addition.

Using that $-(r \cdot x) = (-r) \cdot x$ (since $(-r) \cdot x + r \cdot x = (-r + r) \cdot x = 0 \cdot x = 0$), we obtain that

$$-(r_1 \cdot x_1 + \cdots + r_n \cdot x_n) = -r_1 \cdot x_1 - \cdots - r_n \cdot x_n = (-r_1) \cdot x_1 + \cdots + (-r_n) \cdot x_n$$

we see that (additive) inverses of elements in I are in I again.

Finally, note that

$$0 = 0 \cdot x_1 + \cdots + 0 \cdot x_n.$$

Hence $0 \in I$ as well.

The above shows that I is a subgroup of $(R, +)$. Now we check that for any $r \in R$ and any $x \in I$ we have $r \cdot x \in I$. By definition of I, for any $x \in I$ there exist $r_1, \ldots, r_n \in R$ such that $x = r_1 \cdot x_1 + \cdots + r_n \cdot x_n$. Then we have

$$r \cdot x = r \cdot (r_1 \cdot x_1 + \cdots + r_n \cdot x_n) = (r \cdot r_1) \cdot x_1 + \cdots + (r \cdot r_n) \cdot x_n.$$

Hence $r \cdot x \in I$ if $x \in I$.

We have now shown that I is an ideal of R. ∎

The ideal $\langle x_1, \cdots, x_n \rangle$ is called the ideal generated by $x_1, \ldots, x_n$. The elements $x_1, \ldots, x_n$ are called generators of $\langle x_1, \cdots, x_n \rangle$.

Example 131 Let $R = \mathbb{Z}$ and consider the ideal $\langle 5, 17 \rangle$. We claim that $1 \in I$. Indeed $1 = 3 \cdot 17 - 10 \cdot 5$, which is an element of $\langle 5, 17 \rangle$. But then by the defining properties of an ideal, any element $r \in \mathbb{Z}$ is in the ideal, since the element $r = r \cdot 1$ is in the ideal again by the second defining property of an ideal. This means that $\langle 5, 17 \rangle = \mathbb{Z}$.

More generally, if $I \subset R$ is an ideal in a commutative ring $(R, +_R, \cdot_R)$ then $1_R \in I$ if and only if $I = R$. Indeed $I \subset R$ always holds. Conversely $R \subset I$ if $1_R \in I$, since then for any $r \in R$ we have $r = r \cdot_R 1_R \in I$.

Example 132 Let $R = (\mathbb{Z} \mod 6)[X]$, be the set of polynomials in one variable with coefficients in $\mathbb{Z} \mod 6$. Then we claim that the set

$$I := \{2 \cdot s + X \cdot t \,|\, s, t \in (\mathbb{Z} \mod 6)[X]\}$$

is an ideal of $(\mathbb{Z} \mod 6)[X]$.

First we check that I is a subgroup of $((\mathbb{Z} \mod 6)[X], +)$. Since

$$0 = 2 \cdot 0 + X \cdot 0,$$

we have $0 \in I$. Also if $2 \cdot s + X \cdot t \in I$, then

$$-(2 \cdot s + X \cdot t) = -2 \cdot s - X \cdot t = 2 \cdot (-s) + X \cdot (-t)$$

is in I. If we pick two elements from I, say $2 \cdot s_1 + X \cdot t_1$ and $2 \cdot s_2 + X \cdot t_2$, then

$$(2 \cdot s_1 + X \cdot t_1) + (2 \cdot s_2 + X \cdot t_2) = 2 \cdot (s_1 + s_2) + X \cdot (t_1 + t_2) \in I.$$

Finally if $r \in (\mathbb{Z} \mod 6)[X]$ and $2 \cdot s + X \cdot t \in I$, then

$$r \cdot (2 \cdot s + X \cdot t) = 2 \cdot (r \cdot s) + X \cdot (r \cdot t) \in I.$$

The ideal I in a sense contains half of all polynomials in $(\mathbb{Z} \mod 6)[X]$. Polynomials of the form $2 \cdot s + X \cdot t$ are namely precisely those polynomials that have an even constant term. Using this description, one can also see that I is a subgroup of $((\mathbb{Z} \mod 6)[X], +)$ and that any multiple of such a polynomial again is in I.

7.2 Quotient rings

Keywords: quotient rings

We will now use ideals to construct so-called quotient rings. The situation is analogous to groups, where one could use normal subgroups to construct quotient groups. First of all, since I is a subgroup of $(R, +)$, we can define cosets of I in R by r as:

$$r + I := \{r + x \,|\, x \in I\}.$$

Since $(R, +)$ is an abelian group, the left coset $r + I$ is the same as the right coset $I + r$ for any $r \in R$. This implies that I in fact is a normal subgroup of R and therefore that we can define the quotient group $(R/I, +)$ with addition defined by

$$(r + I) + (s + I) := (r + s) + I.$$

It turns out that we can give R/I the structure of a ring by defining the following multiplication on cosets:

Lemma 133 *Let $I \subset R$ be an ideal of a commutative ring $(R, +, \cdot)$. Then the following operation on cosets is well defined:*

$$(r + I) \cdot (s + I) := (r \cdot s) + I.$$

Proof. To show that the multiplication is well defined, we need to show the following: If $r_1, r_2, s_1, s_2 \in R$ are chosen such that $r_1 + I = r_2 + I$ and $s_1 + I = s_2 + I$, then $(r_1 \cdot s_1) + I = (r_2 \cdot s_2) + I$. Equivalently, we need to show that $(r_1 \cdot s_1) - (r_2 \cdot s_2) \in I$. However, we have

$$(r_1 \cdot s_1) - (r_2 \cdot s_2) = (r_1 \cdot s_1) - (r_1 \cdot s_2) + (r_1 \cdot s_2) - (r_2 \cdot s_2) = r_1 \cdot (s_1 - s_2) + (r_1 - r_2) \cdot s_2. \quad (7.1)$$

Since $s_1 + I = s_2 + I$, we have $s_1 - s_2 \in I$ and since I is an ideal, we can deduce that $r_1 \cdot (s_1 - s_2) \in I$. Similarly, one obtains that $(r_1 - r_2) \cdot s_2 = s_2 \cdot (r_1 - r_2) \in I$. Therefore, equation (7.1) implies that $(r_1 \cdot s_1) - (r_2 \cdot s_2) \in I$, which was what we wanted to show. ■

Now we can give the set R/I a ring structure:

Theorem 134 *Let $I \subset R$ be an ideal of a commutative ring $(R, +, \cdot)$. Then $(R/I, +, \cdot)$ with addition and multiplication defined by*

$$(r + I) + (s + I) := (r + s) + I$$

and

$$(r + I) \cdot (s + I) := (r \cdot s) + I$$

is a commutative ring with zero element $0 + I$ and identity element $1 + I$.

Proof. We check that for example the distributive law is satisfied.

$$(r + I) \cdot ((s + I) + (t + I)) = (r + I) \cdot ((s + t) + I) = (r \cdot (s + t)) + I = ((r \cdot s) + (r \cdot t)) + I$$

$$= ((r \cdot s) + I) + ((r \cdot t) + I) = ((r + I) \cdot (s + I)) + ((r + I) \cdot (t + I)).$$

We have used the addition and multiplication of cosets of I, but also the fact that the distributive law is satisfied in R. Similarly all ring axioms for R/I, including commutativity, follow from the fact that they are satisfied in R. ■

Example 135 Consider the ring $(\mathbb{Z}, +, \cdot)$ and choose $I = n\mathbb{Z}$, the ideal of $\mathbb{Z}$ generated by n. Then $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is a ring whose elements consist of cosets of $n\mathbb{Z}$ in $\mathbb{Z}$. If $n \neq 0$, there are n such cosets, namely the cosets $a + n\mathbb{Z}$ with $a \in \{0, 1, \ldots, n-1\}$. For if $m + n\mathbb{Z}$ is any coset, then using division with remainder, one can write $m = r + n \cdot q$ for $q, r \in \mathbb{Z}$ and $0 \leq r \leq n - 1$. But then $m + n\mathbb{Z} = r + n\mathbb{Z}$, since $m - r = n \cdot q \in n\mathbb{Z}$.

The ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ can in fact be identified with the ring $(\mathbb{Z} \pmod n, +_n, \cdot_n)$ by identifying a coset $a + n\mathbb{Z}$ with $a \in \mathbb{Z} \pmod n$. Indeed, since $a +_n b$ and $a + b$ differ by a multiple of n we then get

$$(a + n\mathbb{Z}) + (b + n\mathbb{Z}) = (a + b) + n\mathbb{Z} = (a +_n b) + n\mathbb{Z}.$$

Similarly

$$(a + n\mathbb{Z}) \cdot (b + n\mathbb{Z}) = (a \cdot b) + n\mathbb{Z} = (a \cdot_n b) + n\mathbb{Z}.$$

Example 136 Consider the ring $(\mathbb{R}[X], +, \cdot)$ and choose $I = \langle X^2 \rangle$, the ideal of $\mathbb{R}[X]$ generated by X^2. Then $(\mathbb{R}[X]/\langle X^2 \rangle, +, \cdot)$ is a ring whose elements consist of cosets of $\langle X^2 \rangle = X^2\mathbb{R}[X]$ in $\mathbb{R}[X]$.

For any polynomial $p(X) = p_0 + p_1 X + p_2 X^2 + \cdots$, we have

$$p(X) + \langle X^2 \rangle = p_0 + p_1 X + \langle X^2 \rangle,$$

since

$$p(X) - (p_0 + p_1 X) = X^2 \cdot (p_2 + p_3 X + \cdots) \in \langle X^2 \rangle.$$

Therefore only the linear approximation of the polynomial $p(X)$ matters. This means that all cosets of $\langle X^2 \rangle$ in $\mathbb{R}[X]$ can be described in the form $p_0 + p_1 X + \langle X^2 \rangle$ for certain $p_0, p_1 \in \mathbb{R}$. Using this description, we can explicitly describe addition and multiplication of cosets in the quotient ring $\mathbb{R}[X]/\langle X^2 \rangle$. Addition of cosets is given by

$$(\ p_0 + p_1 X + \langle X^2 \rangle\ ) + (\ q_0 + q_1 X + \langle X^2 \rangle\ ) = (p_0 + q_0) + (p_1 + q_1)X + \langle X^2 \rangle$$

and multiplication by

$$(\ p_0 + p_1 X + \langle X^2 \rangle\ ) \cdot (\ q_0 + q_1 X + \langle X^2 \rangle\ ) = (p_0 \cdot q_0) + (p_0 \cdot p_1 + p_1 \cdot q_0)X + \langle X^2 \rangle.$$

In the multiplication, we used that the quadratic term $p_1 \cdot q_1 X^2$ the occurs when multiplying $(p_0 + p_1 X)(q_0 + q_1 X)$ can be left out, since it is contained in the ideal $\langle X^2 \rangle$.

If we write $\epsilon = X + \langle X^2 \rangle$, then we can describe the quotient ring $\mathbb{R}[X]/\langle X^2 \rangle$ as a polynomial ring $\mathbb{R}[\epsilon]$, but with the rule that $\epsilon^2 = 0$, since

$$\epsilon^2 = (X + \langle X^2 \rangle) \cdot (X + \langle X^2 \rangle) = X^2 + \langle X^2 \rangle = 0 + \langle X^2 \rangle.$$

With this more compact notation, we can write the addition and multiplication in the quotient ring more compactly as

$$(p_0 + p_1\epsilon) + (q_0 + q_1\epsilon) = (p_0 + q_0) + (p_1 + q_1)\epsilon$$

and

$$(p_0 + p_1\epsilon) \cdot (q_0 + q_1\epsilon) = (p_0 \cdot q_0) + (p_0 \cdot q_1 + p_1 \cdot q_0)\epsilon.$$

This ring is sometimes called the ring of dual numbers and rings of this type are used to describe and investigate linear approximations of functions.

7.3 Principal ideal domains

Keywords: PIDs,

In this section we will consider special rings where one can show that any ideal is a principal ideal. That is to say, we will consider rings $(R, +, \cdot)$ such that for any ideal $I \subset R$, there exists $x \in R$ such that $I = xR$. If moreover, the ring is a domain, such a ring is called a principal ideal domain (PID).

First of all, we show that the ring $(\mathbb{Z}, +, \cdot)$ is a PID.

Proposition 137 *The ring $(\mathbb{Z}, +, \cdot)$ is a principal ideal domain.*

Proof. Let I be an ideal of $\mathbb{Z}$. We claim that $I = d\mathbb{Z}$ for a suitably chosen integer d (which will depend on I). If $I = \{0\}$, we can choose $d = 0$, so we will suppose from now on that $I \neq \{0\}$. Then let d be the smallest positive integer occurring in I. We claim that $I = d\mathbb{Z}$. Since $d \in I$ and I is an ideal, it is clear that $d\mathbb{Z} \subset I$. Conversely suppose that $n \in I$. Using division with remainder, we may write $n = r + d \cdot m$ for integers r and m such that $0 \leq r \leq d - 1$. Since $r = n - d \cdot m$ and I is an ideal, we see that $r \in I$. On the other hand we assumed that d was the smallest positive element of I. Therefore we may conclude that $r = 0$. This means that $n = d \cdot m \in d\mathbb{Z}$, implying that $I \subset d\mathbb{Z}$.

All in all we may conclude that $I = d\mathbb{Z}$. ■

Example 138 This is a continuation of Example 135. From the above proposition we see that any ideal of $\mathbb{Z}$ is of the form $n\mathbb{Z}$. This means that all possible quotient rings of $\mathbb{Z}$ are of the form $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$. As remarked in Example 135, such a quotient ring gives a more algebraic way to define addition and multiplication modulo n.

The ring $(\mathbb{Z}, +, \cdot)$ is apparently an example of a PID. A second one is given below:

Proposition 139 *Let $(\mathbb{F}, +, \cdot)$ be a field. Then the polynomial ring $(\mathbb{F}[X], +, \cdot)$ is a principal ideal domain.*

Proof. Let I be an ideal of $\mathbb{F}[X]$. As for the ring of integers, we claim that $I = d(X)\mathbb{F}[X] = \langle d(X)\rangle$ for a suitably chosen polynomial $d(X)$. If $I = \{0\}$, we can choose $d(X) = 0$. Now suppose that $I \neq \{0\}$. Then let $d(X)$ be a polynomial of smallest degree among all non-zero polynomials occurring in I. We claim that $I = d(X)\mathbb{F}[X]$. The inclusion $d(X)\mathbb{F}[X] \subset I$ holds, since $d(X) \in I$ and I is an ideal. Conversely suppose that $n(X) \in I$. Using division with remainder, but now of polynomials with coefficients in a field, we may write $n(X) = r(X)+d(X)\cdot m(X)$ for polynomials $r(X)$ and $m(X)$ such that $r(X) = 0$ or $\deg r(X) < \deg d(X)$. Since $r(X) = n(X) - d(X)\cdot m(X)$, we see that $r(X) \in I$. On the other hand we assumed that $d(X)$ was a polynomial in I of smallest degree. Apparently, $r(X) = 0$ and we may conclude that $n(X) \in d(X)\mathbb{F}[X]$, implying that $I \subset d(X)\mathbb{F}[X]$.

Combining the above, we conclude that $I = d(X)\mathbb{F}[X]$. ■

It will be useful for us also to study quotient rings of $(\mathbb{F}[X], +, \cdot)$. Since any ideal of this ring is of the form $d(X)\mathbb{F}[X]$, it will be practical to be able to describe all cosets in $\mathbb{F}[X]/d(x)\mathbb{F}[X]$. We do this in the following lemma.

Lemma 140 *Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X) \in \mathbb{F}[X]$ be a polynomial of degree at least one. Then any coset of the ideal $I := f(X)\mathbb{F}[X]$ can uniquely be described in the form $r(X) + I$, with $r(X) = 0$ or $\deg r(X) < \deg f(X)$.*

Proof. We start by showing that any coset of I can be described in the desired form. Given a coset $g(X) + I$, we can find, using the division algorithm, polynomials $r(X)$ and $q(X)$ such that $g(X) = r(X) + f(X)\cdot q(X)$ and $r(X) = 0$ or $\deg r(X) < \deg f(X)$. Since $g(X) - r(X) = f(X)\cdot q(X) \in I$, we have $g(X) + I = r(X) + I$.

Now we show uniqueness. Suppose that for both $i = 1$ and $i = 2$ it holds that $r_i(X) = 0$ or $\deg r_i(X) < \deg f(X)$. Additionally assume that $r_1(X) \neq r_2(X)$. Note that this implies that $\deg(r_1(X) - r_2(X)) < degf(X)$. We need to show that $r_1(X) + I \neq r_2(X) + I$. Assume on the contrary that $r_1(X) + I = r_2(X) + I$. Then $r_1(X) - r_2(X) \in I$, which means that we can find a polynomial $q(X)$ such that $r_1(X) - r_2(X) = f(X)\cdot q(X)$. But then $\deg(r_1(X) - r_2(X)) \geq \deg f(X)$, a contradiction. Apparently the only possibility is that $r_1(X) + I \neq r_2(X) + I$. ■

Example 141 Let $R = \mathbb{R}[X]$ and $I := (X^2 + 1)\mathbb{R}[X]$. The cosets of this particular ideal I can all be described as $a + bX + I$, with $a, b \in \mathbb{R}$. The multiplication in the quotient ring $\mathbb{R}[X]/(X^2 + 1)\mathbb{R}[X]$ can be described explicitly using this description: First of all by definition we have

$$((a + bX) + I)\cdot((a + bX) + I) = (a + bX)(c + dX) + I = ac + (ad + bc)X + bdX^2 + I.$$

Now since $X^2 + 1 \in I$, we have that $X^2 + I = -1 + I$ and therefore

$$(a + bX + I)\cdot(c + dX + I) = ac - bd + (ad + bc)X + I.$$

This is similar to the formula one would obtain when multiplying the two complex numbers $a+bi$ and $c+di$ with each others. We see that the quotient ring $\mathbb{R}[X]/(X^2+1)\mathbb{R}[X]$ can be identified with the complex numbers $\mathbb{C}$ by identifying $a + bX + (X^2 + 1)\mathbb{R}[X]$ and $a + bi$.

7.4 Prime ideals and maximal ideals

Keywords: prime ideals, maximal ideals

Quotient rings can be used to construct new rings from a given ring. As we have seen before, domains are somewhat similar to the ring of integers in the sense that a product of two non-

zero elements cannot be zero. Therefore, ideals I of a commutative ring $(R,+,\cdot)$ such that the quotient ring $(R/I,+,\cdot)$ is a domain are of special interest.

Definition 142 *Let $(R,+,\cdot)$ be a commutative ring. An ideal $I \subset R$ is called a prime ideal if it satisfies*

$$r \cdot s \in I \Rightarrow r \in I \text{ or } s \in I.$$

Example 143 Let $R = \mathbb{Z}$ and consider $I = p\mathbb{Z}$ for some prime number p. Then I is a prime ideal. Indeed, if $n \cdot m \in p\mathbb{Z}$, then there exists $k \in \mathbb{Z}$ such that $n \cdot m = p \cdot k$. Since p is a prime number, it divides either n or m. This means that either $n \in p\mathbb{Z}$ or $m \in p\mathbb{Z}$. Let us for the sake of completion prove the claim that $n \cdot m \equiv 0 \bmod p$ implies that $n \equiv 0 \bmod p$ or $m \equiv 0 \bmod p$. First we use division with remainder on both n and m and write $n = p \cdot q_1 + r_1$ and $m = p \cdot q_2 + r_2$, with $0 \leq r_1 < p$ and $0 \leq r_2 < p$. Then $n \cdot m \equiv r_1 \cdot r_2 \bmod p$. If p would divide neither n nor m, then r_1 and r_2 are both units in $\mathbb{Z} \mod p$, whose inverses can be found using the extended Euclidean algorithm. Since units are not zero-divisors, this would imply that $r_1 \cdot r_2 \equiv 0 \bmod p$, a contradiction.

The importance of prime ideals lies in the following theorem:

Theorem 144 *Let $(R,+,\cdot)$ be a commutative ring and $I \subset R$ an ideal. Then $(R/I,+,\cdot)$ is a domain if and only if $I \subset R$ is a prime ideal.*

Proof. First assume that I is a prime ideal. We will show that $(R/I,+,\cdot)$ is a domain. If $(r+I)\cdot(s+I) = 0+I$, then, since $(r+I)\cdot(s+I) = r\cdot s + I$, this implies that $r \cdot s \in I$. Since I is a prime ideal, we obtain that either $r \in I$ or $s \in I$. If $r \in I$, then $r + I = 0 + I$, while, similarly, if $s \in I$, then $s + I = 0 + I$. Since $0 + I$ is the zero element in R/I, this implies the theorem.

Conversely assume that $(R/I,+,\cdot)$ is a domain and that $r \cdot s \in I$. We need to show that $r \in I$ or $s \in I$. Since $r \cdot s \in I$, we have that $r \cdot s + I = 0 + I$. This implies that $(r+I)\cdot(s+I) = 0+I$. Since $(R/I,+,\cdot)$ is a domain, we may conclude that $r + I = 0 + I$ or $s + I = 0 + I$. In the former case we obtain $r \in I$ in the latter $s \in I$. This shows that I is a prime ideal. ■

We see that prime ideals can be used to construct domains from commutative rings. Another type of ideals give rise to fields.

Definition 145 *Let $(R,+,\cdot)$ be a commutative ring and let $I \subset R$ be an ideal such that:*

1. *$I \neq R$*
2. *If J is an ideal of R such that $I \subset J \subset R$, then $J = I$ or $J = R$.*

Then I is called a maximal ideal of R.

Example 146 Let $R = \mathbb{Z}$ and consider $I = p\mathbb{Z}$ for some prime number p. Then I is a maximal ideal. Indeed, let J be an ideal of $\mathbb{Z}$ containing I. We will show that $J = I$ or $J = \mathbb{Z}$. If $J = I$, we are done, so we will assume that $J \neq I$. In that case there exists $k \in J$ such that $k \not\equiv 0 \bmod p$. This means that p and k are relatively prime, since p is a prime number. By the extended Euclidean algorithm, this implies that there exist r and s such that $r \cdot p + s \cdot k = 1$. Since J contains both k and p, it will then also contain the element 1. This implies that $J = \mathbb{Z}$.

The importance of maximal ideals is that they give rise to fields:

Theorem 147 *Let $(R,+,\cdot)$ be a commutative ring and $I \subset R$ an ideal. Then $(R/I,+,\cdot)$ is a field if and only if $I \subset R$ is a maximal ideal.*

Proof. First assume that $I \subset R$ is a maximal ideal. We need to show that any coset $r + I$ with $r \notin I$ has a multiplicative inverse in R/I. We consider the ideal $J := \{s \cdot r + x \,|\, s \in R; x \in I\}$. It contains I, but is not equal to it, since $r \notin I$, but $r \in J$. Since I is a maximal ideal, we have $J = R$. In particular we have $1 \in J$, which implies that there exist $s \in R$ and $x \in I$ such that $s \cdot r + x = 1$. This implies that

$$(s + I) \cdot (r + I) = (s \cdot r) + I = (1 - x) + I = 1 + I.$$

Conversely assume that $(R/I, +, \cdot)$ is a field. Since the ring with one element is not a field, we see that $I \neq R$. Now assume that $J \subset R$ is an ideal containing I. If $J = I$ we are done, so we may assume that J is not equal to I. In this case we need to show that $J = R$. Let us choose $x \in J \backslash I$. Then $x + I$ is a unit in R/I, since $(R/I, +, \cdot)$ is a field. Therefore, there is $y \in R$ such that $x \cdot y + I = (x + I) \cdot (y + I) = 1 + I$, which implies that $x \cdot y - 1 \in I$. Since $I \subset J$ and $x \in J$, this implies that $1 \in J$. However, the only ideal of R containing 1 is the ideal R. Therefore we obtain that $J = R$ as desired. ■

Example 148 Let p be a prime number. Then $(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$ is a field. It is called a finite field with p elements. This gives a more algebraic construction of a finite field. Alternatively, the ring $(\mathbb{Z} \mod p, +_p, \cdot_p)$ is also a finite field with p elements. One can show that these two (and indeed any finite field with p elements) are essentially the same, more precisely are isomorphic as fields. Therefore one can talk about *the* finite field with p elements, which is commonly denoted by $(\mathbb{F}_p, +, \cdot)$.

7.5 Exercises

1. We consider the ring $(\mathbb{Z}, +, \cdot)$ of integers and the ideal $I := \langle 6, 15 \rangle$. Find $n \in \mathbb{Z}$ such that $I = n\mathbb{Z}$.

2. Suppose that $R = \mathbb{R}[X]$ with the usual addition and multiplication of polynomials as ring operators. Further suppose that an ideal $I \subset \mathbb{R}[X]$ contains the polynomials $X^2 + 1$ and $X^5 + X + 1$. Show that $I = \mathbb{R}[X]$. Hint: according to the extended Euclidean algorithm, one can find polynomials r and s such that $r \cdot (X^2+1) + s \cdot (X^5+X+1) = \gcd(X^2+1, X^5+X+1)$.

3. Let I and J be two ideals of a commutative ring R. Show that the intersection $I \cap J$ also is an ideal of R. Also find an example of a ring R and ideals I and J showing that the union $I \cup J$ in general is not an ideal of R.

4. Let $(\mathbb{F}, +, \cdot)$ be a field. Show that there are only two possible ideals $I \subseteq \mathbb{F}$.

5. Let $(\mathbb{F}_2[X], +, \cdot)$ be the polynomial ring over the finite field $\mathbb{F}_2$ and let $I := \langle X^3 + X + 1 \rangle$. Show, using the extended Euclidean algorithm on $X^3 + X + 1$ and $X^2 + X + 1$ that $X^2 + X + 1 + I$ is a unit in $\mathbb{F}_2[X]/I$.

6. Show that the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ contains zero divisors if and only if n is a composite number (that is to say, $n = k \cdot \ell$ for numbers k and ℓ both not equal to ± 1).

7. Let $R = \{0, 1, a, b\}$ consist of four elements. Find a multiplication and addition on R such that $(R, +, \cdot)$ is a finite field with four elements.

8. Find an example of a ring R and an ideal $I \neq \{0\}$ such that I is not a maximal ideal of R. Hint: Try a ring of the form $(\mathbb{Z} \mod n, +_n, \cdot_n)$ for a suitably chosen n.

Chapter 8

Finite fields

In this chapter we will use the theory of quotient rings to construct fields, especially finite fields (that is to say fields with a finite number of elements).

8.1 Construction of fields using irreducible polynomials

Keywords: irreducible polynomials, construction of new fields

We have seen that the quotient ring $(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$ is a finite field if p is a prime number. We have seen this field before in another description, namely as $(\mathbb{Z} \mod p, +_p, \cdot_p)$ and usually denote this field by $(\mathbb{F}_p, +, \cdot)$. Another procedure to produce field is by considering quotient rings of $\mathbb{F}[X]$, the polynomial ring with coefficients in a field $\mathbb{F}$. As we saw in Example 141, the complex numbers $\mathbb{C}$ be obtained in this way. As we will see, what we will need is an analogue of a prime number in a polynomial ring. We will define this in the following:

Definition 149 *Let $(\mathbb{F}, +, \cdot)$ be a field. A polynomial $f(X) \in \mathbb{F}[X]$ is called irreducible if it has positive degree and if $f(X) = g(X) \cdot h(X)$ for $g(X), h(X) \in \mathbb{F}[X]$ implies that $\deg f(X) = \deg g(X)$ or $\deg f(X) = \deg h(X)$.*

Since $\deg(g(X) \cdot h(X)) = \deg g(X) + \deg h(X)$, an equivalent definition of a polynomial $f(X)$ to be irreducible is the following: $f(X) = g(X) \cdot h(X)$ implies $\deg g(X) = 0$ or $\deg h(X) = 0$. Since the only polynomials of degree zero are the non-zero constants, this means that $g(X)$ or $h(X)$ is a non-zero constant.

Example 150 Let $\mathbb{F} = \mathbb{R}$, the real numbers. Then the polynomial $X^2 + 1 \in \mathbb{R}[X]$ is irreducible, since $X^2 + 1$ has no roots in $\mathbb{R}$. A factorization $X^2 + 1 = g(X) \cdot h(X)$ with either $g(X)$ or $h(X)$ of degree one, would give rise to a root of $X^2 + 1$ in $\mathbb{R}$.

Example 151 Let $\mathbb{F} = \mathbb{F}_5$, the finite field with 5 elements. Then the polynomial $X^2 + 1 \in \mathbb{F}_5[X]$ is reducible, since $X^2 + 1 = (X + 2)(X + 3)$ (check yourself).

Warning: a polynomial $f(X) \in \mathbb{F}[X]$ of degree 4 or larger can be reducible even though it does not have any roots in $\mathbb{F}$. For example the polynomial $X^4 + 2X^2 + 1 \in \mathbb{R}[X]$ does not have any roots in $\mathbb{R}$, but it is reducible since $X^4 + 2X^2 + 1 = (X^2 + 1)^2$. Irreducible polynomials are useful, as it turns out that they can be used to define maximal ideals:

Theorem 152 *Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X) \in \mathbb{F}[X]$ be a polynomial. Then the ideal $\langle f(X)\rangle = f(X)\mathbb{F}[X]$ generated by $f(X)$ is a maximal ideal if and only if $f(X)$ is an irreducible polynomial.*

Proof. First assume that $\langle f(X)\rangle$ is maximal. We wish to show that $f(X)$ is irreducible. Let us assume that $f(X)$ is not. Then we can find polynomials $g(X)$ and $h(X)$ such that $f(X) = g(X)h(X)$ and such that neither $g(X)$ nor $h(X)$ is a constant. This implies that the ideal $J := \langle g(X)\rangle$ contains $\langle f(X)\rangle$ (since $f(X)$ is a multiple of $g(X)$), but $I \neq J$ (since $\deg g(X) < \deg f(X)$). Also since $g(X)$ is not a constant, we get $1 \notin J$, implying that $J \neq \mathbb{F}[X]$. Hence $\langle f(X)\rangle$ is not a maximal ideal. This gives a contradiction and therefore the assumption that $f(X)$ was reducible cannot be valid. We conclude that $f(X)$ is irreducible.

Now assume that $f(X)$ is irreducible. We wish to show that $\langle f(X)\rangle$ is a maximal ideal. Let J be an ideal of $\mathbb{F}[X]$ that contains the ideal $\langle f(X)\rangle$. We know from Proposition 139 that there exists $d(X) \in \mathbb{F}[X]$ such that $J = \langle d(X)\rangle$. Since $f(X) \in \langle f(X)\rangle$ and $\langle f(X)\rangle \subseteq \langle d(X)\rangle$, we see that we can write $f(X)$ as a multiple of $d(X)$. That is to say, there exists a polynomial $q(X)$ such that $f(X) = q(X)d(X)$. Since $f(X)$ is irreducible, this implies that $\deg d(X) = \deg f(X)$ or $\deg d(X) = 0$ or in other words: $d(X) = \alpha f(X)$ for some $\alpha \in \mathbb{F}^*$ or $d(X) = \alpha \in \mathbb{F}^*$. If $d(X) = \alpha f(X)$, then $J = \langle d(X)\rangle = \langle f(X)\rangle$, while if $d(X) = \alpha$, then $\langle d(X)\rangle = \mathbb{F}[X]$. By the definition of a maximal ideal, we may conclude that $\langle f(X)\rangle$ is a maximal ideal. ■

The above theorem gives rise to an abundance of maximal ideals in polynomial rings. Using Theorem 147, we can construct fields using these maximal ideals. One such field we have already encountered in Example 141. There we considered the quotient ring $(\mathbb{R}[X]/\langle X^2+1\rangle, +, \cdot)$, which could be identified with the field of complex numbers $(\mathbb{C}, +, \cdot)$. In the next section we will encounter more fields that are very useful in several areas of discrete mathematics, but for now we give a different example:

Example 153 Let $\mathbb{F} = \mathbb{F}_2$ and let $f(X) = X^3 + X + 1 \in \mathbb{F}_2[X]$. First we show that $f(X)$ is irreducible in $\mathbb{F}_2[X]$. If the polynomial $f(X) = X^3 + X + 1$ would be reducible, say $f(X) = g(X)\cdot h(X)$ with $g(X), h(X) \in \mathbb{F}_2[X]$. Since $\deg g(X) + \deg h(X) = \deg f(X) = 3$, the polynomial $f(X)$ would have a factor of degree one. Therefore it would have a root in $\mathbb{F}_2$. However, since $f(0) = 1$ and $f(1) = 1$, neither 0 or 1 is actually a root of $f(X)$. We conclude that $f(X)$ cannot have a factor of degree one. But in that case it needs to be irreducible. Now we may conclude that the ring $(\mathbb{F}_2[X]/\langle X^3 + X + 1\rangle, +, \cdot)$ is a field.

In general, we see that if a polynomial $f(X) \in \mathbb{F}[X]$ is irreducible, we can construct a field $\mathbb{E} := \mathbb{F}[X]/\langle f(X)\rangle$. The original field $\mathbb{F}$ can be seen as a subset of $\mathbb{E}$ by identifying $a \in \mathbb{F}$ with $a + \langle f(X)\rangle$. One says that $\mathbb{F}$ is a subfield of $\mathbb{E}$ and conversely that $\mathbb{E}$ is an extension field of $\mathbb{F}$. We can therefore interpret a polynomial in $\mathbb{F}[X]$ as a polynomial with coefficients in $\mathbb{E}$. It turns out that $f(X)$ has a root in $\mathbb{E}$, when seen as a polynomial with coefficients in $\mathbb{E}$. We have seen this in one example where $f(X) = X^2 + 1$ and $\mathbb{F} = \mathbb{R}$. There we have seen that the quotient ring $(\mathbb{R}[X]/\langle X^2 + 1\rangle, +, \cdot)$ can be identified with the complex numbers. The polynomial $X^2 + 1$ has a root (in fact two roots) in $\mathbb{C}$. We now show the general case:

Lemma 154 *Let $f(X) \in \mathbb{F}[X]$ be an irreducible polynomial and define $\mathbb{E} := \mathbb{F}[X]/\langle f(X)\rangle$. Then the element $X + \langle f(X)\rangle \in \mathbb{E}$ is a root of $f(X)$ when seen as element of $\mathbb{E}[X]$.*

Proof. As remarked before, we can see $\mathbb{F}$ as a subset of $\mathbb{E}$ by identifying elements $a \in \mathbb{F}$ with the coset $a + \langle f(X)\rangle$. To avoid confusing elements from $\mathbb{E}$ we use T for the polynomial variable.

Then writing

$$f(T) = \sum_{i=0}^{d} a_i T^i \in \mathbb{F}[T],$$

we can write

$$f(T) = \sum_{i=0}^{d} (a_i + \langle f(X)\rangle) T^i \in \mathbb{E}[T].$$

We claim that $X + \langle f(X)\rangle$ is a root of $f(T)$. Indeed, we have

$$f(X + \langle f(X)\rangle) = \sum_{i=0}^{d} (a_i + \langle f(X)\rangle)(X + \langle f(X)\rangle)^i$$

$$= \sum_{i=0}^{d} a_i \cdot X^i + \langle f(X)\rangle = f(X) + \langle f(X)\rangle = 0 + \langle f(X)\rangle,$$

which is what we wanted to show. ■

The above lemma shows that given an irreducible polynomial $f(X)$ with coefficients in a field $\mathbb{F}$, one can construct an extension field $\mathbb{E}$ (that is to say, a field $\mathbb{E}$ containing $\mathbb{F}$) in which $f(X)$ has a root. Also seen in this light, it makes sense that $\mathbb{R}[X]/\langle X^2 + 1\rangle$ could be identified with $\mathbb{C}$. The field $\mathbb{R}[X]/\langle X^2 + 1\rangle$ namely has to be an extension field of $\mathbb{R}$ in which the polynomial $X^2 + 1$ has a root and $\mathbb{C}$ has exactly these properties.

8.2 Finite fields

Keywords: construction of finite fields

So far we have seen finite fields with p elements: $(\mathbb{F}_p, +, \cdot)$, also see Example 148. Another common notation for this field is GF(p). Here GF stands for Galois field, in honour of Évariste Galois, who was one of the first scientists to work with such fields.

We will now introduce more finite fields using the construction from the previous section. In fact we have already seen on such a field in Example 153. As we will see, it turns out that one can construct finite fields with $q = p^d$ elements for any prime number p and positive integer d. We will use the notation $\mathbb{F}_q$ for a field with q elements (another common notation is GF(q)). Warning: If $d > 1$, it is **not** true that $\mathbb{F}_{p^d} = \mathbb{Z} \bmod p^d$.

Theorem 155 *Let $f(X) \in \mathbb{F}_p[X]$ be an irreducible polynomial of degree d. Then the quotient ring $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$ is a finite field with p^d elements.*

Proof. By Theorem 152, the ideal $\langle f(X)\rangle \subset \mathbb{F}_p[X]$ is a maximal ideal. Therefore, by Theorem 147, we can conclude that $(\mathbb{F}_p[X]/\langle f(x)\rangle, +, \cdot)$ is a field. By Lemma 140, any element in $\mathbb{F}_p[X]/\langle f(x)\rangle$ can be described uniquely in the form $r(X) + \langle f(X)\rangle$ with $r(X) + \langle f(X)\rangle$ with $r(X) = a_0 + a_1 X + \cdots a_{d-1} X^{d-1}$ with $a_0, a_1, \ldots, a_{d-1} \in \mathbb{F}_p$. There are p^d possibilities for choosing $a_0, \ldots, a_{d-1}$ and hence exactly p^d elements in $\mathbb{F}_p[X]/\langle f(X)\rangle$. ■

One can show that for any prime p and positive integer d there exists an irreducible polynomial in $\mathbb{F}_p[X]$ of degree d. Proving this requires a little more theory than we have available at the moment, so we will not do that here. From the previous theorem, we see that we can construct a finite field with p^d elements for any prime p and positive integer d.

Example 156 We construct a finite field with 4 elements. We start by finding an irreducible polynomials in $\mathbb{F}_2[X]$ of degree two. A polynomial of degree two is reducible if and only if it has a factor of degree one. Therefore, a polynomial of degree two in $\mathbb{F}_2[X]$ is irreducible if and only if it has no roots in $\mathbb{F}_2$. Note that this also holds for degree three polynomials, but no longer for degree four or higher. Since the polynomial $X^2 + X + 1$ has no roots in $\mathbb{F}_2$, it is irreducible. We can now construct a finite field with four elements as the quotient ring $(\mathbb{F}_2[X]/\langle X^2+X+1\rangle, +, \cdot)$. Similarly, the field constructed in Example 153 is a finite field with 8 elements.

Finite fields have a very rich structure. One that is important for applications concerns the multiplicative structure of a finite field. We have seen that in general the units of a ring give rise to a group $(R^*, \cdot)$. In the case of a finite field much more can be said:

Theorem 157 *Let $(\mathbb{F}_q, +, \cdot)$ be a finite field with q elements. Then the group $(\mathbb{F}_q^*, \cdot)$ of units is a cyclic group.*

Proof. We need to find $c \in \mathbb{F}_q^*$ such that any element $h \in \mathbb{F}_q^*$ can be written as a power of c. That is to say: we need to find c such that

$$\mathbb{F}_q^* = \{1, c, c^2, \ldots, c^{q-2}\}.$$

Using the notion of the order of an element in a group, we can also say that we need to find $c \in \mathbb{F}_q^*$ such that $\mathrm{ord}(c) = q - 1$.

Now suppose that $h \in \mathbb{F}_q^*$ has order m. First of all we know from Proposition 62 that m divides $q - 1$. We will first count the possible number of elements of order m. We know that h is a root of the polynomial $X^m - 1$. In fact all elements $1, h, h^2, \ldots, h^{m-1}$ will be distinct roots of this polynomial. By Corollary 125 no other element of $\mathbb{F}_q^*$ can be a root of the polynomial $X^m - 1$. This means that all elements in $\mathbb{F}_q^*$ of order m are among the elements $1, h, h^2, \ldots, h^{m-1}$. On the other hand, if $\gcd(e, m) > 1$, then $\mathrm{ord}(h^e) = m/\gcd(e, m) < m$. Therefore among the elements $1, h, h^2, \ldots, h^{m-1}$, at most $\phi(m)$ elements of H have order m (where $\phi(m)$ is the Euler totient function from Corollary 63). This means that for any divisor m of $q-1$ there are at most $\phi(m)$ elements of order m. All in all we have shown that

$$q - 1 \leq \sum_{m \mid q-1} \phi(m).$$

On the other hand, for any natural number n, an element i between 0 and $n - 1$ has a certain greatest common divisor d with n and we have

$$\#\{i \mid 0 \leq i < n,\ \gcd(i, n) = d\} = \#\{j \mid 0 \leq j < n/d,\ \gcd(j, n/d) = 1\} = \phi(n/d)$$

This shows that

$$\sum_{d \mid n} \phi(d) = \sum_{d \mid n} \phi(n/d) = n.$$

Combining the above equations, we see that

$$q - 1 \leq \sum_{m \mid q-1} \phi(m) = q - 1.$$

Apparently, equality holds and for any m dividing $q - 1$ there exist elements in $\mathbb{F}_q^*$ of order m. In particular, there exists an element $c \in \mathbb{F}_q^*$ having order $q - 1$. ■

An element $c \in \mathbb{F}_q^*$ with multiplicative order $q - 1$ is called a *primitive element* of $\mathbb{F}_q$. Given such a primitive element c and an arbitrary element $h \in \mathbb{F}_q^*$, we now know that there exist $n \in \mathbb{Z}$

such that $c^n = h$. Finding n given c and h is in general a computationally hard problem called the discrete logarithm problem. The hardness of the discrete logarithm problem is the basis of applications of finite fields in cryptography. Also in other areas of discrete mathematics, finite fields are used with great success. For example the well-known Reed–Solomon codes, used to achieve reliable communication over a noisy channel, make extensive use of finite fields. These applications are the subject of other courses.

Example 158 We consider again the finite field with 4 elements constructed in the previous example. A primitive element is given by the element $c := X + \langle X^2 + X + 1\rangle$. Indeed, a direct calculation shows that $c^2 = c+1$ and $c^3 = 1$. Also c^2 is a primitive element. Indeed, $c^4 = c^3 \cdot c = c$ and $c^6 = c^3 \cdot c^3 = 1$.

8.3 Exercises

1. Is $(\mathbb{F}_3[X]/\langle X^3 + X + 1\rangle, +, \cdot)$ a field?
2. Check that $(\mathbb{F}_2[X]/\langle X^4 + X + 1\rangle)$ is a field and find a primitive element of it.
3. Give all the necessary ingredients to construct a finite field with 32 elements explicitly. Why is it easy in this field to find a primitive element?
4. Let $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$ be a finite field with $q = p^d$ elements and let $a, b \in \mathbb{F}_p[X]/\langle f(X)\rangle$. Show that $(a + b)^p = a^p + b^p$. You may assume the fact that p divides the binomial coefficients $\binom{p}{i}$ for $1 \leq i \leq p - 1$.
5. Let $(\mathbb{F}_2, +, \cdot)$ be the finite field with two elements and let us write $\mathbb{F}_2 = \{0, 1\}$. We define $R := \mathbb{F}_2[X]/\langle X^3\rangle$.
   (a) How many elements does R have?
   (b) Find a zero-divisor in the ring $(R, +, \cdot)$.
   (c) Find the multiplicative inverse of the element $X + 1 + \langle X^3\rangle \in R$.
   (d) How many elements does the set R^* contain? (Recall that R^* is the set of units in R).