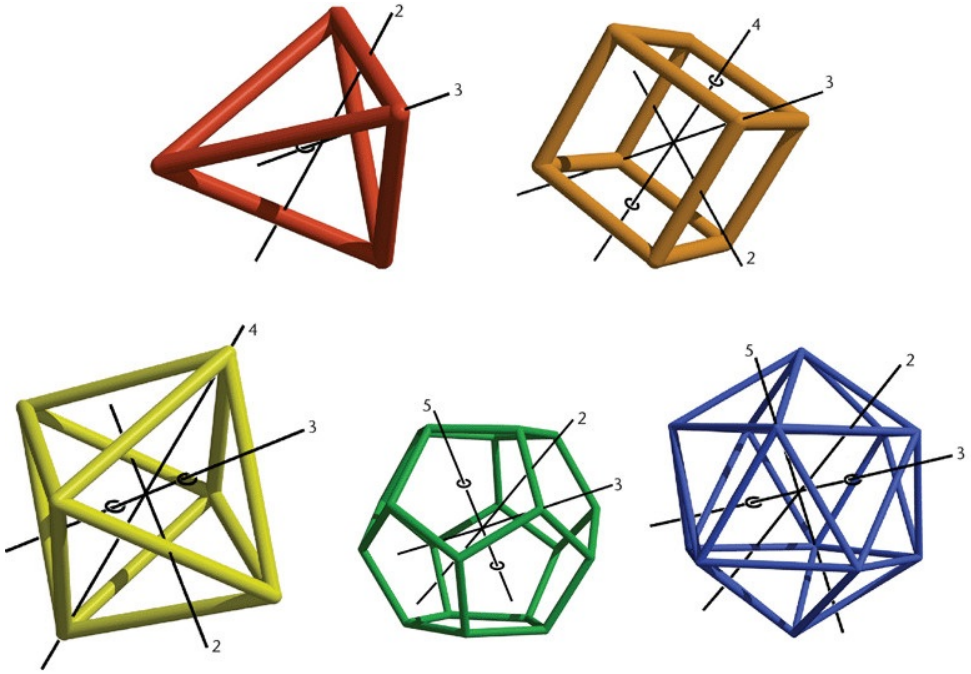
Course material for 01018



Peter Beelen and Maria Montanucci

Fall 2025

Contents

List of symbols

Chapter 1

Symbol	Symbol meaning
$\in$	is an element of
$\notin$	is not an element of
$\approx$	approximately equal to
$\implies$	implies
$\iff$	if and only if
$\forall$	for all
$\exists$	there exist(s)
$\exists!$	there exists a unique
$\neg$	not
∞	infinity
π	ratio of a circle's circumference to its diameter (180 degree)
$\wedge$	logical operator and
$\vee$	logical operator or
$f : A \to B$	f is a function from A to B
$f(a)$ or $f[a]$	image of the element a through f
$a \mapsto f[a]$	the function f maps a to $f[a]$
$\mathrm{im} f$	image of the function (or homomorphism) f
$\ker f$	kernel of the function (or homomorphism) f
f^{-1}	inverse of the function f
$g \circ f$	composition of the functions g and f (g after f)
$\{x \mid \lambda(x)\}$	set of all x such that the logical expression $\lambda(x)$ is true
$\subseteq$	is a subset of
$\supseteq$	contains
$\subsetneq$	is a proper subset of

$\supsetneq$	contains properly
$\not\subseteq$	is not a subset of
$\not\supseteq$	does not contain
$\emptyset$	empty set
$\cap$	set intersection
$\cup$	set union
$\cap_{i \in I} A_i$	intersection of the sets A_i where i is in the index set I
$\cup_{i \in I} A_i$	union of the sets A_i where i is in the index set I
$\setminus$	set difference
$\times$	Cartesian product
A^n	Cartesian product of the set A with itself n times
$\mathbb{N}$	set of natural numbers (zero included)
$\mathbb{Z}$	ring of integers
$\mathbb{Z}_{\geq a}$	set of integers which are larger than or equal to a
$n\mathbb{Z}$	set of integers that are multiples of n
$\mathbb{Q}$	field of rational numbers
$\mathbb{R}$	field of real numbers
$\mathbb{R}_{\geq a}$	set of real numbers which are larger than or equal to a
$\mathbb{C}$	field of complex numbers
$a \equiv b \pmod{n}$	a is congruent to b modulo n
$a \bmod n$	remainder of the division of a by n
$a \operatorname{quot} n$	quotient of the division of a by n
$a + n\mathbb{Z}$	congruence class of a modulo n
$\sim$	is related to/relation
$\not\sim$	is not related to
$[a]_\sim$	equivalence class of a with respect to the relation $\sim$
$\mathbb{Z}_n$	ring of integers modulo n $\{0, \ldots, n-1\}$
$+_n$	addition modulo n
$\cdot_n$	multiplication modulo n
$\gcd(a, b)$	greatest common divisor of a and b
$\lfloor \cdot \rfloor$	floor function
$\mathrm{Mat}_{m \times n}(\mathbb{R})$	set of $m \times n$ matrices with coefficients in $\mathbb{R}$
$\mathrm{Mat}(n, \mathbb{R})$	set of $n \times n$ matrices with coefficients in $\mathbb{R}$
A^{-1}	(multiplicative) inverse of the element A
$-A$	additive inverse of the element A

Chapter 2

Symbol	Symbol meaning
log	natural logarithm
e^x	exponential function
id_A	identity permutation on the set A
$\|A\|$ or $\#A$	cardinality of the set A
S_A	set of all permutations $f : A \to A$
S_n	symmetric group on n letters, i.e. S_A where $A = \{1, \ldots, n\}$
ord(f)	order of the element f
$(a_0 a_1 \ldots a_{m-1})$	m-cycle
$\prod_{i=a}^{b}$	product for i from a to b
$\sum_{i=a}^{b}$	summation for i from a to b
$n!$	n factorial, i.e. $\prod_{i=1}^{n} i$, if $n \geq 1$ and 1, if $n = 0$
$\sqrt{n}$	square root of n
M_f	permutation matrix of the permutation f
$M_{i,j}$	entry in the i-th row and j-th column of the matrix M
det(M)	determinant of the square matrix M
sign(f)	sign of the permutation f
$\lim_{n\to\infty} f(n)$	limit of the function $f(n)$ as n goes to infinity

Chapter 3

Symbol	Symbol meaning
(a, b)	generic element of the cartesian product $A \times B$ where $a \in A$ and $b \in B$
$(\mathbb{Z}_n)^*$	set of integers $k \in \{1, \ldots, n-1\}$ such that $\gcd(k, n) = 1$
R^*	set of units in the ring $(R, +, \cdot)$
$\phi(n)$	Euler's totient function $\phi(n) = \|\{a \in \mathbb{Z}_n : \ \gcd(a, n) = 1\}\|$
$GL(2, \mathbb{R})$	set (group) of all invertible 2×2 matrices with coefficients in $\mathbb{R}$
$SL(2, \mathbb{R})$	set (group) of all 2×2 matrices with coefficients in $\mathbb{R}$ and determinant 1
ord(g)	order of the element g in the group G
C_n	cyclic group $C_n = \{e, r, r^2, \ldots, r^{n-1}\}$ of order n
D_n	dihedral group $D_n = \{e, r, r^2, \ldots, r^{n-1}, s, rs, \ldots, r^{n-1}s\}$ of order $2n$
Q_8	quaternion group
A_n	alternating group on n letters $A_n = \{f \in S_n : \ \text{sign(f)} = 1\}$

Chapter 4

Symbol	Symbol meaning
$\langle g \rangle$	cyclic group generated by g $\langle g \rangle = \{g^i \mid i \in \mathbb{Z}\}$
$M \cdot N$	product of subsets M, N of a group $(G, \cdot)$, $M \cdot N = \{f \cdot g \mid f \in M,\ g \in N\}$
$f \cdot H$	left coset of the subgroup H in the group G by $f \in G$, $f \cdot H = \{f \cdot h \mid h \in H\}$
$H \cdot f$	right coset of the subgroup H in the group G by $f \in G$, $H \cdot f = \{h \cdot f \mid h \in H\}$
$\sim_H$	equivalence relation induced by a subgroup H: $f \sim_H g \iff f^{-1} \cdot g \in H$
${}_H\sim$	equivalence relation induced by a subgroup H : $f \sim_H g \iff g \cdot f^{-1} \in H$
$[G : H]$	index of the subgroup H of a group G

Chapter 5

Symbol	Symbol meaning
φ_g	given an action $\varphi : G \to S_A$, $\varphi_g = \varphi[g] \in S_A$
O_a	orbit of an element a in the action φ of a group G on a set A
G_a	stabilizer of an element a in the action φ of a group G on a set A
$\mathrm{Fix}(g)$	set of fixed elements of g in a group G with an action φ on a set A, $\mathrm{Fix}(g) = \{a \in A \mid \varphi_g[a] = a\}$

Chapter 6

Symbol	Symbol meaning
$Z_g(t_1, \ldots, t_{\#X})$	cycle index of the map s_g (see Definition 6.1.2)
$Z_G(t_1, \ldots, t_{\#X})$	cycle index of the group action . (see Definition 6.1.2)
$Z_{n_G}(t_1, \ldots, t_{\#X})$	normalized cycle index of the group action ., that is, $Z_G(t_1, \ldots, t_{\#X})/\lvert G\rvert$
C^X	set of all maps from X to C
w_c	weight of an element c in a set C (see Definition 6.2.1)
$f_C(f)$	weight generating function of a set C (see Definition 6.2.1)

Chapter 7

Symbol	Symbol meaning
$GL(n, \mathbb{R})$	set (group) of all invertible $n \times n$ matrices with coefficients in $\mathbb{R}$
I_n	$n \times n$ identity matrix
G/H	quotient group of G by the normal subgroup H, $G/H = \{g \cdot H \mid g \in G\}$

Chapter 8

Symbol	Symbol meaning
$+_R$	addition operation in the ring R
$\cdot_R$	multiplication operation in the ring R
0_R	identity element for the operation $+_R$
1_R	identity element for the operation $\cdot_R$
i	imaginary element in $\mathbb{C}$, $i^2 = -1$
$\mathbb{Z}[i]$	set of Gaussian integers $a + bi$, $a, b \in \mathbb{Z}$
$p(X)$	polynomial with coefficients in a ring R (and indeterminate X)
$\deg(p(X))$	degree of the polynomial $p(X)$
$R[X]$	set (ring) of polynomials with coefficients in a ring R
$p(a)$	evaluation of the polynomial $p(X) \in R[X]$ in the element $a \in R$
$\max\{a, b\}$	maximum of the values $a, b \in \mathbb{N}$
$(D, +, \cdot)$	integral domain
$q(X)$	quotient of a polynomial division
$r(X)$	remainder of a polynomial division
$\mathbb{Z}[\sqrt{2}]$	set of all elements $a + b\sqrt{2}$, $a, b \in \mathbb{Z}$
$\mathbb{H}$	ring of quaternions

Chapter 9

Symbol	Symbol meaning
I	ideal of a ring R
$I = \langle x \rangle = xR$	principal ideal generated by $x \in R$, $\langle x \rangle = \{x \cdot r \mid r \in R\}$
$I = \langle x_1, \ldots, x_n \rangle$	finitely generated ideal, $\langle x_1, \ldots, x_n \rangle = \{x_1 \cdot r_1 + \ldots x_n \cdot r_n \mid r_1, \ldots, r_n \in R\}$
$(\mathbb{F}, +, \cdot)$	field
$r + I$	coset of the ideal I by r in the ring R, $r + I = \{r + x \mid x \in I\}$
R/I	quotient ring of R by the ideal I
$\mathbb{F}_p$	for a prime p, the finite field with p elements $\mathbb{Z}_p$
$\mathbb{F}_q$ or $GF(q)$	for a prime power $a = p^n$, denotes a finite field with q elements
$I_d(p)$	number of distinct, monic and irreducible polynomials of degree d in $\mathbb{F}_p[X]$

Chapter 1

Equivalence relations

1.1 Prelude: about sets

Keywords: set, subset, intersection, union, partition, Cartesian product of sets.

The notion of a *relation* is very fundamental in mathematics and builds on the theory of sets. Therefore let us start by briefly introducing some terminology and notation concerning sets. Some of it, perhaps even most of it, is probably familiar to you already, but it is a good idea to make sure by reading this section anyway.

A lot could be said about what a set really is and how to define them. Indeed, a precise definition of what a set is and what properties it has, would take us to the very foundation of mathematics, such as the Zermelo–Fraenkel axioms. However, for the purposes of these notes, we do not need to go to such lengths and an intuitive description of sets with some of its basic properties will suffice.

Basically, a set A consists of elements and stating that a is an element of A is expressed as: $a \in A$. Some authors prefer to write the set first and then the element, writing $A \ni a$ instead of $a \in A$. If an element a is not in the set A, one writes $a \notin A$ or alternatively $A \not\ni a$. A set is determined by its elements, meaning that two sets A and B are equal, $A = B$, if and only if they contain the same elements, $\forall a\,(a \in A \iff a \in B)$. If A and B are sets, then B is called a *subset* of A, if any element of B is also an element of A. A common notation for this is $B \subseteq A$. In other words, the statement $B \subseteq A$ is by definition true if and only if the statement $\forall a\ a \in B \Rightarrow a \in A$ is true. In particular $A \subseteq A$, since for all a the implication $a \in A \Rightarrow a \in A$ is true. The *empty set* is the set not containing any elements at all. It is commonly denoted by $\emptyset$, inspired by the letter Ø in the Danish and Norwegian alphabet. Some authors use $\{\}$ for the empty set, but we will always use the notation $\emptyset$ for it. Since $\emptyset$ does not contain any element, the logical statement $a \in \emptyset$ is false for all a. This implies that the implication $a \in \emptyset \Rightarrow a \in A$ is true for all a and any set A. Hence $\emptyset \subseteq A$ is true for any set A. Instead of writing $B \subseteq A$, one may also write $A \supseteq B$. A common way to construct subsets of a set A is by defining a logical expression that for any $a \in A$ is true or false. For the sake of notation, let us denote this logical expression by $\lambda(a)$. Then $\{a \in A \mid \lambda(a)\}$ denotes the subset of A consisting of precisely those elements $a \in A$ for which the logical expression $\lambda(a)$ is true. For example, if $\mathbb{N} = \{0, 1, 2, 3, \dots\}$, the set of natural numbers, then $\{a \in \mathbb{N} \mid 1 \leq a \leq 3\} = \{1, 2, 3\}$.

If one wants to stress that a set B is a subset of A, but not equal to all of A, one writes $B \subsetneq A$ or alternatively $A \supsetneq B$. Finally, if you want to express in a formula that B is not a subset of A, it is possible to use the logical negation symbol $\neg$ and write that $\neg(B \subseteq A)$, but it is more customary to write $B \not\subseteq A$ or alternatively $A \not\supseteq B$. We now state a helpful result that

we will use quite often without explicitly mentioning it.

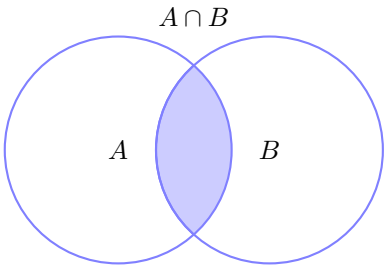
Lemma 1.1.1 *Let A and B be two sets. Then $A = B$ if and only if $A \subseteq B$ and $A \supseteq B$.*

Proof. The statement $A = B$, is logically equivalent to the statement $\forall a\ (a \in A \Leftrightarrow a \in B)$. Splitting the biimplication up in two implications, we can reformulate this as: $\forall a\ (a \in A \Rightarrow a \in B)$ and $\forall a\ (a \in A \Leftarrow a \in B)$. But then we can equivalently state that $A \subseteq B$ and $A \supseteq B$. ■
This lemma will often be used in proving that two particular sets A and B are equal. Instead of somehow proving $A = B$ directly, the lemma implies that it is equivalent to show that $A \subseteq B$ and $A \supseteq B$.

Warning 1.1.2 *Instead of $\subseteq$ and $\supseteq$, some authors prefer the symbols $\subset$ and $\supset$. However, yet other authors, use the symbols $\subset$ and $\supset$ in the meaning of $\subsetneq$ and $\supsetneq$, inspired by a similar use of the inequality ($\leq$ and $\geq$) and strict inequality ($<$ and $>$) symbols. To avoid confusion, we will from now on never use the symbols $\subset$ or $\supset$ again.*

There are several basic definitions and operations involving sets that we will use later on. The definitions can be a bit formal and are therefore illustrated in Example 1.1.4. If A and B are two sets, we define the *intersection* of A and B, denoted by $A \cap B$, to be the set consisting of all elements that are both in A and in B. In other words:

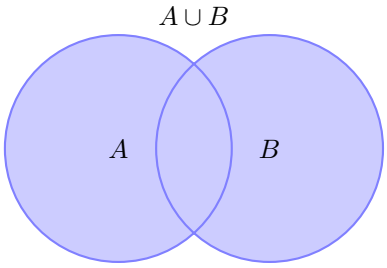
$$A \cap B := \{a \mid a \in A \text{ and } a \in B\}.$$



Two sets A and B are called *disjoint*, if $A \cap B = \emptyset$.

The *union* of A and B is defined as:

$$A \cup B := \{a \mid a \in A \text{ or } a \in B\}.$$



The union $A \cup B$ is called a *disjoint union* if $A \cap B = \emptyset$.

Similarly, for any positive integer n, one can define the intersection and union of any finite number of sets $A_1, \ldots, A_n$:

$$\cup_{i=1}^{n} A_i := \{a \mid \exists i \in \{1, \ldots, n\}\, a \in A_i\} \quad \text{and} \quad \cap_{i=1}^{n} A_i := \{a \mid \forall i \in \{1, \ldots, n\}\, a \in A_i\}.$$

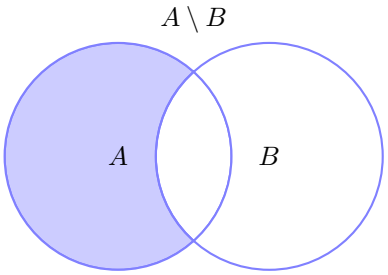
This can even be generalized further to (possibly) infinite families of sets. Given a family of sets $A_i, i \in I$, where I is some set containing all the used indices (often called an *index set*), we can define

$$\cup_{i \in I} A_i := \{a \mid \exists i \in I\, a \in A_i\} \quad \text{and} \quad \cap_{i \in I} A_i := \{a \mid \forall i \in I\, a \in A_i\}$$

The case of finitely many sets $A_1, \ldots, A_n$ is captured by choosing $I = \{1, \ldots, n\}$. A family of sets $A_i, i \in I$ is called a *partition* of A, if $A = \cup_{i \in I} A_i$, none of the sets A_i is empty, and for any distinct $i, j \in I$ the sets A_i and A_j are disjoint. In other words: a family of sets $A_i, i \in I$ is a partition of A, if for each $a \in A$, there exists exactly one set A_i in the family containing a.

The *set difference* of A and B, often pronounced as A minus B, is defined to be:

$$A \setminus B := \{a \mid a \in A \text{ and } a \notin B\}.$$



Finally, the *Cartesian product* of A and B is the set:

$$A \times B := \{(a, b) \mid a \in A \text{ and } b \in B\}.$$

In other words, the Cartesian product of two sets A and B, is simply the set of all pairs (a, b), whose first coordinate is from A and whose second coordinate is from B. The Cartesian product of a set A with itself is sometimes denote as A^2, in other words $A^2 = A \times A$.

Later on we will mainly use the Cartesian product of two sets, but it is not hard to define the Cartesian product of more than two sets. One simply uses more coordinates, one for each set in the Cartesian product. For example $A \times B \times C = \{(a, b, c) \mid a \in A, b \in B, \text{ and } c \in C\}$. More generally, if n is a positive integer and $A_1, \ldots, A_n$ are sets, then

$$A_1 \times \cdots \times A_n = \{(a_1, \ldots, a_n) \mid a_1 \in A_1, \ldots, a_n \in A_n\}.$$

If all sets are equal, say $A_1 = A, \ldots, A_n = A$, then one often writes A^n for their Cartesian product.

Aside 1.1.3 *The Cartesian product is named after René Descartes (latinized to Renatus Cartesius) who lived 1596–1650. Among his contributions to mathematics is the introduction of the Cartesian coordinate system. Denoting by $\mathbb{R}$ the set of real numbers, his coordinate system for the plane is nothing but the set $\mathbb{R} \times \mathbb{R}$, the Cartesian product of $\mathbb{R}$ and $\mathbb{R}$. Descartes is also famous in philosophy. The statement "cogito ergo sum" ("I think, therefore I am") is from him.*

Let us illustrate the introduced concepts for sets in an example.

Example 1.1.4 Let $1, 2, 3$, and 4 be the first four positive integers. Then:

1. $\{1, 2\} \subseteq \{1, 2, 3\}$,
2. $\{1, 2\} \supseteq \{2\}$,
3. $\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}$,
4. $\{1, 2\}$ and $\{3\}$ are disjoint sets,
5. $\{1, 2, 3\} \cup \{2, 3, 4\} = \{1, 2, 3, 4\}$,
6. $\{1, 2, 3, 4\}$ is the disjoint union of $\{1, 2\}$ and $\{3, 4\}$,
7. The family of sets $\{1, 3\}$, $\{2\}$, $\{4\}$ form a partition of $\{1, 2, 3, 4\}$,
8. $\{1, 2, 3\} \setminus \{2, 3, 4\} = \{1\}$,
9. $\{2, 3, 4\} \setminus \{1, 2, 3, 4\} = \emptyset$,
10. $\{1, 2\} \times \{2, 3\} = \{(1, 2), (1, 3), (2, 2), (2, 3)\}$.

1.2 Some properties of congruences modulo an integer

Keywords: congruences modulo an integer, congruence classes, representatives.

In this section we recall what it means to work modulo an integer n. For good measure, let us first introduce some notation involving integers. First of all, with $\mathbb{N} = \{0, 1, 2, \dots\}$ one denotes the set of *natural numbers*, while $\mathbb{Z} = \{\dots, -2, -1, 0, 1, 2, \dots\}$ denotes the set of *integers*. Be warned that in some books, 0 is not included in the set $\mathbb{N}$. If we want to avoid any doubt, we will write $\mathbb{Z}_{\geq 0} = \{0, 1, 2, \dots\}$ and more general $\mathbb{Z}_{\geq a} = \{a, a + 1, \dots\}$. Other commonly used "blackboard font" symbols are $\mathbb{Q}$ for the set of *rational numbers* (all fractions a/b with $a \in \mathbb{Z}$ and $b \in \mathbb{Z} \setminus \{0\}$), $\mathbb{R}$ for the set of *real numbers*, and $\mathbb{C}$ for the set of *complex numbers*.

Returning to working modulo an integer, let us define what this entails:

Definition 1.2.1 *Let $n \in \mathbb{Z}$ be an integer. Given $a, b \in \mathbb{Z}$, we say that a and b are congruent modulo n if $a - b$ is a multiple of n. In other words, a and b are congruent modulo n precisely if there exists an integer k such that $a - b = k \cdot n$. A commonly used notation for the statement that a and b are congruent modulo n is the following:*

$$a \equiv b \pmod{n}.$$

For example, one has $-101 \equiv -11 \pmod{10}$, since $-101 - (-11) = -90 = -9 \cdot 10$.

Now we will for a fixed a and nonzero n, determine the set of all integers congruent to a modulo n.

Definition 1.2.2 *For $a, n \in \mathbb{Z}$, we define*

$$a + n\mathbb{Z} := \{a + k \cdot n \mid k \in \mathbb{Z}\}.$$

Lemma 1.2.3 *Let $a, b, n \in \mathbb{Z}$ be given. Then $b \in a + n\mathbb{Z}$ if and only if $a \equiv b \pmod{n}$.*

Proof. If $b \in a+n\mathbb{Z}$, then there exists $k \in \mathbb{Z}$ such that $b = a+k\cdot n$. This implies that $b-a = k\cdot n$ and hence that $a \equiv b \pmod{n}$. Conversely, if $a \equiv b \pmod{n}$, then there exists $k \in \mathbb{Z}$ such that $a - b = k \cdot n$. This implies that $b = a - k \cdot n = a + (-k) \cdot n$ and hence that $b \in a + n \cdot Z$. ■

In words, the set $a + n\mathbb{Z}$ consists precisely of those integers that are congruent to a modulo n. For this reason, $a + n\mathbb{Z}$ is called the *congruence class* of a modulo n. Any element from a congruence class $a + n\mathbb{Z}$ is called a *representative* of the class $a + n\mathbb{Z}$. In particular, $a \in a + n\mathbb{Z}$, so that a is always a representative of the congruence class $a + n\mathbb{Z}$.

For example the congruence class $0 + 2\mathbb{Z}$ consists of all even integers and 0, 12 are possible representative of this class. The congruence class $1 + 2\mathbb{Z}$ consists of all odd integers and -1 is a possible representative of that class. Since congruence classes $a + n\mathbb{Z}$ with $n \neq 0$, contain infinitely many integers, one cannot write down all of its elements. Typically, one therefore writes for example

$$1 + 5\mathbb{Z} = \{\ldots, -14, -9, -4, 1, 6, 11, 16, \ldots\}.$$

Any of the seven listed integers is a possible representative of the congruence class $1 + 5\mathbb{Z}$. The other congruence classes modulo 5 can similarly be listed. See Figure 1.1 for an illustration.

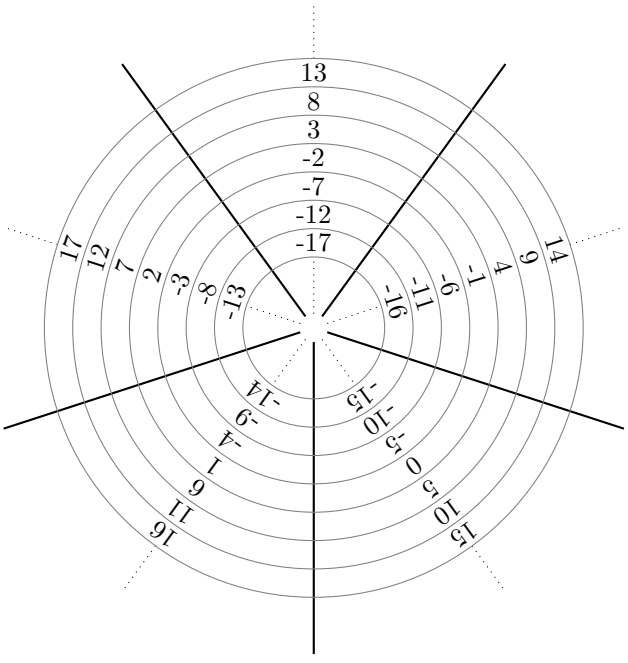


Figure 1.1: Each segment contains a congruence class modulo 5.

At this point it would be natural to show various properties of congruence classes, for example that if r is a representative of $a + n\mathbb{Z}$, then $r + n\mathbb{Z} = a + n\mathbb{Z}$. From the examples, it also seems reasonable to state that two congruence classes $a+n\mathbb{Z}$ and $b+n\mathbb{Z}$ are either identical, or disjoint (have no elements in common). However, it turns out to be more convenient to approach these statements from another point of view. At first sight, this approach will seem more abstract and more general than needed, but later in these notes we will reap the benefits of this generality.

Aside 1.2.4 *The notation $a \equiv b \pmod{n}$ was introduced by the German mathematician and physicist Carl Friedrich Gauss in his mathematical diary 'Disquisitiones Arithmeticae' (arithmetical investigations), which he wrote when he was 21. Gauss was an incredibly productive scientist who made profound contributions to various areas of mathematics and natural sciences of his time.*

1.3 Equivalence relations

Keywords: equivalence relation, equivalence class, representative.

One very useful tool that occurs very often in mathematics is the notion of an *equivalence relation*. Let us briefly define what a relation is first. Let us start with an example: if $A = \mathbb{R}$, the set of real numbers, 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 under $\leq$, namely all elements in $\mathbb{R}$ less than or equal to 2. Some elements are not related to 2 under $\leq$, namely all elements of $\mathbb{R}$ strictly greater than 2. Basically, to describe the relation $\leq$ on $\mathbb{R}$, all we need to know is for which pairs (a, b) of real numbers $a \leq b$ is true.

In general for a relation $\sim$ on a set A, we require a similar property: for any pair $(a, b) \in A \times A$ either a is related to b (in which case we say that $a \sim b$ is true, or simply that $a \sim b$) or a is not related to b (in which case we say that $a \sim b$ is false or simply that $a \not\sim b$). In other words, formally speaking, a relation on A is just a way to assign to each pair $(a, b) \in A \times A$, the value true or false. Therefore we can describe a relation $\sim$ completely using the set

$$R := \{(a, b) \in A \times A \mid a \sim b\}.$$

Indeed, R is the subset of the Cartesian product $A \times A$ consisting precisely of all pairs (a, b) for which $a \sim b$ is true. Going back to the example where $A = \mathbb{R}$, the set of real numbers, and $\sim$ is the inequality relation $\leq$, we see that $R = \{(a, b) \in \mathbb{R} \times \mathbb{R} \mid a \leq b\}$.

We can also turn this description around and use a subset R of $A \times A$ to define a relation on A. Indeed, given $R \subseteq A \times A$, we can define a relation $\sim$ by stating that $a \sim b$ is true if and only if $(a, b) \in R$. For this reason, many books formally define a relation on A as a subset R of $A \times A$. For example, if $A = \mathbb{R}$ and $R = \{(a, b) \in \mathbb{R} \times \mathbb{R} \mid a = b\}$, then the resulting relation satisfies that $a \sim b$ precisely if $a = b$. Hence in this example, the relation $\sim$ is the usual equality $=$.

Strictly speaking we have discussed what is known as a *binary relation* on A. A ternary relation on A would be a subset of $A \times A \times A$ and similarly, one can define for any positive integer n, an n-ary relation on A as a subset of $A \times \cdots \times A$, the n-fold Cartesian product of

A with itself. However, in the remainder of these notes, we will only need binary relations. Whenever we use the word relation, we will mean a binary relation.

Relations on A defined using a subset $R \subseteq A \times A$ are very general. We now look at a special kind of relations satisfying several additional properties:

Definition 1.3.1 *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).*

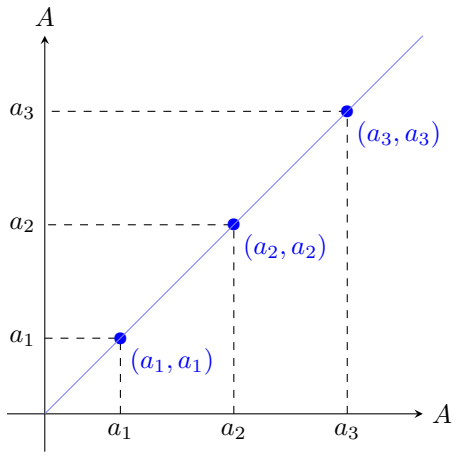


Figure 1.2: An equivalence relation $R \subseteq A \times A$ is reflexive.

Note that the relation $\leq$ on the real numbers $\mathbb{R}$ 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.

The congruence relation modulo n defined in Definition 1.2.1 is also an equivalence relation. Showing that this indeed holds, is an exercise.

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:

$$[a]_\sim := \{b \in A \mid a \sim b\}. \tag{1.1}$$

An element $r \in [a]_\sim$ is called a *representative* of the equivalence class $[a]_\sim$. Since any equivalence relation is symmetric, we know that $a \sim b$ if and only if $b \sim a$. Therefore it always holds that $[a]_\sim := \{b \in A \mid b \sim a\}$.

As an example of equivalence classes, let $A = \mathbb{R}$ and use the usual $=$ as equivalence relation. Then a is only related to itself, and hence $[a]_=$ is simply $\{a\}$. As a second example, we consider the congruence relation modulo n from Definition 1.2.1:

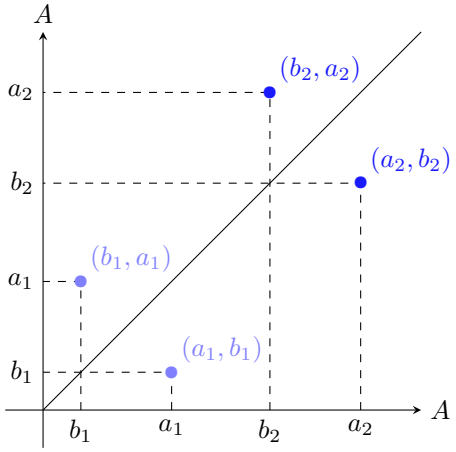


Figure 1.3: An equivalence relation $R \subseteq A \times A$ is symmetric.

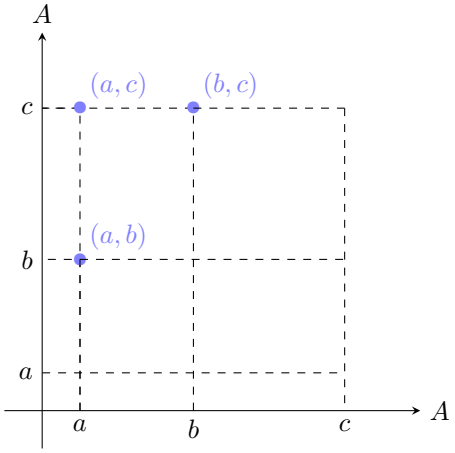


Figure 1.4: An equivalence relation $R \subseteq A \times A$ is transitive.

Example 1.3.2 Let $A = \mathbb{Z}$, the set of integers. Let n be a positive integer. As before, for $a, b \in \mathbb{Z}$ we say that a and b are congruent modulo n, in notation $a \equiv b \pmod{n}$, if n divides $a - b$. Then according to Lemma 1.2.3, the equivalence class of an integer a under the congruent modulo n relation, is precisely the congruence class $a + n\mathbb{Z} := \{a + kn \mid k \in \mathbb{Z}\}$.

Equivalence classes have several nice properties. We collect some of them in the following theorem.

Theorem 1.3.3 *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 the equivalence classes: $\cup_{a \in A}[a]_\sim = A$.*

3. *For any* $a, b \in A$ *we have that either* $[a]_\sim \cap [b]_\sim = \emptyset$ *or* $[a]_\sim = [b]_\sim$.

4. *For any* $a, b \in A$ *we have that* $[a]_\sim = [b]_\sim$ *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 at least one equivalence class, namely 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 $[a]_\sim \subseteq [b]_\sim$ and $[b]_\sim \subseteq [a]_\sim$.

$[a]_\sim \subseteq [b]_\sim$: 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.

$[b]_\sim \subseteq [a]_\sim$: 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 $[a]_\sim = [b]_\sim$. ■

We see that the equivalence classes form a partition of A, see Section 1.1. One also says that the set A is *partitioned* into equivalence classes. The word partitioned (meaning divided into parts) is appropriate, since the set A is divided into mutually disjoint subsets, namely the various equivalence classes. One can think of this as dividing a cake into (not necessarily equal sized) pieces, each piece being an equivalence class (see Figure 1.5).

For the equivalence relation "congruence modulo n", the above theorem specializes to the following:

Corollary 1.3.4 *Let* a, n *be integers. Then we have:*

1. *For any* $a \in \mathbb{Z}$ *we have* $a \in a + n\mathbb{Z}$.

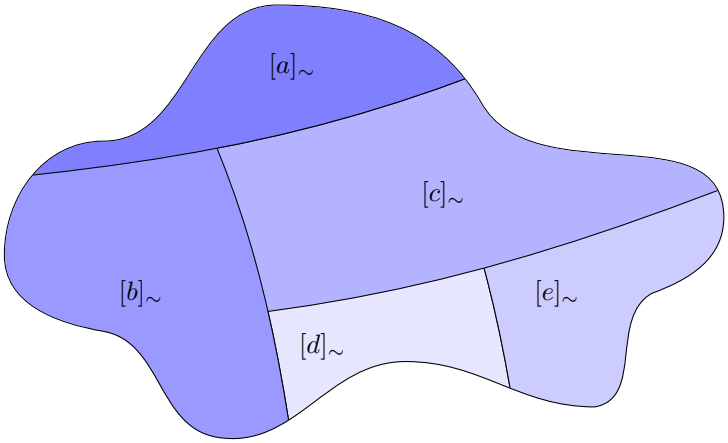


Figure 1.5: Equivalence classes form a partition of the set A.

2. *The set of integers $\mathbb{Z}$ is covered by the equivalence classes: $\cup_{a\in\mathbb{Z}}(a+n\mathbb{Z})=\mathbb{Z}$.*

3. *For any $a,b\in\mathbb{Z}$ we have $(a+n\mathbb{Z})\cap(b+n\mathbb{Z})=\emptyset$ or $a+n\mathbb{Z}=b+n\mathbb{Z}$.*

4. *For any $a,b\in\mathbb{Z}$ we have $a+n\mathbb{Z}=b+n\mathbb{Z}$ if and only if $a\equiv b \pmod n$.*

This summarizes the main properties of the congruence classes themselves. We now study special representatives of the congruence classes. To describe them, we use division with remainder for integers.

Fact 1.3.5 (Division with remainder) *Let $a,n\in\mathbb{Z}$ and assume that $n>0$. Then there exists precisely one pair of integers $(q,r)\in\mathbb{Z}^2$, satisfying that*

1. $a=q\cdot n+r$,

2. $0\le r\le n-1$.

The value q is called the *quotient* of the division of a by n, while r is called the *remainder* of the division. A common notation for the remainder is $a \bmod n$ and we will use this notation in these notes. Some books also use a notation like a quot n for the quotient of the division of a by n, but we will not use this very often. For example, since $-101=-11\cdot 10+9$, we have -101 quot $10=-11$ and $-101 \bmod 10=9$. We will not prove Fact 1.3.5, but see Exercise 21.

Practically by definition of the remainder, we obtain the following:

Lemma 1.3.6 *Let $a,n\in\mathbb{Z}$ and assume that $n>0$. Then*

$$a\equiv (a \bmod n) \pmod n.$$

Proof. All we need to show is that $a - (a \bmod n)$ is a multiple of n. However, the division algorithm guarantees this, since $a = (a \text{ quot } n) \cdot n + (a \bmod n)$. ■

In view of Lemma 1.2.3, this means that $a \bmod n$ is a representative of the congruence class $a + n\mathbb{Z}$. It is called the *standard representative* of the congruence class $a + n\mathbb{Z}$. The word "the" suggests that it is unique in some way. This is indeed the case as we show now.

Theorem 1.3.7 *Let n be a positive integer and a an arbitrary integer. Then the only representative of the congruence class $a + n\mathbb{Z}$ in $\{0, 1, \ldots, n-1\}$ is $a \bmod n$. In particular, $a + n\mathbb{Z} = (a \bmod n) + n\mathbb{Z}$ and there are exactly n different congruence classes modulo n, namely $0 + n\mathbb{Z}, 1 + n\mathbb{Z}, \ldots, n - 1 + n\mathbb{Z}$.*

Proof. Let $a \in \mathbb{Z}$. Since $a \equiv (a \bmod n) \pmod n$ by Lemma 1.3.6, item 4 from Corollary 1.3.4 implies that $a \bmod n + n\mathbb{Z} = a + n\mathbb{Z}$.

Now suppose that $a + n\mathbb{Z} = b + n\mathbb{Z}$ for $b \in \mathbb{Z}$ satisfying $0 \leq b < n$. We wish to show that $b = a \bmod n$. Since $a \bmod n + n\mathbb{Z} = a + n\mathbb{Z} = b + n\mathbb{Z}$, item 4 of Corollary 1.3.4 implies $b \equiv a \bmod n \pmod n$. This implies that $b - (a \bmod n)$ is a multiple of n. However, since $a \bmod n$ and b both lie between 0 and $n-1$, we have $-n+1 \leq b-(a \bmod n) \leq n-1$. Hence $b-(a \bmod n)$ can only be a a multiple of n if $b = a \bmod n$.

Finally, since $0 \leq a \bmod n < n$ and the union of the congruence classes $0+n\mathbb{Z}, 1+n\mathbb{Z}, \ldots, n-1+n\mathbb{Z}$ equals $\mathbb{Z}$, we conclude that there are precisely n different congruence classes modulo n. ■

For a given positive n, the various congruence classes modulo n, have standard representatives $0,\ 1, \ldots, n-1$. Therefore the set $\mathbb{Z}_n := \{0, \ldots, n-1\}$ is in this context called the set of standard representatives. Using standard representatives, one can prove several useful facts about remainders modulo n. Two central ones involving addition and multiplication, we collect in the following corollary.

Corollary 1.3.8 *Let $a, b, n \in \mathbb{Z}$ and assume that n is positive. then*

$$(a + b) \bmod n = ((a \bmod n) + (b \bmod n)) \bmod n$$

and

$$(a \cdot b) \bmod n = ((a \bmod n) \cdot (b \bmod n)) \bmod n.$$

Proof. We know by Theorem 1.3.7 that $(a + b) \bmod n$ is the standard representative of the congruence class $(a + b) + n\mathbb{Z}$. However, using Lemma 1.3.6, we see that

$$((a \bmod n) + (b \bmod n)) \bmod n \equiv (a \bmod n) + (b \bmod n) \equiv a + b \pmod n.$$

Hence $((a \bmod n) + (b \bmod n)) \bmod n$ is a representative of $a + b + n\mathbb{Z}$. Since $((a \bmod n) + (b \bmod n)) \bmod n$ is in $\mathbb{Z}_n$, it is in fact the standard representative of $a + b + n\mathbb{Z}$. We conclude that $(a + b) \bmod n = ((a \bmod n) + (b \bmod n)) \bmod n$. The identity involving multiplication can be shown in a similar way. ■

In the next section, we will use Corollary 1.3.8 to study addition and multiplication modulo n, but for now we illustrate its use for computing modular exponentiations, that is to say, expressions of the form $a^e \bmod n$.

Example 1.3.9 In this example, we compute $37^{60} \bmod 1001$. The direct approach, first calculating 37^{60}, then computing its remainder modulo 1001 is not very efficient. Indeed, one can compute that the number 37^{60} itself contains 95 decimal digits. While still manageable using a computer, larger exponents would quickly lead to impossibly large number to deal with even for a computer.

The trick is to use Corollary 1.3.8 in such a way that dealing with very large integers becomes unnecessary. First of all, we write the exponent in base two: $60 = 2^5 + 2^4 + 2^3 + 2^2$. This implies that

$$37^{60} = 37^{2^5} \cdot 37^{2^4} \cdot 37^{2^3} \cdot 37^{2^2}.$$

Now $(37 \cdot 37) \bmod 1001 = 37^2 \bmod 1001 = 368$. Hence, by Corollary 1.3.8, we have

$$37^{2^2} \bmod 1001 = (37^2 \bmod 1001)^2 \bmod 1001 = 368^2 \bmod 1001 = 289.$$

Similarly, we can deduce that:

$$\begin{array}{lcl} 37^{2^3} \bmod 1001 & = & 289^2 \bmod 1001 \;=\; 438, \\ 37^{2^4} \bmod 1001 & = & 438^2 \bmod 1001 \;=\; 653, \\ 37^{2^5} \bmod 1001 & = & 653^2 \bmod 1001 \;=\; 984. \end{array}$$

Now to compute $37^{60} \bmod 1001$, we use the above calculations and, again, Corollary 1.3.8 to conclude that:

$$37^{60} \bmod 1001 = 37^{2^5} \cdot 37^{2^4} \cdot 37^{2^3} \cdot 37^{2^2} \bmod 1001 = 984 \cdot 653 \cdot 438 \cdot 289 \bmod 1001.$$

Finally, we obtain that

$$37^{60} \bmod 1001 = ((984 \cdot 653 \bmod 1001) \cdot (438 \cdot 289 \bmod 1001)) \bmod 1001 = 911 \cdot 456 \bmod 1001 = 1.$$

The point of this approach is that in order to compute $37^{60} \bmod 1001$, we did not have to multiply integers with each other larger than 1000, computing the remainder modulo 1001 after each product or square.

The outlined algorithm for computing $37^{60} \bmod 1001$, has the disadvantage that one needs to store the values of $37^{2^i} \bmod 1001$ for $i = 1, 2, 3, 4, 5$ before being able to compute the final result. However, this was only done to ease the description and could easily have been avoided using small variations. The interested reader is encouraged to consult the literature on repeated squaring algorithms for modular exponentiation.

Aside 1.3.10 *Given any set A and some equivalence relation $\sim$ on it, one may wonder if it is always possible to find a subset C of A, containing precisely one representative for each equivalence class. If $A = \mathbb{Z}$ and $\sim$ is congruence modulo n, for some positive integer n, we can simply choose $C = \mathbb{Z}_n = \{0, 1, \ldots, n-1\}$, the set of standard representatives. To show the existence of a*

similar set C for an arbitrary equivalence relation on a set A, it turns out one needs the famous axiom of choice (in fact it is equivalent to it). The axiom of choice states that for every set $\mathcal{A}$ of nonempty sets there exists a function $f : \mathcal{A} \to \cup_{A\in\mathcal{A}} A$ such that for all $A \in \mathcal{A}$, $f(A) \in A$. The function f is called a choice function. Applying this to the set of all equivalence classes, we see that the image of f is exactly the subset C of A that we were looking for. The axiom of choice was formulated in 1904 by Ernst Zermelo as a part of the now standard axiomatic setup of the foundation of mathematics.

1.4 Modular arithmetic

Keywords: modular arithmetic, adding and multiplying modulo n.

Now we arrive at a main topic of this chapter: modular arithmetic, that is to say, addition and multiplication modulo n, where n denotes a positive integer. As before, we will use the notation $\mathbb{Z}_n := \{0, \ldots, n-1\}$ for the set of standard representatives of the n congruence classes modulo n.

Definition 1.4.1 *Let n be a positive integer and choose $a, b \in \mathbb{Z}_n$ arbitrarily. Then we define*

$$a +_n b := (a + b) \bmod n \quad \textit{and} \quad a \cdot_n b := (a \cdot b) \bmod n.$$

The operation $+_n$ is called addition modulo n, while $\cdot_n$ is called multiplication modulo n.

Example 1.4.2 We have $7 +_{10} 9 = 16 \bmod 10 = 6$ and $8 \cdot_{10} 9 = 72 \bmod 10 = 2$.

In view of Theorem 1.3.7, we can reformulate multiplication and addition modulo n in the following way: for $a, b \in \mathbb{Z}_n$, multiply (or add) a and b in the usual way as integers, then find the standard representative of the congruence class $ab+n\mathbb{Z}$ in case of multiplication and of $a+b+n\mathbb{Z}$ in case of addition.

Addition and multiplication modulo n have several intuitive properties that often are used without even mentioning them. Let us mention a few of these explicitly in the following theorem:

Theorem 1.4.3 *Let n be a positive integer and choose $a, b, c \in \mathbb{Z}_n$ arbitrarily. Then*

1. $a +_n b = b +_n a$,
2. $(a +_n b) +_n c = a +_n (b +_n c)$,
3. $a \cdot_n b = b \cdot_n a$,
4. $(a \cdot_n b) \cdot_n c = a \cdot_n (b \cdot_n c)$,
5. $a \cdot_n (b +_n c) = a \cdot_n b +_n a \cdot_n c$.

Proof. Items 1 and 3 follow directly:

$a +_n b = (a+b) \bmod n = (b+a) \bmod n = b +_n a$ and $a \cdot_n b = (a \cdot b) \bmod n = (b \cdot a) \bmod n = b \cdot_n a$,

using that $a + b = b + a$ and $a \cdot b = b \cdot a$ for any two integers a and b.

For the remaining items, Corollary 1.3.8 should be applied in an appropriate way. First of all, if $a \in \mathbb{Z}_n$, then $a \bmod n = a$. Then for example for Item 2, one can argue as follows:

$$\begin{aligned}
(a +_n b) +_n c &= ((a +_n b) + c) \bmod n \\
&= ((a + b) \bmod n + (c \bmod n)) \bmod n \\
&= ((a + b) + c) \bmod n \\
&= (a + (b + c)) \bmod n \\
&= ((a \bmod n) + (b + c \bmod n)) \bmod n \\
&= (a + (b +_n c)) \bmod n \\
&= a +_n (b +_n c).
\end{aligned}$$

Here we used that for any integers a, b, c, the equation $(a + b) + c = a + (b + c)$ holds. Items 4 and 5 can be proven similarly. The details are left to the reader as an exercise. ■

One may wonder if division modulo n is possible as well. It turns out that this cannot always be defined in a meaningful way, but it is possible to divide modulo n by an integer a as long as a and n are relatively prime, that is to say, as long as $\gcd(a, n) = 1$. Here gcd stands for: greatest common divisor. If $\gcd(a, n) = 1$, one can namely use the extended Euclidean algorithm for integers to find integers r and s such that $r \cdot a + s \cdot n = \gcd(a, n) = 1$. Then r can be thought of as the multiplicative inverse of a modulo n, since

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

1.5 The extended Euclidean algorithm for integers: a brush up

The usual (non-extended) version of the Euclidean algorithm has as purpose to compute the greatest common divisor (gcd) of two integers a and b fast. First of all, if a is negative, then $\gcd(a, b) = \gcd(-a, b)$. Therefore, we may always assume that a is not negative. A similar remark holds for b. Therefore, we will only describe a way to compute the gcd of two integers in case they both are not negative.

The starting point of the Euclidean algorithm is to use the following simple observation:

For natural numbers N and M such that $N \geq M$, it holds that $\gcd(N, M) = \gcd(N - M, M)$.

This identity may seem uninteresting, but it is in fact very practical. It has for example the

following consequence:

$$\begin{aligned}
\gcd(904,339) &= \gcd(565,339) && \text{since } 904-339=565 \text{ and } \gcd(N,M)=\gcd(N-M,M)\\
&= \gcd(226,339) && \text{since } 565-339=226 \text{ (and } \gcd(N,M)=\gcd(N-M,M))\\
&= \gcd(339,226) && \text{since } \gcd(N,M)=\gcd(M,N)\\
&= \gcd(113,226) && \text{since } 339-226=113\\
&= \gcd(226,113) && \text{since } \gcd(N,M)=\gcd(M,N)\\
&= \gcd(113,113) && \text{since } 226-113=113\\
&= \gcd(0,113) && \text{since } 113-113=0\\
&= 113 && \text{since } \gcd(0,M)=M.
\end{aligned}$$

In practice, we could have skipped a few steps and conclude directly that $\gcd(226,339)=113$, since $226=2\cdot 113$ and $339=3\cdot 113$. Another variant uses division with remainder: if $N\geq M$ and $N=q\cdot M+r$, then $\gcd(N,M)=\gcd(M,r)$. Using this saves several steps as well. The point is though that the above simple procedure always works regardless of the size of the integers N and M and without having to factor any integer at all into a product of prime numbers.

This procedure can be formalized using recursive definitions: the input in the algorithm consists of two natural numbers N and M. Now recursively define the sequence of tuples of natural numbers $(a_0,b_0),(a_1,b_1),\ldots$ as follows (we will write a tuple as a vector for convenience):

$$\begin{bmatrix} a_n \\ b_n \end{bmatrix} := \begin{cases} \begin{bmatrix} N \\ M \end{bmatrix} & \text{if } n=0 \\ \begin{bmatrix} a_{n-1}-b_{n-1} \\ b_{n-1} \end{bmatrix} & \text{if } n\geq 1 \text{ and } a_{n-1}\geq b_{n-1} \\ \begin{bmatrix} b_{n-1} \\ a_{n-1} \end{bmatrix} & \text{if } n\geq 1 \text{ and } a_{n-1}< b_{n-1} \end{cases} \tag{1.2}$$

For the numbers $N=904$ and $M=339$ one obtains for example

n	0	1	2	3	4	5	6	7	8	≥ 9
a_n	904	565	226	339	113	226	113	0	113	113
b_n	339	339	339	226	226	113	113	113	0	0

One can now show the following lemma:

Lemma 1.5.1 *Let a_n and b_n be defined as in Definition 1.2.*

1. *For all natural numbers n it holds that $\gcd(a_n,b_n)=\gcd(N,M)$.*
2. *For all natural numbers $n\geq 1$ it holds that $b_n\leq b_{n-1}$.*

3. If $b_m > 0$, then there exists a natural number k such that $b_{m+k} < b_m$.

Proof.

1. We show the claim using induction on n. If $n = 0$, then $a_0 = N$ and $b_0 = M$. Therefore $\gcd(a_0, b_0) = \gcd(N, M)$ and the induction basis is shown. Now we show the induction step. Assume as induction hypothesis that $\gcd(a_{n-1}, b_{n-1}) = \gcd(N, M)$. If $a_{n-1} \geq b_{n-1}$, then according to equation (1.2), it holds that $a_n = a_{n-1} - b_{n-1}$ and $b_n = b_{n-1}$. Therefore $\gcd(a_n, b_n) = \gcd(a_{n-1} - b_{n-1}, b_{n-1})$. However, the set of all common divisors of the numbers $a_{n-1} - b_{n-1}$ and b_{n-1} is exactly the same as the set of all common divisors of the numbers a_{n-1} and b_{n-1}. Therefore $\gcd(a_n, b_n) = \gcd(a_{n-1}, b_{n-1})$, still assuming that $a_{n-1} \geq b_{n-1}$. From the induction hypothesis, we deduce that $\gcd(a_n, b_n) = \gcd(a_{n-1}, b_{n-1}) = \gcd(N, M)$. If $a_{n-1} < b_{n-1}$, then according to Definition 1.2 we have $a_n = b_{n-1}$ and $b_n = a_{n-1}$. In this case, the induction hypothesis immediately implies that $\gcd(a_n, b_n) = \gcd(b_{n-1}, a_{n-1}) = \gcd(a_{n-1}, b_{n-1}) = \gcd(N, M)$.
2. If $a_{n-1} \geq b_{n-1}$, then $b_n = b_{n-1}$. In particular $b_n \leq b_{n-1}$ in this case. If $a_{n-1} < b_{n-1}$, then $b_n = a_{n-1}$. Therefore $b_n < b_{n-1}$ in this case and a fortiori $b_n \leq b_{n-1}$.
3. Assume that $b_m > 0$ for a certain natural number m. We will show that there exists k such that $b_{m+k} < b_m$. Using division with remainder modulo b_m, we can find a natural number q such that $a_m = qb_m + r$, where $q \geq 0$ and $0 \leq r < b_m$. Running q steps of the Euclidean algorithm in a row, we obtain $a_{m+q} = a_m - qb_m = r$ and $b_{m+q} = b_m$. In the next step of the algorithm, one will obtain $a_{m+q+1} = b_m$ and $b_{m+q+1} = r < b_m$. We see that provided $b_m > 0$, it is possible to find k such that $b_{m+k} < b_m$ (namely $k = q + 1$).

■

Theorem 1.5.2 *Let N and M be natural numbers. Let a_n and b_n be defined recursively as in equation (1.2). Then there exists a natural number m such that $b_m = 0$ and $a_m = \gcd(N, M)$.*

Proof. We first show that there exists a natural number m such that $b_m = 0$. Let us write $d := \min\{b_0, b_1, b_2, \dots\}$, that is to say that d is the least natural number among all numbers in the sequence $b_0, b_1, b_2, \dots$ (note that each decreasing sequence of natural numbers has a minimum). If $d = 0$, then we are done, since in that case there exists a natural number m such that $b_m = 0$. We can now ask ourselves: Can d be positive? This can in fact not happen: By contradiction, assume that $d > 0$. Then there is a natural number m such that $b_m = d$. According to part three of Lemma 1.5.1, there exists a natural number k such that $b_{m+k} < b_m$. But then $\min\{b_0, b_1, b_2, \dots\}$ is strictly less than d, which is a contradiction with the choice of d.

Therefore $b_m = 0$ for some natural number m. But then according to part one of Lemma 1.5.1, we have $\gcd(a_m, b_m) = \gcd(N, M)$. On the other hand, since $b_m = 0$, we conclude that $\gcd(a_m, b_m) = \gcd(a_m, 0) = a_m$. Hence $a_m = \gcd(N, M)$. ■

In some applications, it is not enough to compute $\gcd(N, M)$, but is it also important to express $\gcd(N, M)$ in N and M. More precisely, to find integers r and s such that

$$\gcd(N, M) = rN + sM.$$

The fact that this always can be done is often called Bézout's identity. The extended Euclidean algorithm not only computes $\gcd(N, M)$, but also the integers r and s from Bézout's identity.

We first consider the example with $N = 904$ and $M = 339$. The idea is to start with the equations

$$1 \cdot 904 + 0 \cdot 339 = 904 \tag{1.3}$$

and

$$0 \cdot 904 + 1 \cdot 339 = 339. \tag{1.4}$$

Taking the difference, we obtain that

$$1 \cdot 904 - 1 \cdot 339 = 904 - 339 = 565.$$

Subtracting equation (1.4) once more, we obtain

$$1 \cdot 904 - 2 \cdot 339 = 565 - 339 = 226 \tag{1.5}$$

Subtracting equation (1.5) from equation (1.4), we obtain

$$-1 \cdot 904 + 3 \cdot 339 = 113.$$

The numbers r and s can apparantly be chosen to be -1 and 3 in this case. To find r and s in general, we extend the first version of the Euclidean algorithm. A sequence of tuples of integers $(a_0, b_0), (a_1, b_1), \ldots$ are defined as in equation (1.2), but moreover two further sequences of tuples of integers $(r_0, s_0), (r_1, s_1), \ldots$ and $(t_0, u_0), (t_1, u_1), \ldots$ are defined recursively in the following way:

$$\begin{bmatrix} a_n & r_n & s_n \\ b_n & t_n & u_n \end{bmatrix} := \begin{cases} \begin{bmatrix} N & 1 & 0 \\ M & 0 & 1 \end{bmatrix} & \text{if } n = 0, \\ \begin{bmatrix} a_{n-1} - b_{n-1} & r_{n-1} - t_{n-1} & s_{n-1} - u_{n-1} \\ b_{n-1} & t_{n-1} & u_{n-1} \end{bmatrix} & \text{if } n \geq 1 \text{ and } a_{n-1} \geq b_{n-1}, \\ \begin{bmatrix} b_{n-1} & t_{n-1} & u_{n-1} \\ a_{n-1} & r_{n-1} & s_{n-1} \end{bmatrix} & \text{if } n \geq 1 \text{ and } a_{n-1} < b_{n-1}. \end{cases} \tag{1.6}$$

The point of the two additional tuples of integers is described in the following:

Lemma 1.5.3 *For all natural numbers n it holds that $r_n a_0 + s_n b_0 = a_n$ and $t_n a_0 + u_n b_0 = b_n$.*

Proof. Induction basis: Left to the reader.

Induction step: Assume that $r_{n-1}a_0 + s_{n-1}b_0 = a_{n-1}$ and $t_{n-1}a_0 + u_{n-1}b_0 = b_{n-1}$. If $a_{n-1} \geq b_{n-1}$, then

$$\begin{aligned} r_n a_0 + s_n b_0 &= (r_{n-1} - t_{n-1})a_0 + (s_{n-1} - u_{n-1})b_0 \text{ according to equation (1.6)} \\ &= (r_{n-1}a_0 + s_{n-1}b_0) - (t_{n-1}a_0 + u_{n-1}b_0) \\ &= a_{n-1} - b_{n-1} \text{ according to the induction hypothesis} \\ &= a_n \text{ according to equation (1.2).} \end{aligned}$$

Similarly, one obtains that $t_n a_0 + u_n b_0 = b_n$. Also in case $a_{n-1} < b_{n-1}$ the induction step can be carried out in a similar way. ■

Now we return to the original problem, namely to find integers r and s such that $r \cdot N + s \cdot M = \gcd(N, M)$ given any two integers N and M. First of all, we may assume that N and M are not negative. If for example N is negative and M is positive, we could first find integers r and s such that $r \cdot (-N) + s \cdot M = \gcd(-N, M)$. Once we have done this and using that by definition $\gcd(-N, M) = \gcd(N, M)$, we simply rewrite the identity as $(-r) \cdot N + s \cdot M = \gcd(N, M)$. Therefore we may assume that N and M both are not negative. Given $a_0 = N$ and $b_0 = M$, we can find d such that $a_d = \gcd(N, M)$ and $b_d = 0$ by Theorem 1.5.2. Now Lemma 1.5.3 implies that $r_d \cdot N + s_d \cdot M = \gcd(N, M)$. Hence the we can choose $r = r_d$ and $s = s_d$.

Equation (1.6) may seem complicated at first sight, but gives a fast and transparent algorithm to compute both $\gcd(N, M)$ as well as integers r and s, satisfying $rN + sM = \gcd(N, M)$. This algorithm begins with the following 2 by 3 matrix:

$$\begin{bmatrix} N & 1 & 0 \\ M & 0 & 1 \end{bmatrix}.$$

This matrix is gradually modified using row operations (more details in a moment), till it has the form:

$$\begin{bmatrix} \gcd(N, M) & r & s \\ 0 & * & * \end{bmatrix}.$$

The first row then contains all the information we need: $\gcd(N, M)$ and the numbers r and s. To modify the matrix only two row operations are permitted: In the first place interchanging the two rows, in the second place row two can be subtracted from row one. These two cases correspond exactly to the two cases in equation (1.6).

As an example, consider $N = 904$ and $M = 339$.

$$\begin{bmatrix} 904 & 1 & 0 \\ 339 & 0 & 1 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \text{ - } R_2}{} \begin{bmatrix} 565 & 1 & -1 \\ 339 & 0 & 1 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \text{ - } R_2}{} \begin{bmatrix} 226 & 1 & -2 \\ 339 & 0 & 1 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_2 \leftrightarrows R_1}{}$$

$$\begin{bmatrix} 339 & 0 & 1 \\ 226 & 1 & -2 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \text{ - } R_2}{} \begin{bmatrix} 113 & -1 & 3 \\ 226 & 1 & -2 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \leftrightarrows R_2}{} \begin{bmatrix} 226 & 1 & -2 \\ 113 & -1 & 3 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \text{ - } R_2}{}$$

$$\begin{bmatrix} 113 & 2 & -5 \\ 113 & -1 & 3 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \text{ - } R_2}{} \begin{bmatrix} 0 & 3 & -8 \\ 113 & -1 & 3 \end{bmatrix} \overset{}{\longrightarrow} \underset{R_1 \leftrightarrows R_2}{} \begin{bmatrix} 113 & -1 & 3 \\ 0 & 3 & -8 \end{bmatrix}.$$

We conclude that

$$\gcd(904, 339) = 113 \text{ and } -1 \cdot 904 + 3 \cdot 339 = 113.$$

It should be noted that there are many variants of the (extended) Euclidean algorithm. Some are faster than others and more suitable for implementation in a computer than the above version. The advantage of the above version is, that it is simple to state and explain. One improvement is

to use division with remainder. More precisely, for a given pair of integers N, M, the case $n \geq 1$ and $a_{n-1} \geq b_{n-1}$ in equation (1.6) can be replaced by

$$\begin{bmatrix} a_n & r_n & s_n \\ b_n & t_n & u_n \end{bmatrix} := \begin{bmatrix} a_{n-1} - q_{n-1}b_{n-1} & r_{n-1} - q_{n-1}t_{n-1} & s_{n-1} - q_{n-1}u_{n-1} \\ b_{n-1} & t_{n-1} & u_{n-1} \end{bmatrix},$$

where $q_{n-1} = \lfloor a_{n-1}/b_{n-1} \rfloor$. Note that $a_{n-1} - q_{n-1}b_{n-1}$ is simply $a_{n-1} \bmod b_{n-1}$. Compared to the previous version, q_{n-1} steps are taken in one go, costing one division with remainder, making this algorithm faster. The main cost of the algorithm is this number of division steps. Note that in the extreme case that N is arbitrary and $M = 1$, the extended Euclidean algorithm would take $N + 1$ steps using equation (1.3), while with the above modification, only two steps and only one division step.

Aside 1.5.4 *The Euclidean algorithm is named after the Greek mathematician Euclid who lived around 300 BCE. The algorithm is described in his book 'Elements'. Much more recently the running time of the Euclidean algorithm was investigated. The French mathematician Gabriel Lamé showed in 1844 that the maximum number of divisions occurs precisely when N and M are consecutive Fibonacci numbers recursively defined by $F_1 = F_2 = 1$ and $F_{n+1} = F_n + F_{n-1}$ for $n \geq 2$. Note that for $n \geq 3$, the computation of $\gcd(F_n, F_{n-1})$ takes $n - 2$ division steps. Since $F_n = (\tau^n + \overline{\tau}^n)/\sqrt{5}$, where $\tau = (1 + \sqrt{5})/2$ is the golden ratio and $\overline{\tau} = (1 - \sqrt{5})/2$, this means that for given positive integers N and M, the maximum number of division steps is around $\log_\tau(\max N, M)$. In particular, the greatest common divisor of two positive integers can be computed in polynomial time in the number of digits of N and M.*

1.6 Extra: some applications of modular arithmetic

Modular arithmetic is ingrained in daily life. After all, the clock is based on working modulo 12 for a clock with handles, or possibly modulo 24 on a clock with an electronic display. More generally, when a pattern repeats itself, it is natural to use modular arithmetic to describe this. For example the keys on a piano display a regular pattern modulo 12 (taking both white and black keys into consideration). As another example, the decimals of a fraction of natural numbers exhibit repeating patterns. For example $1/7 = 0.142857142857\ldots$, so if d_i denotes the i-th decimal in the decimal expansion of $1/7$, then

$$d_i = \begin{cases} 1 & \text{if } i \bmod 6 = 1, \\ 4 & \text{if } i \bmod 6 = 2, \\ 2 & \text{if } i \bmod 6 = 3, \\ 8 & \text{if } i \bmod 6 = 4, \\ 5 & \text{if } i \bmod 6 = 5, \\ 7 & \text{if } i \bmod 6 = 0. \end{cases}$$

In this section we describe some cases of how modular arithmetic is used in practical settings. Some of these are given as small examples, some of the more involved ones are given in subsections and may involve some new theory as well. Many more examples could be given, but usually require a deeper understanding of the algebraic structures involved. Some of such examples will be given in later chapters.

Example 1.6.1 The International Standard Book Number (ISBN) is a number used to uniquely identify books. For example, the book "Algebraic function fields and codes" by H. Stichtenoth has ISBN number $978-3-540-76878-4$. More properly this is the ISBN-13 code, since older systems had fewer digits. To verify if a given ISBN-number is valid, a *check-sum* involving modular arithmetic is used. More precisely, given the thirteen digits $a_1a_2a_3-a_4-a_5a_6a_7-a_8a_9a_{10}a_{11}a_{12}-a_{13}$ of a valid ISBN-13 number, the following congruence is satisfied:

$$a_1+3a_2+a_3+3a_4+a_5+3a_6+a_7+3a_8+a_9+3a_{10}+a_{11}+3a_{12}+a_{13}\equiv 0 \pmod{10}$$

In other words, the thirteenth digit a_{13} can be computed from the first twelve using the formula

$$a_{13}=(-a_1-3a_2-a_3-3a_4-a_5-3a_6-a_7-3a_8-a_9-3a_{10}-a_{11}-3a_{12}) \bmod 10$$

Applying this to the example of the book given above, we see that

$$(-9-3\cdot 7-8-3\cdot 3-5-3\cdot 4-0-3\cdot 7-6-3\cdot 8-7-3\cdot 8) \bmod 10=(-146) \bmod 10=4,$$

so indeed the given ISBN number was valid. Exactly the same method is used for the 13-digit International Article Number (EAN).

Example 1.6.2 Every resident in Denmark has a ten digit personal identification number of the form ddmmyy-ssss called a CPR-number. The first six digits give the date of birth, while the final four are called sequence numbers. The ten digits $a_1a_2a_3a_4a_5a_6-a_7a_8a_9a_{10}$ ideally satisfy the congruence:

$$4a_1+3a_2+2a_3+7a_4+6a_5+5a_6+4a_7+3a_8+2a_9+a_{10}\equiv 0 \pmod{11}$$

Note however that some birth dates currently occur so frequently, that starting in October 2007 the congruence is not necessarily satisfied anymore.

1.6.1 The Chinese remainder theorem and distributed storage

Many applications of modular arithmetic use a result called the *Chinese remainder theorem*. Let us formulate and prove it for future reference:

Theorem 1.6.3 *Let n_1 and n_2 be positive integers such that $\gcd(n_1,n_2)=1$ and write $n=n_1\cdot n_2$. Then the map $\varphi:\mathbb{Z}_n\to\mathbb{Z}_{n_1}\times\mathbb{Z}_{n_2}$ defined by $\varphi(a):=(a \bmod n_1, a \bmod n_2)$ has an inverse.*

Proof. Since $\gcd(n_1,n_2)=1$, there exist integers r,s such that $rn_1+sn_2=1$. This will be the main ingredient in showing that φ has an inverse. Define the map $\psi:\mathbb{Z}_{n_1}\times\mathbb{Z}_{n_2}\to\mathbb{Z}_n$ as $\psi(a_1,a_2):=(sn_2a_1+rn_1a_2) \bmod n$. Then using that n_1 divides n, we see that

$$\psi(a_1,a_2) \bmod n_1=(sn_2a_1+rn_1a_2) \bmod n_1=sn_2a_1 \bmod n_1=(1-rn_1)a_1 \bmod n_1=a_1$$

and similarly $\psi(a_1,a_2) \bmod n_2=a_2$. Hence $\varphi(\psi(a_1,a_2))=(a_1,a_2)$. Now note that for any $a\in\mathbb{Z}_n$, we have

$$a=a(rn_1+sn_2)=sn_2a+rn_1a\equiv sn_2(a \bmod n_1)+rn_1(a \bmod n_2) \pmod{n},$$

which implies that $a=(sn_2(a \bmod n_1)+rn_1(a \bmod n_2)) \bmod n$. This shows that $\psi(\varphi(a))=a$. Combining all the above, we conclude that ψ is the inverse of φ. ■

Remark 1.6.4 *Using the language from Chapters 7 and 8, one can reformulate this theorem as follows: if* $\gcd(n_1, n_2) = 1$*, then the map* $\varphi : \mathbb{Z}_n \to \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$ *defined by* $\varphi(a) := (a \bmod n_1, a \bmod n_2)$ *is an ring isomorphism between the rings* $(\mathbb{Z}_n, +_n, \cdot_n)$ *and* $(\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}, +_{n_1} \times +_{n_2}, \cdot_{n_1} \cdot_{n_2})$.

If $n = n_1 \cdots n_\ell$ with $\gcd(n_i, n_j) = 1$ for all distinct i and j, one can use the Chinese remainder theorem iteratively and show that the map $\varphi : \mathbb{Z}_n \to \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_\ell}$ given by $\varphi(a) = (a \bmod n_1, \ldots, a \bmod n_\ell)$ has an inverse. The inverse of this map can also be described: define $N_i = n/n_i$. Then $\gcd(N_1, \ldots, N_\ell) = 1$, so there exist integers $r_1, \ldots, r_\ell$ such that $r_1 N_1 + \cdots + r_\ell N_\ell = 1$. The integers $r_1, \ldots, r_\ell$ can be computed using the extended Euclidean algorithm iteratively. Given $(a_1, \ldots, a_\ell) \in \mathbb{Z}_{n_1} \times \cdots \times \mathbb{Z}_{n_\ell}$, one has $\varphi^{-1}(a_1, \ldots, a_\ell) = (r_1 N_1 a_1 + \cdots r_\ell N_\ell a_\ell) \bmod n$. This can be used to store a large number $a \in \mathbb{Z}_n$ as ℓ smaller parts $a_i = a \bmod n_i$ in a natural way. Knowing the parts $(a_1, \ldots, a_\ell)$ is equivalent to knowing a itself. If one now considers a setting of distributed storage, one can think of the information $a \in \mathbb{Z}_n$, being stored by a client on ℓ servers $S_1, \ldots, S_\ell$ by storing a_i on server S_i.

In realistic settings of distributed storage, there is the risk that some of the servers crash or are temporarily unavailable. A small variation of the above technique can handle this situation (as long as not too many servers go down simultaneously!). As an example, let us consider seven servers and assume that no more than one server can be down at the same time. Choose $(n_1, \ldots, n_7) = (11, 13, 17, 19, 23, 29, 31)$. Consider an integer a such that $0 \leq a < n_1 n_2 n_3 n_4 n_5 n_6 = 30808063$, say $a = 10000000$. Storing a on seven servers amounts to storing a_i on server S_i, where $(a_1, \ldots, a_7) = (10, 10, 5, 15, 14, 17, 20)$. Note that the a_i are small, even though a was not. Now suppose that the fifth server goes down, so that the information that is available for the client is $(10, 10, 5, 15, ?, 17, 20)$. Since $a < n_1 n_2 n_3 n_4 n_5 n_6 < n_1 n_2 n_3 n_4 n_6 n_7$, the client can still reconstruct a uniquely! Note that here the assumption that $a < n_1 n_2 n_3 n_4 n_5 n_6$ is used. More concretely, let us for convenience define $M = n_1 n_2 n_3 n_4 n_6 n_7$, $M_1 = n_2 n_3 n_4 n_6 n_7$, $M_2 = n_1 n_3 n_4 n_6 n_7$, $M_3 = n_1 n_2 n_4 n_6 n_7$, $M_4 = n_1 n_2 n_3 n_6 n_7$, $M_5 = n_1 n_2 n_3 n_4 n_7$, and $M_6 = n_1 n_2 n_3 n_4 n_6$. Then using the extended Euclidean algorithm iteratively, one can obtain that

$$-6M_1 + 5M_2 + 3M_3 + 3M_4 + 9M_5 - 15M_6 = 1.$$

Hence

$$a = (-6M_1 10 + 5M_2 10 + 3M_3 5 + 3M_4 15 + 9M_5 17 - 15M_6 20) \bmod M = 10000000,$$

confirming that a can be recovered even if server S_5 goes down. A similar computation can be done if one of the other servers goes down.

1.6.2 Dixon's factorization algorithm

Suppose n is a positive integer. A factorization algorithm for integers is an algorithm that takes n as input and returns prime numbers $p_1, \ldots, p_\ell$ and exponents $e_1, \ldots, e_\ell$ such that $n = p_1^{e_1} \cdots p_\ell^{e_\ell}$. Dixon's factorization algorithm is based on the observation that if $a^2 \equiv b^2 \pmod{n}$, then $a^2 - b^2 = (a-b)(a+b)$ is a multiple of n. Hence any prime number dividing n divides either $a - b$ or $a + b$. If two distinct prime numbers divide n, one can hope that one of them divides $a - b$, but not $a + b$, and the other one $a + b$, but not $a - b$. If that happens, either $\gcd(a - b, n)$ or $\gcd(a + b, n)$ is a proper factor of n.

Since the Euclidean algorithm can be used to compute $\gcd(a-b,n)$ and $\gcd(a+b,n)$ fast, the main difficulty is to find a, b in $\mathbb{Z}_n$ satisfying $a^2 \equiv b^2 \pmod{n}$. Of course one can always choose $a \in \mathbb{Z}_n$ arbitrary and set $b = a$ or $b = -a$, but then $\gcd(a-b,n) = n$ or $\gcd(a+b,n) = n$, which does not give a proper factor of n. The idea is instead to factor $\alpha^2 \bmod n$ for small values of $\alpha \in \mathbb{Z}_n$ and to combine the obtained factorizations to find a and b. More precisely, if we can find $\alpha_1, \ldots, \alpha_m \in \mathbb{Z}_n$ such that $\prod_{i=1}^{m}(\alpha_i^2 \bmod n)$ is a square, say the square of a, then we can choose this a and set $b = \alpha_1 \cdots \alpha_m$, since

$$a^2 = \prod_{i=1}^{m}(\alpha_i^2 \bmod n) \equiv \prod_{i=1}^{m}\alpha_i^2 = b^2 \pmod{n}.$$

To avoid finding $a = \pm b$, one typically considers $\alpha_i > \sqrt{n}$. Let us consider the example $n = 1649$ and $\alpha_i = 40 + i$. Then

α_i	41	42	43	44	45	46	47
$\alpha_i^2 \bmod 1767$	2^5	$5 \cdot 23$	$2^3 \cdot 5^2$	$7 \cdot 41$	$2^3 \cdot 47$	467	$2^4 \cdot 5 \cdot 7$

Then we have $(\alpha_1^2 \bmod 1649)(\alpha_3^2 \bmod 1649) = 2^5 \cdot 2^3 \cdot 5^2 = (2^5 \cdot 5)^2$. Hence we can conclude that $(2^5 \cdot 5)^2 \equiv (41 \cdot 43)^2 \pmod{1649}$. Further $\gcd(1649, 41 \cdot 43 - 2^5 \cdot 5) = 17$ and $\gcd(1649, 41 \cdot 43 + 2^5 \cdot 5) = 97$, giving us two proper divisors of 1649. In fact $1649 = 17 \cdot 97$ and since 17 and 97 are prime numbers, the process of factoring 1649 is complete.

The factorization method based on looking for congruences of the form $a^2 \equiv b^2 \pmod{n}$ was found by a Belgian mathematician, Maurice Borisovich Kraitchik around 1920. The mathematician John D. Dixon took this further and combined this with the idea of not always trying to factor $\alpha_i^2 \bmod n$, but only to check whether or not it is the product of prime numbers all smaller than some chosen bound B, called the *smoothness bound*. If in the example $n = 1649$, we would have chosen $B = 7$, we would only have obtained the entries from the table for $\alpha_i \in \{41, 43, 47\}$, but this would already have been enough to factor 1649. Even $B = 5$ would already have worked in this case.

1.7 Exercises

Multiple choice exercises

1. Let A and B be two sets.
   - A. $A \setminus B$ is the set of elements in A which are not in B
     TRUE ☐ FALSE ☐
   - B. if $A = \{quadrilaterals\}$ and $B = \{polygons\}$ then $A \subseteq B$
     TRUE ☐ FALSE ☐
   - C. If $A \cup B = \emptyset$ then $A = \emptyset$ and $B = \emptyset$
     TRUE ☐ FALSE ☐

D. If $A \cap B = \emptyset$ then $A = \emptyset$ or $B = \emptyset$
TRUE ☐ FALSE ☐

E. $\{a\} \in \{d, e, f, a\}$
TRUE ☐ FALSE ☐

F. Natural numbers $\subseteq$ whole numbers
TRUE ☐ FALSE ☐

G. Integers $\subseteq$ natural numbers
TRUE ☐ FALSE ☐

H. $0 \in \emptyset$
TRUE ☐ FALSE ☐

I. $\emptyset \subseteq \{1, 2, 3\}$
TRUE ☐ FALSE ☐

2. Let $a, b, n \in \mathbb{Z}$. Then $a \equiv b \pmod{n}$ if and only if

A. n divides $a + b$

B. n divides $a - b$

C. both a and b are multiples of n

D. there exists $k \in \mathbb{Z}$ such that $a \cdot b = kn$.

3. Let $a, n \in \mathbb{Z}$. Then the congruence class of a modulo n is

A. $a + n + \mathbb{Z}$

B. $n + a\mathbb{Z}$

C. $a + n\mathbb{Z}$

D. $an\mathbb{Z}$

4. Let $\sim$ be a relation on the set A.

A. $\sim$ is *reflexive* if and only if for all $a \in A$, $a \sim a$
TRUE ☐ FALSE ☐

B. $\sim$ is *symmetric* if and only if for all $a, b \in A$, $a \sim b$ and $b \sim a$
TRUE ☐ FALSE ☐

C. $\sim$ is *transitive* if and only if for all $a, b, c \in A$, $a \sim b$ and $b \sim c$ imply $a \sim c$
TRUE ☐ FALSE ☐

D. if $\sim$ is an *equivalence relation* then $\sim$ is symmetric and transitive
TRUE ☐ FALSE ☐

E. $\sim$ is an *equivalence relation* if and only if $\sim$ is reflexive, symmetric and transitive
TRUE ☐ FALSE ☐

5. Which of the following stamements is true for an equivalence relation $\sim$ on a set A?

A. A is the disjoint union of all the distinct congruence classes $[a]_\sim$ with $a \in A$

B. two congruence classes can intersect without being equal

C. the elements of $[a]_\sim$ are exacly all the elements $b \in A$ such that $b \sim a$

D. to decide whether two congruence classes coincide it is sufficient to decide whether they intersect or not

6. The *standard representative* of the congruence class $a + n\mathbb{Z}$ is
   A. b such that $0 \leq b \leq n-1$ and $a \equiv b \pmod{n}$
   B. $a \bmod n$
   C. $a - n$
7. Let n be a positive integer.
   A. the operation $a +_n b$ is defined as the remainder modulo n of $a + b$
      TRUE ☐ FALSE ☐
   B. the operation $+_n$ is commutative and associative
      TRUE ☐ FALSE ☐
   C. the operation $a \cdot_n b$ is defined as $a \cdot b$
      TRUE ☐ FALSE ☐
   D. the operation $\cdot_n$ is commutative but not associative
      TRUE ☐ FALSE ☐
8. Let a be an integer and $n \in \mathbb{N}$.
   A. The extended Euclidean algorithm gives as output both $\gcd(a, n)$ and $r, s \in \mathbb{Z}$ with $ra + sn = \gcd(a, n)$
      TRUE ☐ FALSE ☐
   B. if $\gcd(a, n) = 1$ then $ra \equiv 0 \pmod{n}$.
      TRUE ☐ FALSE ☐
   C. if $\gcd(a, n) = 1$ then $ra \equiv 1 \pmod{n}$.
      TRUE ☐ FALSE ☐

Comment. If $\gcd(a, n) = 1$ then the integer r given by the extended Euclidean algorithm is called the *multiplicative inverse of a modulo n*, because just as the multiplicative inverse $y = 1/z$ of a real number z satisfies $y \cdot z = 1$, it holds $r \cdot_n a = (r \cdot a) \bmod n = 1$.

Exercises to get to know the material better

9. Determine whether or not the following statements are true:
   (a) $-2 \equiv 97 \pmod{11}$.
   (b) $2 \equiv -97 \pmod{-11}$.
   (c) $10 \equiv 3 \pmod{13}$.
   (d) $101 \equiv 23 \pmod{1}$.
   (e) $12 \equiv 3 \pmod{0}$.
10. Compute quotient and remainder of the division of 101 by 15.

11. Write down the five representatives of the congruence class of 11 modulo 7 of smallest absolute value.

12. Are 4 and 77 representatives of the same congruence class modulo 11?

13. Compute the multiplicative inverse of 37 modulo 101. Hint: Use the extended Euclidean algorithm for integers.

14. Given that $5x \equiv 6 \pmod 8$, find x.

15. Find the last digit of 7^{100}.

16. If $n!$ denotes the product of the integers 1 through n, what is the remainder when $(1! + 2! + 3! + 4! + 5! + 6! + \ldots)$ modulo 9?

17. (a) Compute $1234 \pmod{357}$ using the standard long division with remainder.
    (b) Compute the greatest common divisor of 357 and 1234 using the Euclidean algorithm.
    (c) Compute the multiplicative inverse of 357 modulo 1234 using the extended Euclidean algorithm. Check if $357 \cdot 35^{-1} \equiv 1 \pmod{1234}$ holds.

18. Describe equivalence classes for the following equivalence relations on the given set A.
    (a) $A = \mathbb{R}$, and $a \sim b$ if and only if $a = b$ or $a = -b$.
    (b) $A = \mathbb{R}$, and $a \sim b$ if and only if $a^2 + a = b^2 + b$.
    (c) A is the set of all points in the plane, and $a \sim b$ if and only if a and b are at the same distance from the origin.
    (d) $A = \mathbb{N}$, and $a \sim b$ if and only if ab is a square.
    (e) $A = \mathbb{R} \times \mathbb{R}$, and $(x, y) \sim (a, b)$ if and only if $x^2 + y^2 = a^2 + b^2$.

Exercises to get around in the theory

19. Show that any two integers a and b are congruent modulo 1, but that $a \equiv b \pmod 0$ if and only if $a = b$.

20. Prove that if $a \equiv b \pmod n$, then for all positives integers d that divide both a and b it holds
$$\frac{a}{d} \equiv \frac{b}{d} \pmod{\frac{n}{\gcd(n,d)}}$$

21. The floor function $\lfloor \cdot \rfloor : \mathbb{R} \to \mathbb{Z}$ is defined for $x \in \mathbb{R}$ by
$$\lfloor x \rfloor := \max\{a \in \mathbb{Z} \mid a \leq x\}.$$
In other words, given $x \in \mathbb{R}$, it returns the largest integer less than or equal to x.
    (a) Check that $10 \text{ quot } 3 = \lfloor 10/3 \rfloor = 3$ and $-11 \text{ quot } 5 = \lfloor -11/5 \rfloor = -3$.
    (b) Show that for any $x \in \mathbb{R}$, the following holds: $x - 1 < \lfloor x \rfloor \leq x$. Conclude that $\lfloor x \rfloor$ is the unique integer in the interval $]x - 1, x] := \{a \in \mathbb{R} \mid x - 1 < a \leq x\}$. Hint: to show that $x - 1 < \lfloor x \rfloor$, assume that $\lfloor x \rfloor \leq x - 1$ and derive a contradiction.

(c) Use this to show that $0 \leq a - \lfloor a/n \rfloor \cdot n < n$ and conclude that a quot $n = \lfloor a/n \rfloor$ and $a \bmod n = a - \lfloor a/n \rfloor \cdot n$. Remark: in fact you have now proven Fact 1.3.5. Part c) of this exercise shows that the pair $(q, r) := (\lfloor a/n \rfloor, a - \lfloor a/n \rfloor \cdot n) \in \mathbb{Z}^2$ satisfies the requirements from Fact 1.3.5. Part b) implies uniqueness of a quot n, which in turn implies uniqueness of the remainder $a \bmod n$ as well.

22. Show that the relation $\equiv \pmod n$ is an equivalence relation on the set of integers $\mathbb{Z}$.

23. Prove Items 4 and 5 of Theorem 1.4.3.

24. Let $\mathrm{Mat}_{m\times n}(\mathbb{R})$ be the set of $m \times n$ matrices with coefficients in $\mathbb{R}$. For $M, N \in \mathrm{Mat}_{m\times n}(\mathbb{R})$, we say that M is equivalent to N, if and only if there exists an invertible $n \times n$ matrix P in $\mathrm{Mat}_{n\times n}(\mathbb{R})$ and an invertible matrix Q in $\mathrm{Mat}_{m\times m}(\mathbb{R})$ such that $N = Q^{-1} \cdot M \cdot P$. Show being equivalent is indeed an equivalence relation on the set $\mathrm{Mat}_{m\times n}(\mathbb{R})$. Remark: If M represents a linear map L from $\mathbb{R}^m$ to $\mathbb{R}^n$, then by choosing different bases for $\mathbb{R}^m$ and $\mathbb{R}^n$, the same linear map L is represented by a matrix equivalent to M. Hence the equivalence class of M, just gives all possible matrices to represent the linear map L for all possible basis choices for $\mathbb{R}^m$ and $\mathbb{R}^n$.

25. Let $A = \mathbb{Z} \times \mathbb{Z} \setminus \{0\}$. Consider the following relation on A: $(a, b) \sim (c, d)$ if and only if there exists $e \in \mathbb{Z} \setminus \{0\}$ such that $e \cdot (a \cdot d - b \cdot c) = 0$.

    (a) Show that $\sim$ is an equivalence relation.

    (b) Show that an equivalence class $[(a, b)]_\sim$ can identified with the rational number $a/b \in \mathbb{Q}$. That is to say, show that if (c, d) is a representative of the equivalence class $[(a, b)]_\sim$, then $c/d = a/b$. Hence rational numbers are really equivalence classes.

    (c) Show that any equivalence class $[(a, b)]_\sim$ has unique representative (c, d) satisfying that $\gcd(c, d) = 1$. Remark: identifying rational numbers with equivalence classes of pairs $(a, b) \in \mathbb{Z} \times \mathbb{Z} \setminus \{0\}$ is a common mathematical way to define the rational numbers.

26. Let $n \geq 1$ be an integer and let $S \subseteq \mathbb{Z}_n$ be a set satisfying that for all $b, d \in S$ also $b \cdot_n d \in S$. Now define $A = \mathbb{Z}_n \times S$ and consider the following relation on A: $(a, b) \sim (c, d)$ if and only if there exists $e \in S$ such that $e \cdot_n (a \cdot_n d +_n (-b) \cdot_n c) = 0$.

    (a) Show that $\sim$ is an equivalence relation. Hint: compare to the first item of the previous exercise.

    (b) Let $n = 6$ and $S = \{1, 3\}$. Compute all equivalence classes. Hint: there are only two. Apparently when constructing "fractions" for $\mathbb{Z}_6$ with denominator 1 or 3, we get less fractions than elements we started with!

27. Let A be a set and $\{A_i\}_{i\in I}$ a partition of A. Show that the relation $a \sim b$ if and only if there exists $i \in I$ such that $a \in A_i$ and $b \in A_i$, is an equivalence relations. What are its equivalence classes?

28. Show that $\{(x, y) \in \mathbb{Z} \times \mathbb{Z} \ : \ x \equiv y \pmod 6\} \subseteq \{(x, y) \in \mathbb{Z} \times \mathbb{Z} \ : \ x \equiv y \pmod 3\}$.

Chapter 2

Permutations

2.1 Prelude: a little bit on functions

Keywords: function, injective, surjective, bijective, bijection, cardinality.

Before starting out with the theory of permutations, let us first recall some terminology for functions. For two given sets A and B, a function f from A to B assigns to any $a \in A$ one element of B. The set A is called the *domain* of the function, while the set B is called the *co-domain*. There is a compact notation to capture all this information, namely $f : A \to B$. 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 f maps the value a in A to $f[a]$, one can also briefly write $a \mapsto f[a]$. You may see a different notation in other books. It is for example more common to write $f(a)$ rather than $f[a]$ and we will also do this from time to time. However, to avoid "overloading" the use of the usual parentheses (and), we will stick to $f[a]$ in this chapter. All the notation so far for a function f is often compactly given as follows:

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

For example, the function sending a real number to its square can be given as:

$$\begin{aligned} f : \mathbb{R} &\to \mathbb{R} \\ a &\mapsto a^2 \end{aligned}$$

This example also shows that the co-domain of a function f (in the example $\mathbb{R}$) does not have to be equal to its image (in the example $\mathbb{R}_{\geq 0}$). The image of a function is often denoted by $\mathrm{im}f$ or as $f(A)$. We will mainly use the notation $\mathrm{im}(f)$ or, if no confusion is possible, the shorter notation $\mathrm{im}f$. Let us give a formal definition of the image of a function $f : A \to B$ for good measure:

$$\mathrm{im}f := \{f[a] \mid a \in A\}.$$

Given a function $f : A \to B$ and subset $S \subseteq B$ of its co-domain, one defines the *preimage* of S under f to be the set $\{a \in A \mid f[a] \in S\}$.

Two functions $f : A \to B$ and $g : C \to D$ are equal precisely if $A = C$, $B = D$, and they assign the same values to each of the elements of A. In formulas:

$$f = g \iff A = C, B = D \text{ and } \forall a \in A\ f[a] = g[a].$$

Example 2.1.1 Consider the functions

$$\begin{aligned} f : \mathbb{Z}_2 &\to \mathbb{Z}_2 \\ a &\mapsto 0 \end{aligned}$$

and

$$\begin{aligned} g : \mathbb{Z}_2 &\to \mathbb{Z}_2 \\ a &\mapsto a^3 +_2 a^2 \end{aligned}$$

The functions f and g have the same domain and co-domain. Moreover, $f[0] = 0$, $f[1] = 0$, while $g[0] = 0^3 +_2 0^2 = 0$ and $g[1] = 1^3 +_2 1^2 = 0$. Hence $f = g$.

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

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

The reason that in this definition the co-domain of the function f needs to be the same as the domain of the function g, is to guarantee that $g[f[a]]$ is always defined: for any $a \in A$, we know that $f[a] \in B$, so that it indeed makes sense to use $f[a]$ as input for the function g, since the domain of g is assumed to be B.

This function is usually denoted by $g \circ f$ (pronounce: g **after** f) and is called the composition of g and f. Hence we have $(g \circ f)[a] = g[f[a]]$ by definition. For example, if $f : \mathbb{R} \to \mathbb{R}_{>0}$ is defined by $f[x] = x^4 + 1$ and $g : \mathbb{R}_{>0} \to \mathbb{R}$ is defined by $g[x] = \log[x]$, then $g \circ f : \mathbb{R} \to \mathbb{R}$ is the function sending $x \in \mathbb{R}$ to $\log[x^4 + 1]$.

Lemma 2.1.2 *Let A, B, C, and D be sets and suppose that we are given functions $h : A \to B$, $g : B \to C$, and $f : C \to D$. Then we have $(f \circ g) \circ h = f \circ (g \circ h)$.*

Proof. First of all note that both $(f \circ g) \circ h$ and $f \circ (g \circ h)$ are functions from A to D, so they have the same domain and codomain. To prove the lemma it is therefore enough to show that for all $a \in A$, we have $((f \circ g) \circ h)[a] = (f \circ (g \circ h))[a]$. By definition of the composition $\circ$, we have

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

while

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

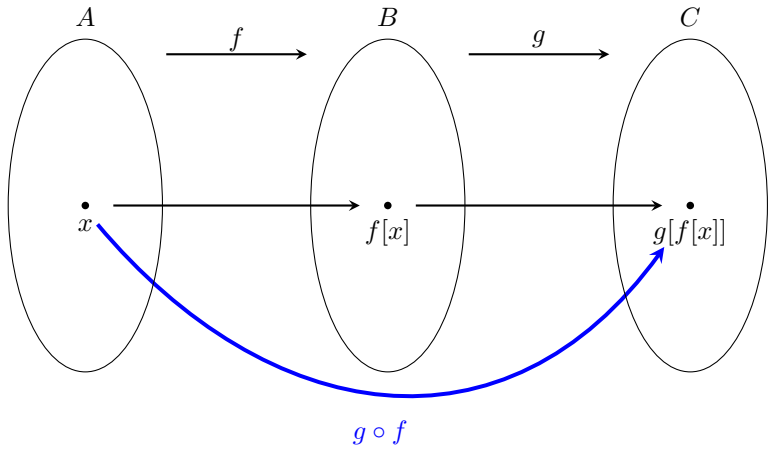


Figure 2.1: composition of the functions $f : A \to B$ and $g : B \to C$

We conclude that for any $a \in A$ it holds that $(f \circ (g \circ h))[a] = ((f \circ g) \circ h)[a]$, which is what we needed to show. ■

The result from this lemma is usually stated as: composition of functions is an *associative* operation. Because of Lemma 2.1.2, it is common to simplify formulas involving composition of several functions, by leaving out the parentheses. For example, one simply writes $f \circ g \circ h$, when taking the composite of three functions.

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

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

A standard 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:

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

This reformulation can be convenient in practice.

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

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

Using the notation $\mathrm{im} f$ for the image of f, this can compactly be restated as: a function $f : A \to B$ is called surjective precisely if $\mathrm{im} f = B$. An example of a function that is injective, but not surjective, is $f : \mathbb{R} \to \mathbb{R}$ given by $f[x] = e^x$, where e here denotes the base of the natural logarithm. Indeed, this function is not surjective, since its image is $\mathbb{R}_{>0}$, while its co-domain was defined to be $\mathbb{R}$. An example of a function that is surjective, but not injective is $g : \mathbb{R} \to [-1, 1]$

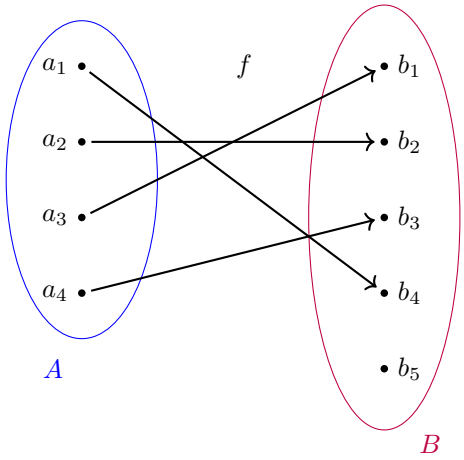


Figure 2.2: injective function $f : A \to B$

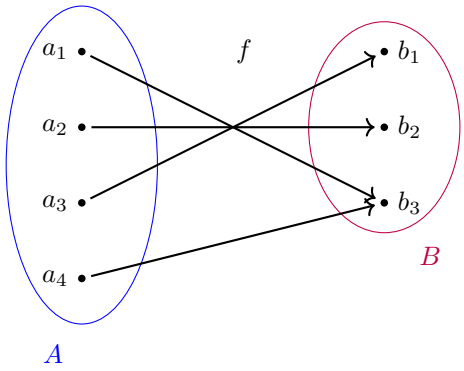


Figure 2.3: surjective function $f : A \to B$

given by $g[x] = \sin[x]$. This function is not injective, since for example 0 and π are both mapped to 0 by the sine function.

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 $h[x] = x^3 + 1 \bmod 3$. Note that $h[0] = 1$, $h[1] = 2$ and $h[2] = 9 \bmod 3 = 0$, 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 2.1.3 *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 g is the inverse function of f. In fact we show the even stronger following result.

Lemma 2.1.4 *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 described that in that case one can define the function $g : B \to A$ 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 2.1.3 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 for checking if 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 precisely n elements, for some natural number n, we have $|A| = n$ and for a set containing infinitely many elements, we have $|A| = \infty$. Another notation that is used quite often is $\#A$ instead of $|A|$. We will stick to the notation $|A|$ for the cardinality of a set. 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 2.1.5 *Suppose that A and B are 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 $\mathrm{im} f = \{f[a] \mid a \in A\}$ contains exactly $|A|$ elements. Since $\mathrm{im} f$ is a subset of B, we see that B contains at least as many elements as A.

Similarly, if f is surjective, then $\mathrm{im} f = B$ 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 sets with finite cardinality, that is to say for finite sets.

Lemma 2.1.6 *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 bijective. Similarly, if f is surjective, it is bijective.*

Proof. Suppose that f is injective, but not surjective. Since f is injective, we have $|\mathrm{im}f| = |A| = |B|$, where the last equality follows by the assumption that $|A| = |B|$. On the other hand, since f is not surjective, we have $\mathrm{im}f \subsetneq B$ and hence $|\mathrm{im}f| < |B|$. 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 $\mathrm{im}f = B$ and hence $|\mathrm{im}f| = |B| = |A|$, since we assumed that $|A| = |B|$. On the other hand, if f is not injective, we have $|A| > |\mathrm{im}f|$. Again we arrive at a contradiction. Apparently, if f is surjective, it needs to be injective as well. ■

The assumption that A and B are finite sets, is important and without it, the conclusion may be false. Consider for example the function $f : \mathbb{N} \to \mathbb{N}$ defined by $n \mapsto n + 1$. Then f is injective, but not surjective.

Aside 2.1.7 *If A is not a finite set, one says that the set A has infinite cardinality, $|A| = \infty$. The mathematician Georg Cantor (1845-1918), the inventor of set theory, famously showed that there are "different kinds" of infinite cardinalities. The idea starts with the observation that two finite sets A and B have the same cardinality if and only if there exists a bijection between them. If one writes for general sets that $A \sim B$ if there exists a bijection $f : A \to B$, then it is not hard to show that $\sim$ is an equivalence relation. Now any two sets in the same equivalence class are said to have the same cardinality. If a set A is in the equivalence class of $\mathbb{N}$, that is to say, if there exists a bijection $f : \mathbb{N} \to A$, the set A is called* countably infinite *and one writes $|A| = \aleph_0$ (the symbol $\aleph_0$ is pronounced as aleph-nought or aleph-zero).* Countable *sets are defined to be those sets that are either finite or countably infinite. In 1874, Cantor published an article in which he showed that there does not exist a bijection between $\mathbb{N}$ and $\mathbb{R}$ and hence that $\mathbb{R}$ is* uncountable*. Later, in 1891, he published a second proof, using what is now known as Cantor's diagonal argument. An interested reader is encouraged to look for this beautiful argument in the literature.*

2.2 Definition of permutations

Keywords: permutation, set of permutations, composition, permutation group.

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. Hence a permutation of A is simply a bijective function from a set A to the same set A. 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 always has an inverse by Lemma 2.1.4. If the permutation is denoted by f, its inverse is, just as for functions in general, denoted by f^{-1}.

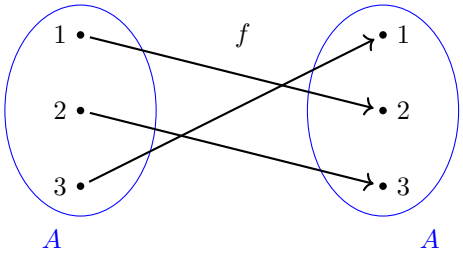


Figure 2.4: picture for Example 2.2.2

Let us from now on assume that the set A is finite, say containing precisely n elements for some natural number n. A compact notation for this is $|A| = n$ and one says that the set A has *cardinality* n.

Definition 2.2.1 *Let A be a set. The set of all permutations $f : A \to A$ is denoted by S_A. In case $A = \{1, 2, \ldots, n\}$, it is customary to write 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 $2 \times n$ 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 2.2.2 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. Similarly, if A is a set with cardinality n, then there are exactly $n!$ possible permutations of A. In other words: if $|A| = n < \infty$, then $|S_A| = n!$.

Example 2.2.3 This is a continuation of Example 2.2.2. The set S_3 contains $3! = 6$ permutations. More precisely, we have:

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

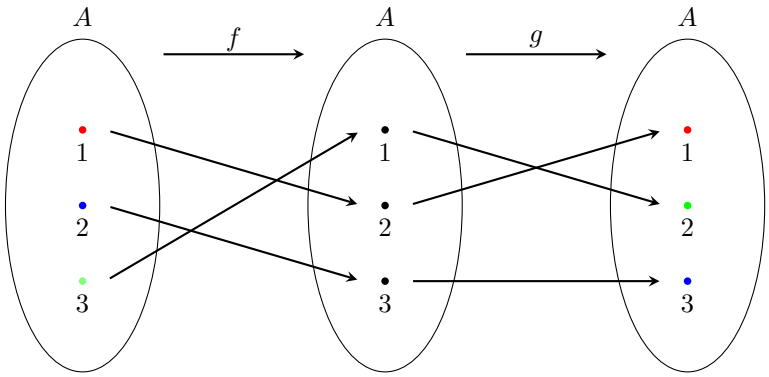


Figure 2.5: the composition $g \circ f$ in Example 2.2.4

If we have two permutations f and g of the same set A, then we can compose these permutations to a new one $f \circ g$ (as for functions in general, one pronounces this as "f after g" or "f composed with g") defined by $(f \circ g)[a] := f[g[a]]$. One should make sure that $f \circ g$ really is a permutation of A. According to the definition, we should check that $f \circ g$ is a function from A to A, but this is clear from the definition of the composition operator $\circ$, and that $f \circ g$ is a bijection. It is an exercise to show that indeed the function $f \circ g : A \to A$ is a bijection.

Example 2.2.4 This is a continuations of Example 2.2.3. 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 of A and is denoted by id_A. Then we have $f \circ f^{-1} = \mathrm{id}_A$ and $f^{-1} \circ f = \mathrm{id}_A$. If the set A is clear from the context, one often writes id instead of id_A.

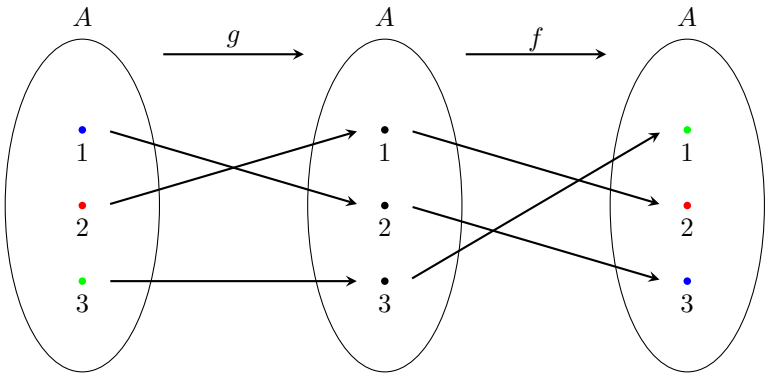


Figure 2.6: the composition $f \circ g$ in Example 2.2.4

Example 2.2.5 This is a continuations of Example 2.2.4. The identity element in S_3 is give by

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

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 $f^{-i} = (f^{-1})^i$. Finally, it is customary to define $f^0 = \mathrm{id}_A$. If there exists a positive integer i such that $f^i = \mathrm{id}_A$, then the *order* of a permutation $f \in S_A$, denoted by $\mathrm{ord}(f)$, is defined to be the *smallest* positive integer i such that $f^i = \mathrm{id}_A$. If for all positive integers i, $f^i \neq \mathrm{id}_A$, one says that f has infinite order: $\mathrm{ord}(f) = \infty$. If the order of a permutation is finite, it is equal to the minimum number of times we have to apply that permutation till all elements of the set A are mapped to themselves again.

Example 2.2.6 This is a continuation of Example 2.2.5. Let f and g be the permutations in S_3 as in Example 2.2.5. Then $g^2 = g \circ g = \mathrm{id}$. Since $g \neq \mathrm{id}$, this implies that $\mathrm{ord}(g) = 2$. Similarly, one can show that $\mathrm{ord}(f) = 3$.

With these notations in place, we collect the central properties for the composition of permutations in the following theorem:

Theorem 2.2.7 *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*

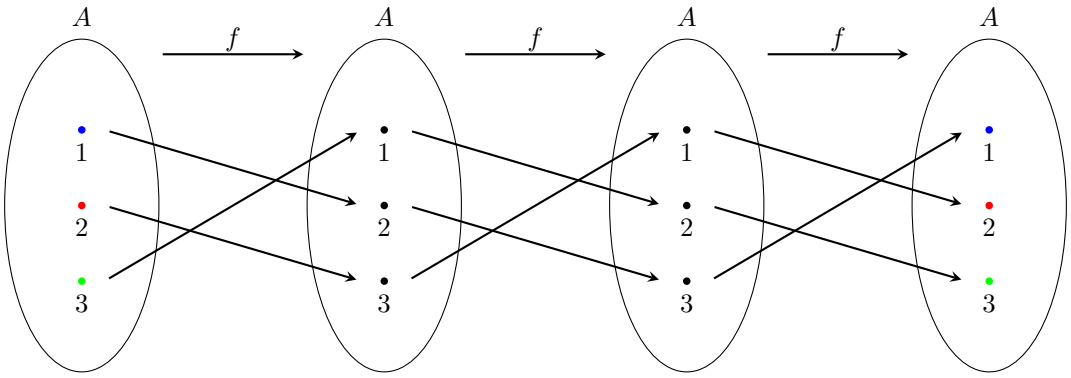


Figure 2.7: $f^3 = \mathrm{id}$ in Example 2.2.6

- $\forall f, g, h \in S_n$ *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. The first item of the theorem follows directly from Lemma 2.1.2 (note that $A = B = C = D$ in the setting of the theorem). We leave the proof of the remaining items to the reader. ■

Definition 2.2.8 *The pair* $(S_A, \circ)$ *is called the* symmetric group *on the set* A*. In case* $A = \{1, 2, \ldots, n\}$ *one says that* $(S_n, \circ)$ *is the symmetric group on* n *letters.*

As one might guess from the name "symmetric group", the S in the notation S_A is historic and comes from the word symmetric.

2.3 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 2.3.1 *Let* $m \geq 1$ *be an integer. A permutation* $f \in S_A$ *is called an* m-cycle *if there exist* m *distinct elements* $a_0, a_1, \ldots, a_{m-1}$ *in* A *such that*

$$f[a_i] = a_{(i+1) \bmod m} \text{ for } i = 0, \ldots, m-1 \text{ and } f[a] = a \text{ for } a \in A \setminus \{a_0, \ldots, a_{m-1}\}.$$

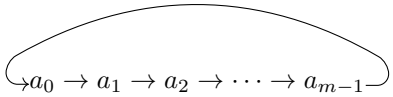


Figure 2.8: picture of an m-cycle

That is to say, f sends a_0 to a_1, a_1 to a_2, ..., a_{m-2} to a_{m-1} and a_{m-1} to a_0, while f fixes all other elements of A. One writes $f = (a_0\, a_1 \ldots a_{m-1})$.

A 1-cycle (a_0) is actually just the identity permutation id_A. Indeed, by definition it fixes any element from $A \setminus \{a_0\}$ and sends a_0 to $a_{(0+1) \bmod 1} = a_0$. A 2-cycle $(a_0\, a_1)$ is sometimes called a *transposition*, since all it does is interchange (transpose) a_0 and a_1.

Lemma 2.3.2 *Let $m \geq 1$ be an integer and $a_0, a_1, \ldots, a_{m-1}$ distinct elements of A. Then the m-cycle $(a_0\, a_1\, \ldots\, a_{m-1})$ has order m.*

Proof. Let us write $f = (a_0\, a_1\, \ldots\, a_{m-1})$. First of all, by definition of a cycle, f fixes any element in $A \setminus \{a_0, \ldots, a_{m-1}\}$. Then also f^n fixes any element in $A \setminus \{a_0, \ldots, a_{m-1}\}$ for any positive integer n. Note that for any integer n and i between 0 and $m-1$, we have $f^n[a_i] = a_{(i+n) \bmod m}$. Hence if $0 < n < m$, then $f^n[a_0] = a_{n \bmod m} = a_n \neq a_0$ implying that $f^n \neq id_A$. On the other hand, $f^m[a_i] = a_{(i+m) \bmod m} = a_i$. This shows that m is the smallest positive integer such that $f^m = \mathrm{id}_A$. ■

Two cycles $(a_0\, a_1\, \ldots\, a_{m-1})$ and $(b_0\, b_1\, \ldots\, b_{\ell-1})$ are called mutually disjoint, if the sets $\{a_0, a_1, \ldots, a_{m-1}\}$ and $\{b_0, b_1, \ldots, b_{\ell-1}\}$ are disjoint, that is to say, have no elements in common. Even though we have seen in Example 2.2.4 that in general it does not hold that $f \circ g = g \circ f$, this does hold if f and g are mutually disjoint cycles. If for two permutations $f, g \in S_A$ it holds that $f \circ g = g \circ f$, one says that f and g *commute*. The claim is therefore that mutually disjoint cycles commute. To show this is an exercise posed at the end of this chapter. More generally, we say that a collection of cycles is mutually disjoint, if any two of them are mutually disjoint.

Example 2.3.3 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. More generally one has

$$(a_0\, a_1\, \ldots\, a_{m-1}) = (a_1\, \ldots\, a_{m-1}\, a_0) = \cdots = (a_{m-1}\, a_0 \ldots\, a_{m-2}).$$

Hence one can "cycle around" the entries in an m-cycle, without changing the permutation.

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.

The cycles (1 2 3) and (4 5) are mutually disjoint. Indeed, in matrix form both (1 2 3) ∘ (4 5) and (4 5) ∘ (1 2 3) is given by

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

We can depict this permutation as follows:

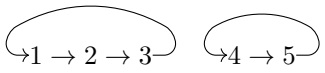


Figure 2.9: an example of two disjoint cycles

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

Example 2.3.4 We consider the permutation $f \in S_{11}$ defined by

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

If f can be written as a composition of mutually disjoint cycles, one of the cycles, say c, should contain the number 1. We may assume that this cycle starts with 1 by "cycling around" the entries in the cycle if necessary, that is to say $c = (1 \ldots)$. 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 $c = (1\,2 \ldots)$. Similarly, since $f[2] = 6$, we see that $c = (1\,2\,6 \ldots)$. 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

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

The 1-cycle (10) expresses that f *fixes* 10, that is to say, that $f[10] = 10$. It is customary not to write 1-cycles such as (10) in the disjoint cycle decomposition. If an element does not occur in the disjoint cycle decomposition of a permutation f, the rule is that it is 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 it matters in which order cycles are written. For example (1 2)(2 3) = (1 2 3), while (2 3)(1 2) = (1 3 2). However, as we already mentioned before, for mutually disjoint cycles the order does not matter. Therefore, we have for example:

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

Now we show that any permutation on a finite set A can be written as a composition of mutually disjoint cycles.

Theorem 2.3.5 *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 a composition of mutually disjoint cycles $c_1, \ldots, c_\ell$.*

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 strong induction on n. First we look at the case $n = 1$. In that case, the only permutation in S_A is the identity permutation id_A. Therefore the theorem is true for $n = 1$.

Suppose that $n \geq 2$ and that the theorem is true for S_B if B is a set such that $|B| \leq n - 1$. Let $f \in S_A$ be a permutation, where A is a set of cardinality n. Now choose $a \in A$. If $f[a] = a$, we can interpret f as a permutation of the set $A \setminus \{a\}$ and the induction hypothesis applies. We claim that the sequence $a, f[a], f^2[a], f^3[a], \ldots$ contains the element a more than once. Indeed, if this would not be the case, we could create an injective function from $\mathbb{Z}_{\geq 0}$ to A by sending i to $f^i[a]$. However, since $\mathbb{Z}_{\geq 0}$ is an infinite set and A is a finite set, this is not possible by Lemma 2.1.5. Therefore $f^i[a] = f^j[a]$ for some i and j with $i < j$. Composing with f^{-i}, we see that $a = f^{j-i}[a]$. This shows that the element a occurs more than once in the sequence, as was claimed.

Now let m be the smallest positive integer such that $f^m[a] = a$. Then we can define the m-cycle $c_1 = (a\, f[a] \ldots f^{m-1}[a])$. The permutation $c_1^{-1} \circ f$ fixes the element a, since $(c_1^{-1} \circ f)[a] = c_1^{-1}[f[a]] = a$. In fact by a similar reasoning, we see that $c_1^{-1} \circ f$ fixes all the elements $a, f[a], f^2[a], \ldots, f^{m-1}[a]$. If $m = n$, we have $c_1^{-1} \circ f = \mathrm{id}_A$, whence $f = c_1$ and we are done. Otherwise, we interpret $c_1^{-1} \circ f$ as a permutation of the set $A \setminus \{a, f[a], \ldots, f^{m-1}[a]\}$ consisting of $n - m$ elements. Using the induction hypothesis, we can write $c_1^{-1} \circ f$ as a product of mutually disjoint cycles, say

$$c_1^{-1} \circ f = c_2 \circ \cdots \circ c_\ell,$$

for certain mutually disjoint cycles $c_2, \ldots, c_\ell$. Since none of the cycles $c_2, \ldots, c_\ell$ contain any of the elements $a, f[a], \ldots, f^{m-1}[a]$, they are all mutually disjoint with c_1. Therefore

$$f = \mathrm{id} \circ f = (c_1 \circ c_1^{-1}) \circ f = c_1 \circ (c^{-1} \circ f) = c_1 \circ (c_2 \circ \cdots c_\ell),$$

gives the required decomposition of f in mutually disjoint cycles. This concludes the induction step. By the induction principle, we conclude that the theorem is true. ■

This theorem assures us that any permutation can be written as a composition of mutually disjoint cycles. We have already remarked that it is customary to refrain from writing 1-cycles. If the 1-cycles are removed, there is essentially only one way to write f as a composition of mutually disjoint cycles. The "essentially" in this statement just means that the only freedom one has is to change order of the cycles in the composition, which does not really matter since disjoint cycles commute. This is called the *disjoint cycle decomposition* of a permutation. The case of the identity permutation id_A is a bit special. In that case the disjoint cycle decomposition is simply taken to be id_A itself. The prove of the uniqueness of the disjoint cycle decomposition is not hard and follows from the proof of Theorem 2.3.5

Corollary 2.3.6 *Let n be a natural number, A be a set of cardinality n, and $f \in S_A$ distinct from the identity permutation. Suppose $f = c_1 \circ \cdots c_\ell = d_1 \circ \cdots d_k$ are two ways to write f as a composition of mutually disjoint cycles and also suppose that each cycle has order at least two. Then $k = \ell$ and after relabelling the cycles if necessary, we have $c_i = d_i$ for all i between 1 and ℓ.*

Proof. Like Theorem 2.3.5, we will use strong induction[1] on n. The basis case $n = 1$ is clear, since there is not permutation $f \in S_A$ distinct from the identity permutation id_A.

Now let $n > 1$ and assume the corollary holds for set of cardinality at most $n - 1$. Further, choose $a \in A$. If $f[a] = a$, we can interpret f as a permutation of the set $A \setminus \{a\}$ and apply the induction hypothesis. If $f[a] \neq a$, as in the proof of Theorem 2.3.5, let m be the smallest positive integer such that $f^m[a] = a$. There exists cycles c_r and d_s in which a occurs. Relabelling the cycles if necessary, we may assume that $r = s = 1$. Then $f[a] = c_1[a] = d_1[a]$ and applying f iteratively, we can conclude that for all integers $i \geq 0$ we have $f^i[a] = c_1^i[a] = d_1^i[a]$. But then both c_1 and d_1 are equal to the m-cycle $(a\, f[a] \ldots f^{m-1}[a])$, using that $f^m[a] = a$ and that m was the smallest positive integer with this property. Hence $c_1 \circ \cdots c_\ell = d_1 \circ \cdots d_k$ implies $c_2 \circ \cdots c_\ell = d_2 \circ \cdots d_k$ and the induction hypothesis applies, since we can interpret $c_2 \circ \cdots c_\ell$ as a permutation of $A \setminus \{a, f[a], \ldots, f^{m-1}[a]\}$. This concludes the induction step. By the induction principle, we conclude that the corollary is true. ■

Aside 2.3.7 *Inductive proofs are applied when proving a statement of the form $\forall n \in \mathbb{Z}_{\geq a}\, P(n)$, where $P(n)$ is some logical statement involving the parameter n and a is some base value, typically $a = 0$ or $a = 1$. The most common inductive proof is to prove $P(a)$ (the base case) and then to prove that the implication $P(n-1) \Rightarrow P(n)$ is true for all $n > a$ (the induction step). In this setting $P(n-1)$ is called the induction hypothesis. At first sight, the inductive proof used in Theorem 2.3.5 does not follow this pattern. Let us denote by $P(n)$ the statement that any $f \in S_A$ where A is a set of cardinality n, can be written as a composition of mutually disjoint cycles. Then the structure of the proof was as follows: in the base case $P(1)$ was showed, just as expected, but the induction hypothesis was not $P(n-1)$, but the stronger hypothesis that $P(\ell)$ is true for all $\ell < n$. In other words, what was shown was that the implication $(P(1) \wedge \cdots \wedge P(n-1)) \Rightarrow P(n)$ is true for all $n > 1$. This form of using induction is called* strong induction. *It turns out that the validity of strong induction actually can be shown using the "common" form of induction. Hence strong induction is in fact not stronger than common induction. The trick is to consider the logical statement $Q(n) := P(1) \wedge \cdots \wedge P(n)$. Then $Q(1) = P(1)$, so that proving $P(1)$ is the same as proving $Q(1)$. Further, the implication $Q(n-1) \Rightarrow Q(n)$ is logically equivalent to the implication $Q(n-1) \Rightarrow P(n)$, since $Q(n) = Q(n-1) \wedge P(n)$. Hence the proof of Theorem 2.3.5 can actually be viewed as a "usual" induction proof of the logical statement that $Q(n)$ is true for all positive integers n. Since $Q(n)$ implies $P(n)$, after all $P(n)$ is a part of $Q(n)$ by definition of $Q(n)$, we then know that $P(n)$ is true for all positive integers n as well.*

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:

[1] In strong induction, the induction hypothesis used in the induction step does not only assume the to be proven result for $n - 1$, but for all integers up till and including $n - 1$.

Example 2.3.8 Let us consider the composition $f \circ g$ with $f = (1\,2\,6)(3\,7)(4\,8\,11)(5\,9)$ and $g = (1\,7\,10\,2)(3\,9\,5)$, both permutations in S_{11}. We wish to compute the disjoint cycle decomposition of $f \circ g$. We simply proceed as in the proof of Theorem 1.3.7. In this case $A = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11\}$, so let us choose the element $1 \in A$ and determine the cycle of the disjoint cycle decomposition of $f \circ g$ containing the 1. Since

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

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

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

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

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

$$(f \circ g)^6[1] = (f \circ g)[(f \circ g)^5[1]] = (f \circ g)[6] = f[g[6]] = f[6] = 1,$$

we see that the first cycle in the disjoint cycle description of $f \circ g$ will be $(1\,3\,5\,7\,10\,6)$. Next we consider the element $2 \in A$, since 2 does not occur in $(1\,3\,5\,7\,10\,6)$. Since $(f \circ g)[2] = 2$, it is a fixed point corresponding to a 1-cycle (2). We do not write 1-cycles down explicitly though. Next we choose the element 4 and find that $(f \circ g)[4] = 8$, $(f \circ g)[8] = 11$ and $(f \circ g)[11] = 4$, obtaining the 3-cycle $(4\,8\,11)$. The only element in A, we have not considered is 9, and we can compute that $(f \circ g)[9] = 9$, so 9 is another fixed point. Therefore, we find that

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

is the disjoint cycle decomposition of $f \circ g$.

Now that we know that a permutation $f \in S_A$ for a finite set A has a unique disjoint cycle decomposition, the following definition makes sense.

Definition 2.3.9 *Let A be a finite set of cardinality n. Suppose that $f \in S_A$ has t_1 fixed points and that for each $m \geq 2$ exactly t_m cycles of length m occur in the disjoint cycle decomposition of f. Then the n-tuple $(t_1, t_2, \ldots, t_n)$ is called the* cycle type *of f.*

Example 2.3.10 Let us consider the permutation $f \circ g \in S_{11}$ from the previous example. We will determine its cycle type. First of all $t_1 = 2$, since $f \circ g$ fixes precisely the values 2 and 9. Further the disjoint cycle decomposition of $f \circ g$ is the composition of a 3-cycle and a 6-cycle. Therefore the cycle type of $f \circ g$ is $(2, 0, 1, 0, 0, 1, 0, 0, 0, 0, 0)$. Note that $2 + 3 \cdot 1 + 6 \cdot 1 = 11$.

Note that a permutation in S_A with $|A| = n$ cannot contain an m-cycle with $m > n$, simply because there are only n elements to permute. That is the reason t_m only was defined for $m \leq n$. In fact, if $f \in S_A$ has cycle type $(t_1, t_2, \ldots, t_n)$, then since there are n elements in A, we have $t_1 + 2t_2 + \cdots + nt_n = n$. This puts a strong restriction on which cycle types are possible.

Example 2.3.11 For small n it is not so hard to make a table of possible cycle types and the elements in S_n with that cycle type. For $n = 1, 2, 3$:

cycle type	elements in S_1	cycle type	elements in S_2	cycle type	elements in S_3
(1)	id	$(2,0)$	id	$(3,0,0)$	id
		$(0,1)$	$(1\,2)$	$(1,1,0)$	$(1\,2),(1\,3),(2\,3)$
				$(0,0,3)$	$(1\,2\,3),(1\,3\,2)$

For $n = 4$:

cycle type	elements in S_4
$(4,0,0,0)$	id
$(2,1,0,0)$	$(1\,2),(1\,3),(1\,4),(2\,3),(2\,4),(3\,4)$
$(1,0,1,0)$	$(1\,2\,3),(1\,3\,2),(1\,2\,4),(1\,4\,2),(1\,3\,4),(1\,4\,3),(2\,3\,4),(2\,4\,3)$
$(0,0,0,1)$	$(1\,2\,3\,4),(1\,2\,4\,3),(1\,3\,2\,4),(1\,3\,4\,2),(1\,4\,2\,3),(1\,4\,3\,2)$
$(0,2,0,0)$	$(1\,2)(3\,4),(1\,3)(2\,4),(1\,4)(2\,3)$

Previously, we introduced the order of a permutation and showed in Lemma 2.3.2 that an m-cycle has order m. Using the disjoint cycle decomposition of a permutation, it is easy to determine its order.

Proposition 2.3.12 *Let A be a finite set of cardinality n and let $f \in S_A$ have disjoint cycle decomposition $f = c_1 \circ \cdots \circ c_\ell$, where $c_1, \ldots, c_\ell$ are mutually disjoint cycles of lengths $m_1, \ldots m_\ell$. Then the order of f is the least common multiple of $m_1, \ldots, m_\ell$, that is*

$$\operatorname{ord}(f) = \operatorname{lcm}(m_1, \ldots, m_\ell).$$

Proof. Suppose that $f = c_1 \circ \cdots \circ c_\ell$, where $c_1, \ldots, c_\ell$ are mutually disjoint cycles of lengths $m_1, \ldots, m_\ell$. Since mutually disjoint cycles commute, for any integers $i \geq 0$ we have $f^i = c_1^i \circ \cdots \circ c_\ell^i$. Let us denote by $A_k \subset A$ the set of elements occurring in cycle c_k. Then the permutation f^i fixes all elements in the set A_k if and only if c_k^i fixes all elements in the set A_k, which in turn is the case if and only if i is a multiple of m_k, the length of the cycle c_k. Hence f^i is the identity if and only if i is a common multiple of $m_1, \ldots, m_\ell$. This implies that the smallest positive integer i such that $f^i = \mathrm{id}_A$ is equal to the least common multiple of $m_1, \ldots, m_k$. ■

Note that to determine the order of a permutation, it is sufficient to know its cycle type. After all, the cycle type of a permutation f tells you precisely which lengths the cycles have that occur in the disjoint cycle decomposition of f. In particular, we can conclude that any two permutations that have the same cycle type, have the same order.

Example 2.3.13 The number of permutations in S_5 is $5! = 120$. This number is already large enough to make it cumbersome to write a table such as in Example 2.3.11. The number of cycle types is still rather modest though. Using Proposition 2.3.12, we can also determine their orders quite easily:

cycle type	shape of permutation	order of permutation
$(5,0,0,0,0)$	id	1
$(3,1,0,0,0)$	$(a\,b)$	2
$(2,0,1,0,0)$	$(a\,b\,c)$	3
$(1,2,0,0,0)$	$(a\,b)(c\,d)$	2
$(1,0,0,1,0)$	$(a\,b\,c\,d)$	4
$(0,1,1,0,0)$	$(a\,b)(c\,d\,e)$	6
$(0,0,0,0,1)$	$(a\,b\,c\,d\,e)$	5

Apparently, the largest possible order an element of S_5 can have is 6.

Aside 2.3.14 *Let n be a positive integer, A a set with cardinality n, and $(t_1,\ldots,t_n)\in\mathbb{Z}^n_{\geq 0}$ an n-tuple satisfying $\sum_{i=1}^n it_i = n$. Then there are exactly $n!\prod_{i=1}^n i^{-t_i}/\prod_{i=1}^n t_i!$ many permutations in S_A with cycle type $(t_1,\ldots,t_n)$. This formula can be understood for example in the following way. If the n elements of A are given in some order, say $a_1,\ldots,a_n$, then a permutation with cycle type $(t_1,\ldots,t_n)$ can be obtained by taking the first t_1 elements as 1-cycles, the next $2t_2$ elements as entries in consecutive 2-cycles, the next $3t_3$ as entries in consecutive 3-cycles, and so on. Then $n!$ in the formula simply comes from the total ways to order the n elements of A. However, distinct orderings can give rise to the same permutation, since the consecutive i-cycles commute with each other and since the elements within each i-cycle can be shifted in a cyclic way. The term i^{-t_i} takes care of the cyclic shifts within the t_i many i-cycles, while dividing by $t_i!$ takes care of the i-cycles commuting with each other.*

Another question is to determine the number of possible cycle types in S_A if $|A| = n$. A cycle type then corresponds to what is known as a partition of the number n. In this context, a partition is a sequence of positive integers $s_1 \leq \cdots \leq s_\ell$ such that $s_1 + \cdots s_\ell = n$. Given a cycle type $(t_1,\ldots,t_n)$, we immediately find a partition of n, by choosing the first t_1 values of s_i equal to 1, the next t_2 values of s_i equal to 2, and so on. Then the sum of all s_i is equal to n, since $t_1 + 2t_2 + \cdots + nt_n = n$. For example the seven cycle types found in Example 2.3.13 for $n = 5$, correspond to the following seven partitions $1+1+1+1+1, 1+1+1+2, 1+1+3, 1+2+2, 1+4, 2+3, 5$ Hence the number of possible cycle types in S_A is the same as the number of partitions of n, often denoted by $p(n)$. No closed formula is known for $p(n)$, but G.H. Hardy and S. Ramanujan published in 1918 a good approximation for large n. Their results imply that $p(n)$ grows asymptotically as $\frac{1}{4n\sqrt{3}}\exp\left(\pi\sqrt{\frac{2n}{3}}\right)$ as n tends to infinity.

2.4 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 $\text{sign} : S_A \to \{-1, 1\}$ 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 2.4.1 *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:

$$M_{g\circ f} = M_f \cdot M_g, \text{ where } \cdot \text{ denotes matrix multiplication.} \tag{2.1}$$

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

$$\operatorname{sign}(f) := \det(M_f). \tag{2.2}$$

Equation (2.1) implies that

$$\operatorname{sign}(f)\operatorname{sign}(g) = \det(M_f)\det(M_g) = \det(M_f M_g) = \det(M_{g\circ f}) = \operatorname{sign}(g \circ f). \tag{2.3}$$

This is a very useful property when investigating or computing the sign of a permutation. Let us for example show that $\operatorname{sign}(f)$ only can take the values ± 1.

Lemma 2.4.2 *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 (2.3) 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 $\operatorname{sign}(f) \in \{-1, 1\}$. ∎

As a consequence from the proof of this lemma, we also obtain that

$$\operatorname{sign}(f^{-1}) = \operatorname{sign}(f). \tag{2.4}$$

Indeed, the proof showed that $\operatorname{sign}(f^{-1}) = 1/\operatorname{sign}(f)$, but since the sign can be 1 or -1 only, $1/\operatorname{sign}(f) = \operatorname{sign}(f)$. Then equation (2.4) follows.

Permutations with sign equal to 1 are called *even* permutations. Those with sign equal to -1 are called odd permutations. The reason for this terminology is that an even permutation can be written as a composition of an even number of 2-cycles, while an odd permutation can be written as a composition of an odd number of 2-cycles. That this is true, will be shown in the exercises.

Example 2.4.3 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.

Equation (2.3) can be used to compute the sign of permutations without using equation (2.2) directly. For example, we have

Lemma 2.4.4 *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 four 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 (2.3), 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 2.3.5 any permutation can be written as the composite of (mutually disjoint) cycles. Using the above lemma and equation (2.3), it is now easy to compute the sign of a permutation.

Example 2.4.5 This example is a continuation of Example 2.3.8. 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 (2.3) an obtain that: $\text{sign}(f \circ g) = \text{sign}(f)\text{sign}(g) = -1$. Just to

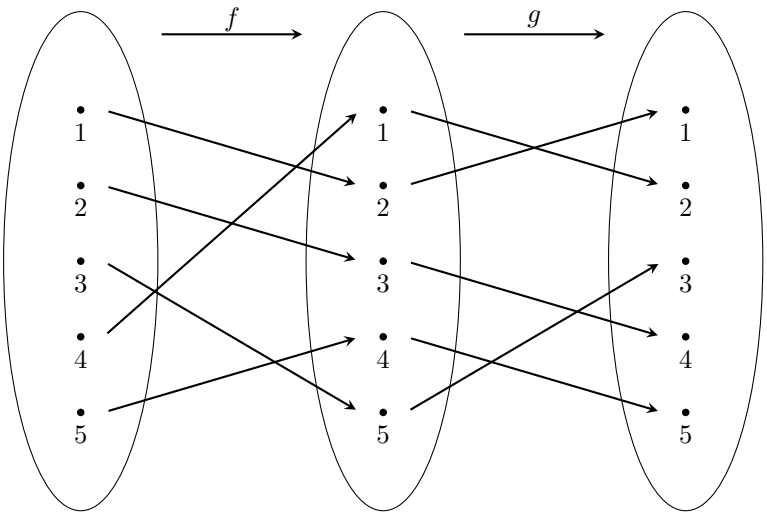


Figure 2.10: even and odd permutations

check, let us use the computation from Example 2.2.6 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.

Remark 2.4.6 *A visual intuition regarding even and odd permutations can be given as follows. To describe two permutations f and g on the set $\{1, 2 \ldots, n\}$ we can draw the static arrow diagram where the numbers in the domain and codomain match up horizontally. It is then likely that some arrows will need to cross. It is not difficult to arrange the crossings so that no three or more arrows cross at the same point. Then it turns out that if the number of crossings is an even number then the permutation is even, while if the number of crossings is odd then the permutation is odd. For an illustration, see Figure 2.10 where the given permutation f is even, while g is odd.*

Aside 2.4.7 *It is possible to give a compact formula for the determinant of a square matrix A 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

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

For example for $n = 2$:

$$\det \begin{pmatrix} a_{1\,1} & a_{1\,2} \\ a_{2\,1} & a_{2\,2} \end{pmatrix} = \text{sign}(\text{id})a_{1\,1}a_{2\,2} + \text{sign}(\,(1\,2)\,)a_{1\,2}a_{2\,1} = a_{1\,1}a_{2\,2} - a_{1\,2}a_{2\,1}.$$

Proving that his formula holds for general n is easiest done using an abstract description of the determinant of $n \times n$ matrices: it is the unique alternating n-linear form that takes the value 1 on the identity matrix. What the "n-linear form" means is that the determinant is a linear map, when viewed as a map in each row separately. Alternating means that if one interchanges two rows, the sign of the determinant changes. Once one has shown that the determinant is the only map that can satisfy all these requirements, it is not so hard anymore to show that the given formula for the determinant is correct. The most tricky is to verify the alternating property. If rows i and j are interchanged, equation (2.3) *implies that the determinant simply changes sign.*

2.5 Extra: problems involving permutations

Permutations play an important role in the study of combinatorial problems and algorithms. In this section, we will illustrate applications of permutations in this area. There are many possible, but we have chosen two problems: one is the 100 prisoners problem, illustrating the use of the cycle structure of a permutation, one is the famous 15-puzzle, illustrating the use of the sign of a permutation.

2.5.1 The 100 prisoners problem

The following problem is a variation on a problem posed by the Danish computer scientist Peter Bro Miltersen in 2003. Imagine a prison with very cruel wardens and a hundred prisoners. Each prisoner has an identifying number between 1 and 100. One day, the wardens come up with a game.

The wardens made a hundred pieces of paper, each piece having a number between 1 and 100 written on it. No two pieces of paper have the same number written on it, so each number between 1 and 100 occurs on exactly one piece. Further, they procure a hundred identical looking boxes and just like the prisoners, each box has an identifying number between 1 and 100 written on it. Now the wardens arbitrarily place in each box one of the pieces of paper. After closing all the boxes, they place them on the courtyard.

Next they explain a game to the prisoners. Each prisoner is to go to the courtyard, one at the time. A prisoner is then allowed to open up to fifty boxes in order to find the piece of paper with the prisoner's own identifying number. If successful, the prisoner shows this paper to the wardens, otherwise the wardens declare a failure. Before the next prisoner comes in, all pieces of paper are placed were they were and all opened boxes are closed again. After a prisoner is done, no communication with the other prisoners is allowed, though they are allowed to agree upon a joint strategy before the game starts. So far so good, but the catch is that if at least one of the prisoners fails, all prisoners will be shot.

The challenge is to come up with a strategy that gives the prisoners at least a reasonable chance to survive. A first idea might be the following. If one of the prisoners arbitrarily opens 50 boxes, there is a fifty percent chance that that prisoner's number is found, but all the other

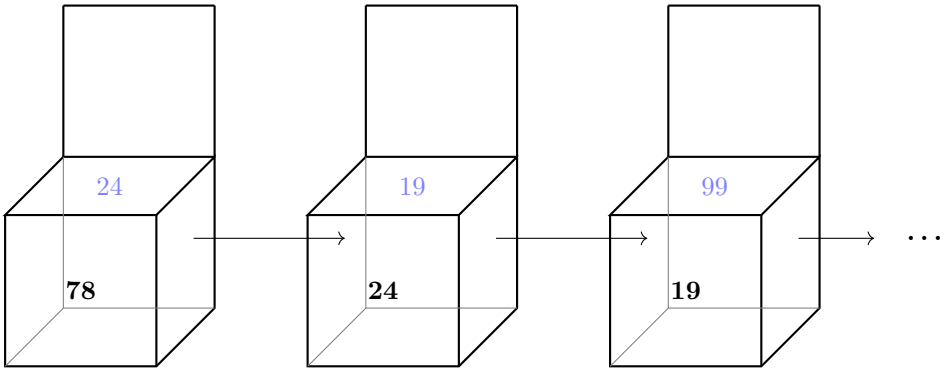


Figure 2.11: strategy for the 100 prisoners problem

prisoners also have to find their numbers if they are to survive. This strategy therefore has a success rate of only $(1/2)^{100}$.

A fascinating alternative strategy is for each prisoner to go to the box with the prisoner's number on it, open that box, read the number on the piece of paper in it, then go toward the box with that number on it and to repeat that process till either fifty boxes have been opened or the prisoner found the box with the right number in it. The key insight behind this strategy is that the hundred boxes and the hundred pieces of paper in it, really describe a permutation f in S_{100}. More precisely, box i contains the number $f[i]$. What is the success rate of this strategy?

Answer: prisoner i is then looking for box number $f^{-1}[i]$. Now suppose the the disjoint cycle decomposition of f is $c_1 \circ \cdots \circ c_\ell$ and that these cycles have lengths $m_1, \ldots, m_\ell$. If the outlined alternative strategy is followed, then prisoner i will enter the cycle containing the number i, say cycle c_k. Then $f^{-1}[i] = c_k^{-1}[i] = c_k^{m_k-1}[i]$ so that prisoner i in principle will find the right box after m_k turns. As long as $m_k \leq 50$ the outlined strategy will therefore work for prisoner i. Consequently, the outlined strategy will work for all prisoners, if and only if all cycles $c_1, \ldots, c_\ell$ have length 50 or less.

To compute the success rate now boils down to finding out how many permutations in S_{100} have no cycles of length 51 or more in their disjoint cycle decompositions. First of all, note that there are exactly $100!/((100-m)!m)$ m-cycles in S_{100}. Now note that the cycle type of $f \in S_{100}$ can contain only one cycle of length 51 or more. Therefore, if $51 \leq m \leq 100$, there are exactly $100!/m$ permutations whose disjoint cycle decomposition contains a cycle of length m. This shows that the probability that an arbitrarily chosen permutation does not contain a cycle of length at least 51 is:

$$1 - \sum_{m=51}^{100} \frac{1}{m} \approx 0.31183$$

This success rate is surprisingly large.

In fact even more surprisingly, the success rate hardly changes if the same game is played with $2n$ prisoners, a prisoner is allowed to open n boxes in each round, and n tends to infinity. Using that $\lim_{n\to\infty} -\log(n) + \sum_{m=1}^{n} \frac{1}{m} = \gamma$, where γ denotes the Euler–Mascheroni constant

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

Figure 2.12: target constellation of the 15-puzzle

and the base of the logarithm is e, the success rate is

$$\lim_{n\to\infty} 1 + \sum_{m=1}^{n} \frac{1}{m} - \sum_{m=1}^{2n} \frac{1}{m} = 1 - \log(2) \approx 0.30685$$

2.5.2 The 15-puzzle

The 15-puzzle is a game where the goal is to rearrange 15 squares from any given constellation into the target constellation depicted in Figure 2.12

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. The empty square is denoted as square 16. The possible constellations can be described using permutations in S_{16}. More precisely, a permutation $f \in S_{16}$ corresponds to the constellation in Figure 2.13

In particular, the target constellation corresponds to the identity permutation $\mathrm{id} \in S_{16}$. The question is now whether or not the initial constellation given by $f = (14\,15)$, that is the configuration in Figure 2.14, can be rearranged into the target constellation using legal moves. In other words, 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, one studies the effect of a "legal move" in terms of permutations. First of all, if a constellation C is described by the permutation $f \in S_{16}$, then $a := f^{-1}[16]$ describes the position of the empty square, since $f[a] = 16$. Hence the legal moves can give rise to constellations of the form $f \circ (a\,b)$, where b is a neighbouring square of square a.

Now suppose that a constellation C is changed in such a way that the empty square ends up in the same position again. If the empty square originally was in position (i, j), then one legal move changes this position to one of the four positions $(i\pm1, j\pm1)$. Note that not all four positions may

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

Figure 2.13: constellation corresponding to a permuation f

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

Figure 2.14: an impossible configuration?

be possible, since the empty square may be on the border. Regardless, each legal move changes the parity of the sum of the coordinates of the position of the empty square. Therefore, to end up in the same position, necessarily an even number of moves was used. This means that if two constellations described by two permutations $f, g \in S_{16}$ can be changed from the one to the other using legal moves, then $f = g \circ (a_1\, b_1) \circ \ldots \circ (a_\ell\, b_\ell)$ for certain integers a_i, b_i and *even* integer ℓ. In particular $\text{sign}(f) = \text{sign}(g)$. This implies immediately that the constellation from Figure 2.14, corresponding to the odd permutation (14 15), cannot be transformed to the one from Figure 2.12 corresponding to the even permutation id, using legal moves. Playing with the puzzle a bit, it is not so hard to see that any constellation corresponding to a 3-cycle can be transformed to Figure 2.14 with legal moves. Hence a constellation corresponding to a permutation $f \in S_{16}$ can be transformed to Figure 2.14 with legal moves if and only if $\text{sign}(f) = 1$.

2.6 Exercises

Multiple choice exercises

1. Let $f : A \to B$ and $g : B \to C$ be two functions.
   - A. the domain and co-domain of f are B and A respectively
     TRUE ☐ FALSE ☐
   - B. it is always true that $B = \text{im} f$
     TRUE ☐ FALSE ☐
   - C. $\text{im} f$ is the subset of elements $b \in B$ that can be written as $b = f[a]$ for some $a \in A$
     TRUE ☐ FALSE ☐
   - D. $g \circ f$ is defined as $(g \circ f)[a] = g[f[a]]$ for all $a \in A$
     TRUE ☐ FALSE ☐

2. Let $f : A \to B$ be a function. Then f is *surjective* if and only if
   - A. ☐ $B = \text{im} f$, that is, for all $b \in B$ there exists $a \in A$ such that $a = f[b]$
   - B. ☐ $B = \text{im} f$, that is, for all $b \in B$ there exists $a \in A$ such that $b = f[a]$
   - C. ☐ A is the domain of f
   - D. ☐ $f[a_1] = f[a_2]$ implies $a_1 = a_2$

3. Let $f : A \to B$ be a function. Then f is *injective* if and only if
   - A. ☐ for all $b \in B$ there exists $a \in A$ such that $a = f[b]$
   - B. ☐ whenever $f[a_1] = f[a_2]$ one has $a_1 = a_2$
   - C. ☐ $|B| \geq |A|$
   - D. ☐ every element in A has exactly one image in B

4. A. A *permutation* of a set A is a bijective function $f : A \to A$
   TRUE ☐ FALSE ☐

B. S_n is the set of permutations on $\{1, \ldots, n+1\}$
TRUE ☐ FALSE ☐

C. if
$$f = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 4 & 1 \end{pmatrix}$$
then $f[3] = 1$
TRUE ☐ FALSE ☐

D. if
$$f = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 3 & 2 & 4 \end{pmatrix}$$
then $f \in S_4$
TRUE ☐ FALSE ☐

E. if
$$f = \begin{pmatrix} a_1 & a_2 & \ldots & a_n \\ b_1 & b_2 & \ldots & b_n \end{pmatrix}.$$
then $b_i = f[a_i]$ for all $i = 1, \ldots, n$
TRUE ☐ FALSE ☐

F. $|S_n| = n!$
TRUE ☐ FALSE ☐

5. A permutation $f \in S_A$ is called an m-cycle if and only if there exist distinct elements $a_0, \ldots, a_{m-1} \in A$ such that

A. ☐ $f[a_i] = a_{i+1}$ for all $i = 0, \ldots, m-1$,

B. ☐ $f[a_i] = a_{(i+1) \pmod m}$ for all $i = 0, \ldots, m-1$ and $f[a] = a$ for all $a \in A \setminus \{a_0, \ldots, a_{m-1}\}$,

C. ☐ $f[a_i] = a_{i \pmod m}$ for all $i = 0, \ldots, m-1$ and $f[a] = a$ for all $a \in A \setminus \{a_0, \ldots, a_{m-1}\}$,

D. ☐ f fixes all elements in A but $a_0, \ldots, a_{m-1}$

E. ☐ f has order m.

6. A. The cycles $(123), (14) \in S_4$ are disjoint
TRUE ☐ FALSE ☐

B. Two permutations $f, g \in S_A$ commute if and only if $f[g[a]] = g[f[a]]$ for some $a \in A$
TRUE ☐ FALSE ☐

7. Let $f \in S_n$ be a permutation of cycle type $(t_1, \ldots, t_n)$. Then

A. if $t_1 = 0$ then f does not fix any value in $\{1, \ldots, n\}$
TRUE ☐ FALSE ☐

B. t_i is the number of i-cycles appearing in the disjoint cycles decomposition of f
TRUE ☐ FALSE ☐

C. if $t_n = 1$ then f is an n-cycle
TRUE ☐ FALSE ☐

D. to understand whether a permutation with cycle type $(t_1, \ldots, t_n)$ can exist on can first check that $\sum_{i=1}^{n} it_i = n$
TRUE ☐ FALSE ☐

8. Which of the following statements is true for a permutation $f \in S_n$?
   A. ☐ $\text{sign}(f) \in \{-1, 1\}$
   B. ☐ $\text{sign}(f \circ g) = \text{sign}(f) + \text{sign}(g)$
   C. ☐ if $(M_f)_{i,j} = 1$ then $f[i] = j$
   D. ☐ if f is an m-cycle then $\text{sign}(f) = (-1)^m$
   E. ☐ $\text{sign}(f) = \text{sign}(f^{-1})$

Exercises to get to know the material better

9. Compute the disjoint cycle decomposition of the permutation
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 \\ 2 & 3 & 7 & 5 & 4 & 1 & 6 \end{pmatrix}.$$

10. Given the permutations $f = (1\,4)(2\,3\,6)$ and $g = (1\,3\,2)(4\,5\,6)$ from S_6, compute the disjoint cycle decomposition of $f \circ g$ and $g \circ f$.

11. Show that the sign of a 2-cycle is -1 using equation (2.2) directly.

12. Determine which of the following functions are surjective, injective, bijective or permutations.
   (a) $f : \mathbb{Z}_6 \to \mathbb{Z}_3$ defined by $x \mapsto x \pmod 3$,
   (b) $f : \mathbb{Z}_4 \to \mathbb{Z}_4$ defined by $x \mapsto x^2 \pmod 4 = x \cdot_4 x$,
   (c) $f : \mathbb{Z}_6 \to \mathbb{Z}_6$ defined by $x \mapsto (x + 3) \pmod 6 = x +_6 3$,
   (d) $f : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto 3x + \sqrt{2}$,
   (e) $f : \mathbb{Z} \to \mathbb{Z}$ defined by $x \mapsto |x|$.

13. Determine which of the following functions are surjective, injective, bijective or permutations.
   (a) $f : \mathbb{R}^2 \to \mathbb{R}^2$ define by $v \mapsto v + (1, 1)$,
   (b) $f : \mathbb{R}^2 \to \mathbb{R}^2$ define by $v \mapsto Av$, where A is a 2×2 matrix with coefficients in $\mathbb{R}$ and rank 1.

14. Write the following permutations as a product of disjoint cycles
   (a)
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 5 & 1 & 2 & 4 & 3 \end{pmatrix},$$
   (b)
$$\begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 8 & 2 & 3 & 5 & 6 & 7 & 1 & 4 \end{pmatrix},$$

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

(d) $(1\,2)(3\,5\,1)$,

(e) $(1\,6)(1\,5)(1\,4)(1\,3)(1\,2)$.

15. (a) Show that there exists an element of order 6 in S_5,

(b) show that there does not exist an element of order 8 in S_5

(c) find all the possible orders of the elements in S_5.

16. For which of the following values of n and cycle types $(t_1, ..., t_n)$ there exists a permutation $\sigma \in S_n$ with cycle type $(t_1, ..., t_n)$?

(a) $n = 5$, $(1, 2, 3, 2, 0)$,

(b) $n = 6$, $(0, 2, 2, 0, 0, 0)$,

(c) $n = 9$, $(1, 1, 2, 0, 0, 0, 0, 0, 0)$,

(d) $n = 7$, $(0, 2, 1, 0, 0, 0, 0)$.

17. Compute the order of the permutations in Exercise 14.

18. Which is the largest order that an element $f \in S_5$ can have?

Exercises to get around in the theory

19. 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 $f \circ g$, is a permutation of the set A.

(b) Check the second item in Theorem 2.2.7: the identity permutation $\mathrm{id} \in S_A$ satisfies $\mathrm{id} \circ f = f$ and $f \circ \mathrm{id} = f$ for any $f \in S_A$.

20. 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 all 12 of its rotational symmetries. Now enumerate the vertices of a tetrahedron from 1 up to 4. Any rotational symmetry of the tetrahedron permutes these vertices and therefore gives rise to a permutation from S_4. Describe the permutations one can obtain from the rotational symmetries of the tetrahedron. Are all elements of S_4 obtained?

21. (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.

(d) Going back to symmetries of the tetrahedron as in the previous exercise: what mirror symmetries does the tetrahedron have? Show that any of its rotational symmetries can be written as the composite of such mirror symmetries.

22. Let f and g in S_A be two disjoint cycles. Show that f and g commute, that is to say, show that $f \circ g = g \circ f$.

23. Show equation (2.1): $M_{g \circ f} = M_f \cdot M_g$, 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 $(A \cdot B)_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}$.

24. 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.

25. 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.

26. For those wanting to try a case of strong induction as described in Remark 2.3.7: prove that any integer $n > 1$ can be written as the product of prime numbers.

27. Let $n \in \mathbb{N}$ with $n \geq 20$. Let us call a cycle in S_n non-trivial if it is not the identity mapping.

    (a) Find a cycle α such that α^3 is the product of 3 non-trivial disjoint cycles.

    (b) Find a cycle β such that β^{12} is the product of 4 non-trivial disjoint cycles.

28. Prove that if $f, g \in S_n$ have the same cycle type that f and g have the same order.

29. Prove that if $f \in S_n$ is a cycle of odd length, then f^2 is also a cycle. Show that this is not true for cycles of even length by giving a counterexample.

30. Let $f : A \to B$ be a function. A function $g : B \to A$ is called a left-inverse of f if $g \circ f = \mathrm{id}_A$. Similarly, g is called a right-inverse of f if $f \circ g = \mathrm{id}_B$.

    (a) Prove that f has a left inverse if and only if f is injective.

    (b) Prove that f has a right inverse if and only if f is surjective.

Chapter 3

Groups

3.1 Abstract groups

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

In the previous chapter, we studied permutations. The essential ingredients were S_A, the set of permutations of a set A, and the composition operator $\circ$. In Chapter 1, we considered the set $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ of remainders modulo n and saw that it is possible to define the operator $+_n$ on it. In this chapter, we will capture some common ingredients of these examples and define a more abstract structure called a group . More precisely:

Definition 3.1.1 *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 group 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}).*

If the group operator $\cdot : G \times G \to G$ is clear from the context, one sometimes calls G itself a group, rather than $(G, \cdot)$.

Example 3.1.2 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 2.2.7.

Example 3.1.3 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 3.1.1 was f^{-1}, but when the group operation is an addition, it is common to write $-f$ for the inverse of f, just as we did for integers. It is possible to prove that the associative law $f + (g + h) = (f + g) + h$ holds for integers, but we will not do this. A possible proof would involve induction on f, g and h. When solving exercises, you may freely use that addition is associative for integers.

The group $(\mathbb{Z}, +)$ is an example of a 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 $f, g \in G$, the group is called an *abelian* group, after the famous mathematician Niels Henrik Abel. Other examples of abelian groups are $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$. As before, $\mathbb{Q}$ denote the set of rational numbers, $\mathbb{R}$ the set of real numbers, and $\mathbb{C}$ the set of complex numbers.

For a group $(G, \cdot)$, the group operator $\cdot$, assigns to any pair $(f, g) \in G \times G$ a group element $f \cdot g \in G$. Therefore it was in Definition 3.1.1 specified as a function $\cdot : G \times G \to G$. Sometimes it is more convenient to specify a group operation using a table. Depending on the concrete group operation, such a table is called a *multiplication table* or an *addition table*. Such a table has the form:

$\cdot$	$\cdots$	g	$\cdots$
$\vdots$	$\ddots$	$\vdots$	
f	$\cdots$	$f \cdot g$	$\cdots$
$\vdots$		$\vdots$	$\ddots$

For example, for $(S_3, \circ)$, we obtain:

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

Note that the multiplication table for $(S_3, \circ)$ is not symmetric, which illustrates that it is not an abelian group. For groups $(G, \cdot)$ where $|G| = \infty$, multiplication/addition tables are of course not a practical way to specify the group operation, but for groups where $|G| < \infty$, it can be very convenient. All four of the groups $(\mathbb{Z}, +)$, $(\mathbb{Q}, +)$, $(\mathbb{R}, +)$, and $(\mathbb{C}, +)$, have infinitely many elements. Such groups are said to have infinite order. More generally, we have the following:

Definition 3.1.4 *The order of a group $(G, \cdot)$ is defined as the number of elements in G, that is to say as $|G|$. If $|G| = n < \infty$, the group is called* a finite group of order n *or* a group with n elements.

Example 3.1.5 Let n be a positive integer, then $(\mathbb{Z}_n, +_n)$ is a group. The associativity of the operator $+_n$ was shown in Theorem 1.4.3. The identity element of this group is 0, since $a +_n 0 = a = 0 +_n a$ for all $a \in \mathbb{Z}_n$. If $a \in \mathbb{Z}_n$, the inverse of $a \in \mathbb{Z}_n$ with respect to $+_n$ is given by $(-a) \bmod n$, which is equal to $n - a$ if $a \neq 0$ and 0 if $a = 0$. Indeed $a +_n (n - a) = 0 \bmod n = 0$ and likewise $(n - a) +_n a = 0 \bmod n = 0$. This group is an abelian group, since $a +_n b = b +_a$ for all $a, b \in \mathbb{Z}_n$, see part 1 of Theorem 1.4.3. The group $(\mathbb{Z}_n, +_n)$ is a finite group of order n, since $|\mathbb{Z}_n| = n$.

Example 3.1.6 Let $n > 1$ be an integer, then $(\mathbb{Z}_n, \cdot_n)$ is **not** a group, though it comes close. The operator $\cdot_n$ is associative according to part 4 of Theorem 1.4.3 and 1 is its identity element. The problem is, that not all elements have a multiplicative inverse modulo n, that is to say, it is not true in general that for any element $a \in \mathbb{Z}_n$ there exists $b \in \mathbb{Z}_n$ such that $a \cdot_n b = 1$ and $b \cdot_n a = 1$. For example, 0 does not have a multiplicative inverse modulo n, since $0 \cdot_n b = 0$ for all $b \in \mathbb{Z}_n$.

As a concrete example, consider $n = 6$. The elements 1 and 5 in $\mathbb{Z}_6$ have a multiplicative inverse modulo 6, since $1 \cdot_6 1 = 1$ and $5 \cdot_6 5 = 25 \bmod 6 = 1$. The elements 0, 2, 3, and 4 do not have a multiplicative inverse modulo 6. To see why, one can simply try all possibilities and observe that a product of one of these elements with any other element in $\mathbb{Z}_6$ never gives 1. For example, for the element 2, we have $2 \cdot_6 0 = 0$, $2 \cdot_6 1 = 2$, $2 \cdot_6 2 = 4$, $2 \cdot_6 3 = 0$, $2 \cdot_6 4 = 2$, and $2 \cdot_6 5 = 4$. Observe, that the elements in $\mathbb{Z}_6$ having a multiplicative inverse, are exactly those elements that have greatest common divisor 1 with 6.

Let us now investigate for general n, which elements in $\mathbb{Z}_n$ have a multiplicative inverse modulo n. First of all, if a does have such an inverse, then $a \cdot_n b = 1$, meaning that $a \cdot b \equiv 1 \pmod{n}$. Therefore there exists $k \in \mathbb{Z}$ such that $a \cdot b + k \cdot n = 1$. This in turn implies that any common divisor of a and n also divides 1. But then the greatest common divisor of a and n is 1, that is $\gcd(a, n) = 1$. Conversely, if $\gcd(a, n) = 1$, then using the extended Euclidean algorithm, one can find integers b and k such that $a \cdot b + k \cdot n = 1$. We may even choose b such that $0 \leq b < n$, because $a \cdot (b + \ell n) + (k - \ell a) \cdot n = 1$ for any $\ell \in \mathbb{Z}$. Let us therefore assume that $b \in \mathbb{Z}_n$. The equation $a \cdot b + k \cdot n = 1$ then implies that $b \cdot_n a = a \cdot_n b = 1$ and hence that b is the multiplicative inverse of a with respect to the operation $\cdot_n$. All in all we have shown that an element $a \in \mathbb{Z}_n$ has a multiplicative inverse modulo n if and only if $\gcd(a, n) = 1$.

Now define

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

Then $((\mathbb{Z}_n)^*, \cdot_n)$ is a group. Part 3 of Theorem 1.4.3 implies that this group is abelian. If $n = 15$, we obtain for example that $(\mathbb{Z}_{15})^* = \{1, 2, 4, 7, 8, 11, 13, 14\}$. It is clear that $((\mathbb{Z}_n)^*, \cdot_n)$ is a finite group, but it is not so clear what the order of this group is. Let us define

$$\phi(n) := |\{a \in \mathbb{Z}_n \mid \gcd(a, n) = 1\}|. \tag{3.1}$$

Then $((\mathbb{Z}_n)^*, \cdot_n)$ is a finite group of order $\phi(n)$. We have just seen that $|(\mathbb{Z}_{15})^*| = 8$, so we have for example $\phi(15) = 8$. The function $\phi : \mathbb{Z}_{\geq 1} \to \mathbb{Z}_{\geq 1}$ is called *Euler's totient function*. It will come up several times later on. To simplify that notation, sometimes we will write $\mathbb{Z}_n^*$ instead of $(\mathbb{Z}_n)^*$.

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 and standard facts about sets and functions. 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. An example of this is the following lemma.

Lemma 3.1.7 *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:

$$e_1 \cdot f = f \text{ and } f \cdot e_1 = f \text{ for any } f \in G \tag{3.2}$$

and

$$e_2 \cdot f = f \text{ and } f \cdot e_2 = f \text{ for any } f \in G. \tag{3.3}$$

Choosing f equal to e_2 in equation (3.2), we find $e_1 \cdot e_2 = e_2$, but choosing f equal to e_1 in equation (3.3), 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 3.1.8 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 3.1.9 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.

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 used by defining $f^{-m} = (f^{-1})^m$. This notation is practical because the common rule $f^n f^m = f^{n+m}$ then holds for any $f \in G$ and any $n, m \in \mathbb{Z}$. This exponential notation is commonly used in proofs and definitions involving abstract groups, but also in concrete groups such as the permutation group. However, if the group operation is something like an addition, the exponential notation is confusing. For example if $(G, \cdot) = (\mathbb{Z}, +)$, it would be very confusing to write 7^3 if what is meant is $7 + 7 + 7$. Therefore,

the exponential notation is not used if the group operation resembles addition, simply to avoid confusion. In such cases the notation nf (or $n \cdot f$) is often used instead of f^n, since one can think of "f added n times to itself" as "n times f". Just as for permutations, one can define the order of an element.

Definition 3.1.10 *Let $(G, \cdot)$ be a group and $g \in G$. The smallest positive natural number i (if it exists) such that $g^i = e$ is called the order of g . If for all positive natural numbers i it holds that $g^i \neq e$, the order of g is said to be infinite. We will use the notation* $\mathrm{ord}(g)$ *for the order of g.*

Example 3.1.11 In this example, we consider the group $(\mathbb{Z}_{15}, +_{15})$. Let us study the order of the element 2 in this group. Since the group operation is $+_{15}$, we have $2 \cdot 2 = 2 +_{15} 2 = 4$, $3 \cdot 2 = 2 +_{15} 2 +_{15} 2 = 6$, $4 \cdot 2 = 2 +_{15} 2 +_{15} 2 + 152 = 8$, $5 \cdot 2 = 10$, etcetera. Here we have used the convention we mentioned before, that 2 added n times to itself modulo 15, can compactly be written as $n \cdot 2$. Continuing like this, one finds $15 \cdot 2 = 0$ and that 15 is the order of 2 in the group $(\mathbb{Z}_{15}, +_{15})$. Similarly, one can compute the order of all elements given in the following table

$a \in \mathbb{Z}_{15}$	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$\mathrm{ord}(a)$	1	15	15	5	15	3	5	15	15	5	3	15	5	15	15

Summing up we apparently have 1 element of order 1, 2 elements of order 3, 4 elements of order 5, and 8 elements of order 15.

Example 3.1.12 Consider the group $((\mathbb{Z}_{15})^*, \cdot_{15})$ and let us determine the order of the element 2. We have $2 \cdot_{15} 2 = 4$, $2^3 = 2 \cdot_{15} 2 \cdot_{15} 2 = 8$, and $2^4 = 2 \cdot_{15} 2 \cdot_{15} 2 \cdot_{15} 2 = 8 \cdot_{15} 2 = 1$. Therefore 2 has order 4 when viewed as element of the group $((\mathbb{Z}_{15})^*, \cdot_{15})$.

We now give a lemma that can be very useful while determining the order of an element.

Lemma 3.1.13 *Let $(G, \cdot)$ be a group and $g \in G$ an element. Suppose that for some positive integer i on has $g^i = e$. Then the order of g divides i. Conversely, if i is a multiple of* $\mathrm{ord}(g)$*, then $g^i = e$.*

Proof. By definition of the order of an element, it is clear that $\mathrm{ord}(g) \leq i$. Now use division with remainder to write $i = q \cdot \mathrm{ord}(g) + r$ for certain integers q and r, such that $0 \leq r \leq \mathrm{ord}(g) - 1$. Then

$$g^r = g^{i - q \cdot \mathrm{ord}(g)} = g^i (g^{\mathrm{ord}}(g))^{-q} = e,$$

where the last equality follows since $g^i = e$ and $g^{\mathrm{ord}(g)} = e$. However, since $r < \mathrm{ord}(g)$ and g^r, the definition of the order of an element implies that $r = 0$. Hence $\mathrm{ord}(g)$ divides i.

The converse is easier: if i is a multiple of $\mathrm{ord}(g)$, say $i = k \cdot \mathrm{ord}(g)$, then $g^i = (g^{\mathrm{ord}(g)})^k = e^k = e$. ■

Example 3.1.14 In Example 1.3.9, we showed that $37^{60} \bmod 1001 = 1$. Then Lemma 3.1.13 implies that the order of the element 37, when viewed as an element of the group $((\mathbb{Z}_{1001})^*, \cdot_{1001})$, divides 60. We claim that actually the order is equal to 60. One can show using similar techniques as in Example 1.3.9, that $37^{30} \bmod 1001 \neq 1$, $37^{20} \bmod 1001 \neq 1$, and $37^{12} \bmod 1001 \neq 1$. This is enough to conclude that $\text{ord}(g) = 60$. Indeed, if $\text{ord}(37) < 60$, Lemma 3.1.13 would imply that $\text{ord}(g)$ would have to divide either $60/2 = 30, 60/3 = 20$, or $60/5 = 12$, since 2, 3, and 5 are the only prime numbers dividing 60. But then we arrive at a contradiction, since then at least one of $37^{30} \bmod 1001$, $37^{20} \bmod 1001$, or $37^{12} \bmod 1001$ would have to be equal to 1.

3.2 Cyclic groups

Keywords: cyclic group, generator, Euler's totient function

Let us consider the rotational symmetries of a regular n-gon, where n is a positive integer. For $n = 3$ one considers a regular triangle, for $n = 4$ a square, etc. A rotational symmetry of a regular n-gon then is a rotation over a suitable angle and having as center the midpoint of the n-gon. The word "suitable" then means that the vertices of the regular n-gon should be mapped to other vertices of this n-gon by the rotation.

Now let us denote by r, the counterclockwise rotation over the angle $2\pi/n$ radians with center the midpoint of the regular n-gon. The rotation r is a rotational symmetry of the regular n-gon. If we enumerate the vertices of the n-gon counterclockwise as $v_0, v_1, \ldots, v_{n-1}$, then the effect of r on the vertices is that v_k is mapped to v_{k+1} for $k = 0, \ldots, n-2$, while v_{n-1} is mapped to v_0. Hence we can compactly write: $r[v_k] = v_{(k+1) \bmod n}$ for all k. Using the composition operator $\circ$, we can define $r^0 = e$, the identity, $r^1 = r$, $r^2 = r \circ r$, $r^3 = r \circ r \circ r$, and more generally r^i as the i-fold composition of r with itself. The rotation r^i corresponds to another rotational symmetry of the n-gon, namely rotation over the angle $2i\pi/n$. We can make a group out of the rotations using $\circ$, composition, as group operation. More precisely, let us define $C_n := \{e, r, \ldots, r^{n-1}\}$, then we have the following:

Lemma 3.2.1 *Let n be a positive integer, then $(C_n, \circ)$ is a group. The following identities hold:*

1. $r^n = e$,
2. *for any i between 0 and $n-1$,* $(r^i)^{-1} = r^{(-i) \bmod n}$,
3. *for any i and j between 0 and $n-1$,* $r^i \circ r^j = r^{(i+j) \bmod n} = r^{i +_n j}$.

Proof. First of all, since e is the identity function fixing all vertices of the n-gon, we have $r^i \circ e = r^i = e \circ r^i$ for all i. Hence C_n contains an identity element.

We have already seen that $r[v_k] = v_{(k+1) \bmod n}$ for all i between 0 and $n-1$. This implies that $r^i[v_k] = v_{(i+k) \bmod n}$. Hence $r^n[v_k] = v_{(k+n) \bmod n} = v_k$. This implies that r^n is the identity element e. Consequently, the element r^{n-i} is the inverse of r^i, since $r^{n-i} \circ r^i = r^n = e$ and $r^i \circ r^{n-i} = r^n = e$. Hence for $1 \leq i \leq n-1$, we have $(r^i)^{-1} = r^{n-i}$, while $(r^0)^{-1} = e^{-1} =) e = r^0$.

Combining this, we see that for any i between 0 and $n-1$, we have $(r^i) = r^{(-i) \bmod n}$. Hence any element in C_n has an inverse in C_n.

For an arbitrary integer i, we have $r^i = r^{i \bmod n}$, since $r^n = e$. Indeed, if $i = qn + (i \bmod n)$, then $r^i = r^{qn+(i \bmod n)} = (r^n)^q \circ r^{i \bmod n} = e^q \circ r^{i \bmod n} = r^{i \bmod n}$. This implies that for any i and j between 0 and $n-1$, we have $r^i \circ r^j = r^{i+j} = r^{(i+j) \bmod n} = r^{i+_n j}$. In the last equality, we used the definition of $+_n$. This makes $\circ$ into an operation $\circ : C_n \times C_n \to C_n$. That $\circ$ is an associative operation follows from the fact that more generally, composition of functions is an associative operation, see Lemma 2.1.2. ■

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

Definition 3.2.2 *A group $(G, \cdot)$ is called cyclic if there exists a $g \in G$ such that $G = \{g^i \mid i \in \mathbb{Z}\}$. In other words, if any element in G can be written as a power of g. The element g is called a* generator *of the cyclic group.*

Examples of cyclic groups are $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +_n)$, n a positive integer. In both examples, the group operation is addition, which means that one should interpret g^i as g added i times to itself, or in other words as $i \cdot g$. This means that the integer 1 is a generator of both $(\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +_n)$. In the case of $\mathbb{Z}_n$, there is exactly no need for negative values of i. Indeed, if $g = 1$, then $\mathbb{Z}_n = \{0, g, 2 \cdot g, \ldots, (n-1) \cdot g\}$. Another example of a finite cyclic group is $(C_n, \circ)$. In this case r is a generator and we have $G = \{e, r, \ldots, r^{n-1}\}$. In both cases, there exist elements of order n. These examples generalize to the following result that holds for any finite cyclic group.

Lemma 3.2.3 *Let $(G, \cdot)$ be a finite group of order n. Then $(G, \cdot)$ is cyclic if and only if there exists an element $g \in G$ such that* $\mathrm{ord}(g) = n$. *For a finite cyclic group $(G, \cdot)$ of order n with generator g, one has $G = \{e, g, \ldots, g^{n-1}\}$.*

Proof. Suppose that $(G, \cdot)$ is any group and $g \in G$ some element of finite order. Further, let i, j be two integers such that $g^i = g^j$ and suppose that $j \leq i$. Then $g^{i-j} = g^i \cdot g^{-j} = e$, implying that $\mathrm{ord}(g)$ divides $i - j$ by Lemma 3.1.13. This implies that $\{\ldots, g^{-2}, g^{-1}, g^0, g, g^2, \ldots\} = \{e, g, \ldots, g^{\mathrm{ord}(g)-1}\}$ and that all elements in the set $\{e, g, \ldots, g^{\mathrm{ord}(g)-1}\}$ are distinct.

Now let $(G, \cdot)$ be a finite group of order n. If $(G, \cdot)$ is in fact a finite cyclic group of order n and g a generator, the above implies that $G = \{e, g, \ldots, g^{\mathrm{ord}(g)-1}\}$ and hence that $\mathrm{ord}(g) = n$. Conversely, if G contains an element of order n, we obtain that $\{e, g, \ldots, g^{n-1}\}$ is a subset of G containing n elements. But since $|G| = n$, this implies $G = \{e, g, \ldots, g^{n-1}\}$ and hence that $(G, \cdot)$ is a cyclic group. ■

Example 3.2.4 Continuing with Example 3.1.11, we see that $(\mathbb{Z}_{15}, +_{15})$ is a cyclic group, since it contains an element of order 15. In fact any of the eight elements $1, 2, 4, 7, 8, 11, 13, 14$ could be used as generator.

In Example 3.1.11, we determined the order of all elements of the group $(\mathbb{Z}_{15}, +_{15})$. For general finite cyclic groups this is possible as well. The following theorem will allow us to do that and we will investigate its consequences for finite cyclic groups directly afterwards.

Theorem 3.2.5 *Let $(G,\cdot)$ be a group and $g \in G$ an element of finite order n. Further, let $i \in \mathbb{Z}$ be a nonnegative integer. Then* $\operatorname{ord}(g^i) = n/\gcd(i,n)$.

Proof. For convenience, let us write $d = \gcd(i,n)$. Note that by definition of a common divisor, both n/d and i/d are integers. We have $(g^i)^{n/d} = (g^n)^{i/d} = e^{i/\gcd(i,n)} = e$, so the order g^i is at most $n/\gcd(i,n)$. Now suppose that $(g^i)^m = e$ for some positive integer m. Then Lemma 3.1.13 implies that n, being the order of g, divides im. This implies that n/d divides $(i/d)m$. Since $\gcd(i,n) = d$, we have $\gcd(i/d, n/d) = 1$. Hence n/d divides $(i/d)m$ implies that n/d divides m. This shows that $m \geq n/d$ and hence that $\operatorname{ord}(g^i) = n/d$. ■

Corollary 3.2.6 *Let $(G,\cdot)$ be a finite cyclic group of order n. Then for any $g \in G$,* $\operatorname{ord}(g)$ *divides n. Moreover, if d divides n, there exist precisely $\phi(d)$ elements of order d, where ϕ denotes Euler's totient function.*

Proof. Denote by $f \in G$ a generator of G. Then $\operatorname{ord}(f) = |G| = n$. Since $(G,\cdot)$ is cyclic, any element in G can be expressed as f^i for some nonnegative integer. If we assume that $0 \leq i < n$, each element of G can uniquely be written as f^i. Now the first part of the theorem follows, since by Theorem 3.2.5, we have $\operatorname{ord}(f^i) = n/\gcd(n,i)$, which is a divisor of n.

Now let d be a divisor of n. To make sure that f^i has order d, all we have to do it to make sure that $\gcd(n,i) = n/d$. For convenience, let us write $D = n/d$ so that $n = dD$. Further denote by M_d the set of all elements of G of order d. Remembering that we also have posed the condition $0 \leq i < n$, we obtain that

$$\{i \in \mathbb{Z} \mid 0 \leq i < n \text{ and } \operatorname{ord}(f^i) = d\} = \{i \in \mathbb{Z} \mid 0 \leq i < n \text{ and } \gcd(i,n) = D\}$$

The condition $\gcd(i,n) = D$ implies that $\gcd(i/D, n/D) = 1$, so introducing $j := i/D$, we can find a bijection between the set M_d and the set $\{j \in \mathbb{Z} \mid 0 \leq j < n/D \text{ and } \gcd(j, n/D) = 1\}$. Since $n/D = d$, we then obtain:

$$|M_d| = |\{i \in \mathbb{Z} \mid 0 \leq i < n \text{ and } \operatorname{ord}(f^i) = d\}| = |\{j \in \mathbb{Z} \mid 0 \leq j < d \text{ and } \gcd(j,d) = 1\}| = \phi(d).$$

The last equality follows directly from the definition of Euler's totient function, see equation (3.1). ■

In particular, this corollary implies that a finite cyclic group of order n has precisely $\phi(n)$ elements of order n. In other words: a finite cyclic group of order n has $\phi(n)$ generators. Finally, we state one other consequence for Euler's totient function that will come in handy later on.

Corollary 3.2.7 *Let n be a positive integer. Then*

$$n = \sum_{d \text{ divides } n} \phi(d).$$

Proof. Consider a cyclic group $(G,\cdot)$ of order n. We know from Corollary 3.2.6 that the order of any element of G is some divisor of n and that for a given divisor d of n, there exist exactly $\phi(d)$ elements of order d. Hence

$$n = |G| = \sum_{d \text{ divides } n} |\{g \in G \mid \operatorname{ord}(g) = d\}| = \sum_{d \text{ divides } n} \phi(d).$$

■

Example 3.2.8 The last corollary gives a recursive, though somewhat slow, way to compute Euler's totient function. The recursion starts with $\phi(1) = |\{i \in \mathbb{Z}_1 | \gcd(i,1) = 1\}| = 1$. Let us calculate $\phi(15)$. First of all by Corollary 3.2.7, we have

$$15 = \phi(15) + \phi(5) + \phi(3) + \phi(1) = \phi(15) + \phi(5) + \phi(3) + 1,$$

whence

$$\phi(15) = 14 - \phi(5) - \phi(3).$$

To calculate $\phi(5)$ and $\phi(3)$, we use Corollary 3.2.7 again and obtain $\phi(5) = 5 - 1 = 4$ and $\phi(3) = 3 - 1 = 2$. Hence we see that $\phi(15) = 14 - 4 - 2 = 8$. This fully confirms the results in Example 3.1.11, where we showed by direct computation that the cyclic group $(\mathbb{Z}_{15}, +_{15})$ contains $\phi(15) = 8$ elements of order 15, $\phi(5) = 4$ elements of order 5, $\phi(3) = 2$ elements of order 3, and $\phi(1) = 1$ element of order 1.

Aside 3.2.9 *Using Corollary 3.2.7, one can show the following: suppose that the factorization of a number $n \geq 1$ into prime powers is given by $n = \prod_{i=1}^{r} p_i^{e_i}$. Here $p_1, \ldots, p_r$ denote mutually distinct prime numbers and the exponents e_i are assumed to be positive. Then $\phi(n) = \prod_{i=1}^{r}(p_i^{e_i} - p_i^{e_i-1})$. In particular, $\phi(p) = p-1$ if p is a prime. This formula gives a good way to compute $\phi(n)$, but requires the factorization of n into prime numbers. It is believed that this is a computationally hard problem in general. In fact, the safety of the RSA cryptosystem is based on this. In the RSA cryptosystem, the case $n = p_1 \cdot p_2$ is used, where p_1 and p_2 are two distinct prime numbers of size around 2^{1024}. In this setting, n is publicly known, but p_1 and p_2 are not. For such numbers, one has $\phi(n) = \phi(p_1 \cdot p_2) = (p_1 - 1) \cdot (p_2 - 1) = n - (p_1 + p_2) + 1$. If one could compute $\phi(n)$, one could break the cryptosystem, but knowing $\phi(n)$ is equivalent to knowing $p_1 + p_2$ (since n is publicly known). However, knowing $p_1 + p_2$ makes it easy to factor n into $p_1 \cdot p_2$, for example using that p_1 and p_2 are roots of the polynomial $X^2 - (p_1 + p_2)X + n$. Hence in the RSA setting, computing $\phi(n)$ is as hard as factoring n into a product of prime numbers. More generally, it is known that factoring n into a product of prime powers, is as hard as computing $\phi(n)$.*

3.3 The dihedral groups

Keywords: symmetries of a regular n-gon, dihedral groups.

We started in the previous section to study rotational symmetries of the regular n-gon. Now we continue that study, but also include reflection symmetries. Again we enumerate the vertices of the n-gon by $v_0, \ldots, v_{n-1}$. We have already seen that $r[v_k] = v_{(k+1) \bmod n}$.

Now if we also allow reflection symmetries, we can observe that the reflection with reflection axis through v_0, is such a reflection symmetry. We will denote it by s. Also s has a rather simple effect on the vertices: it fixes v_0, and for k between 1 and $n-1$ interchanges v_k and v_{n-k}. Note that if n is even, this actually means that s apart from v_0 also fixes the vertex $v_{n/2}$. All in all, we can conveniently describe this by the equation $s[v_k] = v_{(-k) \bmod n}$.

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 $\circ$ in the notation as usual for convenience), but we can use the following lemma to simplify such expressions:

Lemma 3.3.1 *Let r and s be the symmetries of the regular n-gon as described above. Then we have*

1. *$s^{-1} = s$ or equivalently $s^2 = e$.*
2. *$sr = r^{-1}s$ and more generally $sr^i = r^{-i}s$ for all i between 0 and $n-1$.*

Proof. Since s is a reflection, we have $s^2 = e$. This is the same as saying that s is its own inverse. This shows the first item.

To prove the second item, recall that $r^i[v_k] = v_{(i+k) \bmod n}$ for all integers i and k between 0 and $n-1$. Since $s[v_k] = v_{(-k) \bmod n}$, we see that

$$sr^i[v_k] = s[r^i[v_k]] = s[v_{(i+k) \bmod n}] = v_{(-(i+k)) \bmod n}.$$

Here we used that in the last equality that $-((i+k) \bmod n) \bmod n = (-(i+k)) \bmod n$, since both elements of $\mathbb{Z}_n$ and equivalent to $-(i+k)$ modulo n. Similarly, we obtain that

$$r^{-i}s[v_k] = r^{-i}[s[v_k]] = r^{-i}[v_{(-k) \bmod n}] = v_{((-i)+(-k)) \bmod n},$$

using that $((-i) + ((-k) \bmod n)) \bmod n = ((-i) + (-k)) \bmod n$. Since $(-i) + (-k) = -(i+k)$ for all k, the second item follows. ■

The above lemma can be used to simplify the expression $srsrr$. Strictly speaking we should have put parentheses, for example $(((sr)s)r)r$, but we know from the associative law that the choice of parentheses does not change the final outcome. With this choice of parentheses and the above lemma we obtain using Lemma 3.3.1:

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

With this lemma we can also show the following:

Theorem 3.3.2 *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.*

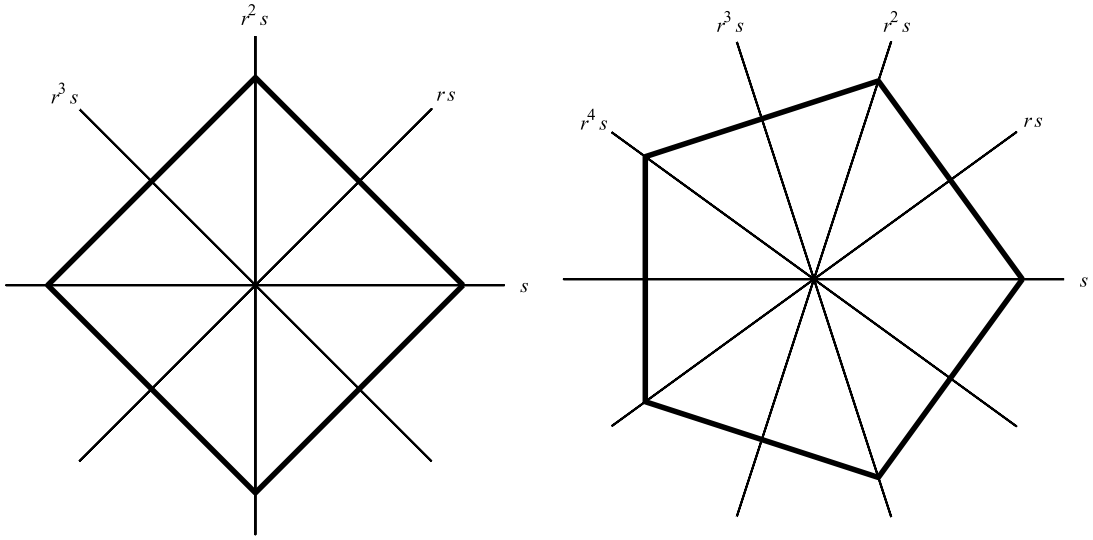
Proof. We first need to show that any composition of two elements in D_n is again in D_n. This means that we need to show that for any pair of natural numbers i and j (both between 0 and $n-1$) that $r^i \circ r^j \in S$, $r^i s \circ r^j \in S$, $r^i \circ r^j s \in S$, and $r^i \circ r^j s \in S$. Using that $r^n = e$, we have already seen in Lemma 3.2.1 that $r^i r^j = r^{i +_n j} \in D_n$. This implies that $r^i r^j s = r^{i +_n j} s \in D_n$.

Using Lemma 3.3.1, we derive that $r^i s r^j = r^i r^{-j} s = r^{i-j} s = r^{(i-j) \bmod n} s \in D_n$. Similarly one shows that $r^i s r^j s = r^{i-j} s^2 = r^{(i-j) \bmod n} \in D_n$. This makes $\circ : D_n \times D_n \to D_n$ an operation on D_n. Associativity follows, like in the case of the group $(C_n, \circ)$, from Lemma 2.1.2. The element e is the identity, which leaves the existence of inverses to be checked. We already

know that $(r^i)^{-1} = r^{(-i) \bmod n}$. On the other hand $(r^i s) \circ (r^i s) = r^i r^{-i} s^2 = e$ so that for all i, the element $r^i s$ is its own inverse. ■

The elements $\{e, r, \ldots, r^{n-1}\}$ correspond to the rotational symmetries of the regular n-gon as we have seen. The elements $\{s, rs, \ldots, r^{n-1}s\}$ correspond to its reflection symmetries. More precisely, the line of reflection of $r^k s$ is the line passing through the midpoint of the n-gon having an angle of $k \cdot \pi/n$ radians with the right-hand part of line of reflection of s. In Figure 3.1 the reflection symmetries of the square and the regular pentagon are indicated by drawing the corresponding lines of reflection.

Figure 3.1: The reflection symmetries of a square and regular 5-gon.



The group in the above theorem is called the *dihedral* group and is denoted by $(D_n, \circ)$. It is a group of order $2n$, since it includes n rotations (including e) and n reflections. Because of this, 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.

3.4 Products of groups and examples of groups of small order

One way to construct groups from previously known ones is the following:

Theorem 3.4.1 *Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be groups. Define $G := G_1 \times G_2 = \{(g_1, g_2) \mid g_1 \in G_1, g_2 \in G_2\}$ and $\cdot : G \times G \to G$ as $(f_1, f_2) \cdot (g_1, g_2) := (f_1 \cdot_1 g_1, f_2 \cdot_2 g_2)$. Then $(G, \cdot)$ is a group.*

Proof. We start by showing that $\cdot$ is an associative operator. If f_1, g_1, h_1 are elements of G_1 and f_2, g_2, h_2 are elements of G_2, then

$$(f_1, f_2) \cdot ((g_1, g_2) \cdot (h_1, h_2)) = (f_1, f_2) \cdot (g_1 \cdot_1 h_1, g_2 \cdot_2 h_2) = (f_1 \cdot_1 (g_1 \cdot_1 h_1), f_2 \cdot_2 (g_2 \cdot_2 h_2))$$

and

$$((f_1, f_2) \cdot (g_1, g_2)) \cdot (h_1, h_2) = (f_1 \cdot_1 g_1, f_2 \cdot_2 g_2) \cdot (h_1, h_2) = ((f_1 \cdot_1 g_1) \cdot_1 h_1, (f_2 \cdot_2 g_2) \cdot_2 h_2)$$

Since $\cdot_1$ and $\cdot_2$ are associative operators, we can conclude that $(f_1, f_2) \cdot ((g_1, g_2) \cdot (h_1, h_2)) = ((f_1, f_2) \cdot (g_1, g_2)) \cdot (h_1, h_2)$.

Next, observe that if e_1 is the identity element of $(G_1, \cdot_1)$ and e_2 that of $(G_2, \cdot_2)$, then (e_1, e_2) is an identity element for $(G, \cdot)$, since $(e_1, e_2) \cdot (g_1, g_2) = (e_1 \cdot g_1, e_2 \cdot g_2) = (g_1, g_2)$ and $(g_1, g_2) \cdot (e_1, e_2) = (g_1 \cdot e_1, g_2 \cdot e_2) = (g_1, g_2)$.

Finally, inverses are simply taken coordinate-wise: $(g_1, g_2)^{-1} = (g_1^{-1}, g_2^{-1})$. ■

The construction in this theorem is called the *direct product* construction and the group $(G, \cdot)$ is called the direct product of the groups $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$. Note that as a set G is simply the Cartesian product of G_1 and G_2. In particular, if $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ are finite groups of orders, say, n_1 and n_2 respectively, the order of their direct product is $n_1 n_2$. The group operator $\cdot$ in the direct product construction is sometimes written as $\cdot_1 \times \cdot_2$. Using the direct product construction, we are able to construct many groups from the ones we already know.

Example 3.4.2 Let us construct various groups of order 8. In the first place, we can construct $(\mathbb{Z}_8, +_8)$, but also $(\mathbb{Z}_4 \times \mathbb{Z}_2, +_4 \times +_2)$ and $(\mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2, +_2 \times +_2 \times +_2)$. The third group has elements of the form $(a, b, c) \in \mathbb{Z}_2^3$, while the group operation $+_2 \times +_2 \times +_2$ simply adds such three-tuples coordinatewise: $(a_1, b_1, c_1) +_2 \times +_2 \times +_2 (a_2, b_2, c_2) = (a_1 +_2 a_2, b_1 +_2 b_2, c_1 +_2 c_2)$.

Other groups of order 8 are $(D_4, \circ)$ and the following group:

Definition 3.4.3 *Let $Q_8 := \{1, -1, i, -i, j, -j, k, -k\}$ and let $\cdot$ be defined using the following multiplication table:*

$\cdot$	1	-1	i	$-i$	j	$-j$	k	$-k$
1	1	-1	i	$-i$	j	$-j$	k	$-k$
-1	-1	1	$-i$	i	$-j$	j	$-k$	k
i	i	$-i$	-1	1	k	$-k$	$-j$	j
$-i$	$-i$	i	1	-1	$-k$	k	j	$-j$
j	j	$-j$	$-k$	k	-1	1	i	$-i$
$-j$	$-j$	j	k	$-k$	1	-1	$-i$	i
k	k	$-k$	j	$-j$	$-i$	i	-1	1
$-k$	$-k$	k	$-j$	j	i	$-i$	1	-1

This group is called the quaternion group. *Note that -1 commutes with any other element: for all $a \in Q_8$, $(-1) \cdot a = -a = a \cdot (-1)$ and that $(-1)^2 = 1$. Using this, the multiplication table can*

then in principle be derived from the relations $i^2 = j^2 = k^2 = ijk = -1$. For example, $ij = k$, since $ij = ij(-1)(-1) = ijkk(-1) = (-1)k(-1) = (-1)^2k = k$.

Aside 3.4.4 *The elements of the quaternion group are actually special cases of quaternions: expressions of the form $a + bi + cj + dk$ with $a, b, c, d \in \mathbb{R}$. They can be multiplied using the same relations $i^2 = j^2 = k^2 = ijk = -1$ if we assume that the distributive law holds, so that for example $(1 + 2i + 3j + 4k) \cdot k = k + 2ik + 3jk + 4k^2 = -4 + 3i - 2j + k$. Quaternions were discovered by the Irish mathematician William Rowan Hamilton in 1843, while he was taking a walk. When resting on Broughham Bridge, he carved the defining relations into a stone of this bridge. It is common to write $\mathbb{H} = \{a + bi + cj + dk \mid a, b, c, d \in \mathbb{R}\}$ for the set of all quaternions in honour of Hamilton. The complex numbers $\mathbb{C}$ can then be viewed as a subset of $\mathbb{H}$, namely as quaternions of the form $a + bi + 0j + 0k$. In analogy with complex numbers, one defines the real part of $a + bi + cj + dk \in \mathbb{H}$ to be $a \in \mathbb{R}$. A quaternion with real part zero, that is to say, of the form $bi + cj + dk$, is called a pure quaternion. It is quite standard in physics to denote by $\mathbf{i}, \mathbf{j}, \mathbf{k}$ a positively oriented orthonormal basis for $\mathbb{R}^3$, so that vectors in $\mathbb{R}^3$ can be written in the form $a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ for $a_1, a_2, a_3 \in \mathbb{R}$. This looks remarkably much like a pure quaternion! There is a good reason for this: if $\mathbf{v} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ and $\mathbf{w} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$ are two vectors in $\mathbb{R}^3$, their inner product $\mathbf{v} \cdot \mathbf{w}$ and cross product $\mathbf{v} \times \mathbf{w}$ equal*

$$\mathbf{v} \cdot \mathbf{w} = a_1b_1 + a_2b_2 + a_3b_3 \quad \text{and} \quad \mathbf{v} \times \mathbf{w} = (a_2b_3 - a_3b_2)\mathbf{i} + (a_3b_1 - a_1b_3)\mathbf{j} + (a_1b_2 - a_2b_1)\mathbf{k},$$

while at the level of pure quaternions

$$(a_1i+a_2j+a_3k)(b_1i+b_2j+b_3k) = -(a_1b_1+a_2b_2+a_3b_3)+(a_2b_3-a_3b_2)i+(a_3b_1-a_1b_3)j+(a_1b_2-a_2b_1)k.$$

Hence in this sense, inner and cross products of vectors in $\mathbb{R}^3$ are involved when multiplying (pure) quaternions.

3.5 Extra: applications of group theory

In this section, we give some applications of group theory. First we give a way to describe the group of rotational symmetries of the dodecahedron. After this a brief description is given of Cayley graphs and a connection is given with the famous Rubik's cube.

3.5.1 Rotational symmetries of a dodecahedron

Using the sign function from the previous section, one can define the group whose elements are the even permutations of S_n:

Definition 3.5.1 *Let $A_n := \{f \in S_n \mid \text{sign}(f) = 1\}$, the set of all even permutations. Then $(A_n, \circ)$ is called the alternating group on n letters.*

Note that $(A_n, \circ)$ is indeed a group: $\circ$ is an associative operation, id is an even permutation, if f is an even permutation, then so is f^{-1}, and if f and g are two even permutations, then so

is $f \circ g$ using equations (2.3) and (2.4). This means that the composition operator restricted to the set of even permutations gives A_n a group structure. Later we will introduce the notion of a subgroup of a group. Using that terminology, we will say that A_n is a subgroup of $(S_n, \circ)$.

Example 3.5.2 Let $n = 4$, then

$$A_4 = \{\mathrm{id}, (1\,2\,3), (1\,3\,2), (1\,2\,4), (1\,4\,2), (1\,3\,4), (1\,4\,3), (2\,3\,4), (2\,4\,3), (1\,3)(2\,4), (1\,4)(2\,3), (1\,2)(3\,4)\}. \tag{3.4}$$

The composition is the usual one for permutations, and we already proved that the composition of two even permutations is even. For example, we have: $(1\,2\,3)(1\,2)(3\,4) = (1\,3\,4)$.

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

Theorem 3.5.3 *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 on 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. ∎

Other proofs of this theorem are possible: we know that any permutation can be written as the composite of 2-cycles. For an even permutation, the number of 2-cycles in such a composition is necessarily even. Therefore, an even permutation f can be written in the form:

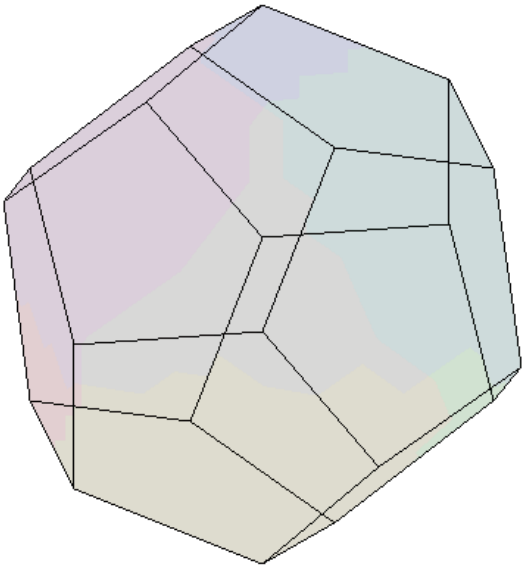
$$f = (a_1\,b_1)(c_1\,d_1)\cdots(a_\ell\,b_\ell)(c_\ell\,d_\ell).$$

This means that the theorem follows if we can show that a permutation of the form $(a\,b)(c\,d)$ can be written as a composite of 3-cycles. There are three cases: Case 1. $\{a, b\} = \{c, d\}$. In this case $(a\,b)(c\,d) = \mathrm{id}$, which we can see as the composite of zero many 3-cycles. Case 2. $\{a, b\}$ and $\{c, d\}$ have precisely one element in common. Since $(a\,b) = (b\,a)$ and $(c\,d) = (d\,c)$, it is no restriction to assume that a is the common element and that $a = c$. Then $(a\,b)(c\,d) = (a\,b)(a\,d) = (a\,d\,b)$ and we are done. Case 3.$\{a, b\}$ and $\{c, d\}$ have no elements in common. In this case $(a\,b)(c\,d) = (a\,b\,c)(b\,c\,d)$ and we are done as well.

Example 3.5.4 Most of the elements in A_4 are 3-cycles, as can be seen from equation (3.4), with four exceptions. Let us write the permutation $(1\,2)(3\,4)$ as composition of 3-cycles, following the recipe given in the proof of Theorem 3.5.3. Since $f = (1\,2)(3\,4)$ 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 $g = (2\,1\,3)(1\,2)(3\,4) = (2\,3\,4)$. Therefore $(1\,2)(3\,4) = (2\,1\,3)^{-1}(2\,3\,4) = (1\,2\,3)(2\,3\,4)$.

We now turn our attention to a description of the group of rotation symmetries of a regular dodecahedron, see Figure 3.2. It turns out that the alternating group on five letters, $(A_5, \circ)$ comes in very handy in this context.

Figure 3.2: The regular dodecahedron.



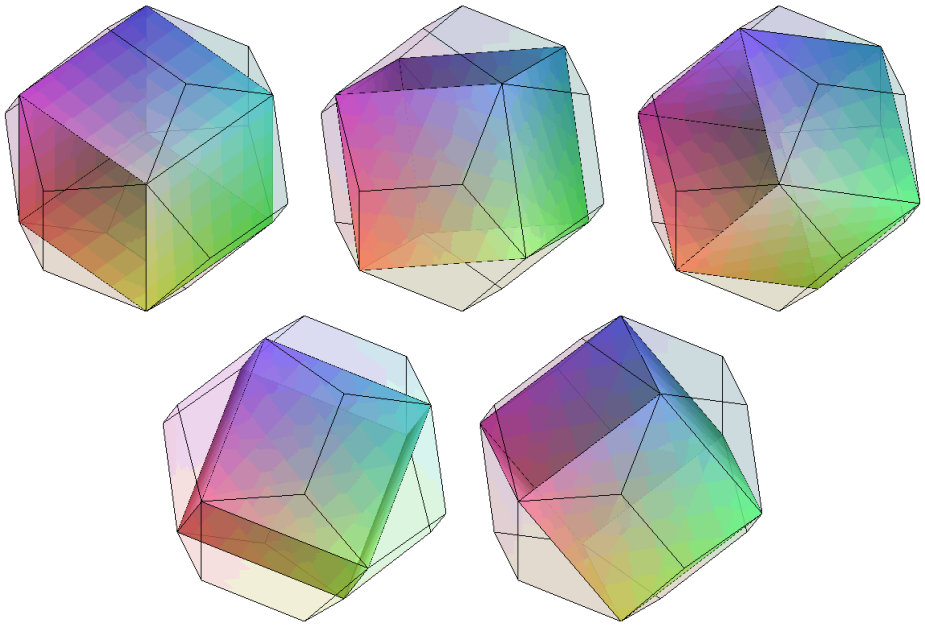
One type of rotation symmetry consists of a rotation over $2\pi/3$ or $4\pi/3$ radians with a rotation axis connecting two opposite vertices of the dodecahedron. Since a regular dodecahedron has 20 vertices, this gives rise to 20 rotation symmetries.

Now consider five cubes inside the dodecahedron as in Figure 3.3. 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.

The number of 3-cycles in A_5 can be computed as follows: first consider all strings of the form $a\,b\,c$, with $a, b, c \in \{1, \ldots, 5\}$. There are $5 \cdot 4 \cdot 3 = 60$ of them and each string $a\,b\,c$ gives rise to a 3-cycle $(a\,b\,c)$. However, since $(a\,b\,c) = (c\,a\,b) = (b\,c\,a)$, we do not obtain 60, but $60/3 = 20$ 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 S_5. Since any 3-cycle is an even permutation, all 20 of them lie 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 3.5.3, 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.

Figure 3.3: Five cubes inside the regular dodecahedron.



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

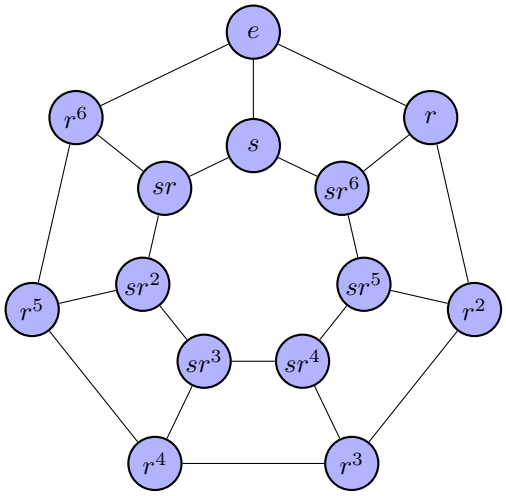


Figure 3.4: Cayley graph of the dihedral group of order 14.

3.5.2 Cayley graphs and Rubik's cube

Let $(G, \cdot)$ be a group and $T \subseteq G$ a finite subset of elements such that

1. T generates the group, i.e., for any element of G, there exist $n \in \mathbb{Z}_{>0}$ and $t_1, \ldots, t_n \in T$ such that $g = t_1 \cdot \cdots \cdot t_n$,
2. the set T does not contain the identity element of G, and
3. if $t \in T$, then also $t^{-1} \in T$.

Given such a set T, one can define a graph (V, E), where $V = G$, so the set of vertices are simply the elements of the group, and E consists of all edges between group elements f and g for which there exists $t \in T$ such that either $f = g \cdot t$ or $f = t \cdot g$. Graphs of this type are called *Cayley graphs*. Considering for example the dihedral group $(D_7, \circ)$ and $T = \{s, r, r^{-1}\}$, one obtains the graph given in Figure 3.4.

Cayley graphs can be used to see how "complicated" it is to express a group element $g \in G$ in the elements of the set T. More precisely, if there exists a path in the Cayley graph from the identity element e to g of length n, then there exist elements $t_1, \ldots, t_n \in T$ such that g can be written as $t_{\sigma[1]} \cdots t_{\sigma[n]}$, where $\sigma \in S_n$ is a permutation. The permutation comes from the fact that if there is an edge between group elements f and g, then either $f = g \cdot t$ or $f = t \cdot g$. Therefore the length of the shortest path from e to g is the smallest number of elements in T needed to express g in them.

This description has some hands-on applications. The famous Rubik's cube is a puzzle where the goal is to restore a "messed up" cube to its initial constellation using only simple moves. Here a simple move is applying quarter turns (both clockwise and counter clockwise) of one face

of the cube. Since a cube has six faces, this gives rise to a set T consisting of 12 simple moves. These 12 simple moves form a natural set T of generators of the group of all possible moves you can achieve in a Rubik's cube, which is called the Rubik's cube group. It turns out that this group has a large order, namely 42252003274489856000. The Cayley graph corresponding to the set T gives another way of thinking about what is going on when you are trying to solve the cube. When you start, the cube is in some constellation corresponding to some element g from the Rubik's cube group. This element is a vertex in the Cayley graph. Performing a quarter turn, then means that you move from that vertex to an adjacent one. The goal is to find a path from g to e, the identity element, which corresponds to the initial configuration. It turns out that no matter what g is, there always exists a path from g to e of length at most 26. Determining that the worst case requires 26 quarter turns, is not an easy task, since the Rubik's cube group is so large. Indeed, this required massive computer calculations and clever algorithms. If T is also allowed to contain all six half turns of faces, then the maximal length of a path from g to e, is a bit smaller, namely 20. Hence no matter how cleverly you mess up the Rubik's cube, it is in principle always possible to restore it in its initial state in at most 20 simple moves (using quarter turns and half turns of the faces only).

3.6 Exercises

Multiple choice exercises

1. Let $(G, \cdot)$ be a group. Then
   A. for every $f, g, h \in G$ it holds $f \cdot (g \cdot h) = (f \cdot g) \cdot h$
   TRUE ☐ FALSE ☐
   B. for every $f, g \in G$ one has $f \cdot g = g \cdot f$
   TRUE ☐ FALSE ☐
   C. the identity element $e \in G$ satisfies $e \cdot f = f \cdot e = e$ for all $f \in G$
   TRUE ☐ FALSE ☐
   D. if $f \in G$ then f admits an inverse $f^{-1} \in G$
   TRUE ☐ FALSE ☐

2. Let $(G, \cdot)$ be a group with identity element e and let $g \in G$. Then the *order* of g is $n \in \mathbb{N}$ if and only if
   A. ☐ $g^n = e$
   B. ☐ $g^n = e$ and $g^i \neq e$ for all positive $i < n$
   C. ☐ n is the smallest positive natural number such that $g^n = e$

3. Let $n \in \mathbb{N}_{>0}$ and let C_n be a cyclic group of order n. Then
   A. there exists $r \in C_n$ such that $C_n = \{e, r, \ldots, r^{n-1}\}$
   TRUE ☐ FALSE ☐
   B. there exists $r \in C_n$ such that r has order exactly n
   TRUE ☐ FALSE ☐

C. $(r^i)^{-1} = r^{i \pmod n}$ for all $i = 0, \dots n-1$
TRUE ☐ FALSE ☐

D. $r^i \cdot r^j = r^{i+j \pmod n}$ for all $i, j = 0, \dots, n-1$
TRUE ☐ FALSE ☐

4. Let $(G, \cdot)$ be a group of order n with identity element e. Then G is a *cyclic group* if and only if

A. ☐ there exists $g \in G$ such that $G = \{g^i \mid i \in \mathbb{N}\}$

B. ☐ there exists $g \in G$ such that $g^n = e$

C. ☐ G admits an element of order exactly n

5. Let $(G, \cdot)$ be a group.

A. If $g \in G$ has order n, then g^i has also order n
TRUE ☐ FALSE ☐

B. If $g \in G$ has order n, then g^i has order $n/\gcd(n, i)$
TRUE ☐ FALSE ☐

C. If G has order n and d divides n, then G does not necessarily contain elements of order d
TRUE ☐ FALSE ☐

D. If G has order n then G contains exactly $\phi(d)$ elements of order d for all d dividing n
TRUE ☐ FALSE ☐

6. Let $(D_n, \cdot)$ be a dihedral group of order $2n$, that is, the group of symmetries of the regular n-gon.

A. There exist $r, s \in D_n$ such that $r^n = e$ and $s^{-1} = s$
TRUE ☐ FALSE ☐

B. r is a reflection symmetry and s is a rotational symmetry
TRUE ☐ FALSE ☐

C. $sr^i = r^{-i}s$ for all $i = 0, \dots, n-1$
TRUE ☐ FALSE ☐

D. $D_n = \{e, r, r^2, \dots, r^{n-1}, s, rs, \dots, r^{n-1}s\}$
TRUE ☐ FALSE ☐

7. Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be two groups. The direct product $(G_1 \times G_2, \cdot)$ is defined as

A. ☐ $G_1 \times G_2 = \{g_1 g_2 : g_i \in G_i,\ i = 1, 2\}$

B. ☐ $G_1 \times G_2 = \{(g_1, g_2) : g_i \in G_i,\ i = 1, 2\}$ and $(f_1, f_2) \cdot (g_1, g_2) = (f_1 \cdot_1 f_2, g_1 \cdot_2 g_2)$

C. ☐ $G_1 \times G_2 = \{(g_1, g_2) : g_i \in G_i,\ i = 1, 2\}$ and $(f_1, f_2) \cdot (g_1, g_2) = (f_1 \cdot_1 g_1, f_2 \cdot_2 g_2)$

Exercises to get to know the material better

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

9. 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.

10. Is $(\{1, 2, 3, 4, 5\}, \cdot_6)$ a group?

11. Let f, g and h be elements of a group $(G, \cdot)$.
    (a) Show that f and f^{-1} have the same order.
    (b) Show that $f \cdot g$ and $g \cdot f$ have the same order. Use this to show that $f \cdot g \cdot h$ and $g \cdot h \cdot f$ have the same order.
    (c) Find three elements f, g, h in the symmetric group S_3, such that $f \cdot g \cdot h$ and $g \cdot f \cdot h$ have different orders.

12. (a) Let r be a generator of the cyclic group $(C_{91}, \circ)$. Is the element r^{98} a generator as well?
    (b) Let $(D_{91}, \circ)$ be the dihedral group of order 182. Write $r^{98}sr^{-17}s^3$ in the form $r^i s^j$ with $0 \leq i \leq 90$ and $0 \leq j \leq 1$.

13. Suppose that $(G, \cdot)$ is a finite abelian group of order n, where $G = \{a_1, a_2, \ldots, a_n\}$. Let $c = a_1 \cdot a_2 \cdot \ldots \cdot a_n$ be the product of all the elements in G. Show that $c^2 = e$ [Hint: first of all, you may use that any $g \in G$ has a unique inverse. Then it may be tempting to say that $c = e$, but what happens if for some $g \in G \setminus \{e\}$ one has $g^{-1} = g$?].

14. Let D_∞ be the set $\mathbb{Z} \times \{-1, 1\}$ (that is, the set of all pairs (x, s) where x is an integer and s is either -1 or 1). Define a multiplication on D_∞ as follows:
$$(x, s) \cdot (y, t) = (x + sy, st)$$
(note that this makes sense since if s and t are either -1 or 1, the same is true for st).
    (a) Show that $(D_\infty, \cdot)$ is a group. What is the identity element of this group? [Note: you can use that multiplication and addition of integers are associative and that the distributive law holds, that is, $a(b + c) = ab + ac$ for all integers a, b, c].
    (b) Show that any element of the form $(x, 1)$ with $x \neq 0$ has infinite order, while any element of the form $(x, -1)$ has order 2.

15. Show that the direct product of two abelian groups is abelian.

16. Recall that for a positive integer n, $(\mathbb{Z}_n, +_n)$ is a group. Hence fixing two positive integers n_1 and n_2 we can construct the direct product $(\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}, +)$ of the groups $(\mathbb{Z}_{n_1}, +_{n_1})$ and $(\mathbb{Z}_{n_2}, +_{n_2})$ where $(a_1, a_2) + (b_1, b_2) = (a_1 +_{n_1} b_1, a_2 +_{n_2} b_2)$ for $a_1, b_1 \in \mathbb{Z}_{n_1}$ and $a_2, b_2 \in \mathbb{Z}_{n_2}$. In this exercise we work with computing the order of elements in this type of direct product for specific values of n_1 and n_2. Find the orders of the given elements:
    (a) $(2, 6) \in \mathbb{Z}_4 \times \mathbb{Z}_{12}$,
    (b) $(8, 10) \in \mathbb{Z}_{12} \times \mathbb{Z}_{18}$.

17. Consider the multiplication table of a group $(G, \cdot)$. Show that each element of G occurs exactly once in each row and in each column.

18. Which of the 4 multiplication tables defined on a set $G = \{a, b, c, d\}$ and listed at the beginning of page 79 form a group? Support your answer in each case.

·	a	b	c	d
a	a	c	d	a
b	b	b	c	d
c	c	d	a	b
d	d	a	b	c

·	a	b	c	d
a	a	b	c	d
b	b	a	d	c
c	c	d	a	b
d	d	c	b	a

·	a	b	c	d
a	a	b	c	d
b	b	c	d	a
c	c	d	a	b
d	d	a	b	c

·	a	b	c	d
a	a	b	c	d
b	b	a	c	d
c	c	b	a	d
d	d	d	b	c

Exercises to get around in the theory

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

20. Let $(G,\cdot)$ be a group. Suppose that $f^2 \cdot g^2 = (f \cdot g)^2$ for all $f, g \in G$. Show that $(G,\cdot)$ is abelian (that is $f \cdot g = g \cdot f$ for all $f, g \in G$).

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

22. * 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 3.3.

    (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.

23. * Let G be an infinite cyclic group with generator g, that is, $G = \{g^i \mid i \in \mathbb{Z}\}$ and G is not of finite order. Show that the only other generator of G is g^{-1}. This implies that infinite cyclic groups have exactly 2 generators.

24. Let $(G_1 \times G_2, \cdot)$ be the direct product of the groups $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$. Let $g := (g_1, g_2) \in G$. Show that the order of g is the least common multiple of the orders of g_1 and g_2.

25. Define an equivalence relation on the elements of a group $(G, \cdot)$ by

$$f \sim g \iff \langle f \rangle = \langle g \rangle.$$

Here $\langle f \rangle$ is the cyclic group generated by f. Show that the equivalence class of an element $f \in G$ is always finite [Hint: Use Exercise 23 if f has infinite order].

26. Let G be the group $\{e, a, b, b^2, ab, ab^2\}$ where $a^2 = e$, $b^3 = e$, and $ba = ab^2$. Show that these three equations determine the multiplication table of G.

27. Use Corollary 3.2.7 to show that $\phi(p) = p - 1$, $\phi(p^2) = p(p-1)$ and $\phi(pq) = (p-1)(q-1)$ for p and q distinct primes.

Chapter 4

Subgroups and cosets

4.1 Subgroups

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

Definition 4.1.1 *Let $H \subseteq G$ 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 $H \subseteq G$ 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. The three conditions can be combined to only one as shown in the following lemma, which sometimes makes it easier to show that a subset is a subgroup.

Lemma 4.1.2 *Let $(G, \cdot)$ be a group and H a nonempty subset of G. Then H is a subgroup of G if and only if for all $f, g \in H$ it holds that $f \cdot g^{-1} \in H$.*

Proof. The proof of the lemma is posed as an exercise at the end of this chapter. ■

Example 4.1.3 Let us consider the set of permutations H given by

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

Example 4.1.4 The set of even integers $2\mathbb{Z} = \{\dots, -4, -2, 0, 2, 4, \dots\}$ is a subgroup of $(\mathbb{Z}, +)$. Indeed, let us choose $k, \ell \in 2\mathbb{Z}$ arbitrarily. Since k and ℓ are even integers, so is $k - \ell$. Hence Lemma 4.1.2 implies that $2\mathbb{Z}$ is a subgroup of $(\mathbb{Z}, +)$.

Example 4.1.5 Let $G = \mathbb{R}^3$ and denote by $+$ the vector addition. Further denote by $V \subseteq \mathbb{R}^3$ the linear subspace defined by

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

Then V is a subgroup of $(\mathbb{R}^3, +)$. Indeed, if $v, w \in V$ are two arbitrarily chosen vectors of V, then $v - w \in V$, since V is a linear subspace. Hence by Lemma 4.1.2, W is a subgroup of $(V, +)$.

Example 4.1.6 We claim that the cyclic group C_n is a subgroup of the dihedral group $(D_n, \circ)$. Since $C_n = \{e, r, \dots, r^{n-1}\}$, it contains the identity. Further if $r^i \in C_n$, then $(r^i)^{-1} = r^{(-i) \bmod n}$. Similarly, if $r^i, r^j \in C_n$, then $r^i \circ r^j = r^{i +_n j}$. Hence by Definition 4.1.1, C_n is a subgroup of $(D_n, \circ)$.

Definition 4.1.7 *Let $(G, \circ)$ be a group and let $g \in G$ be a group element. The set $\langle g \rangle := \{g^i \mid i \in \mathbb{Z}\}$ 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. We use the definition of a subgroup to show this: $e = g^0 \in \langle g \rangle$; 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 $g^i \cdot g^j = g^{i+j} \in \langle g \rangle$.

Lemma 4.1.8 *Let $(G, \cdot)$ be a group and let $g \in G$ be a group element. Then the order of the subgroup $\langle g \rangle$ 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.

What 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. ■

4.2 Cosets of a subgroup

Keywords: multiplication of subsets, cosets of a subgroup.

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

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

Example 4.2.2 If we consider the group $(\mathbb{Z}, +)$ and choose

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

then we obtain

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

The numbers in $M + N$ are in fact exactly the amounts one could pay for using exactly two Danish banknotes. If we want to determine the amounts one could pay for using one or two Danish banknotes, we can simply add 0 to the set N. Then paying with exactly one banknote amounts to choosing 0 from N. In other words, changing N to $L = \{0, 50, 100, 200, 500, 1000\}$, we get

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

Definition 4.2.3 *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

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

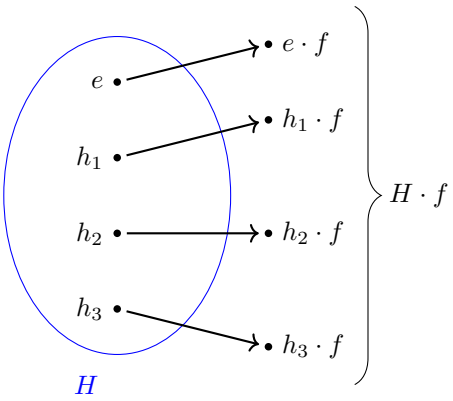


Figure 4.1: The right coset $H \cdot f$

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*. We will encounter these kind of subgroups again in Chapter 6.

Aside 4.2.4 *A commonly used notation for the set of all left cosets of H in G is G/H. We will use this notation in later chapters when discussing quotient groups. The set of all right cosets of H in G is often denoted by $H\backslash G$. However, we will not need this notation later on. For a normal subgroup H of $(G, \cdot)$, the sets G/H and $H\backslash G$ are the same.*

Example 4.2.5 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 4.1.3. 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 4.2.6 Let $G = \mathbb{R}^3$ and $H = V$ be as in Example 4.1.5. The coset of V in $\mathbb{R}^3$ by $(5, 6, -4) \in \mathbb{R}^3$ is given by

$$\begin{aligned} (5,6,-4) + V &= \{(5,6,-4) + (v_1, v_2, v_3) \,|\, (v_1, v_2, v_3) \in V\} \\ &= \{(5,6,-4) + (v_1, v_2, 0) \,|\, v_1, v_2 \in \mathbb{R}\} \\ &= \{(5 + v_1, 6 + v_2, -4) \,|\, v_1, v_2 \in \mathbb{R}\} \\ &= \{(w_1, w_2, -4) \,|\, w_1, w_2 \in \mathbb{R}\} \end{aligned} \tag{4.1}$$

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

4.3 Cosets as equivalence classes

Keywords: cosets and equivalence classes.

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

Definition 4.3.1 *Let $(G, \cdot)$ be a group and $H \subseteq G$ 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 $g \cdot f^{-1} \in H$.*

A beautiful property of these relations is that they in fact are equivalence relations.

Lemma 4.3.2 *Let $(G, \cdot)$ be a group and $H \subseteq G$ a subgroup. Then $\sim_H$ and ${}_H\!\sim$ are equivalence relations on G.*

Proof. We will 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. ■

As we have seen in Chapter 1, equivalence relations on a set A partition that set into equivalence classes. The point of the equivalence relations $\sim_H$ and ${}_H\!\sim$ is that their equivalence classes actually are cosets. Hence the properties of equivalence classes in general as derived in Chapter 1 will imply several nice properties of cosets.

Lemma 4.3.3 *Let $(G, \cdot)$ be a group and $H \subseteq G$ 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 $[f]_{\sim_H} \subseteq f \cdot H$: 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 $[f]_{\sim_H} \subseteq f \cdot H$.

Now we show the converse inclusion, namely that $f \cdot H \subseteq [f]_{\sim_H}$: 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 $f \cdot H \subseteq [f]_{\sim_H}$.

Combining that we now know that $[f]_{\sim_H} \subseteq f \cdot H$ and $f \cdot H \subseteq [f]_{\sim_H}$, 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 1.3.3 to these equivalence relations. The result is the following:

Theorem 4.3.4 *Let $(G, \cdot)$ be a group and $H \subseteq G$ 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 that either $f \cdot H \cap g \cdot H = \emptyset$ or $f \cdot H = g \cdot H$. Similarly we have that either $H \cdot f \cap H \cdot g = \emptyset$ 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 4.3.3 and Theorem 1.3.3. ■

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 $f \in H$. Following the terminology in Chapter 1, we call an element from a coset a *representative* of that coset.

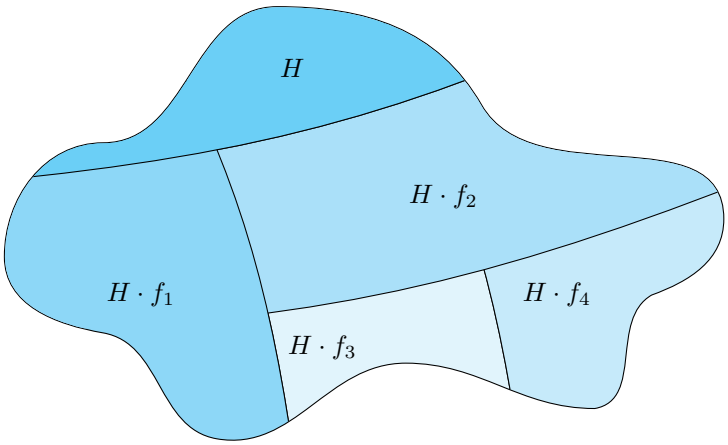


Figure 4.2: Cosets $H \cdot f$ form a partition of the group G.

Example 4.3.5 One can get a good intuitive understanding of Theorem 4.3.4 by considering Examples 4.1.5 and 4.2.6 again. First of all, part one of Theorem 4.3.4 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 exactly part four of Theorem 4.3.4). 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 4.3.4 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 4.3.4 boils down to in this example.

Example 4.3.6 Let us consider the group $(D_4, \circ)$. We have $D_4 = \{e, r, r^2, r^3, s, rs, r^2s, r^3s\}$. Let us consider the subgroup $H = \{e, rs\}$. That H indeed is a subgroup is not so hard to check: the key is that rs is one of the reflection symmetries of the square so that $rsrs = e$). The left and right cosets of H are given in the table given below. This table also illustrates Theorem 4.3.4.

left cosets of H	right cosets of H
$H = rsH = \{e, rs\}$	$H = Hrs = \{e, rs\}$
$rH = r^2sH = \{r, r^2s\}$	$Hr = Hs = \{r, s\}$
$r^2H = r^3sH = \{r^2, r^3s\}$	$Hr^2 = Hr^3s = \{r^2, r^3s\}$
$r^3H = sH = \{r^3, s\}$	$Hr^3 = Hr^2s = \{r^3, r^2s\}$

We see for example that D_4 is partitioned into the four left cosets $\{e, rs\}$, $\{r, r^2s\}$, $\{r^2, r^3s\}$, and $\{r^3, s\}$. This example also illustrates that left and right cosets in general are distinct from each other (for example the left coset rH is not equal to Hr or to any of the other right cosets of H). For special cases it might happen that they are equal: for example $r^2H = Hr^2$.

Aside 4.3.7 *In some books on group theory, the notion of a* double coset *is introduced: given a group $(G, \cdot)$, two of its subgroups K and H, and an element $f \in G$, one defines the double coset*

$$KfH := \{k \cdot f \cdot h \mid k \in K, h \in H\}.$$

One can show that like cosets, double cosets are equivalence classes for the equivalence relation $_K \sim_H$ defined by $f_K \sim_H g$ if there exist $k \in K$ and $h \in H$ such that $f = kgh$. In particular, double cosets partition the set of group elements G. The set of double cosets is sometimes denoted by $K\backslash G/H$.

4.4 The order of a subgroup and of an element

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

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

Theorem 4.4.1 *Let $(G, \cdot)$ be a finite group and let $H \subseteq G$ be a subgroup. Then the order of the subgroup H divides the order of the group G, that is to say: $|H|$ divides $|G|$.*

Proof. By Theorem 4.3.4, one can write G as the disjoint union of left cosets of H (also as right cosets of H for that matter). Since G is finite, we see that we can write G as the disjoint union of finitely many left cosets of H, say

$$G = \cup_{i=1}^{n} f_i \cdot H. \tag{4.2}$$

We claim that each left coset $f \cdot H$ of H in G contains the same number of elements as H itself. Consider the function $\ell : H \to f \cdot H$, defined by $\ell(h) = f \cdot h$. By definition of a coset, this function is surjective. The function ℓ is also injective: if $f \cdot h_1 = f \cdot h_2$, then by multiplying with f^{-1} from the left, we obtain that $h_1 = h_2$. We can conclude that the function $\ell : H \to f \cdot H$ is a bijection. But then $|(f \cdot H)| = |H|$, which is what we wanted to show. Since this holds for any $f \in G$, equation (4.2) now implies that $|G| = n|H|$. ■

This theorem is known as *Lagrange's theorem*. The following notation concerning the number of cosets is often used:

Definition 4.4.2 *Let $(G, \cdot)$ be a group and H a subgroup. The number of left (or right) cosets of H in G is denoted by $[G : H]$ and called the index of H in G.*

A priori, the number of left cosets of H in G could have been different from the number of right cosets of H in G. However, if G is a finite group, we see from the proof of Theorem 4.4.1 that $[G : H] = |G|/|H|$ and since this proof works equally well using left as using right cosets, the number of left or right cosets is actually the same. For infinite groups this is actually true as well, see for example Exercise 28.

Example 4.4.3 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, $rH = \{r, r^3, r^5, r^7\}$, $sH = \{s, r^6 s, r^4 s, r^2 s\}$ and $rsH = \{rs, r^7 s, r^5 s, r^3 s\}$.

A further consequence of Theorem 4.4.1 is the following:

Proposition 4.4.4 *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 4.1.8 and Theorem 4.4.1. 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 $g \in G$. ■

It is not true that for any group $(G, \cdot)$ and any divisor d of $|G|$, there exists an element $g \in G$ such that $\mathrm{ord}(g) = d$. For example, if we consider the group $(S_3, \circ)$, then $|S_3| = 6$, but no element in S_3 has order six.

Aside 4.4.5 *The French scientist Augustin-Louis Cauchy proved in 1845 that if d is a prime number dividing the order of a group $(G, \cdot)$, then there exists $g \in G$ such that* $\mathrm{ord}(g) = d$. *To illustrate this, consider again the group $(S_3, \cdot)$. Then the only prime numbers dividing $|S_3|$ are 2 and 3. Indeed there exist elements in S_3 having order 2, for example* $(1\,2)$, *and order 3, for example* $(1\,2\,3)$.

Proposition 4.4.4 has a number of interesting consequences for specific groups.

Corollary 4.4.6 *Let $d, n \in \mathbb{Z}$ be two integers and assume that* $\gcd(d, n) = 1$. *Then* $d^{\phi(n)} \equiv 1 \pmod{n}$. *As before, $\phi : \mathbb{Z}_{\geq 1} \to \mathbb{Z}_{\geq 1}$ denotes Euler's totient function defined by $\phi(n) := |(\mathbb{Z}_n)^*|$.*

Proof. The corollary follows by applying Proposition 4.4.4 to the group $((\mathbb{Z}_n)^*, \cdot_n)$. ■

This corollary is known as Euler's theorem in group theory. In case n is a prime it is known as Fermat's little theorem.

Corollary 4.4.7 *Let p be a prime number and d a natural number such that p does not divide d. Then* $d^{p-1} \equiv 1 \pmod{p}$.

Proof. This corollary follows from Corollary 4.4.6 once we show that $\phi(p) = p - 1$. However,

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

since p is a prime any positive integer strictly smaller than p is relatively prime to p, so we find that

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

■

Aside 4.4.8 *Fermat's little theorem should not be confused with Fermat's last theorem, which states that for any integers $n \in \mathbb{Z}_{\geq 3}$, the equation $x^n + y^n = z^n$ has no solutions within the integers, apart from trivial solutions where at least one of the variables is zero.*

Remark 4.4.9 *Corollary 4.4.6 is used in the RSA cryptosystem. In this cryptosystem, a natural number n is chosen that is the product of two distinct prime numbers p and q. The number n is public, the prime numbers p and q are not. Further an encryption exponent $e \in \mathbb{Z}> 1$ is chosen satisfying* $\gcd(e, \varphi(n)) = 1$ *and made public. A (secret) message $m \in (\mathbb{Z}_n)^*$ is now encrypted as $c := m^e \bmod n$ and sent to a receiver. The receiver, who does know the prime numbers p and q, knowns $\varphi(n) = (p-1)(q-1)$ and can therefore compute $d \in \mathbb{Z}$ such that* $de \equiv 1 \pmod{\varphi(n)}$. *Note that d exists, since we chose e such that* $\gcd(e, \gcd(n)) = 1$. *Then Euler's theorem implies that* $m^{de} \equiv 1 \pmod{n}$, *so that* $c^d = m^{de} \equiv m \pmod{n}$. *Hence the receiver can retrieve the original message from the encrypted message c. In the above formulation, we assumed for simplicity that the message m lies in $(\mathbb{Z}_n)^*$. This is in fact not necessary. Indeed, since $\varphi(n) = (p-1)(q-1)$, the integer $p-1$ divides $de-1$. Hence for any $m \in \mathbb{Z}_n$ that is not a multiple of p, Fermat's little theorem implies that* $m^{de-1} \equiv 1 \pmod{p}$. *However, if p does divide m, then it also divides $m^{de} \bmod p$. Hence for any $m \in \mathbb{Z}_n$ we have* $m^{de} \equiv m \pmod{p}$. *Similarly one shows that* $m^{de} \equiv m \pmod{q}$. *But then p as well as q divide $m^{de} - m$. Since p and q are*

distinct prime numbers, this implies that pq divides $m^{de} \equiv m$. Using that $n = pq$, we conclude that $m^{de} \equiv m \pmod{n}$. In practice some modifications are made to make en- and decryption faster. Also the primes p and q have to be chosen carefully, since for some choices the system can be broken.

4.5 Extra: using cosets for error correction

The fact that Lagrange's theorem can be proven so elegantly using cosets, already motivates their usefulness in group theory itself, but cosets appear in different contexts as well. In this section, we will explain how they can be used in communication to correct errors. The classical setting where error correction is needed is that of information transmission: a message needs to be sent from a "sender" to a "receiver". However, while the message is on its way, it may be slightly altered due to "noise", so that the version of the message that the receiver obtains can differ from the message that was sent. Examples of this are phone calls, data transmission using satellites, or pictures sent to earth by a space exploration vessel. In these examples, there is a clear sender and receiver (though in a phone call their roles are reversed all the time). Other examples are a BluRay-player reading the information (the message) from a BluRay-disc and converting it to signals for your TV. In this case, the BluRay-disc can be seen as the sender, while the BlueRay-player is the receiver. Errors in the message can occur due to various factors, one of them being fabrication mistakes (errors in the information on the BluRay-disc itself), but also scratches on the disc. Hence the word "noise" that we used previously should not be taken literally. In the BluRay-case the "noise" that changes the message, comes from fabrication errors and scratches, not actual sound. Without a way to correct errors, none of these technologies would work properly!

For simplicity, we will assume that a message $\mathbf{m}$ is an element of $\mathbb{Z}_2^k$ for some positive integer k. Hence a message in this model, consists of k bits (zeros or ones). Before sending information, it is often compressed. Therefore we will assume that all bits in $\mathbf{m}$ are equally important and that when one bit is lost or changed, there is no a priori way for the receiver to detect this, let alone to undo these changes. The solution to this problem in the theory of error-correcting codes, is to encode the message by adding cleverly chosen redundancy to it. In other words: the message $\mathbf{m} \in \mathbb{Z}_2^k$ is not sent directly, but instead a *codeword* $\mathbf{c} \in Z_2^n$, where $n \geq k$ and where $\mathbf{c}$ depends on $\mathbf{m}$. More precisely, the sender uses an encoding-function $E : \mathbb{Z}_2^k \to \mathbb{Z}_2^n$, so that $\mathbf{c} = \phi(\mathbf{m})$. Since it is necessary to be able to reconstruct the original message $\mathbf{m}$ from its associated codeword $\mathbf{c}$, the encoding-function needs to be injective. The image of E is called an *error-correcting code* and elements in it are called *codewords*.

The redundancy in $\mathbf{c}$ can then be used by the receiver to detect and correct the errors that occurred during transmission. We will make the simplifying assumption that the only changes that can occur during transmission are bit-flips: any of the coordinates of $\mathbf{c}$ can change from a 0 to a 1 or from a 1 to a 0 with a certain (small) probability. This means that the number of coordinates does not change during transmission (there are no insertions or deletions). Now let us denote by $\mathbf{r} \in \mathbb{Z}_2^{\mathbf{n}}$ the *received word*. Using addition modulo 2, we can describe the relation between the sent codeword $\mathbf{c}$ and the received word $\mathbf{r}$ using addition modulo two: $\mathbf{r} = \mathbf{c} +_2 \mathbf{e}$. Here $\mathbf{e} \in \mathbb{Z}_2^n$ is called the *error vector*. Writing $\mathbf{e} = (e_1, \ldots, e_n)$, we see that if $e_i = 0$, then no error occurred in the i-th coordinate during transmission, while if $e_i = 1$, the i-th coordinate of

the sent codeword $\mathbf{c}$ was flipped during transmission.

As an example, let us consider the case where one bit needs to be transmitted, so that $\mathbf{m} = 0$ or $\mathbf{m} = 1$. If the probability of a bit flip equals 0.01, the probability that the bit arrives unchanged at the receiver equals $1 - 0.01 = 0.99$. In some situations this may not be good enough. Now the sender decides to send $\mathbf{c} = (0, 0, 0) \in \mathbb{Z}_2^3$ if $\mathbf{m} = 0$ and $\mathbf{c} = (1, 1, 1) \in \mathbb{Z}_2^3$ if $\mathbf{m} = 1$. In other words: the used encoding function $E : \mathbb{Z}_2 \to \mathbb{Z}_2^3$ is defined by $E(0) = (0, 0, 0)$ and $E(1) = (1, 1, 1)$. Now if the receiver obtains the sent word $\mathbf{r} = (1, 0, 1)$, it is immediately clear that an error occurred. After all, $(1, 0, 1)$ is not one of the possible codewords $(0, 0, 0)$ or $(1, 1, 1)$. Also, since we are assuming that the probability for bit flips to occur is rather small, it is most likely that the sent codeword was $(0, 0, 0)$. Indeed if $(0, 0, 0)$ was the sent codeword, only one bit flip occurred during transmission, while if $(1, 1, 1)$ would have been the sent codeword, two bit flips would have occurred. Hence a simple majority-vote can be used to *decode*, that is, to reconstruct the most likely sent codeword from the received word. The following table illustrates the decoding:

received word	most likely sent codeword	most likely message
$(0, 0, 0), (1, 0, 0), (0, 1, 0),$ or $(0, 0, 1)$	$(0, 0, 0)$	0
$(1, 1, 1), (0, 1, 1), (1, 0, 1),$ or $(1, 1, 0)$	$(1, 1, 1)$	1

The sent codeword is not the same as the most likely sent codeword if two or three bit flips occurred during transmission. If for example, the probability of a bit flip equals 0.1, then the correct message is reconstructed with probability $1 - \binom{3}{2}(1 - 0.1)(0.1)^2 - \binom{3}{3}(0.1)^3 \approx 0.9997$, which is a nice improvement compared to not using coding theory.

There is another way to view this example. To determine the most likely sent codeword from the received vector $\mathbf{r}$, it also suffices to find a most likely error vector for a given received word. Since more bit flips are less likely than few bit flips, this amounts to finding an error vector $\mathbf{e}$ with fewest number of nonzero coordinates, such that $\mathbf{r}$-$\mathbf{e}$ is a codeword. A table giving a most likely error vector for the possible received words is the following:

received word	most likely error vector
$(0, 0, 0)$ or $(1, 1, 1)$	$(0, 0, 0)$
$(1, 0, 0)$ or $(0, 1, 1)$	$(1, 0, 0)$
$(0, 1, 0)$ or $(1, 0, 1)$	$(0, 1, 0)$
$(0, 0, 1)$ or $(1, 1, 0)$	$(0, 0, 1)$

Now note that in this example, the error-correcting code we use is $\{(0, 0, 0), (1, 1, 1)\}$, which is a subgroup of the group $(\mathbb{Z}_2^3, +_2)$. It is called the 3-fold repetition code. Using the given tables, it is easy to check that all received words with a given most likely error vector $\mathbf{e}$ form a coset of this subgroup, namely $\mathbf{e} + \{(0, 0, 0), (1, 1, 1)\}$. Moreover, for all received words within a given coset, a most likely error vector is simply given by an element of this coset that has fewest number of nonzero coordinates. We wrote "a most likely error vector", since in principle there might have been several words in a given coset with fewest number of nonzero coordinates. In the given example though, there is actually no choice and we could have written "the most likely error vector" instead.

We now generalize this example as follows:

Definition 4.5.1 *A binary, linear, error-correcting code C is a subgroup of the group $(\mathbb{Z}_2^n, +_2)$.*

By Lagrange's theorem, $|C| = 2^k$ for some integer k satisfying $0 \leq k \leq n$. This integer k is called the dimension of the code. One can think of C as a vector space of dimension k over the field $\mathbb{Z}_2$, but the terminology "field" will first be explained in a later chapter. Strictly speaking we also need to define an encoding function. Using the language of vector spaces, this can be done simply by choosing an injective linear map $E : \mathbb{Z}_2^k \to C$ of vector spaces over $\mathbb{Z}_2$. The question is now: how to correct errors using a given binary, linear error-correcting code C?

The key to answering this question is to consider cosets of the code C in $\mathbb{Z}_2^n$. It is common for $\mathbf{r} = (r_1, \ldots, r_n) \in \mathbb{Z}_2^n$ to define $w_H(\mathbf{r}) := |\{i \mid r_i \neq 0\}|$. This is called the Hamming weight of $\mathbf{r}$. Now for a given received word $\mathbf{r}$ a most likely error vector is simply a word $\mathbf{e} \in \mathbb{Z}_2^n$ such that in the first place $\mathbf{r} - \mathbf{e} \in C$ and in the second place $w_H(\mathbf{e})$ is as small as possible. The condition $\mathbf{r} - \mathbf{e} \in C$ is equivalent to the statement that $\mathbf{r} + C = \mathbf{e} + C$, so the sought error vector $\mathbf{e}$ can be characterized as follows: for a given received word $\mathbf{r} \in \mathbb{Z}_2^n$, a most likely error vector $\mathbf{e} \in \mathbb{Z}_2^n$ is a representative of the coset $\mathbf{r} + C$ of smallest possible Hamming weight. Such an error vector is called a *coset leader* of the coset $\mathbf{r} + C$.

The idea of *syndrome decoding* is to combine this insight with some computational tricks. There are two computational problems to consider: first of all, for each coset of C in $\mathbb{Z}_2^n$, we need to compute a coset leader. This is in general a difficult problem though. A way out is to precompute a table, listing for each coset a coset leader. Since there are $[\mathbb{Z}_2^n : C] = |\mathbb{Z}_2^n|/|C| = 2^{n-k}$ distinct cosets of C in $\mathbb{Z}_2^n$, this table has a reasonable size as long as $n - k$ is not too large. The second computational problem is that of identifying the right coset in the table from a received word $\mathbf{r}$. Theoretically, the coset is just $\mathbf{r} + C$, but the representative used in the table may be another one. An elegant way out of this second problem is to use a little bit of linear algebra. It is possible to find a full rank matrix $(n - k) \times n$ matrix H such that $H \cdot \mathbf{r} = \mathbf{0}$ if and only if $\mathbf{r} \in C$. Here all multiplications and additions in $H \cdot \mathbf{r}$ should be performed modulo two. For a received word $\mathbf{r}$, the word $H \cdot \mathbf{r} \in \mathbb{Z}_2^{n-k}$ is called the *syndrome* of $\mathbf{r}$. The point of introducing these syndromes is that they make it easy to determine if two words $\mathbf{r}, \mathbf{s} \in \mathbb{Z}_2^n$ lie in the same coset of C or not. Indeed, we have $\mathbf{r} + C = \mathbf{s} + C$ if and only if $\mathbf{r} - \mathbf{s} \in C$ if and only if $H \cdot (\mathbf{r} - \mathbf{s}) = \mathbf{0}$ if and only if $H \cdot \mathbf{r} = H \cdot \mathbf{s}$. Hence, rather than listing all 2^{n-k} possible cosets of C in $\mathbb{Z}_2^n$ in the table with their chosen coset leaders, one can list all possible syndromes in $\mathbb{Z}_2^{n-k}$ with the chosen coset leaders of the corresponding cosets.

In the previous example $C = \{(0,0,0), (1,1,1)\}$, one could have chosen $H_1 = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$, so that the table in terms of syndromes would become

syndrome received word	coset leader
$(0,0)$	$(0,0,0)$
$(1,0)$	$(1,0,0)$
$(1,1)$	$(0,1,0)$
$(0,1)$	$(0,0,1)$

Now let us consider a slightly more complicated example for which, like in the previous example, one can correct up to one error: the binary Hamming code of length seven. Consider the matrix

$$H_2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

The binary Hamming code then consists of the 16 codewords in $\mathbb{Z}_2^7$ in the kernel of this matrix. In this case there are eight possible syndromes, namely all words in $\mathbb{Z}_2^3$. A quick calculation shows that the eight words $(0,\ldots,0),(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,\ldots,0,1)$ all have distinct syndromes, meaning they have to lie in distinct cosets. Since these are precisely all words in $\mathbb{Z}_2^7$ of Hamming weight zero or one, these are the coset leaders we are looking for. In particular, as in the previous example, this example is special since we can talk about the coset leader and the most likely error vector. Then we obtain the following table:

syndrome received word	coset leader
$(0,0,0)$	$(0,0,0,0,0,0,0)$
$(1,0,0)$	$(1,0,0,0,0,0,0)$
$(0,1,0)$	$(0,1,0,0,0,0,0)$
$(1,1,0)$	$(0,0,1,0,0,0,0)$
$(0,0,1)$	$(0,0,0,1,0,0,0)$
$(1,0,1)$	$(0,0,0,0,1,0,0)$
$(0,1,1)$	$(0,0,0,0,0,1,0)$
$(1,1,1)$	$(0,0,0,0,0,0,1)$

If for example the received word is $\mathbf{r}=(1,0,1,0,1,1,1)$, then $H\cdot\mathbf{r}=(0,1,1)$ so that the most likely error vector is $\mathbf{e}=(0,0,0,0,0,1,0)$. Hence the most likely sent codeword is $\mathbf{r}-\mathbf{e}=(1,0,1,0,1,0,1)$ in this case.

To compare the performance of the Hamming code with the 3-fold repetition codes, let us compute the probability that a codeword is decoded to the sent codeword. This will happen when no or one error occurs. If as before, we assume that the bit flip probability is 0.01, this means that correct decoding occurs with probability $(1-0-01)^7+\binom{7}{1}(1-0-01)^6(0.01)\approx 0.9226$. At first sight, this looks worse than the 3-fold repetition code, but one should realize that to transmit one bit, one needs to send three bits using the repetition code, while one can send four bits of information using seven bits with the Hamming code. Hence the Hamming code is much more efficient. This tradeoff between efficiency on the one hand and error-correcting capability on the other hand, is a common phenomenon in the theory of error-correcting codes.

The error-correcting codes used in practice are much more advanced than the two examples given above. However, with the theory developed in the later chapters (specifically: polynomial rings and finite fields), one could study a class of algebraic codes called *Reed-Solomon* codes. These codes are widely used in many applications, including the BluRay-disc.

4.6 Exercises

Multiple choice exercises

1. Let $(G,\cdot)$ be a group and let $H\subseteq G$ be a subgroup. Then
   A. the identity element $e\in G$ is not necessarily contained in H
      TRUE ☐ FALSE ☐
   B. $f^{-1}\in H$ for all $f\in H$
      TRUE ☐ FALSE ☐

C. for all $f, g \in H$ it must hold that $f \cdot g \in H$
TRUE ☐ FALSE ☐

2. Let $(G, \cdot)$ be a group and let $g \in G$. Then the subgroup generated by g is denoted by $\langle g \rangle$ and

A. ☐ is given by all g^i with $i \in \mathbb{Z}$

B. ☐ has order $n + 1$ if g has order n

C. ☐ is a cyclic group

3. Let $(G, \cdot)$ be a group and let $M, N \subseteq G$ be subsets. Then

A. $M \cdot N$ is the set of all $f \cdot g$ where $f \in M$ and $g \in N$
TRUE ☐ FALSE ☐

B. if M is a subgroup of G and $f \in G$ then $f \cdot M$ is $M \cdot \{f\}$
TRUE ☐ FALSE ☐

C. $f \cdot M$ and $M \cdot f$ do not necessarily coincide
TRUE ☐ FALSE ☐

D. $M \cdot f$ is the set of all $m \cdot f$ where $m \in M$
TRUE ☐ FALSE ☐

4. Let $(G, \cdot)$ be a group and let $H \subseteq G$ be a subgroup. Then

A. H is normal if and only if $f \cdot H = H \cdot f$ for all $f \in G$
TRUE ☐ FALSE ☐

B. $f \sim_H g$ if and only if $g \cdot f \in H$
TRUE ☐ FALSE ☐

C. $\sim_H$ is an equivalence relation
TRUE ☐ FALSE ☐

D. $[f]_{\sim_H} = f \cdot H$ for all $f \in G$
TRUE ☐ FALSE ☐

E. $f \notin [f]_{\sim_H}$
TRUE ☐ FALSE ☐

F. $|H|$ divides $|G|$
TRUE ☐ FALSE ☐

5. Let $(G, \cdot)$ be a group, $f, g \in G$ and $H \subseteq G$ be a subgroup. Then

A. ☐ if $f \cdot H \cap g \cdot H \neq \emptyset$ then $f \cdot H = g \cdot H$

B. ☐ $g \in f \cdot H$ is not enough to ensure $f \cdot H = g \cdot H$

C. ☐ to check whether $f \cdot H = g \cdot H$ I can check that $f \sim_H g$

Exercises to get to know the material better

6. As before we denote by $(S_n, \circ)$ the permutation group on a set with n elements. Is $\{\mathrm{id}, (1\,2), (1\,2\,3), (1\,3\,2)\}$ a subgroup of $(S_3, \circ)$?

7. Show that any element in the group $((\mathbb{Z}_8)^*, \cdot_8)$ has order 1 or 2 and use this to determine all possible subgroups (see the examples in the first section of Chapter 3 for a definition of this group).

8. The aim of this exercise is to describe the subgroup $H \subseteq \mathrm{GL}(2, \mathbb{R})$ 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

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

9. Let $G = S_4$ and H the subgroup of $(S_4, \circ)$ defined in Examples 4.1.3 and 4.2.5. Compute $(1\,2\,3)H$.

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

11. 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 can show that the cycle types of g and fgf^{-1} are the same for any $f, g \in S_n$ and that for a given g, all permutations with that cycle type are obtained as f varies among the elements of S_n.

12. Show that $H := \{\mathrm{id}, (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3)\}$ is a normal subgroup of S_4. Hint: First show that $H \subseteq S_4$ is a subgroup of $(S_4, \circ)$. To show normality, use the remark from the previous exercise.

13. Find the left cosets of $H = \{0, 3\}$ in the group $(\mathbb{Z}_6, +_6)$.

14. Let $G = \{\mathrm{id}, (1\,3), (2\,4), (1\,2)(3\,4), (1\,3)(2\,4), (1\,4)(2\,3), (1\,2\,3\,4), (1\,4\,3\,2)\}$ and $N = \{\mathrm{id}, (1\,3)(2\,4)\}$. Show that N is a normal subgroup of G.

15. Determine all possible orders of elements in S_6. [Hint: use the structure of all possible cycle types].

16. Given a finite group $(G, \cdot)$, the exponent of the group is defined as the smallest positive natural number n such that $g^n = e$ for all $g \in G$. Compute the exponent of S_6. Hint: consider the possible cycle types of elements in S_6.

Exercises to get around in the theory

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

18. Let n be a prime number and let $C_n \subseteq D_n$ 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 $an + bm = 1$.)

19. Show that if $H \subseteq G$ and $K \subseteq G$ 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 $H \cup K$?

20. Show that the elements of finite order in an abelian group $(G, \cdot)$ form a subgroup of G.

21. Let $(A, \cdot)$ be a group and consider the direct product $G = A \times A$. Show that the diagonal subset $H = \{(a, a) \mid a \in A\}$ is a subgroup of G.

22. Let f and g be elements of a finite group. Show that the order of fgf^{-1} is the same as the order of g.

23. Recall that a finite 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.

24. Let $H \subseteq G$ 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.

25. Show that if $H \subseteq G$ and $K \subseteq G$ are two normal subgroups of a group $(G, \cdot)$, then $H \cap K$ is also a normal subgroup of $(G, \cdot)$. [Hint: in a previous exercise you already showed that $H \cap K$ is a subgroup of $(G, \cdot)$.]

26. Let H and K be subgroups of a group $(G, \cdot)$, each of finite index. Prove that $H \cap K$ has finite index as well. [Hint: recall that the index is the number of left cosets of $H \cap K$ in G. Prove that this number is at most the product of the number of left cosets of H and K (which is finite by hypothesis) by proving that the map $x(H \cap K) \mapsto (xH, xK)$ is well-defined and injective].

27. Suppose that $H \subseteq G$ is a subgroup of $(G, \cdot)$ and that the product of any two left cosets of H is a left coset of H. Prove that H is normal in G [Hint: prove that for all $f \in G$, $fH \cdot f^{-1}H = H$ and use this to show that $fH = Hf$].

28. Let $(G, \cdot)$ be a group and H one of its subgroups. Denote by G/H the set of left cosets of H in G and by $H\backslash G$ the set of right cosets of H in G. The goal of this exercise is to show that the number of left cosets of H in G is the same as the number of right cosets of H in G, or in other words, that $|G/H| = |H\backslash G|$.

    (a) Suppose that $fH = gH$ for certain $f, g \in H$. Show that $Hf^{-1} = Hg^{-1}$ and conclude that one can define a map $\sigma : G/H \to H\backslash G$ as $\sigma(fH) := Hf^{-1}$.

(b) Show that the map σ from the previous part if a bijection and hence that $|G/H| = |H\backslash G|$.

Comment: If $(G, \cdot)$ is a group of finite order, we had already seen after Definition 4.4.2, that the number of left and right cosets of H in G is the same, namely $|G|/|H|$.

Chapter 5

Group actions and Burnside's lemma

5.1 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 inner diagonals. In this example we can say that the group of symmetries of a cube gives rise to a group of permutations of the elements of the set A containing the four inner diagonals. Similarly, given a set A, elements in S_A give by their very nature rise to a permutation of the elements of A. In this chapter we capture these type of examples into one concept: group actions.

Definition 5.1.1 *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 inconvenient to write $\varphi(f)$, since on occasion we would like to evaluate the permutation $\varphi(f)$ in an element a of A. This would give rise to a notation such as $\varphi(f)[a]$. To avoid having to use parentheses and brackets in this way, we will often write φ_f instead of $\varphi(f)$. In this notation, 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 5.1.2 *Let a group $(G, \cdot)$, a set A and a group action $\varphi : G \to S_A$ be given. 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 5.1.3 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. We claim that the resulting map $\varphi : \mathrm{GL}(2, \mathbb{R}) \to S_{\mathbb{R}^2}$ is 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

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

which means that $\varphi_{M \cdot N} = \varphi_M \circ \varphi_N$.

Example 5.1.4 The symmetric group $(S_n, \circ)$ acts on the set $\{1, 2, \ldots, n\}$ by defining $\varphi_f := f$. In other words, the map $\varphi : S_n \to S_n$ just sends any permutation f to itself. Then $\varphi_{\mathrm{id}} = \mathrm{id}$ and $\varphi_{f \circ g} = f \circ g$, so indeed we have a group action.

Example 5.1.5 The rotation symmetry group of a cube acts on the set of 4 diagonals of the cube. The identity element fixes all diagonals and hence gives rise to the identity permutation. If two rotations r_1 and r_2 give rise to the permutations f_1 and f_2, then $r_1 \circ r_2$ gives rise to the permutation $f_1 \circ f_2$. Hence we indeed obtain a group action.

Similarly, the rotation symmetry group of a cube acts on the set of 12 edges, the set of 6 faces, or the set of 8 vertices.

Group actions do not only occur in applications of group theory, but also in the abstract setting. We consider one example:

Example 5.1.6 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 $\varphi_f[g] := fgf^{-1}$.

Aside 5.1.7 *There is another commonly used way to describe group actions, namely as a map $\alpha : G \times A \to A$ satisfying that $\alpha(f \cdot g, a) = \alpha(f, \alpha(g, a))$ and $\alpha(\mathrm{id}, a) = a$ for all $f, g \in G$ and $a \in A$. This approach is equivalent to ours. Given such a map $\alpha : G \times A \to A$, one can define $\varphi_f : A \to A$ by $\varphi_f[a] := \alpha(f, a)$. Then the map $\varphi : G \to S_A$ sending f to φ_f is a group action. Conversely, given a group action $\varphi : G \to S_A$, one can define the map $\alpha : G \times A \to A$ by $\alpha(f, a) := \varphi_f[a]$. Then $\alpha(f \cdot g, a) = \varphi_{f \cdot g}[a] = \varphi_f[\varphi_g[a]] = \alpha(f, \alpha(g, a))$ and $\alpha(\mathrm{id}, a) = \varphi_{\mathrm{id}}[a] = a$ for all $f, g \in G$ and $a \in A$.*

Aside 5.1.8 *Strictly speaking, the group action from Definition 5.1.1 is called a left group action. For a right group action, condition 2. in Definition 5.1.1 is replaced by*

2.′ $\psi_{f\cdot g} = \psi_g \circ \psi_f$ *for all* $f, g \in G$.

An example of a right group action is obtained when acting on the elements of the group itself by multiplying from the right. Then for any $h \in G$, $\psi_{f\cdot g}[h] = h \cdot f \cdot g = \psi_g[h \cdot f] = \psi_g[\psi_f[h]]$ *so that indeed* $\psi_{f\cdot g} = \psi_g \circ \psi_f$. *Since any right group action* $\psi : G \to S_A$ *gives rise to a left group action* $\varphi : G \to S_A$ *by defining* $\varphi_f := \psi_{f^{-1}}$, *the theory in this chapter also applies to left group actions.*

5.2 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. We will mainly use group action to solve several problems in combinatorics.

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

Definition 5.2.1 *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 $a \in A$.

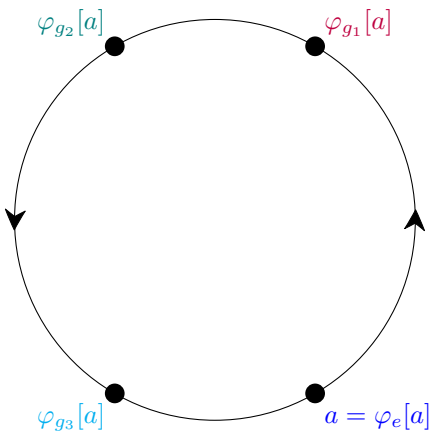


Figure 5.1: The orbit O_a with $G = \{e, g_1, g_2, g_3\}$

Example 5.2.2 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 5.2.3 A more abstract example of a group action is given by the following: Let $(G, \cdot)$ be a group and $H \subseteq G$ 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 $h = e$.

Aside 5.2.4 *If in Example 5.2.3 we choose $H = G$, we obtain a group action $\varphi : G \to S_G$. It turns out that the map φ is injective (one says that the group action is faithful). Using φ one can identify the group $(G, \cdot)$ with a subgroup of $(S_G, \circ)$, namely the image of φ. This result is known as Cayley's theorem. When group theory first emerged, one typically thought of groups as subgroups of symmetric groups, while later a more abstract, axiomatic approach was developed. Cayley's theorem assures that one can think of any group as a subgroup of a symmetric group without losing any generality.*

As we have seen in Chapter 4, cosets are equivalence classes of a certain equivalence relation. It turns out that this is true for orbits as well as we will show now.

Lemma 5.2.5 *Let $\varphi : G \to S_A$ be a group action. Define the relation $\sim_\varphi$ on A as:*

$$a \sim_\varphi b \text{ if and only if there exists } f \in G \text{ such that } b = \varphi_f[a].$$

Then $\sim_\varphi$ is an equivalence relation. Moreover for any $a \in A$, we have $O_a = [a]_{\sim_\varphi}$.

Proof. It is true that $a \sim_\varphi a$, since $\varphi_e[a] = a$. Further if $a \sim_\varphi b$, then for some $f \in G$ we have $b = \varphi_f[a]$. Using Lemma 5.1.2, we obtain that $a = \varphi_{f^{-1}}[b]$ implying that $b \sim_\varphi a$. Finally, if

$a \sim_\varphi b$ and $b \sim_\varphi c$, that is to say if $b = \varphi_f[a]$ and $c = \varphi_g[b]$ for certain $f, g \in G$, then we have

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

where in the last equality we used the second defining property of a group action. Hence $a \sim_\varphi c$. This shows that $\sim_\varphi$ is an equivalence relation on A.

By the definition of equivalence class (see equation (1.1)) and the definition of an orbit, we obtain:

$$[a]_{\sim_\varphi} = \{b \in A \mid \exists f \in G : \ a \sim_\varphi b\} = \{b \in A \mid \exists f \in G : \ b = \varphi_f[a]\} = O_a.$$

■

Just as we did for cosets, we can now apply the theory of equivalences classes as described in Theorem 1.3.3 to orbits. We obtain:

Proposition 5.2.6 *Let a group $(G, \cdot)$, a set A, and a group action $\varphi : G \to S_A$ be given. Then*

1. *For any $a \in A$ we have $a \in O_a$.*
2. *The set A is covered by orbits: $\cup_{a \in A} O_a = A$.*
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 $b = \varphi_f[a]$.*

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

Lemma 5.2.7 *Let $\varphi : G \to S_A$ be a group action of a group $(G, \cdot)$ on a set A. Further let $a \in A$ be given. Then G_a is a subgroup of $(G, \cdot)$. If $b \in O_a$ and $f \in G$ satisfies $b = \varphi_f[a]$, then the left coset $f \cdot G_a$ consists precisely of those elements $g \in G$ such that $b = \varphi_g[a]$. In other words:*

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

Proof. First we show that G_a is a subgroup. We use Lemma 4.1.2 for this. If $f, g \in G_a$, then using the defining property of a group action and Lemma 5.1.2, we see that $\varphi_{fg^{-1}} = \varphi_f \circ \varphi_g^{-1}$. Since $\varphi_f[a] = a$ and $\varphi_g[a] = a$, we obtain from this that $\varphi_{fg^{-1}}[a] = \varphi_f[\varphi_g^{-1}[a]] = \varphi_f[a] = a$. Hence $fg^{-1} \in G_a$. This implies that G_a is a subgroup.

Now we show the statement $f \cdot G_a = \{g \in G \,|\, b = \varphi_g[a]\}$. 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 $\varphi_g[a] = \varphi_{f \cdot h}[a] = \varphi_f[\varphi_h[a]] = \varphi_f[a] = b$. Hence $f \cdot G_a \subseteq \{g \in G \,|\, b = \varphi_g[a]\}$.

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 $g \in f \cdot G_a$. We can conclude that $\{g \in G \mid b = \varphi_g[a]\} \subseteq f \cdot G_a$. ■

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

Example 5.2.8 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 5.2.9 *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 5.2.7, 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).

5.3 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 (also called faces or facets) of the cube a colour of our choice. Let us assume that we can choose between 2 colours. Then 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. Then 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 5.3.1 (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 5.2.6 we obtain that $O_a = O_b$. By Theorem 5.2.9, 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 5.2.9.

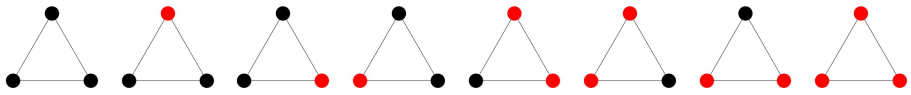
Now we count $|M|$ in another way. Given $g \in G$, there are $|\mathrm{Fix}(g)|$ possibilities for $a \in A$ satisfying $\varphi_g[a] = a$. Therefore

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

The theorem now follows. ■

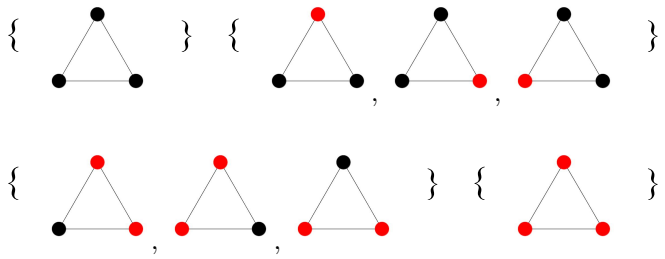
Example 5.3.2 As an example, let us consider an equilateral triangle (that is, a regular 3-gon). Let us suppose that the vertices of the triangle can be coloured using either black or red. Then we obtain 8 possible colourings of the triangle, see Figure 5.2.

Figure 5.2: Eight possible colourings of the vertices of a triangle with colours black and red.



Let us denote the set of such coloured triangles by A. The cyclic group $(C_3, \circ)$ acts on the set A of coloured triangles. Indeed, a coloured triangle can be rotated and in this way mapped to another coloured triangle. This group action gives rise to four orbits are given in Figure 5.3.

Figure 5.3: The four orbits under the action of C_3.



Example 5.3.3 We return to the problem posed in the beginning of the section and consider colourings of the six sides of a cube using two colours. The group of 24 rotation symmetries acts on the set of 2^6 colourings of the cube. We can count the number of distinct colourings by computing the number of orbits under this action. 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.

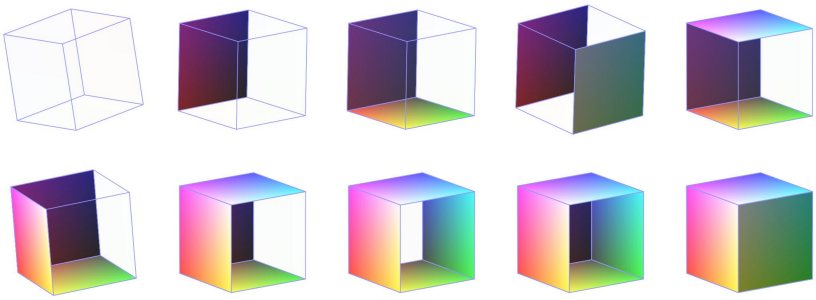
5. Rotations with rotation axis through opposite vertices and rotation angle $2\pi/3$ or $-2\pi/3$. There are 8 such rotations. Each of them fixes 2^2 colourings.

Using Theorem 5.3.1 we can compute the total number of distinct colourings as follows:

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

A list of representatives for each of the ten orbits is given in Figure 5.4. In this figure one of the used the colours is transparent white, in order to make the colours used for the faces at the back of the cube visible. If the number of colours is increased, more orbits will occur. If we

Figure 5.4: Representatives for each of the ten orbits of coloured cubes using two colours.



have m colours, where m is some positive integer, one can still use Burnside's lemma to compute the number of distinct colourings. With a very similar reasoning as for the case of two colours, Burnside's lemma implies that the number of distinct colourings of the sides of the cube using m colours, is given by:

$$\frac{1}{24}\left(m^6 + 6m^3 + 3m^4 + 6m^3 + 8m^2\right). \tag{5.1}$$

This formula shows for example, that if we have three colours at our disposal, there are precisely 57 distinct ways to colour the sides of the cube.

5.4 Extra: Polya theory

So far, we have looked at group actions and seen how they can be used to solve combinatorial problems. In this section we will develop this theory further.

5.4.1 The cycle index of a group action

The first key idea is to describe a group action in an algebraic way, namely using a polynomial in several variables. Given a group action $\varphi : G \to S_A$ on a finite set A, we know from Theorem 2.3.5 that for any $f \in G$, the permutation $\varphi_g \in S_A$ can be written as the composition of mutually disjoint cycles. The cycle type then describes the lengths of these cycles. This information can also be captured using a specific polynomial:

Definition 5.4.1 *Let $(G,\cdot)$ be a finite group and $\varphi : G \to S_A$ a group action on a finite set A, say $|A| = n$. Further, let $g \in G$ and suppose that φ_f has cycle type $(t_1,\ldots,t_n)$. Then the* cycle index *of g, denoted by $Z_g(X_1,\ldots,X_n)$, is defined as*

$$Z_g(X_1,\ldots,X_n) := \prod_{i=1}^{n} X_i^{t_i}.$$

The cycle index of the group action φ is defined as

$$Z_G(X_1,\ldots,X_n) := \sum_{g\in G} Z_g(X_1,\ldots,X_n).$$

Hence the cycle index of $g \in G$ is a polynomial (actually a monomial) in the n variables $X_1,\ldots,X_n$, while the cycle index of the group action is a polynomial in $X_1,\ldots,X_n$ with coefficients from $\mathbb{Z}$. The dependency of the cycle index on the group action is suppressed from the notation Z_g and Z_G. For convenience, we have used variables $X_1,\ldots,X_n$, where $n = |A|$, but some of the variables may not occur in the cycle index. Indeed, by the orbit-stabilizer theorem (Theorem 5.2.9), the length of a cycle occurring in the disjoint cycle decomposition of φ_g has to be a divisor of $|G|$, so that in fact we could suffice with the variables X_d for all divisors $d \le n$ of $|G|$. To simplify notation, we will sometimes in concrete examples leave out in the cycle index all variables that cannot actually occur.

In the literature, the cycle index of a group action is sometimes defined as

$$Z_G^{(\mathrm{norm})}(X_1,\ldots,X_n) := \frac{1}{|G|}\sum_{g\in G} Z_g(X_1,\ldots,X_n).$$

We will call this the *normalized* cycle index of a group action. The cycle index of $e \in G$ is the easiest to compute, since $\varphi_e = \mathrm{id}_A$. Hence $Z_e(X_1,\ldots,X_n) = X_1^n$.

Example 5.4.2 Consider the group $(G,\cdot) = (D_4,\circ)$ of symmetries of the square and the natural group action it has on the set of four vertices of the square. Then the permutations corresponding to the elements $e, r, r^2, r^3, s, rs, r^2s, r^3s$ are respectively id, $(1\,2\,3\,4)$, $(1\,3)(2\,4)$, $(1\,4\,3\,2)$, $(2\,4)$, $(1\,2)(3\,4)$, $(1\,3)$, $(1\,4)(2\,3)$. Hence we find in this case that

$$Z_{D_4}(X_1,X_2,X_4) = X_1^4 + X_4 + X_2^2 + X_4 + X_1^2X_2 + X_2^2 + X_1^2X_2 + X_2^2 = X_1^4 + 2X_1^2X_2 + 3X_2^2 + 2X_4.$$

Example 5.4.3 We consider the group $(G,\cdot)$ of 24 rotation symmetries of the cube and its natural group action on the six sides of the cube. We have seen in Example 5.3.3 that there are five types of rotation symmetries. For each of these five types we determine its cycle index:

1. The identity symmetry e. It has cycle index X_1^6.
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 has cycle index $X_1^2X_4$.
3. Rotations with rotation axis through the midpoints of opposite sides and rotation angle π. There are 3 such rotations. Each of them has cycle index $X_1^2X_2^2$.

4. Rotations with rotation axis through the midpoints of opposite edges and rotation angle π. There are 6 such rotations. Each of them has cycle index X_2^3.

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 has cycle index X_3^2.

In total we see that the cycle index of the action is:

$$Z_G(X_1, X_2, X_3, X_4) = X_1^6 + 6X_1^2X_4 + 3X_1^2X_2^2 + 6X_2^3 + 8X_3^2.$$

Previously, we considered several "colouring problems". In these cases, the elements of the set A could be coloured using a number of colours. More formally, we can see such a colouring as a map colour : $A \to C$, where C denotes the set of colours, say $C = \{C_1, \ldots, C_m\}$. Given such a colouring map, one can then think of colour(a) as the colour that and element $a \in A$ is given. The set of all maps from A to C (this set is often denoted by C^A) then describes all possible colourings of A with colours from C. Note that $|C^A| = |C|^{|A|}$, motivating the notation C^A. Any group action $\varphi : G \to A$ of a group $(G, \cdot)$ on a set A, gives rise to a group action of the same group G on all colourings of A: given an element from C^A, that is given a map colour : $A \to C$, the map colour $\circ\, \varphi_g$ is also a map from A to C and hence an element of C^A. In this way, the original group action on A is said to *induce* a group action on C^A. We have in Example 5.3.3 essentially counted orbits of this induced group actions in C^A, where C contained two colours and A was the set of six faces of the cube.

It turns out that when studying colourings of a set A on which a group G acts, the cycle index of the action of G on A is a useful tool:

Lemma 5.4.4 *Let $(G, \cdot)$ be a finite group and $\varphi : G \to S_A$ a group action on a finite set A. Suppose that we can colour the elements of A using m distinct colours. Then the number of distinct colourings of A is equal to*

$$\frac{1}{|G|} Z_G(m, m, \ldots, m).$$

Proof. Let $C = \{C_1, \ldots, C_m\}$. According to Burnside's lemma, we need for all group elements $g \in G$, to count the number of fixed colourings in C^A under the induced group action. This number of fixed colourings can be determined from the cycle type of the permutation φ_g. Indeed, the colour of all elements $a \in A$ occurring within the same cycle in the disjoint cycle decomposition of φ_g, has to be the same if the colouring is to be fixed under the induced action of g. Therefore the number of colourings that is fixed under the action of g, only depends on the number of cycles (including one-cycles) in the disjoint cycle decomposition of φ_g.

If the cycle index of g is given by $X_1^{t_1} X_2^{t_2} \cdots X_n^{t_n}$, the number of cycles in the disjoint cycle decomposition is exactly given by $t_1 + t_2 + \cdots + t_n$. Therefore, if one can choose between m distinct colours, the number of colourings fixed by g equals $m^{t_1} m^{t_2} \cdots m^{t_n}$. That is to say that

$$|\mathrm{Fix}(g)| = Z_g(m, m, \ldots, m).$$

The lemma then follows using Burnside's lemma and the definition of the cycle index of G. ■

This lemma can for example be used to compute the number of distinct colourings of the sides of a cube. Indeed, using the cycle index computed in Example 5.4.3, one recovers equation (5.1).

5.4.2 Polya's enumeration theorem

Let us return to the problem of counting the number of colourings of a cube using two colours. We have not considered the possibility that the role of the two colours may be different. For example, one colour may be more "expensive" than the other (one "colour" may simply be the option of not painting a side at all). The way this difference can be modelled is by assigning "weights" or "costs" to each colour. The weight of a colouring is then simply the sum of the weights of all colours occurring in the colouring. The problem is now how to count the number of colourings of a cube of a given cost or weight if, as before, two colourings are identified with each other if one can be obtained from the other by a rotation symmetry of the cube.

It turns out that a variation of Lemma 5.4.4 can take care of this more general counting problem quite elegantly. We will again need the cycle index of a group action, but also a systematic way of keeping track of the weight of colours. This is done in the following definition:

Definition 5.4.5 *Let C be a set and suppose that each $c \in C$ has weight $w_c \in \mathbb{Z}$. We define the formal sum $f_C(x) := \sum_{c \in C} x^{w_c}$ to be the weight generating function of C.*

Formal sums are frequently used in combinatorics. We can even allow infinite sets C, as long as for each weight w only finitely many colours have that weight and as long as only finitely many colours have a negative weight. The reason is that in a formal sum, we do not care about convergence, but we only need to be able to add and multiply them. The generating function $f_C(x)$ should therefore not be seen as a function in the variable x in the usual sense. If we would, it would be important to determine for which values of x, the sum converges. Instead, a generating function is just a way to bookkeep all weights. Now we have all the ingredients that we need to count colourings of a certain weight. The following theorem is known as Polya's enumeration theorem:

Theorem 5.4.6 *Let $(G, \cdot)$ be a finite group and $\varphi : G \to A$ a group action on a finite set A with n elements. Suppose that we can colour the elements of A using colours from a set C with weight generating function $f_C(x)$. Then the number of distinct colourings of A of weight w is given by the coefficient of x^w in the expression*

$$\frac{1}{|G|} Z_G(f_C(x), f_C(x^2), \ldots, f_C(x^n)).$$

Proof. When $g \in G$ acts on a colouring, its weight does not change. Therefore we can use Burnside's lemma for the action of G on the set of all colourings of a certain weight w. We can now mimic the proof of Lemma 5.4.4, but have to take weights into account. Assume that the cycle index of g is given by $X_1^{t_1} X_2^{t_2} \cdots X_n^{t_n}$. If a colour $c \in C$ of weight w_c is used in an ℓ-cycle, then it contributes with ℓw_c to the total weight of the colouring. Now suppose that in a colouring

fixed by g the colours $c_{i,j}$ with weights $w_{i,j}$ are used (with $1 \leq i \leq n$ and $1 \leq j \leq t_i$). Then there are t_1 one-cycles in φ_g for which the colours $c_{1,1}, \ldots, c_{1,t_1}$ are used, there are t_2 two-cycles in s_g for which the colours $c_{2,1}, \ldots, c_{2,t_2}$ are used, etc. Then the total weight w of the colouring is

$$w = (w_{1,1} + \cdots + w_{1,t_1}) + 2(w_{2,1} + \cdots + w_{2,t_2}) + \cdots + n(w_{n,1} + \cdots + w_{n,t_n}).$$

This w is exactly the exponent of x one obtains when multiplying

$$x^{w_{1,1}} \cdots x^{w_{1,t_1}} \cdot ((x^2)^{w_{2,1}} \cdots (x^2)^{w_{2,t_2}}) \cdots ((x^n)^{w_{n,1}} \cdots (x^n)^{w_{n,t_n}}), \tag{5.2}$$

which is exactly one of the terms one obtains when multiplying out the expression

$$Z_g(f_C(x), f_C(x^2), \ldots, f(x^n)).$$

Conversely any term of the form as in equation (5.2) gives rise to a colouring of A of weight w that is fixed by g.

If we wish to know how many possible colourings of weight w are fixed by g, we can therefore simply calculate the coefficient of x^w in $Z_g(f_C(x), f_C(x^2), \ldots, f(x^n))$. Summing over all $g \in G$, we obtain the theorem from Burnside's lemma. ■

If all weights are put to zero and C consists of m colours, we obtain Lemma 5.4.4. Indeed in that case $f_C(x) = |C|x^0 = m$. Therefore

$$\frac{1}{|G|} Z_G(f_C(x), f_C(x^2), \ldots, f_C(x^n)) = \frac{1}{|G|} Z_G(m, m, \ldots, m).$$

According to Theorem 5.4.6, the number we are looking for is the coefficient of x^0. Since $\frac{1}{|G|} Z_G(m, m, \ldots, m)$ does not depend on x at all, this is simply the expression $\frac{1}{|G|} Z_G(m, m, \ldots, m)$ itself.

Example 5.4.7 We have seen in Example 5.4.3 that the cycle index of the rotation symmetries of the cube acting on the sides is equal to

$$Z_G(X_1, X_2, X_3, X_4) = X_1^6 + 6X_1^2X_4 + 3X_1^2X_2^2 + 6X_2^3 + 8X_3^2.$$

Now suppose that $C = \{c_1, c_2\}$, that is we want to colour a side with colours c_1 or c_2. Also suppose that $f_C(x) = x + x^5$, that is: color c_1 has weight 1 and color c_2 has weight 5. What weights are possible and how many colourings of a given weight does the cube have up to rotation symmetry?

The solution is obtained using Theorem 5.4.6, with $f_C(x) = x + x^5$ and the above cycle index. After simplifications one obtains

$$\frac{1}{24} Z_G(f_C(x), f_C(x^2), f_C(x^3), f_C(x^4)) =$$

$$(x + x^5)^6 + 6(x + x^5)^2(x^4 + x^{20}) + 3(x + x^5)^2(x^2 + x^{10})^2 + 6(x^2 + x^{10})^3 + 8(x^3 + x^{15})^2$$

$$= x^{30} + x^{26} + 2x^{22} + 2x^{18} + 2x^{14} + x^{10} + x^6.$$

This means that the only possible weights are $6, 10, 14, 18, 22, 26, 30$ and that there for those weights are $1, 1, 2, 2, 2, 1, 1$ possibilities respectively.

5.5 Exercises

Multiple choice exercises

1. Let $(G, \cdot)$ be a group with identity element e, A be a set and $\varphi : G \to S_A$ be a group action. Then
   A. $\varphi(e) = \mathrm{id}$
   TRUE ☐ FALSE ☐
   B. φ is not necessarily a group homomorphism
   TRUE ☐ FALSE ☐
   C. $\varphi(f^{-1})$ is the inverse of $\varphi(f)$
   TRUE ☐ FALSE ☐
   D. $\varphi(f)$ is an element in A
   TRUE ☐ FALSE ☐

2. Let $(G, \cdot)$ be a group and let $\varphi : G \to S_A$ be a group action of G on the set A. The *orbit* of an element $a \in A$ is
   A. ☐ the set of $g \in G$ such that $\varphi(g)[a] = a$
   B. ☐ the set of all $\varphi(f)[a]$ with $f \in G$
   C. ☐ denoted with O_a

3. Let $(G, \cdot)$ be a group and let $\varphi : G \to S_A$ be a group action of G on the set A. The *stabilizer* of an element $a \in A$ is
   A. ☐ the set of $g \in G$ such that $\varphi(g)[a] = a$
   B. ☐ the set of all $\varphi(f)[a]$ with $f \in G$
   C. ☐ denoted with G_a

4. Let $(G, \cdot)$ be a group, A be a set and $\varphi : G \to S_A$ be a group action. Then
   A. "being in the same orbit" is an equivalence relation, denoted with $\sim_\varphi$
   TRUE ☐ FALSE ☐
   B. two elements $a, b \in A$ can be such that $O_a \cap O_b \neq \emptyset$ but $O_a \neq O_b$
   TRUE ☐ FALSE ☐
   C. if for two elements $a, b \in A$ we see that there exists $g \in G$ with $b = \varphi(g)[a]$ then $O_a = O_b$
   TRUE ☐ FALSE ☐
   D. for all $a \in A$ it holds $O_a = [a]_{\sim_\varphi}$
   TRUE ☐ FALSE ☐
   E. $\bigcup_{a \in A} O_a = A$
   TRUE ☐ FALSE ☐

5. Let $(G, \cdot)$ be a group and let $\varphi : G \to S_A$ be a group action of G on the set A. Let $a \in A$ be fixed. The *orbit-stabilizer theorem* states that
   A. ☐ $|G| = |G_a| \cdot |O_a|$

B. ☐ $[G : G_a] = |O_a|$

C. ☐ the length of the orbit O_a is a divisor of $|G_a|$

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

   A. the number of distinct orbits can be computed using *Burnside's Lemma*
   TRUE ☐ FALSE ☐

   B. if I want to compute the number of orbits of φ I can: compute for all $g \in G$ the number of $a \in A$ such that $\varphi(g)[a] = a$. The sum of all the obtained numbers is the desired result
   TRUE ☐ FALSE ☐

   C. if I want to compute the number of orbits of φ I can: compute for all $g \in G$ the number of $a \in A$ such that $\varphi(g)[a] = a$. The sum of all the obtained numbers divided by $|G|$ is the desired result
   TRUE ☐ FALSE ☐

Exercises to get to know the material better

7. Work out the orbits and stabilizers similarly as in Example 5.3.2, but now for $G = C_4$ and A the set of vertex-coloured squares using two colours.

8. 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. Two 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?

9. Determine all the possible 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.

10. (a) Determine how many different colourings of the vertices a pentagon has using m colours. Two colourings are identified with each other if an element from the dihedral group $(D_5, \cdot)$ maps one colouring to the other.

    (b) Answer the same question as in the first part of the exercise, but now considering colourings of the vertices of a regular 8-gon using m colours. Two colourings are identified if an element from $(D_8, \circ)$ maps one colouring to the other.

11. In this exercise, the conjugation action $\varphi : G \to S_G$ given by $\varphi_f[g] = fgf^{-1}$ as in Example 5.1.6, is studied.

    (a) Show that the conjugation action indeed is a group action.

    (b) Show that $\{f \in G \mid \varphi_f = \mathrm{id}_G\} = \{f \in G \mid \forall g \in G\, f \cdot g = g \cdot f\}$. Comment: this set is called the center of G and denoted by $Z(G)$. The Z in the notation comes from the German word for center: Zentrum.

    (c) Let $g \in Z(G)$. What are the orbit and the stabilizer of g under the conjugation action?

12. Let G be a finite group together with a group action φ of G on a set A. The action φ is called *transitive* if $A = O_a$ for some $a \in A$, that is, for all $a, b \in A$ there exists $g \in G$ such that $b = \varphi_g[a]$. Prove that if φ is transitive then

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

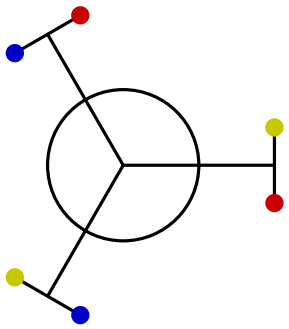
In other words, this exercise says that in a transitive group action, the average number of fixed points of an element is 1.

13. Let G be a non-trivial finite group with a transitive group action on a finite set A (see the previous exercise for the definition of a transitive group action), where $|A| \geq 2$. Show that G must contain at least one element that fixes no element of A.

14. Consider the set A of all 3×3 matrices M, whose entries can be any of the two real numbers $\sqrt{2}$ or π, resulting in 2^9 possible matrices. Suppose that two such matrices M_1 and M_2 are considered to be equivalent if M_2 can be obtained from M_1 by permuting the rows and columns of M_1. How many distinct nonequivalent matrices are there?

15. A decoration, consists of six light bulbs that each independently may have one of the colours red, yellow, or blue. The light bulbs are pairwise attached to iron rods, while the iron rods themselves are attached to the endpoint of one of three spokes of on a wheel (see Figure 5.5).

Figure 5.5: Illustration for Exercise 15



The decoration has several symmetries. First of all, the wheel can be rotated over an angle of $2\pi/3$ or $4\pi/3$, while secondly each rod can be flipped, thus interchanging the light bulbs attached to it. If we identity decorations that can be transformed into one another using combinations of the symmetries described above, how many distinct decorations are there?

Exercises to get around in the theory

16. Let G be a group of order p^n where $n \in \mathbb{N}$ and p is a prime. Suppose that G acts on a finite set A with a group action φ, where $|A|$ satisfies $\gcd(|A|, p) = 1$. Prove that G has a fixed point $a \in A$, that is to say, an element $a \in A$ such that $\varphi(g)[a] = a$ for all $g \in G$.

17. A subgroup H of a group $(G, \cdot)$ is called normal if for all $f \in G$, it holds $fH = Hf$. In this exercise, we relate normal subgroups to the conjugation action $\varphi : G \to S_G$ from Example 5.1.6.

    (a) The orbit of an element $g \in G$ under the conjugation action, is called the *conjugation class* of g and denoted by C_g. Show that if H is a normal subgroup of $(G, \cdot)$ and $g \in H$, then $C_g \subseteq H$.

    (b) Show that any normal subgroup H is a union of conjugation classes.

18. In this exercise, we study conjugation classes of elements in the symmetric group $(S_n, \circ)$.

    (a) Show that for any $g \in S_n$, its conjugation class C_g consists of all permutations having the same cycle type as g. Hint: you may use that $f \circ (a_1 \dots a_m) f^{-1} = (f[a_1] \dots f[a_m])$, see Exercise 11 from Chapter 4.

    (b) Determine all conjugation classes in S_4 and use them to determine all possible normal subgroups of $(S_4, \circ)$. Hint: Exercise 17 and Lagrange's theorem may come in handy.

19. In this exercise we study a transitive group action $\varphi : G \to S_A$ of a finite group G on a finite set A with n elements. See Exercise 12 for the definition of a transitive group action.

    (a) Use the orbit-stabilizer theorem to show that n divides $|G|$.

    (b) The group action φ is called doubly transitive (or 2-transitive) if for any two pairs $(a_1, a_2), (b_1, b_2) \in A^2$, such that $a_1 \neq a_2$ and $b_1 \neq b_2$, there exists $f \in G$ such that $\varphi_f[a_1] = b_1$ and $\varphi_f[a_2] = b_2$. Show that in this case $n(n-1)$ divides $|G|$. Hint: consider a suitable group action on A^2.

    (c) Can you guess what an m-transitive group action is for some positive integer m and generalize the previous?

20. As in Exercise 11, given a group $(G, \cdot)$, define $Z(G) = \{f \in G \mid \forall g \in G\, f \cdot g = g \cdot f\}$.

    (a) Show that $Z(G)$ is a normal subgroup of $(G, \cdot)$.

    (b) Suppose that $|G| = p^n$ for some prime number p and $n \geq 1$. Use the conjugation action of G on itself to show that $|Z(G)| > 1$. Hint: use that $g \in Z(G)$ if and only if $\{g\}$ is an orbit under the conjugation action.

    (c) Conclude that any non-abelian group of order p^n with $n \geq 2$, has a nontrivial normal subgroup. Comment: it turns out that a group of order p^2 always is abelian, but for $n \geq 3$, there exist non-abelian groups of order p^n.

21. Let $n \geq 3$ and $m \geq 1$ be integers. Determine how many different colourings of the vertices a regular n-gon has using m colours. Two colourings are identified with each other if an element from the dihedral group $(D_n, \cdot)$ maps one colouring to the other. Hint: 1) use Corollary 3.2.6, 2) distinguish between n even or odd.

Chapter 6

Maps between groups

6.1 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	$(1\,2\,3)$

Such tables are often called the multiplication tables of a group, even though the actual group operation might be addition. Based on the similarity of the above two 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 6.1.1 *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 *or a* homomorphism of groups*, 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 *or an* isomorphism of groups*.*

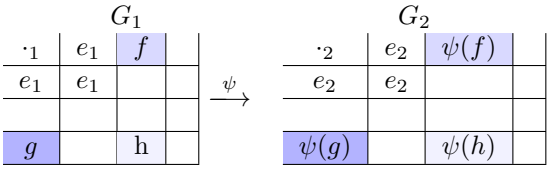


Figure 6.1: The multiplicative tables of two isomorphic groups G_1 and G_2 look the same

Example 6.1.2 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 as can be seen by comparing the tables for the operations $+_3$ and $\circ$. In fact it is a group isomorphism, since ψ is a bijection.

Another example of a group homomorphism comes from Chapter 2:

Example 6.1.3 The sign map $\mathrm{sign} : S_n \to \{1, -1\}$ 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 (2.3), where we saw that $\mathrm{sign}(f \circ g) = \mathrm{sign}(f)\mathrm{sign}(g)$. We also need to check that $\mathrm{sign}(\mathrm{id}) = 1$, 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.

Another example comes from linear algebra:

Example 6.1.4 We denote by $\mathrm{GL}(n, \mathbb{R})$ the set of invertible $n \times n$ matrices with coefficients in $\mathbb{R}$. Let $\det : \mathrm{GL}(n, \mathbb{R}) \to \mathbb{R}\backslash\{0\}$ 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 $\det(I_n) = 1$ (here I_n denotes the $n \times n$ identity matrix) and $\det(A \cdot B) = \det(A)\det(B)$.

If $n = 1$ then $\mathrm{GL}(1, \mathbb{R})$ is just $\mathbb{R}\backslash\{0\}$. 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. For example, we have

$$\det(I) = 1 = \det \begin{pmatrix} -I_2 & 0 \\ 0 & I_{n-2} \end{pmatrix},$$

which shows that the determinant is not an injective function for $n \geq 2$.

Finally, note how similar the definition of a group homomorphism is to our definition of a group action in Definition 5.1.1. Indeed any group action is in fact a group homomorphism:

Example 6.1.5 Let $(G, \cdot)$ be a group, A a set and $\varphi : G \to S_A$ a group action. Since $s_e = \mathrm{id}_A$ 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 5.1.2 can be generalized to any group homomorphism.

Lemma 6.1.6 *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 6.1.7 *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 $\ker(\psi) = \{e_1\}$, we are in a special situation:

Lemma 6.1.8 *Let $\psi : G_1 \to G_2$ be a group homomorphism. Then* $\ker(\psi) = \{e_1\}$ *if and only if ψ is injective.*

Proof. This is one of the exercises posed at the end of this chapter. ■

If ψ is a group isomorphism, it is assumed to be injective. The previous lemma then implies that $\ker(\psi) = \{e_1\}$.

Theorem 6.1.9 *Let $\psi : G_1 \to G_2$ be a group homomorphism. Then* $\ker(\psi) \subseteq G_1$ *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 $\ker(\psi) \subseteq G_1$ 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 $f \cdot_1 (\ker(\psi)) = (\ker(\psi)) \cdot_1 f$. ■

The fact that left and right cosets of the kernel of a group homomorphism are the same is so special that it deserves its own terminology:

Definition 6.1.10 *A subgroup H of a group $(G, \cdot)$ is called* normal, *if $f \cdot H = H \cdot f$ for all $f \in G$.*

Theorem 6.1.9 then simply states that the kernel of a group homomorphism always is a normal subgroup.

Example 6.1.11 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 6.1.12 Given a positive integer n and an integer a, then as explained in Chapter 1, a mod n denotes the remainder of a modulo n. The operation $+_n$ occurring in the group $(\mathbb{Z}_n, +_n)$ is as before explicitly defined by: $a +_n b := (a + b) \bmod n$.

The map

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

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

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

However, this was shown in Corollary 1.3.8. Hence ψ is a group homomorphism.

The kernel of the group homomorphism ψ is given by

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

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

Aside 6.1.13 *For any group $(G, \cdot)$, the subgroups G and $\{e\}$ are normal. A group is called* simple *if these are the only normal subgroups it has. An example of a simple finite group is $(\mathbb{Z}_p, +_p)$ where p is a prime number. Indeed, Lagrange's theorem implies that $\{0\}$ and $\mathbb{Z}_p$ are the only subgroups of $(\mathbb{Z}_p, +_p)$. Simple groups play a similar role in group theory as prime numbers do in the study of integers. A complete description of all possible finite simple groups was undertaken in the second half of the last century, the final details being published in 2004. The entire proof is estimated to be tens of thousands of pages long. It turns out that finite simple groups, when considered up to isomorphism, fall into various infinite families (the cyclic groups of prime order being one of them) and a number of isolated examples (called* sporadic groups*). The largest of these sporadic groups is called the* monster group *$(M, \cdot)$, which has the stunning order*

$$|M| = 808017424794512875886459904961710757005754368000000000.$$

6.2 Quotient groups

Keywords: quotient groups

In this section we will show that a normal subgroup $H \subseteq G$ 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 6.2.1 *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 4.2.1 we can define a multiplication of two cosets. We simply define

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

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 4.2.3. 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 6.2.2 *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. This word was used in Chapter 1 as well: an element of an equivalence class was called a representative of that class. Definition 6.2.2 can then 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 of it. This gives us a way to define a group operation on the set G/H.

Definition 6.2.3 *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 6.2.3 is a group.

Example 6.2.4 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 $(\mathbb{Z}_3, +_3)$ are isomorphic under the isomorphism defined by $\phi(a + 3\mathbb{Z}) = a$ for $a \in \{0, 1, 2\}$. In a moment we will see in Example 6.3.3 that in general $(\mathbb{Z}/n\mathbb{Z}, +)$ is isomorphic to the group $(\mathbb{Z}_n, +_n)$.

6.3 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. 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 6.3.1 *Let $\psi : G_1 \to G_2$ be a homomorphism of groups. Then the map $\overline{\psi} : G_1/\ker(\psi) \to \mathrm{im}(\psi)$ 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 $\psi(f) = \psi(g)$.

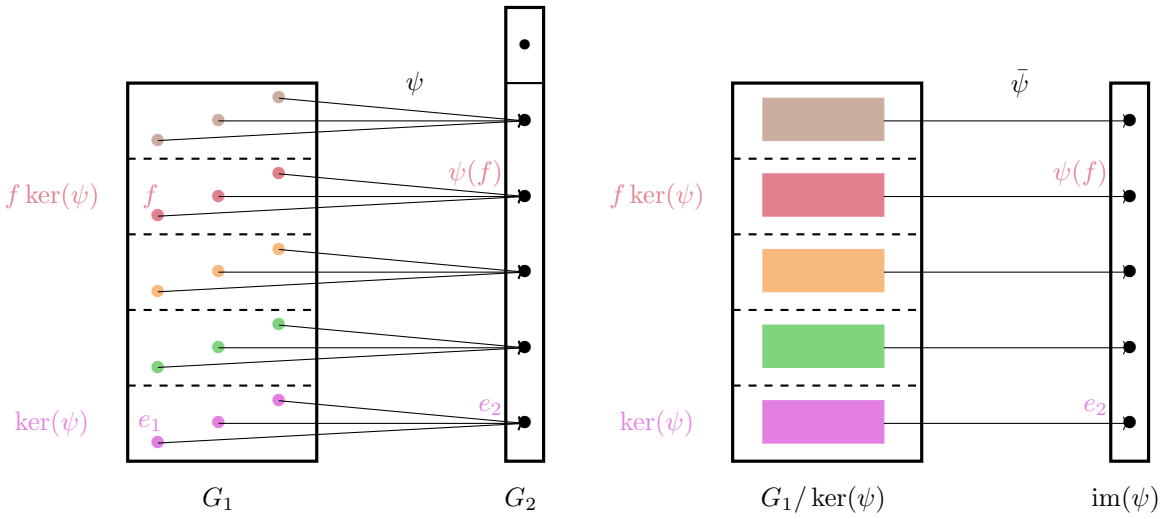
Assuming that $f\ker(\psi) = g\ker(\psi)$, then $f^{-1}g \in \ker(\psi)$. Therefore, we find that $\psi(f)^{-1}\psi(g) = \psi(f^{-1}g) = e_2$, or in other words that $\psi(f) = \psi(g)$. This is exactly what we needed to show.

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. ■

Figure 6.2: Illustration of how a group homomorphism ψ gives rise to a group isomorphism $\overline{\psi}$



Example 6.3.2 The map $\det : \mathrm{GL}(2,\mathbb{R}) \to \mathbb{R}\backslash\{0\}$ is a group homomorphism between the groups $(\mathrm{GL}(2,\mathbb{R}), \cdot)$ 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 $\mathrm{SL}(2,\mathbb{R})$ (all matrices in $\mathrm{GL}(2,\mathbb{R})$ with determinant one). Using Theorem 6.3.1, we see that the groups $(\mathrm{GL}(2,\mathbb{R})/\mathrm{SL}(2,\mathbb{R}),\cdot)$ and $(\mathbb{R}\backslash\{0\},\cdot)$ are isomorphic.

Example 6.3.3 In Example 6.1.12, we saw that the map

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

is a group homomorphism from $(\mathbb{Z},+)$ to $(\mathbb{Z}_n,+_n)$. The image of ψ is all of $\mathbb{Z}_n$ and the kernel of ψ is equal to $n\mathbb{Z}$. Therefore Theorem 6.3.1 gives that the $(\mathbb{Z}/n\mathbb{Z},+)$ is isomorphic to $(\mathbb{Z}_n,+_n)$ with isomorphism $\overline{\psi} : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}_n$ defined by $\overline{\psi}(a+n\mathbb{Z}) = a \bmod n$.

Aside 6.3.4 *In some books, the statement in Theorem 6.3.1 is called the first isomorphism theorem for groups. In such books, the first isomorphism theorem is typically followed by two consequences called the second and third isomorphism theorem (though the numbering is not always the same). The statement of the second and third isomorphism theorem for groups can be found in Exercises 21 and 22.*

6.4 Extra: quotients of vector spaces

Constructions of "quotients" happens for a great deal of mathematical structures and not only for groups. Indeed, later we will discuss quotients of other algebraic structures called rings, see Chapters 7 and 8. In this section, we will discuss quotients of vector spaces and relate it to a well known result from linear algebra: the rank-nullity theorem. To keep the prerequisites for this section to a minimum, we will discuss over the real or complex numbers only. Since the results we want to discuss are equally valid for these two cases, we will simply use the symbol $\mathbb{F}$. Consequently interpreting $\mathbb{F}$ as the real numbers $\mathbb{R}$ is possible, but a reader who would like to think of $\mathbb{F}$ as the complex numbers $\mathbb{C}$, may do so without having to modify any of the given arguments. In Chapter 7 the notion of a *field* will be defined. Vector spaces can be defined over any such field and a reader who is already familiar with the definition of a field, may think of $\mathbb{F}$ as an arbitrary field. With this notation in place, let us define what a vector space over $\mathbb{F}$ is:

Definition 6.4.1 *A vector space over $\mathbb{F}$ is an abelian group $(V,+)$ together with a multiplication map $\cdot : \mathbb{F} \times V \to V$ satisfying:*

1. *For all $a,b \in \mathbb{F}$ and $v \in V$: $a \cdot (b \cdot v) = (a \cdot b) \cdot v$.*
2. *For all $v \in V$: $1 \cdot v = v$.*
3. *For all $a,b \in \mathbb{F}$ and $v \in V$: $(a+b) \cdot v = a \cdot v + b \cdot v$.*
4. *For all $a \in \mathbb{F}$ and $v_1, v_2 \in V$: $a \cdot (v_1 + v_2) = a \cdot v_1 + a \cdot v_2$.*

An element from $\mathbb{F}$ is often called a scalar. In the definition, we have followed the usual convention to use $+$ (respectively $\cdot$) both for addition of vectors and for addition of scalars (respectively multiplication of a scalar and a vector as well as multiplication of two scalars).

A subset W of a vector space V over $\mathbb{F}$ is called a subspace if W is a subgroup of $(V,+)$ that is closed under scalar multiplication.

Lemma 6.4.2 *Let W be a subspace of a vector space V over $\mathbb{F}$. Then V/W is a vector space over $\mathbb{F}$ with $\cdot : \mathbb{F} \times V/W \to V/W$ defined by $a \cdot (v + W) = (a \cdot v) + W$.*

Proof. First of all, we need to check that the suggested scalar multiplication is well defined. In other words, we need to show that if $v_1+W = v_2+W$ and $a \in \mathbb{F}$, then also $(a\cdot v_1)+W = (a\cdot v_2)+W$. However, $v_1+W = v_2+W$ if and only if $v_1-v_2 \in W$. But if this is the case, then $a\cdot(v_1-v_2) \in W$, since W is a subspace of V. Using that $a \cdot (v_1 - v_2) = a \cdot v_1 - a \cdot v_2$, we see that indeed $(a \cdot v_1) + W = (a \cdot v_2) + W$ whenever $v_1 + W = v_2 + W$.

Since V is a vector space over $\mathbb{F}$, the definition of $\cdot : \mathbb{F} \times V/W \to V/W$ directly implies that V/W is a vector space over $\mathbb{F}$ as well. ■

A finite set of vectors $\{v_1, \ldots, v_n\} \subseteq V$ is called linearly independent if for all $a_1, \ldots, a_n \in \mathbb{F}$, a relation of the form $\sum_{i=1}^n a_i \cdot v_i = 0$ only can hold if $a_1 = \cdots = a_n = 0$. An infinite set of vectors $M \subseteq V$ is called linearly independent if for any choice of finitely many vectors $v_1, \ldots, v_n \in M$, a relation of the form $\sum_{i=1}^n a_i \cdot v_i = 0$ only can hold if $a_1 = \cdots = a_n = 0$. A basis of a vector space is then a set of linearly independent vectors $B \subseteq V$ such that for any vector $v \in V$, the set $B \cup \{v\}$ is not a linearly independent set of vectors. This simply means that any vector $v \in V$ can be expressed as a linear combination of a finite number of vectors from B. It is a deep fact from foundational mathematics that any vector space has a basis. Moreover, for a given vector space V over $\mathbb{F}$, any basis of V has the same cardinality. This cardinality is called the dimensions of V. There are often many possible choices for a basis. One can show that if W is a subspace of V and C is a basis of W, then there exists a basis B of V containing C. In particular $\dim W = |C| \leq |B| = \dim V$. This can be used to determine the dimension of V/W.

Lemma 6.4.3 *Let V be a vector space over $\mathbb{F}$ and W a subspace of V. Then* $\dim(V/W) + \dim(W) = \dim(V)$.

Proof. Let C be a basis of W and B a basis of V containing C. We claim that the set $D := \{v + W \mid v \in B \setminus C\}$ is a basis for V/W. First of all, if $v_1, \ldots, v_n \in B \setminus C$ and $\sum_{i=1}^n a_i \cdot (v_i + W) = 0 + W$, then $\sum_{i=1}^n a_i \cdot v_i \in W$. However, since C is a basis for W, this means that there exist $b_j \in \mathbb{F}$ and $w_1, \ldots, w_m \in C$ such that $\sum_{i=1}^n a_i \cdot v_i + \sum_{j=1}^m b_j \cdot w_j = 0$. Now using that B is a basis for V, we see that $a_1 = \cdots = a_n = 0$ (as well as $b_1 = \cdots = b_m = 0$). Hence the vectors in the set D are linearly independent. Further, any element in V/W can be written as a linear combination of some elements in D. Indeed if $v + W \in V/W$, then we can find $v_1, \ldots, v_\ell \in B$ and $a_1, \ldots, a_\ell \in \mathbb{F}$ such that $v = \sum_{i=1}^\ell a_i \cdot v_i$. Rearranging the terms, we may assume that $v_1, \ldots, v_m \in B \setminus C$ and $v_{m+1}, \ldots, v_\ell \in C$. Then $v + W = (\sum_{i=1}^m a_i \cdot v_i) + W$, since $v - \sum_{i=1}^m a_i \cdot v_i = \sum_{i=m+1}^\ell a_i \cdot v_i \in W$. Therefore $v + W = (\sum_{i=1}^m a_i \cdot v_i) + W = \sum_{i=1}^m a_i \cdot (v_i + W)$, meaning that any element of V/W is a linear combination of elements in D. This shows that D is a basis of V/W.

To conclude the proof, we simply observe that we now know that $\dim(V/W) = |D| = |B \setminus C|$, meaning that $\dim(V/W) + \dim(W) = |B \setminus C| + |C| = |B| = \dim(V)$. ■

With these facts established, we can give a proof of the rank-nullity theorem using the isomorphism theorem for groups.

Theorem 6.4.4 *Let V_1 and V_2 be vector spaces over $\mathbb{F}$ and let $L : V_1 \to V_2$ be a linear map of*

vector spaces over $\mathbb{F}$. *Then* $\dim(\ker L) + \dim(\mathrm{im}L) = \dim V_1$.

Proof. We already know from the isomorphism theorem that the groups $(V_1/\ker L, +)$ and $(\mathrm{im}L, +)$ are isomorphic with isomorphism $\overline{L} : V_1/\ker L \to \mathrm{im}L$ sending $v_1 + \ker L$ to $L(v_1)$. Since for any $a \in \mathbb{F}$ and $v_1 \in V_1$, $a \cdot L(v_1) = L(a \cdot v_1)$, the set $\mathrm{im}L$ is actually a subspace of V_2. Moreover, $\overline{L}(a \cdot v_1 + \ker L) = L(a \cdot v_1) = a \cdot L(v_1) = a \cdot \overline{L}(v_1 + \ker L)$, showing that $\overline{L} : V_1/\ker L \to \mathrm{im}L$ is a bijective linear map. In particular, we may conclude that the vector spaces $V_1/\ker L$ and $\mathrm{im}L$ have the same dimension. Now combining this with Lemma 6.4.3, we see that

$$\dim(V_1) = \dim(V_1/\ker L) + \dim(\ker L) = \dim(\mathrm{im}L) + \dim(\ker L).$$

This is precisely what we wanted to show. ■

The name "rank-nullity theorem" comes from the fact the nullity is defined as the dimension of the kernel of a linear map, while the rank of a linear map is the dimension of its image.

6.5 Exercises

Multiple choice exercises

1. Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be groups with identity elements e_1 and e_2, respectively. Suppose that $\psi : G_1 \to G_2$ is a *group homomorphism*. Then
   - A. $\psi(f \cdot_1 g) = \psi(f) \cdot_1 \psi(g)$ for all $f, g \in G_1$
     TRUE ☐ FALSE ☐
   - B. $\psi(f \cdot_1 g) = \psi(f) \cdot_2 \psi(g)$ for all $f, g \in G_1$
     TRUE ☐ FALSE ☐
   - C. $\psi(e_1) = e_2$
     TRUE ☐ FALSE ☐
   - D. $\psi(f^{-1})$ is the inverse of $\psi(f)$ for all $f \in G_1$
     TRUE ☐ FALSE ☐

2. Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be groups with identity elements e_1 and e_2, respectively. Suppose that $\psi : G_1 \to G_2$ is a group homomorphism. Then the *kernel* of ψ is
   - A. ☐ denoted by $\ker(\psi)$
   - B. ☐ the set of elements in G_1 that are mapped to e_2
   - C. ☐ not necessarily containing e_1
   - D. ☐ a normal subgroup of G_1, that is, $f \cdot_1 \ker(\psi) = \ker(\psi) \cdot_1 f$ for all $f \in G_1$

3. Let $(G, \cdot)$ be a group with identity element e and let H be a normal subgroup of G. Then the *quotient group* G/H
   - A. is the set of cosets $fH = \{f \cdot h \mid h \in H\}$ with $f \in G$
     TRUE ☐ FALSE ☐

B. is a group also if we drop the hypothesis of H being normal
TRUE ☐ FALSE ☐

C. is a group with operation $(fH) \cdot (gH) = (f \cdot g)H$
TRUE ☐ FALSE ☐

4. Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be groups with identity elements e_1 and e_2, respectively. Suppose that $\psi : G_1 \to G_2$ is a group homomorphism. Then the *isomorphism theorem* states that

   A. ☐ $G_1/\mathrm{im}(\psi)$ is isomorphic to $\ker(\psi)$

   B. ☐ the quotient group $G_1/\ker(\psi)$ is isomorphic to $\mathrm{im}(\psi)$ and the isomorphism is given by $\bar{\psi} : G_1/\ker(\psi) \to \mathrm{im}(\psi)$, where $\bar{\psi}(g\ker(\psi)) = \psi(g)\ker(\psi)$, for all $g \in G_1$

   C. ☐ the quotient group $G_1/\ker(\psi)$ is isomorphic to $\mathrm{im}(\psi)$ and the isomorphism is given by $\bar{\psi} : G_1/\ker(\psi) \to im(\psi)$, where $\bar{\psi}(g\ker(\psi)) = \psi(g)$, for all $g \in G_1$

Exercises to get to know the material better

5. 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.

6. 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?

7. Work out what the isomorphism theorem implies when considering the group homomorphism from Exercises 5 and 6.

8. Consider the subgroup $H = \{e, (1\,2)\}$ of S_3. Show that it is not true for all $f, g \in S_3$ that $(fH)(gH) = (fg)H$. Definition 6.2.2 can therefore not be extended to the case where H is not a normal subgroup of G.

9. Consider the subset $H = \{e, r^2\}$ of $(D_4, \circ)$, the dihedral group with 8 elements.

   (a) Show that H is a normal subgroup of $(D_4, \circ)$.

   (b) Show that the elements e, r, s, rs form a complete set of representatives of the cosets of H in D_4. In other words: show that the cosets of H in D_4 are eH, rH, sH, and rsH.

   (c) Work out the multiplication table of the quotient group $(D_4/H, \circ)$ using the complete set of representatives given in the previous part.

   (d) Show, without using the isomorphism theorem for groups, that the groups $(D_4/H, \circ)$ and $(\mathbb{Z}_2 \times \mathbb{Z}_2, +_2)$ are isomorphic. Here $+_2$ denotes coordinate-wise addition of tuples in $\mathbb{Z}_2 \times \mathbb{Z}_2$ modulo two.

10. The notation in this exercise is the same as in Exercise 9. Use the isomorphism theorem for groups to show that the groups $(D_4/H, \circ)$ and $(\mathbb{Z}_2 \times \mathbb{Z}_2, +_2)$ are isomorphic.

11. For each of the following maps, determine whether or not it is a homomorphism of groups. If it is, state whether it is injective or surjective (or both, or neither) and find its kernel:

    (a) $\psi(A) = \det(A)$ from the group of 2×2 real matrices under addition into $(\mathbb{R}, +)$,

    (b) $\psi(A) = \det(A)$ from $(GL_2(\mathbb{R}), \cdot)$ into $(\mathbb{R} \setminus \{0\}, \cdot)$,

(c) $\psi(x) = x^2$ from $(\mathbb{Q}, +)$ into $(\mathbb{Q}, +)$,

(d) $\psi(x) = x^5$ from $(\mathbb{Q} \setminus \{0\}, \cdot)$ into $(\mathbb{Q} \setminus \{0\}, \cdot)$.

12. Define $\psi : (\mathbb{Z}_{24}, +_{24}) \to (\mathbb{Z}_8, +_8)$ by $\psi([x]_{24}) = [x]_8$.

    (a) Show that ψ is well defined,

    (b) show that ψ is surjective,

    (c) find the kernel of ψ,

    (d) what does the isomorphism theorem imply?

13. We are given 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. Also show that if ψ_1 and ψ_2 are isomorphisms, then so is $\psi_2 \circ \psi_1$.

Exercises to get around in the theory

14. Prove Lemma 6.1.8, that is to say: let $\psi_1 : G_1 \to G_2$ be a group homomorphism. Show that $\ker(\psi) = \{e_1\}$ if and only if ψ is injective. Hint: to prove the implication from left to right, first show that if $\psi(f) = \psi(g)$, then $f \cdot_1 g^{-1} \in \ker(\psi)$.

15. Given a group homomorphism $\psi : G_1 \to G_2$. Show that $\mathrm{im}(\psi)$, the image of ψ, is a subgroup of G_2.

16. Let $(G, \cdot)$ be a group. Show that the quotient group $(G/\{e\}, \cdot)$ is isomorphic to $(G, \cdot)$ (without using the isomorphism theorem). Conclude that if $\psi : G_1 \to G_2$ is an injective group homomorphism, then the groups $(G_1, \cdot_1)$ and $(\mathrm{im}(\psi), \cdot_2)$ are isomorphic.

17. Let $(Q_8, \cdot)$ be the quaternion group introduced in Definition 3.4.3. Show that all of its subgroups are normal subgroups. Comment: this shows that even if all subgroups of a group $(G, \cdot)$ are normal, the group $(G, \cdot)$ does not have to be abelian.

18. Show that being isomorphic is an equivalence relation on the set of all groups. Hint: transitivity follows from Exercise 13.

19. An automorphism of a group $(G, \cdot)$ is an group isomorphism $\psi : G \to G$. The set of all automorphisms of a group is denoted by $\mathrm{Aut}(G)$. Show that $(\mathrm{Aut}(G), \circ)$ is a group. Here $\circ$ denotes the composition operator. Comment: $(\mathrm{Aut}(G), \circ)$ is called the *automorphism group* of G.

20. Let two groups $(N, \cdot_1)$ and $(H, \cdot_2)$ be given, as well as a group homomorphism $\psi : H \to \mathrm{Aut}(N)$. One then defines the operator $\cdot_\psi$ on the set $N \times H$ as $(n_1, h_1) \cdot_\psi (n_2, h_2) = (n_1 \cdot_1 \psi(h_1)(n_2), h_1 \cdot_2 h_2)$.

    (a) Show that $(N \times H, \cdot_\psi)$ is a group. It is called the *semidirect product* of $(N, \cdot_1)$ and $(H, \cdot_2)$ with respect to ψ and is often denoted as $(N \rtimes_\psi H, \cdot_\psi)$.

    (b) Let $\psi : H \to \mathrm{Aut}(N)$ be the homomorphism sending any $h \in H$ to the identity automorphism $\mathrm{id}_N : N \to N$. Show that in this case $(N \times H, \cdot_\psi)$ is the direct product of $(N, \cdot_1)$ and $(H, \cdot_2)$ as given in Theorem 3.4.1.

    (c) Show that the dihedral groups $(D_n, \circ)$ can be obtained as a semidirect product of $(C_n, \circ)$ and $(\{1, -1\}, \cdot)$.

21. * *The Second Isomorphism Theorem.* Let $(G, \cdot)$ be a group, H a subgroup of G and N a normal subgroup of G. Prove that

    (a) $HN := \{hn \mid h \in H,\ n \in N\}$ is a subgroup of G,

    (b) $H \cap N$ is a normal subgroup of H,

    (c) $H/(H \cap N)$ is isomorphic to HN/N [Hint: prove that $\psi : H \to HN/N$, with $\psi(h) = hN$ is a group homomorphism and then apply the first isomorphism theorem]

22. * *The Third Isomorphism Theorem.* Let $(G, \cdot)$ be a group, N a normal subgroup of G and K a subgroup of G containing N. Prove that

    (a) N is a normal subgroup of $(K, \cdot)$,

    (b) If additionally K is a normal subgroup of $(G, \cdot)$, then K/N is a normal subgroup of $(G/N, \cdot)$,

    (c) If additionally K is a normal subgroup of $(G, \cdot)$, then the groups $((G/N)/(K/N), \cdot)$ and $(G/K, \cdot)$ are isomorphic.

Chapter 7

Rings

7.1 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 7.1.1 *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 identity element is denoted by* 0_R.
2. *There exists an identity 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 (that is to say that $x +_R y = y +_R x$ for all $x, y \in R$). If we would assume that $(R, +_R)$ is a group, not necessarily abelian, it turns out that the other ring axioms imply that the group is abelian after all. For let us assume that $(R, +_R)$ is a group, not necessarily 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 necessarily 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 7.1.2 The set of integers $\mathbb{Z}$ together with the usual addition $+$ and multiplication $\cdot$ 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. All these rings are commutative.

Example 7.1.3 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 by 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 3).

Example 7.1.4 Let $\mathbb{Z}_n$ 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 $(\mathbb{Z}_n, +_n, \cdot_n)$ is a commutative ring. Several of the ring axioms follow from Theorem 1.4.3 from Chapter 1. The distributive laws can be proven in a similar way as there. For example, let us check that $a \cdot_n (b +_n c) = a \cdot_n b +_n a \cdot_n c$. First one shows that

$$a \cdot_n (b +_n c) \equiv a \cdot (b + c) \pmod{n}$$

and

$$a \cdot_n b +_n a \cdot_n c \equiv a \cdot b + a \cdot c \pmod{n}.$$

Because the distributive laws hold for integers, these equations imply that

$$a \cdot_n (b +_n c) \equiv a \cdot_n b +_n a \cdot_n c \pmod{n}.$$

Finally, since both left-hand side and right-hand side of the previous equation are integers between 0 and $n-1$, we obtain that they are equal: $a \cdot_n (b +_n c) = a \cdot_n b +_n a \cdot_n c$.

Definition 7.1.5 *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 x 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 7.1.6 *Let $(R, +_R, \cdot_R)$ be a ring. Then $(R^*, \cdot_R)$ is a group with identity 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 3, 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 an identity element for the operation $\cdot_R$ on R, it is certainly an identity 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 7.1.7 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 $\mathbb{Z}^* \subseteq \{-1, 1\}$. Conversely, $\mathbb{Z}^* \supseteq \{-1, 1\}$, since both -1 and 1 are units (they are their own inverses).

Example 7.1.8 Let $(\mathbb{Q}, +, \cdot)$ be the ring of rational numbers. Then we have $\mathbb{Q}^* = \mathbb{Q}\backslash\{0\}$. Indeed any rational number number a/b different from 0 has a multiplicative inverse, namely b/a. Similarly $\mathbb{R}^* = \mathbb{R}\backslash\{0\}$ and $\mathbb{C}^* = \mathbb{C}\backslash\{0\}$.

Example 7.1.9 Consider the ring $(\mathbb{Z}_6, +_6, \cdot_6)$. 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 $(\mathbb{Z}_6)^*$. This notation was in fact inspired by the more general ring-theoretical notation we just introduced in Definition 7.1.5. As we have seen in Example 3.1.6 from Chapter 3, in general we have

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

Example 7.1.10 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 by n matrices.

Units have the property of being invertible with respect to multiplication. The examples show that in a ring not all elements need to have a multiplicative inverse. In fact, in rings a very different phenomenon may happen: the product of two elements different from zero may be zero. We will see some examples in a moment.

Definition 7.1.11 *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 7.1.12 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 7.1.13 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 7.1.14 Consider again the ring $(\mathbb{Z}_6, +_6, \cdot_6)$. 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 $(\mathbb{Z}_n, +_n, \cdot_n)$ all non-zero elements $a \in \mathbb{Z}_n$ 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

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

Conversely if there exist non-zero $a, b \in \mathbb{Z}_n$ such that $a \cdot_n b = 0$, then $a \cdot b$ is a multiple of n, while both a and b are not multiples of n. This means that there exists a prime number p dividing n as well as a. Therefore $\gcd(a, n) \geq p > 1$. Previously in Example 3.1.6, we have seen that $a \in \mathbb{Z}_n$ is a unit if and only if $\gcd(a, n) = 1$. Therefore we obtain the following result:

Proposition 7.1.15 *Let n be a positive integer and let $a \in \mathbb{Z}_n$ be an element different from 0. Then exactly one of the following holds:*

1. *a is a zero-divisor, which is the case if and only if $\gcd(a, n) > 1$, or*
2. *a is a unit, which is the case if and only if $\gcd(a, n) = 1$.*

7.2 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 7.2.1 *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 7.2.1, 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 7.2.2 The ring $(\mathbb{R}, +, \cdot)$ 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 7.2.3 A Gaussian integer, is a complex number whose real and imaginary parts are both integers. The set of Gaussian integers is traditionally denoted by $\mathbb{Z}[i]$. In other words, one has:

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

The ring $(\mathbb{Z}[i], +, \cdot)$ of Gaussian integers is a domain. Showing this is an exercise at the end of this chapter.

Another important class of rings is given by fields:

Definition 7.2.4 *Let $(R, +, \cdot)$ be a commutative ring such that $R^* = R \backslash \{0\}$. Then R is called a* field.

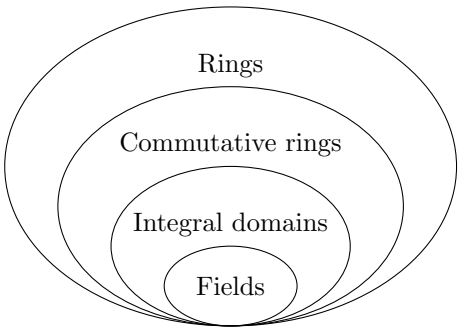


Figure 7.1: The relation between the structures we studied so far

Example 7.2.5 The rings $(\mathbb{Q}, +, \cdot)$, $(\mathbb{R}, +, \cdot)$ and $(\mathbb{C}, +, \cdot)$ are all fields.

Example 7.2.6 Let p be a prime number and a an arbitrary number between 1 and $p-1$. Since p is a prime number, the greatest common divisor of a and p is 1. By Proposition 7.1.15, we see that a is a unit of the ring the ring $(\mathbb{Z}_p, +_p, \cdot_p)$. This means that the ring $(\mathbb{Z}_p, +_p, \cdot_p)$ is a field with p elements. In order to make it clear from the notation that it is a field, one usually writes $\mathbb{F}_p$ instead of $\mathbb{Z}_p$. It is also common to abbreviate the notation and write $+$ an $\cdot$ for the operations $+_p$ and $\cdot_p$. Another commonly used notation for $\mathbb{Z}_p$ is $\mathrm{GF}(p)$. GF stands for Galois Field after the discoverer of finite fields Évariste Galois (1811-1832). He unfortunately died very young at the age of twenty from wounds suffered in a duel.

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: the ring $(\mathbb{Z}_n, +_n, \cdot_n)$ is a field if and only if n is a prime number. We will see later that there exist finite fields with p^e elements (p a prime number, $e > 1$ an integer), but these are constructed in a different way and **cannot** be identified with the ring $(\mathbb{Z}_{p^e}, +_{p^e}, \cdot_{p^e})$.

The reason fields are an important class of rings is that the theory of linear algebra works 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 7.2.6), 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.

7.3 Polynomials with coefficients in a commutative ring

Keywords: polynomials, degree and leading coefficient

In this section we define the ring of polynomials with coefficients in a given ring $(R, +, \cdot)$. We will in this section always assume that the ring $(R, +, \cdot)$ is commutative, that is to say that for all $r, s \in R$ we have $r \cdot s = s \cdot r$.

Definition 7.3.1 *Let $(R, +, \cdot)$ be a ring. A* polynomial $p(X)$ *with coefficients in R, is a formal expression of the form $p(X) = r_0 + r_1 X + \cdots + r_d X^d$, where d is a nonnegative integer and $r_0, \ldots, r_d \in R$. The ring elements $r_0, \ldots, r_d$ are called the coefficients of $p(X)$, while X is called the indeterminate of the polynomial. If $r_d \neq 0$, $\deg(p(X)) = d$ is called the* degree *of $p(X)$ and r_d its leading coefficient. If $r_d = 1$, the polynomial $p(X)$ is called* monic. *Finally, if $p(X) = 0$, the zero polynomial, we define* $\deg(0) = -\infty$, *minus infinity.*

The set of all polynomials with coefficients in R is denoted by $R[X]$. For $a \in R$ and $p(X) = r_0 + r_1 X + \cdots + r_d X^d \in R[X]$, we define the evaluation of $p(X)$ at a, notation $p(a)$, to be the element $p(a) = r_0 + r_1 a + \cdots + r_d a^d$ from R.

Example 7.3.2 Let $R = \mathbb{Z}_2$. Then the polynomials $p(X) = X$ and $q(X) = X^2$ have the same evaluation for all elements from the ring R. Indeed $p(0) = q(0) = 0$ and $p(1) = q(1) = 1$. Nonetheless, we consider them distinct as polynomials. Indeed, two nonzero polynomials $p(X) = r_0 + r_1 X + \cdots + r_d X^d$ and $q(x) = s_0 + s_1 X + \cdots + s_e X^e$ of degree d and e respectively, are equal if and only if $d = e$ and $r_i = s_i$ for all i between 0 and d. Another way of saying this is that a polynomial is uniquely determined by its coefficients.

Remark 7.3.3 *In some books, a polynomial is more formally defined as an infinite sequence* $(r_0, r_1, r_2, \dots)$ *of elements in* R *for which there exists a natural number* $d \in \mathbb{N}$ *such that* $r_n = 0$ *for all* $n > d$*. This has the advantage that the indeterminate* X *does not appear ad hoc as in our definition. Also from this more formal definition it is clear that two polynomials are equal if and only if they have the same coefficients.*

The set of all polynomials $R[X]$ can be given a ring structure by defining addition and multiplication of polynomials.

Definition 7.3.4 *Let* $p(X) = r_0 + r_1X + \cdots + r_dX^d$ *and* $q(X) = s_0 + s_1X + \cdots + s_eX^e$ *be two polynomials in* $R[X]$*. Further define* $r_i = 0$ *for* $i \geq d+1$ *and* $s_j = 0$ *for* $j \geq e+1$*. Then we define*

$$p(X) + q(X) = \sum_{\ell=0}^{\max\{e,d\}} (r_\ell + s_\ell)X^\ell \tag{7.1}$$

and

$$p(X) \cdot q(X) = \sum_{\ell=0}^{d+e} \left(\sum_{i=0}^{\ell} r_i s_{\ell-i} \right) X^\ell \tag{7.2}$$

The multiplication is exactly the one you would expect when multiplying into the parentheses and collecting equal powers of X. For example, $(r_iX^i)\cdot(s_jX^j) = (r_i \cdot s_j)X^{i+j}$ for any nonnegative integers i and j and elements $r_i, s_j \in R$. However, it is not always true that $\deg(p(X) \cdot q(X)) = \deg(p(X)) + \deg(q(X))$. For example, if $R = \mathbb{Z}_4$, then $2X \cdot (2X + 1) = 2X$, since $2 \cdot_4 2 = 0$.

With addition and multiplication of polynomials defined, we proceed to define the ring of polynomials with coefficients in R.

Definition 7.3.5 *Let* $(R, +, \cdot)$ *be a ring and let* $R[X]$ *be the set of all polynomials with coefficients in* R*. Then* $(R[X], +, \cdot)$ *equipped with the operations from equations* (7.1) *and* (7.2) *is a ring, called the polynomial ring in the indeterminate* X *and coefficients in* R*.*

We should spend some time on checking that $(R[X], +, \cdot)$ actually satisfies the ring axioms from Definition 7.1.1. First of all, we use the constant polynomials 0 and 1 as identity 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 $r(X) = r_0 + \cdots + r_dX^d$, $s(X) = s_0 + \cdots + s_eX^e$ and $t(X) = t_0 + \cdots + t_fX^f$, then starting with the ℓ-th coefficient of $(r(X) \cdot s(X)) \cdot t(X)$ and using the distributive and associative law in the ring $(R, +, \cdot)$, we find

$$\sum_{i=0}^{\ell} \left(\sum_{j=0}^{i} r_j \cdot s_{i-j} \right) \cdot t_{\ell-i} = \sum_{i=0}^{\ell} \sum_{j=0}^{i} (r_j \cdot s_{i-j}) \cdot t_{\ell-i} = \sum_{i=0}^{\ell} \sum_{j=0}^{i} r_j \cdot (s_{i-j} \cdot t_{\ell-i}).$$

Interchanging the order of summation, we then find

$$\sum_{i=0}^{\ell} \sum_{j=0}^{i} r_j \cdot (s_{i-j} \cdot t_{\ell-i}) = \sum_{j=0}^{\ell} \sum_{i=j}^{\ell} r_j \cdot (s_{i-j} \cdot t_{\ell-i}) = \sum_{j=0}^{\ell} r_j \sum_{i=j}^{\ell} s_{i-j} \cdot t_{\ell-i}$$

Now using a new summation index $k = i - j$, we obtain

$$\sum_{j=0}^{\ell} r_j \sum_{i=j}^{\ell} s_{i-j} \cdot t_{\ell-i} = \sum_{j=0}^{\ell} r_j \sum_{k=0}^{\ell-j} s_k \cdot t_{\ell-j-k},$$

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

Later on, we will especially consider polynomial rings with coefficients in a field or a domain. The following result should be seen in that light.

Theorem 7.3.6 *Suppose that $(D,+,\cdot)$ is a domain. Then the polynomial ring $(D[X],+,\cdot)$ is a domain as well.*

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

A consequence of the above theorem is the following:

Corollary 7.3.7 *Let $p(X), q(X)$ be two polynomials with coefficients in a domain. Then*

$$\deg(p(X)\cdot q(X)) = \deg p(X) + \deg q(X).$$

Proof. If either $p(X)$ or $q(X)$ is the zero polynomial, both sides become minus infinity. If both $p(X)$ and $q(X)$ are nonzero, the equality follows from the proof of Theorem 7.3.6. ■

Since any field is a domain (see Exercise 4), the conclusion of Theorem 7.3.6 that $(D[X],+,\cdot)$ is a domain is also true if $(D,+,\cdot)$ is a field. Therefore Corollary 7.3.7 is also true for polynomials with coefficients in a field.

7.4 Division with remainder for 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. The same statement turns out to be true for polynomials with coefficients in a domain. A key ingredient in the proof of that fact is the division with remainder algorithm. We start by studying to which extend such an algorithm works over a general commutative ring.

Theorem 7.4.1 *Let $(R,+,\cdot)$ be a commutative ring and let two polynomials $p(X)$ and $d(X)$ in $R[X]$, both different from zero, be given. Assume that the leading coefficient of $d(X)$ is a unit. Then there exist polynomials $q(X)$ and $r(X)$ such that:*

1. $p(X) = d(X) \cdot q(X) + r(X)$, *and*
2. $\deg r(X) < \deg d(X)$.

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

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

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

Then we can write

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

Note that $\deg p_1(X) < n$. Therefore we can conclude from the induction hypothesis that there exist polynomials $q_1(X)$ and $r(X)$ such that $p_1(X) = r(X) + d(X) \cdot q_1(X)$, with $r(X)$ either equal to zero, or $\deg r(X) < \deg d(X)$. In either case $\deg r(X) < \deg d(X)$, since $\deg(0) = -\infty$. Combining the above, we see that

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

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

Recall that we defined the degree of the zero polynomial to be minus infinity, so that if the remainder $r(X)$ happens to be zero, the condition $\deg r(X) < \deg d(X)$ automatically holds. The proof of the theorem gives rise to an algorithm called long division, or simply the division algorithm, to find the quotient $q(x)$ and remainder $r(x)$.

Example 7.4.2 In this example we give an example of long division of polynomials. We work over the field of rational numbers $(\mathbb{Q}, +, \cdot)$ and consider the polynomials $p(X) = 3X^3 + 2X + 1$ and as divisor $d(X) = X + 4$. The first step in the algorithm, following the proof of Theorem 7.4.1, is to multiply $d(X)$ with $3X^2$ and to subtract the result from $p(X)$. As a first remainder one obtains $-12X^2 + 2X + 1$. It is convenient to write this data in the following schematic way:

$$\begin{array}{l|l|l} X+4 & 3X^3 + 2X + 1 & 3X^2 \\ & \underline{3X^3 + 12X^2} & \\ & -12X^2 + 2X + 1 & \end{array}$$

The point is that the intermediate versions of $q(X)$ and $r(X)$ now are $3X^2$ and $-12X^2+2X+1$. Since the degree of the intermediate remainder is still larger than or equal to the degree of $d(X)$, which is 1, we continue. To lower the degree of the intermediate remainder further, we multiply $d(X)$ with $-12X$ and subtract the result. The schematic description is then updated as follows:

$$\begin{array}{r|l|l} X+4 & 3X^3 \quad + \quad 2X+1 & 3X^2-12X \\ & \underline{3X^3+12X^2} & \\ & \quad -12X^2+\ 2X+1 & \\ & \quad \underline{-12X^2-48X} & \\ & \qquad\qquad 50X+1 & \end{array}$$

The intermediate remainder is now $50X + 1$, which still has degree greater than or equal to $\deg(d(X))$. A further step is needed:

$$\begin{array}{r|l|l} X+4 & 3X^3 \quad + \quad 2X+\ \ 1 & 3X^2-12X+50 \\ & \underline{3X^3+12X^2} & \\ & \quad -12X^2+\ 2X+\ \ 1 & \\ & \quad \underline{-12X^2-48X} & \\ & \qquad\qquad 50X+\ \ 1 & \\ & \qquad\quad \underline{50X+200} & \\ & \qquad\qquad -199 & \end{array}$$

Now the algorithm terminates, since $\deg(-199) = 0 < \deg(d(X))$. We can now read the correct values of quotient and remainder, namely $q(X) = 3X^2 - 12X + 50$ and $r(X) = -199$. To check that the calculations were done correctly, one can verify that

$$3X^3 + 2X + 1 = (3X^2 - 12X + 50) \cdot (X + 4) - 199. \tag{7.3}$$

In general, when performing the division algorithm, one ends up with a schematic in the following form:

$$\begin{array}{r|l|l} d(X) & p(X) & q(X) \\ & \underline{\cdots} & \\ & \quad \ddots & \\ & \quad \overline{r(X)} & \end{array}$$

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). We simply say that $a \in R$ a *root* of $p(X)$ if $p(a) = 0$. In other words: $a \in R$ is a root of $p(X)$ if and only if $p(X)$ evaluates to 0 in a. The division algorithm on polynomials relates the notions of evaluation and root to a specific way to write a polynomial:

Proposition 7.4.3 *Let $(R, +, \cdot)$ be a commutative ring and $p(X) \in R[X]$ a non-zero polynomial of degree $n \geq 1$ and let $a \in R$. Then there exists a polynomial $q(X)$ of degree $n - 1$ such that*

$$p(X) = (X - a) \cdot q(X) + p(a). \tag{7.4}$$

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

Proof. Using the division algorithm on $p(X) \in R[X]$ and $X - a$, we can find polynomials $q(X)$ and a constant $r_0 \in R$ such that $p(X) = (X - a) \cdot q(X) + r_0$. 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 $\deg(X - a) \cdot q(X) = 1 + \deg q(X)$. This implies that $\deg q(X) = \deg p(X) - 1 = n - 1$. This proves 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 7.4 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 $p(X)$. ■

Example 7.4.2 illustrates equation (7.4). In that case $p(X) = 3X^3 + 2X + 1$ and $a = -4$. Indeed, the remainder -199 found in equation (7.3) is equal to $p(-4)$. A direct consequence of Proposition 7.4.3 is the following:

Corollary 7.4.4 *Let $(D, +, \cdot)$ be a domain and $p(X) \in D[X]$ a non-zero polynomial of degree n. Then $p(X)$ has at most n roots in D.*

Proof. We prove the corollary with induction on n. If $n = 0$, there is nothing to prove, since then $p = a_0$ for $a_0 \in D\backslash\{0\}$. Assume now that the theorem is true for polynomials of degree $n-1$. If $a \in D$ is a root of $p(X)$, we can write $p(X) = (X-a)\cdot q(X)$. Moreover, $\deg q(X) = n-1$. If $b \in D$ is another root of $p(X)$ distinct from a, then $q(b) \cdot (b - a) = p(b) = 0$. Since $a \neq b$ and D is a domain, the element $b - a$ is not a zero-divisor. Therefore we can conclude that $q(b) = 0$ or in other words that b is a root of $q(X)$. By the induction hypothesis, the polynomial $q(X)$ has at most $n - 1$ roots in D. This implies that $p(X)$ has at most n roots (namely a and the roots of $q(X)$). ■

Since any field is a domain, the corollary in particular holds if $(D, +, \cdot)$ is a field. The corollary is not true in general if $(D, +, \cdot)$ has zero-divisors. It is not hard to check for example that the polynomial $2X \in \mathbb{Z}_4[X]$ has 2 roots in $\mathbb{Z}_4$, namely 0 and 2.

7.5 Extra: the quaternions

In this section, we discuss a famous ring, namely the quaternions. It was already alluded to in Chapter 4, Aside 3.4.4, after the introduction of the quaternion group. A convenient way to introduce quaternions is using two by two matrices with coefficients in the complex numbers. Let us define the matrices $\mathbf{1}, \mathbf{i}, \mathbf{j}$, and $\mathbf{k}$ as follows:

$$\mathbf{1} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{i} := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \qquad \mathbf{j} := \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \qquad \mathbf{k} := \begin{pmatrix} i & 0 \\ 0 & -1 \end{pmatrix},$$

where as usual the complex number i satisfies $i^2 = -1$. It will also be convenient to denote by $\{\mathbf{0}\}$ the two by two zero matrix. A direct computation shows that all of the following relations are satisfied:

$$\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}, \tag{7.5}$$

$$\mathbf{ij} = \mathbf{k}, \qquad \mathbf{jk} = \mathbf{i}, \qquad \mathbf{ki} = \mathbf{j}, \tag{7.6}$$

and

$$\mathbf{ji} = -\mathbf{k}, \qquad \mathbf{kj} = -\mathbf{i}, \qquad \mathbf{ik} = -\mathbf{j}. \tag{7.7}$$

This shows that the set $\{\mathbf{1}, -\mathbf{1}, \mathbf{i}, -\mathbf{i}, \mathbf{j}, -\mathbf{j}, \mathbf{k}, -\mathbf{k}\}$ together with matrix multiplication is simply a way to describe the quaternion group $(Q_8, \cdot)$.

Now define

$$\mathbb{H} := \{a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k} \mid a, b, c, d \in \mathbb{R}\}.$$

The set $\mathbb{H}$ is called the set of *quaternions*. By definition, $\mathbb{H}$ is a subset of the set $M_{2\times 2}(\mathbb{C})$. In particular, we can add and multiply any two quaternions and in fact the result will always be another quaternion, since

$$(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) + (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) = (a_1 + a_2)\mathbf{1} + (b_1 + b_2)\mathbf{i} + (c_1 + c_2)\mathbf{j} + (d_1 + d_2)\mathbf{k}$$

and, using equations (7.5), (7.6), and (7.7)

$$\begin{aligned}(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k})(a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \quad = \quad & (a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2)\mathbf{1} \\ & +(a_1b_2 + b_1a_2 + c_1d_2 - d_1c_2)\mathbf{i} \\ & +(a_1c_2 - b_1d_2 + c_1a_2 + d_1b_2)\mathbf{j} \\ & +(a_1d_2 + b_1c_2 - c_1b_2 + d_1a_2)\mathbf{k}.\end{aligned}$$

Since matrix addition and multiplication satisfy the associative and distributive laws, we can conclude that $(\mathbb{H}, +, \cdot)$ is a ring called the ring of quaternions. It contains the field of complex numbers in a natural way as the set $\{a\mathbf{1} + b\mathbf{i} \mid a, b \in \mathbb{R}\}$. As for the complex numbers, or indeed any field, it is also true for the quaternions that any nonzero element has a multiplicative inverse. We state this as a lemma and subsequently prove it.

Lemma 7.5.1 *Let $(\mathbb{H}, +, \cdot)$ be the quaternion ring. Then $\mathbb{H}^* = \mathbb{H} \setminus \{\boldsymbol{0}\}$.*

Proof. Suppose that $(a, b, c, d) \in \mathbb{R}^4$. Then

$$(a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k})(a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}) = a^2 + b^2 + c^2 + d^2.$$

If $(a, b, c, d) \neq (0, 0, 0, 0)$, then $a^2 + b^2 + c^2 + d^2 > 0$ so that one can define the quaternion

$$A := \frac{a}{a^2 + b^2 + c^2 + d^2}\mathbf{1} - \frac{b}{a^2 + b^2 + c^2 + d^2}\mathbf{i} - \frac{c}{a^2 + b^2 + c^2 + d^2}\mathbf{j} - \frac{d}{a^2 + b^2 + c^2 + d^2}\mathbf{k}.$$

A direct computation then shows that $(a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) \cdot A = \mathbf{1} = A \cdot (a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k})$. ■

In analogy with the complex numbers, the quaternion $a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}$ is called the conjugate of the quaternion $a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$ and a quaternion for which the coefficient of $\mathbf{1}$ is zero is called a pure quaternion. Further $\sqrt{a^2 + b^2 + c^2 + d^2}$ is called the length of the quaternion $a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}$. The ring of quaternions is close to being a field, since any nonzero quaternion is a unit. However the quaternion ring is not commutative, as for example $\mathbf{ij} \neq \mathbf{ji}$. It is an example of a *division ring*, that is to say, a ring $(R, +, \cdot)$ such that $R^* = R \setminus \{0_R\}$. Another common terminology for a noncommutative division ring is *skew field*, so the ring of quaternions is an example of a skew field.

Nowadays, the ring of quaternions are used in the computer graphics industry, since they form a very convenient tool to describe 3D rotations. To illustrate this, let (p_x, p_y, p_z) be a vector in $\mathbb{R}_3$ and denote by R the counter clockwise rotation with angle θ over the rotation axis describe by a vector (a_x, a_y, a_z) of length one. Denoting by (p'_x, p'_y, p'_z) the image of (p_x, p_y, p_z) under R, one can show that

$$p'_x\mathbf{i} + p'_y\mathbf{j} + p'_z\mathbf{k} =$$
$$(\cos(\theta/2)\mathbf{1} + (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\sin(\theta/2))(p_x\mathbf{i} + p_y\mathbf{j} + p_z\mathbf{k})(\cos(\theta/2)\mathbf{1} - (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\sin(\theta/2)). \tag{7.8}$$

In other words, describing a vector $(p_x, p_y, p_z) \in \mathbb{R}_3$ as the pure quaternion $p_x\mathbf{i} + p_y\mathbf{j} + p_z\mathbf{k}$, we can compute the effect of a rotation by multiplying by a suitably chosen, but relatively simple, pair of conjugated quaternions. Let us check this formula in a simple case. If we choose $(a_x, a_y, a_z) = (0, 0, 1)$, then the corresponding rotation R maps (p_x, p_y, p_z) to $(p'_x, p'_y, p'_z) = (p_x\cos(\theta) - p_y\sin(\theta), p_x\sin(\theta) + p_y\cos(\theta), p_z)$. On the other hand, considering one term in $p_x\mathbf{i} + p_y\mathbf{j} + p_z\mathbf{k}$ at the time, we see that

$$\begin{aligned}
(\cos(\theta/2)\mathbf{1} + \mathbf{k}\sin(\theta/2))p_x\mathbf{i}(\cos(\theta/2)\mathbf{1} - \mathbf{k}\sin(\theta/2)) &= p_x(\cos(\theta/2)\mathbf{1} + \mathbf{k}\sin(\theta/2))\mathbf{i}(\cos(\theta/2)\mathbf{1} - \mathbf{k}\sin(\theta/2)) \\
&= p_x(\cos(\theta/2)\mathbf{1} + \mathbf{k}\sin(\theta/2))(\cos(\theta/2)\mathbf{k} + \mathbf{j}\sin(\theta/2)) \\
&= p_x((\cos^2(\theta/2) - \sin^2(\theta/2))\mathbf{i} + 2\cos(\theta/2)\sin(\theta/2)\mathbf{j}) \\
&= p_x(\cos(\theta)\mathbf{i} + \sin(\theta)\mathbf{j}),
\end{aligned}$$

where we used the doubling formulas for cosine and sine in the last equality. Similarly one obtains

$$(\cos(\theta/2)\mathbf{1} + \mathbf{k}\sin(\theta/2))p_y\mathbf{j}(\cos(\theta/2)\mathbf{1} - \mathbf{k}\sin(\theta/2)) = p_y(-\sin(\theta)\mathbf{i} + \cos(\theta)\mathbf{j}),$$

and

$$(\cos(\theta/2)\mathbf{1} + \mathbf{k}\sin(\theta/2))p_z\mathbf{k}(\cos(\theta/2)\mathbf{1} - \mathbf{k}\sin(\theta/2)) = p_z\mathbf{k}.$$

Adding these results together, we see that equation (7.8) indeed is correct in this case.

Note that the key in equation (7.8) is the quaternion $\cos(\theta/2)\mathbf{1} + (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\sin(\theta/2)$. Any such quaternion has length one and the "real part" $\cos(\theta/2)$ is strictly larger than -1, since $0 \leq \theta < 2\pi$. Conversely, any quaternion of length one distinct from $\pm\mathbf{1}$, can be written in that form. Indeed, if $a^2 + b^2 + c^2 + d^2 = 1$, there exist exactly one $\theta \in [0, 2\pi[$ such that $a = \cos(\theta/2)$, since $a \neq -1$. Then $b^2 + c^2 + d^2 = 1 - a^2 = 1 - \cos^2(\theta/2) = \sin^2(\theta/2)$, implies that either $\sin(\theta/2) = 0$, or that $(a_x, a_y, a_z) := (b, c, d)/\sin(\theta/2)$ has length one. Note that since $0 \leq \theta < 2\pi$, the case $\sin(\theta/2) = 0$ only if $\theta = 0$, in which case $\cos(\theta/2)\mathbf{1} + (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\sin(\theta/2) = \mathbf{1}$.

Since the product of two length one quaternions again has length one, this implies for example that the composite of any two rotations with rotation axes both passing through $(0, 0, 0)$ is either the identity, or again a rotation with axis passing through $(0, 0, 0)$. Indeed, such a product is either $\mathbf{1}$ or $-\mathbf{1}$ or in such a form that the recipe described above applies to write it in the form $\cos(\theta/2)\mathbf{1} + (a_x\mathbf{i} + a_y\mathbf{j} + a_z\mathbf{k})\sin(\theta/2)$. Note that the quaternions $\mathbf{1}$ and $-\mathbf{1}$ commute with any other quaternion, so that these cases correspond to the identity rotation (a rotation over angle 0).

As an illustration, consider the rotation R_1 over $2\pi/3$ with axis specified by the vector $(1, 1, 1)/\sqrt{3}$ and the rotation R_2 over $\pi/2$ with axis specified by $(1, 0, 0)$. Then the rotations R_1 and R_2 correspond to the quaternions $(\mathbf{1}+\mathbf{i}+\mathbf{j}+\mathbf{k})/2$ and $(\mathbf{1}+\mathbf{i})/\sqrt{2}$. Hence $R_1 \circ R_2$ corresponds to the quaternion

$$(\mathbf{1} + \mathbf{i} + \mathbf{j} + \mathbf{k})(\mathbf{1} + \mathbf{i})/(2\sqrt{2}) = (\mathbf{i} + \mathbf{j})/\sqrt{2} = \cos(\pi/2)\mathbf{1} + ((\sqrt{2}/2)\mathbf{i} + (\sqrt{2}/2)\mathbf{j} + 0\mathbf{k})\sin(\pi/2).$$

Apparently, $R_1 \circ R_2$ is the rotation over π with axis specified by the vector $(\sqrt{2}/2, \sqrt{2}/2, 0)$.

7.6 Exercises

Multiple choice exercises

1. Let $(R, +_R, \cdot_R)$ be a ring. Then
   A. $(R, +_R)$ is an abelian group with identity element 0_R
   TRUE ☐ FALSE ☐
   B. $(R, \cdot_R)$ is a group
   TRUE ☐ FALSE ☐
   C. $(R^*, \cdot_R)$ is a group
   TRUE ☐ FALSE ☐
   D. the identity element 1_R for $\cdot_R$ satisfies $x \cdot 1_R = 1_R \cdot x = x$ for all $x \in R$
   TRUE ☐ FALSE ☐
   E. the operation $\cdot_R$ is not associative
   TRUE ☐ FALSE ☐

2. Let $(R, +_R, \cdot_R)$ be a ring. An element $x \in R \setminus \{0_R\}$ is a *zero-divisor* if
   A. ☐ $x \cdot_R 0_R = 0_R$
   B. ☐ there exists $y \in R$ such that $x \cdot_R y = 0_R$ or $y \cdot_R x = 0_R$
   C. ☐ there exists $y \in R \setminus \{0_R\}$ such that $x \cdot_R y = 0_R$ or $y \cdot_R x = 0_R$
   D. ☐ there exists $y \in R \setminus \{0_R\}$ such that $x \cdot_R y = 1_R$

3. Let $(R, +_R, \cdot_R)$ be a ring. An element $x \in R$ is a *unit* if
   A. ☐ $x \cdot_R 1_R = x$
   B. ☐ there exists $y \in R$ such that $x \cdot_R y = 0_R$
   C. ☐ there exists $y \in R$ such that $x \cdot_R y = 1_R$
   D. ☐ there exists $y \in R$ such that $y \cdot_R x = x \cdot_R y = 1_R$

4. You are given a set R together with an addition $+_R$ and a multiplication $\cdot_R$. You are asked to decide whether $(R, +_R, \cdot_R)$ is a ring. Then
   A. you check the definition: $(R, +_R)$ is an abelian group, there exists an identity element for $\cdot_R$, the operation $\cdot_R$ is associative and the distributive laws hold. If all the properties are satisfied you can say that $(R, +_R, \cdot_R)$ is a ring
   TRUE ☐ FALSE ☐
   B. if for an element $x \in R \setminus \{0_R\}$ you cannot find $y \in R$ such that $x +_R y = 0_R$ then $(R, +_R, \cdot_R)$ can be a ring
   TRUE ☐ FALSE ☐

5. Let n be a positive integer and $a \in \mathbb{Z}_n$ with $a \neq 0$. Then a is a zero-divisor if and only if

A. ☐ $\gcd(n, a) > 1$

B. ☐ $\gcd(n, a) = 1$

6. Let n be a positive integer and $a \in \mathbb{Z}_n$ with $a \neq 0$. Then a is a unit if and only if

A. ☐ $\gcd(n, a) > 1$

B. ☐ $\gcd(n, a) = 1$

7. A commutative ring $(R, +_R, \cdot_R)$ is an *integral domain* if

A. ☐ $(R, \cdot_R)$ is a group

B. ☐ R does not contain units

C. ☐ the only zero-divisor in R is 0_R

D. ☐ R does not contain zero-divisors

8. Let $(R, +_R, \cdot_R)$ be a *field*. Then

A. $(R, +_R, \cdot_R)$ is not a commutative ring
TRUE ☐ FALSE ☐

B. $R \setminus \{0_R\} = R^*$
TRUE ☐ FALSE ☐

C. every $x \in R \setminus \{0_R\}$ admits a multiplicative inverse
TRUE ☐ FALSE ☐

D. $(R \setminus \{0_R\}, \cdot_R)$ is a group
TRUE ☐ FALSE ☐

9. Let $(R, +_R, \cdot_R)$ be a commutative ring, $a \in R$ and $p(X) \in R[X]$, with $p(X) \neq 0_R$. Then

A. if $p(a) = 0_R$ then a is a root of $p(X)$
TRUE ☐ FALSE ☐

B. if $(X - a)$ does not divide $p(X)$, a can be a root of $p(X)$
TRUE ☐ FALSE ☐

C. if there exists a polynomial $q(X) \in R[X]$ such that $p(X) = (X - a) \cdot_{R[X]} q(X)$ then a is a root of $p(X)$
TRUE ☐ FALSE ☐

D. if $R = \mathbb{Z}_5$ and $p(X) = X^3 - 1$ then 2 is a root of $P(X)$
TRUE ☐ FALSE ☐

E. if $R = \mathbb{Z}_7$ and $p(X) = X^6 - 1$ then 5 is a root of $P(X)$
TRUE ☐ FALSE ☐

Exercises to get to know the material better

10. Determine for the following examples: 1) if it 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.

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

- $(\{0_R\}, +_R, \cdot_R)$. The ring is called *the ring with one element.*

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

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

12. Let $(R, +_R, \cdot_R)$ be a ring. Define the *center* of R to be

$$Z(R) = \{a \in R \mid a \cdot_R r = r \cdot_R a \text{ for all } r \in R\}.$$

    Prove that $(Z(R), +_R, \cdot_R)$ is a commutative ring contained in R.

    [**Hint:** use the previous exercise to show that $Z(R)$ contains the zero element. Then recall that $(R, +_R, \cdot_R)$ satisfies the ring axioms]

13. 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.

14. Find $\lambda, \gamma \in \mathbb{Z}$ such that $13\lambda + 49\gamma = 1$, and show using this that 13 is a unit in $\mathbb{Z}_{49}$. Does $\mathbb{Z}_{49}$ contain any zero-divisor? If yes, give an example.

Exercises to get around in the theory

15. Let $(R, +_R, \cdot_R)$ be a ring with multiplicative identity element 1_R.

    - Show that for all $u \in R$, $-u = (-1_R) \cdot_R u$
      [**Hint:** check that the definition of being the additive inverse of u is satisfied by $(-1_R) \cdot_R u$]
    - show that if $u \in R$ is a unit, so is $-u$
    - prove that if $(R, +_R, \cdot_R)$ is commutative and $u \in R$ is a zero-divisor, then $u \cdot_R x$ is also a zero-divisor or 0_R, for all $x \in R$
    - prove that if $(R, +_R, \cdot_R)$ is an integral domain and $x \in R$ satisfies $x^2 = 1_R$ then $x = \pm 1_R$.

16. If in a ring $(R, +_R, \cdot_R)$ every $r \in R$ satisfies $r^2 = r$, prove that $(R, +_R, \cdot_R)$ must be commutative.

    [**Hint:** prove first that $-1_R = 1_R$. For $x, y \in R$, we need to show $x \cdot_R y = y \cdot_R x$. Compute $(x +_R y)^2$ in two different ways. What is the additive inverse of an element in this ring?]

    Remark 1: A ring $(R, +_R, \cdot_R)$ in which $r^2 = r$ for all $r \in R$ is called a *Boolean ring.* Boolean rings come from *logic.* Elements in a Boolean ring can be identified with logical propositions: the ring multiplication corresponds to conjunction or meet $\wedge$, and ring addition to exclusive disjunction or symmetric difference (not disjunction $\vee$).

    In fact given $x, y \in R$ we can define

$$x \wedge y := x \cdot_R y, \quad x \vee y := x +_R y +_R x \cdot_R y, \quad \neg x := 1_R +_R x.$$

Doing so, note that the condition $r^2 = r$ becomes $r \wedge r = r$ which is clearly true and logically correct.

Remark 2: The *characteristic* of a ring R, often denoted with $\text{char}(R) := n$, is defined to be the smallest number of times one must use the identity element 1_R in a sum to get the zero element 0_R, that is

$$\underbrace{1_R + \ldots + 1_R}_{n \text{ times}} = 0_R.$$

If this sum never reaches the additive identity 0_R the ring is said to have characteristic zero. An example of ring of characteristic zero is $\mathbb{Z}$. In this exercise you showed first $-1_R = 1_R$, that is, $1_R + 1_R = 0_R$ following the hint. Hence a reformulation of what you proved is: *a non-zero Boolean ring has characteristic* 2.

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

18. Let $R = \mathbb{Z}_6$. Find a polynomial $p(X) \in R[X]$ of degree 2 that has more than two roots in R.

19. Let n, a, b be positive integers such that $\gcd(a, n) = d > 1$ and $\gcd(b, d) = 1$. Show that the polynomial $aX - b \in \mathbb{Z}_n[X]$ has no roots in $\mathbb{Z}_n$.

20. This exercise is a generalization of Exercise 14. Let p be a prime number and $\ell > 1$ an integer.

    - Show that the ring $(\mathbb{Z}_{p^\ell}, +_{p^\ell}, \cdot_{p^\ell})$ is not a domain
      [**Hint:** use the characterization of zero-divisors in $\mathbb{Z}_n$ for $n = p^\ell$]
    - show that the number of non-units in $(\mathbb{Z}_{p^\ell}, +_{p^\ell}, \cdot_{p^\ell})$ is $p^{\ell-1}$
      [**Hint:** use the characterization of units in $\mathbb{Z}_n$ for $n = p^\ell$]

21. Let us define the set $\mathbb{Z}[\sqrt{2}] := \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$. Then $(\mathbb{Z}[\sqrt{2}], +, \cdot)$ is a ring, with $+$ and $\cdot$ the usual addition and multiplication of real numbers. 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 distinct units from $1 + \sqrt{2}$]

22. As in Example 139, let us define the set of Gaussian integers $\mathbb{Z}[i] := \{a + bi \mid a, b \in \mathbb{Z}\}$, where i is the imaginary unit in $\mathbb{C}$. Prove that $(\mathbb{Z}[i], +, \cdot)$ is a ring, where $+$ and $\cdot$ are the usual addition and multiplication of complex numbers. The aim of this exercise is to show that the *ring of Gaussian Integers* $(\mathbb{Z}[i], +, \cdot)$ is a domain but not a field.

    - Show that $(a + bi) \cdot (c + di) = e + fi$ implies that $(a^2 + b^2) \cdot (c^2 + d^2) = e^2 + f^2$
    - use this to show that $(\mathbb{Z}[i], +, \cdot)$ is a domain [**Hint:** use that $(\mathbb{Z}, +, \cdot)$ is a domain]
    - show that $(\mathbb{Z}[i])^* = \{1, -1, i, -i\}$.

23. * Let

$$\mathbf{1} := \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{i} := \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$$

$$\mathbf{j} := \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix} \qquad \mathbf{k} := \begin{pmatrix} i & 0 \\ 0 & -1 \end{pmatrix},$$

where $i^2 = -1$. Let $\mathbb{H}$ consist of all elements of the form $a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}$ where $a, b, c, d \in \mathbb{R}$.

- Prove that the following relations are satisfied: $\mathbf{i}^2 = \mathbf{j}^2 = \mathbf{k}^2 = \mathbf{ijk} = -\mathbf{1}$, $\mathbf{ij} = \mathbf{k}$, $\mathbf{jk} = \mathbf{i}$, $\mathbf{ki} = \mathbf{j}$, $\mathbf{ji} = -\mathbf{k}$, $\mathbf{kj} = -\mathbf{i}$ and $\mathbf{ik} = -\mathbf{j}$
- Using the previous part of the exercise and assuming the associativity and distributivity of matrix operations, show that $\mathbb{H}$ is a ring
- Show that $\mathbb{H}^* = \mathbb{H} \setminus \{\mathbf{0}\}$, where $\mathbf{0}$ denotes the zero matrix.
  [**Hint:** Prove and use that $(a\mathbf{1} + b\mathbf{i} + c\mathbf{j} + d\mathbf{k}) \cdot (a\mathbf{1} - b\mathbf{i} - c\mathbf{j} - d\mathbf{k}) = a^2 + b^2 + c^2 + d^2$]
  Remark 1: This ring is called *ring of quaternions*. It was discovered in 1843 by William R. Hamilton who wrote:
  *"I was walking with Lady Hamilton in Dublin, and came up to Brougham Bridge. I there felt the galvanic circuit of thought closed, and the sparks which fell from it were the fundamental equations between **i**, **j** and **k**, exactly as I have used them ever since. I pulled out, on the spot, a pocketbook, which still exists, and made an entry, on which, at the very moment, I felt that it might be worth my while to expend the labour of at least ten years to come."*
  This ring is famous because of the last property you proved: every element in $\mathbb{H}$ but $\mathbf{0} = 0_{\mathbb{H}}$ has a multiplicative inverse. Does it mean that $\mathbb{H}$ is a field? The answer is negative because $\mathbb{H}$ is not commutative (note that $\mathbf{ij} \neq \mathbf{ji}$). A ring $(R, +_R, \cdot_R)$ in which $R^* = R \setminus \{0_R\}$ is called a *division ring*. Hence, a field is a commutative division ring while $\mathbb{H}$ is an example of a division ring which is not a field.

Chapter 8

The theory of ideals

In Chapter 6 we introduced group homomorphisms, which culminated in the isomorphism theorem for groups. In this chapter, we will formulate the ring analogue of this isomorphism theorem. Fortunately, most of the hard work was already done in Chapter 6 and we only need to make some modifications to incorporate the ring multiplication. One important concept that arises from these modifications is the notion of an ideal of a ring. In this chapter, we will only consider commutative rings.

8.1 Ring homomorphisms and ideals

Keywords: ring homomorphism, kernel, ideal, principal ideal

In this section we will develop the theory of ideals of 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 and are used to construct quotient rings. Since we have seen that the kernel of a group homomorphism is always normal, we start by studying ring homomorphisms and their kernels.

Definition 8.1.1 *Let $(R, +_R, \cdot_R)$ and $(S, +_S, \cdot_S)$ be two rings. A function $\psi : R \to S$ is called a* ring homomorphism *or a* homomorphism of rings*, if it satisfies:*

1. *ψ is a group homomorphism between the groups $(R, +_R)$ and $(S, +_S)$. More explicitly: $\psi(0_R) = 0_S$ and $\psi(r +_R s) = \psi(r) +_S \psi(s)$ for all $r, s \in R$,*
2. *$\psi(1_R) = 1_S$,*
3. *for all $r, s \in R$ it holds that $\psi(r \cdot_R s) = \psi(r) \cdot_S \psi(s)$.*

If $\psi : R \to S$ is bijective as well (that is to say both injective and surjective), it is called a ring isomorphism *or an* isomorphism of rings*.*

One rich source of ring homomorphisms can be obtained from polynomial rings. Recall that for $a \in R$ and $p(X) = r_0 + r_1X + \cdots + r_dX^d \in R[X]$, we defined the evaluation of $p(X)$ at a, notation $p(a)$, to be the element $p(a) = r_0 + r_1a + \cdots + r_da^d$ from R. This can be generalized further. We say that a ring $(S, +, \cdot)$ contains R as a subring, if $R \subseteq S$, $0_R = 0_S$, $1_R = 1_S$, and the restriction of the ring operations $+$ and $\cdot$ from S to R make $(R, +, \cdot)$ into a ring. For example, we can see $\mathbb{Z}$ as a subring of $(\mathbb{R}, +, \cdot)$. In such a setting a polynomial in $R[X]$ can also be viewed as a polynomial in $S[X]$. This means in particular, that we can define the evaluation of $p(X)$ at a also for elements $a \in S$. This type of evaluation gives a whole class of ring homomorphisms as we show now:

Proposition 8.1.2 *Let R be a subring of the ring $(S, +, \cdot)$. For any $a \in S$, the map $\psi : R[X] \to S$ defined by $p(X) \mapsto p(a)$ is a ring homomorphism.*

Proof. First of all, we have $\psi(0_R) = 0_R = 0_S$ and $\psi(1_R) = 1_R = 1_S$, where the final equalities hold, since R is a subring of $(S, +, \cdot)$. Now let $p(X) = r_0 + r_1X + \cdots + r_dX^d$ and $q(X) = s_0 + s_1X + \cdots + s_eX^e$ be two polynomials in $R[X]$. Further define $r_i = 0$ for $i \geq d + 1$ and $\psi(0_R) = 0_S$ and $\psi(r +_R s) = \psi(r) +_S \psi(s)$ for all $r, s \in R$. Then

$$\begin{aligned}
\psi(p(X) + q(X)) &= \psi(\sum_{i\geq 0}(r_i + s_i)a^i) \\
&= \psi(\sum_{i\geq 0}(r_ia^i + s_ia^i)) \\
&= \psi(\sum_{i\geq 0} r_ia^i + \sum_{i\geq 0} s_ia^i) \\
&= \psi(p(X)) + \psi(q(X)).
\end{aligned}$$

The only thing left to be verified is that $\psi(p(X)\cdot q(X)) = \psi(p(X))\cdot\psi(q(X))$, which simply means that $p(a) \cdot q(a) = (p \cdot q)(a)$. This follows from equation 7.2:

$$\psi(p(X) \cdot q(X)) = \psi(\sum_{\ell=0}^{d+e}\left(\sum_{i\geq 0} r_is_{\ell-i}\right)X^\ell) = \sum_{\ell\geq 0}\left(\sum_{i=0}^{\ell} r_is_{\ell-i}\right)a^\ell,$$

while

$$\psi(p(X)) \cdot \psi(q(X)) = \left(\sum_{i\geq 0}^{d} r_ia^i\right) \cdot \left(\sum_{j\geq 0} s_ja^j\right) = \sum_{j\geq 0}\sum_{i\geq 0} r_is_ja^{i+j}.$$

After replacing the summation index j by $\ell = i + j$ in the final expression, one can conclude that $\psi(p(X) \cdot q(X)) = \psi(p(X)) \cdot \psi(q(X))$. ■

Example 8.1.3 Let $(S, +, \cdot) = (\mathbb{C}, +, \cdot)$ and $R = \mathbb{Z}$. Choosing $a = \sqrt{2}$, we obtain the ring homomorphism $\psi : \mathbb{Z}[X] \to \mathbb{C}$ defined by $\psi(p(X)) = p(\sqrt{2})$. Note that this is not a ring isomorphism, since it is not injective. Indeed $\psi(2) = 2$, but also $\psi(X^2) = 2$.

Like for group homomorphisms, one can define the kernel of a ring homomorphism:

Definition 8.1.4 *Let $\psi : R \to S$ be a ring homomorphism. Then we define* $\ker(\psi)$ *the kernel of ψ as follows:*

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

Example 8.1.5 In this example, we consider the function

$$\psi : \mathbb{Z} \to \mathbb{Z}_n, \text{ defined by } \psi(a) = a \bmod n.$$

We claim that his is a ring homomorphism. According to Definition 8.1.1, the function ψ needs to satisfy three properties. We check them one at the time:

1. In Example 6.1.12, we have seen that ψ is a group homomorphism from $(\mathbb{Z}, +)$ to $(\mathbb{Z}_n, +_n)$.
2. $\psi(1) = 1 \bmod n = 1$, so the second item is also satisfied.
3. For any $a, b \in \mathbb{Z}$ we have $a \cdot b \bmod n = (a \bmod n) \cdot_n (b \bmod n)$, since both right-hand and left-hand side are numbers between 0 and $n-1$ congruent to $a \cdot b$ modulo n. Hence $\psi(a \cdot b) = \psi(a) \cdot_n \psi(b)$.

The kernel of this ring homomorphism is $n\mathbb{Z}$, which we already saw in Example 6.1.12.

By Theorem 6.1.9, the kernel of a ring homomorphism is a normal subgroup of the group $(R, +_R)$. However, since $(R, +_R)$ is an abelian group, this amounts to saying that the kernel of a ring homomorphism is a subgroup of $(R, +_R)$. More can be said if the multiplication is taken into consideration:

Theorem 8.1.6 *Let $\psi : R \to S$ be a ring homomorphism. Then its kernel* $\ker(\psi)$ *satisfies the following properties:*

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

Proof. The first item was already remarked just before this theorem: since ψ also is a group homomorphism between the groups $(R, +_R)$ and $(S, +_S)$, its kernel is a subgroup of $(R, +_R)$. As for the second item: if $r \in R$ and $x \in \ker(\psi)$, then

$$\psi(r \cdot_R x) = \psi(r) \cdot_S \psi(x) = \psi(r) \cdot_S 0_S = 0_S.$$

Hence $r \cdot_R x \in \ker(\psi)$. ■

The following definition is inspired by the above-mentioned properties of a kernel of a ring homomorphism:

Definition 8.1.7 *Let $(R, +_R, \cdot_R)$ be a commutative ring. A subset $I \subseteq R$ is called an* ideal *if:*

1. *$I \subseteq R$ is a subgroup of the additive group $(R, +_R)$ of the ring, and*

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. In the literature it is customary to drop the suffix R when talking about addition $+_R$, multiplication $\cdot_R$, and the elements 0_R and 1_R. We will from now on do this as well, unless it is needed to use the suffix in order to avoid confusion.

Example 8.1.8 Let $(R, +, \cdot)$ be a commutative ring. Then the sets $\{0\}$ and R are ideals. They are often called the trivial ideals, because any ring has these ideals.

Example 8.1.9 In this example we determine all ideals of the ring $(\mathbb{Z}_6, +_6, \cdot_6)$. Let $I \subseteq \mathbb{Z}$ mod 6 be an ideal. Property 1 of Definition 8.1.7 implies that $I \subseteq \mathbb{Z}_6$ is a subgroup of $(\mathbb{Z}_6, +_6)$. By Lagrange's theorem (see Theorem 4.4.1), we conclude that $|I| \in \{1, 2, 3, 6\}$. In fact, all the subgroups of $(\mathbb{Z}_6, +_6)$ are: $\{0\}, \{0, 3\}, \{0, 2, 4\}$, and $\mathbb{Z}_6$ itself. It is not hard to show that these are in fact ideals of the ring $(\mathbb{Z}_6, +_6, \cdot_6)$ as well. Therefore the ring $(\mathbb{Z}_6, +_6, \cdot_6)$ has exactly 4 ideals.

A special type of ideals are the so-called principal ideals:

Definition 8.1.10 *Let $(R, +, \cdot)$ be a commutative ring and $x \in R$. Define*

$$\langle x \rangle := \{x \cdot r \mid r \in R\}.$$

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

Example 8.1.11 The ideals in Example 8.1.9 are all principal ideals. Indeed, $\{0\} = \langle 0 \rangle$, $\{0, 3\} = \langle 3 \rangle$, $\{0, 2, 4\} = \langle 2 \rangle$, and $\mathbb{Z}_6 = \langle 1 \rangle$. Apparently, any ideal in the ring $(\mathbb{Z}_6, +_6, \cdot_6)$ is a principal ideal.

Example 8.1.12 For the ring of integers $(\mathbb{Z}, +, \cdot)$, we have already encountered a principal ideal in another context. Indeed, the congruence class $n\mathbb{Z}$ from Definition 1.2.2 is simply the principal ideal of $\mathbb{Z}$ generated by n. The notation $n\mathbb{Z}$ for the principal ideal $\langle n \rangle$ is very common when dealing with the ring of integers.

The congruence classes of the form $a + n\mathbb{Z}$ with a an integer satisfying $1 \leq a \leq n - 1$, are not ideals, since they do not contain 0 and hence are not even subgroups of $(\mathbb{Z}, +)$. In fact, these congruence classes are cosets of the subgroup $n\mathbb{Z} \subseteq \mathbb{Z}$.

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

Lemma 8.1.13 *Let $(R, +, \cdot)$ 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 $I = \langle x_1, \ldots, x_n \rangle$. First we check that $I \subseteq R$ is a subgroup of the additive group $(R, +)$ 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 $\langle x_1, \cdots, x_n \rangle$. If $n = 1$, we recover the principal ideals from Definition 8.1.10. Ideals of the form $\langle x_1, \cdots, x_n \rangle$ are called finitely generated ideals. A ring for which each ideal is finitely generated is called a Noetherian ring in honour of the German mathematician Emmy Noether, about whom Albert Einstein wrote shortly after her death: "In the judgment of the most competent living mathematicians, Fräulein Noether was the most significant creative mathematical genius thus far produced since the higher education of women began."

A ring for which each ideal is a principal ideal, is called a principal ideal ring. As Example 8.1.11 shows, the ring $(\mathbb{Z}_6, +_6, \cdot_6)$ is a principal ideal ring.

Example 8.1.14 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 $I \subseteq R$ is an ideal in a commutative ring $(R, +_R, \cdot_R)$ then $1_R \in I$ if and only if $I = R$. Indeed $I \subseteq R$ always holds. Conversely $R \subseteq I$ if $1_R \in I$, since then for any $r \in R$ we have $r = r \cdot_R 1_R \in I$.

In Example 8.1.14, we saw the ideal of $\mathbb{Z}$ generated by 5 and 17, actually was the entire ring $\mathbb{Z}$. This can be generalized to the following:

Lemma 8.1.15 *Let $n, m \in \mathbb{Z}$ be two integers and let $\langle n, m \rangle \subseteq \mathbb{Z}$ be the ideal of the ring of integers generated by n and m. Then $\langle n, m \rangle = \langle \gcd(n, m) \rangle$.*

Proof. Since $\gcd(n, m)$ divides both n and m, any element from the ideal $\langle n, m \rangle$ is a multiple of $\gcd(n, m)$. This implies that $\langle n, m \rangle \subseteq \langle \gcd(n, m) \rangle$.

To show the converse inclusion, the key is the fact that for any two integers n and m, there exist two integers r and s such that $rn + sm = \gcd(n, m)$. For this implies that $\gcd(n, m)$ is an element of the ideal $\langle n, m \rangle$. However, if $\gcd(n, m) \in \langle n, m \rangle$, then any multiple of $\gcd(n, m)$ is in $\langle n, m \rangle$ as well. Therefore $\langle \gcd(n, m) \rangle \subseteq \langle n, m \rangle$. Combining the above, we see that $\langle n, m \rangle = \langle \gcd(n, m) \rangle$. ■

Since 5 and 17 are relatively prime, this lemma implies that $\langle 5, 17 \rangle = \langle 1 \rangle = \mathbb{Z}$, explaining the result from Example 8.1.14. As a matter of fact, it turns out that any ideal in $\mathbb{Z}$ is a principal ideal, but this will be shown in the next section.

8.2 Principal ideal domains

Keywords: principal ideal domain (PID)

In this section we will show that any ideal in the ring $(\mathbb{Z}, +, \cdot)$ is a principal ideal. If $(\mathbb{F}, +, \cdot)$ is a field, we will see that any ideal in the polynomial ring $(\mathbb{F}[X], +, \cdot)$ is a principal ideal as well. This means that if I is an ideal of either of these two rings, then there exists a ring element x such that $I = \langle x \rangle$. A ring $(R, +, \cdot)$ with this property is called a principal ideal ring. If moreover, the ring is a domain, such a ring is called a principal ideal domain (PID). Since we already know that the rings $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{F}[X], +, \cdot)$ are domains, we will therefore be able to conclude that they are PIDs. We start by studying the ring $(\mathbb{Z}, +, \cdot)$.

Proposition 8.2.1 *The ring $(\mathbb{Z}, +, \cdot)$ is a PID.*

Proof. Let I be an ideal of $\mathbb{Z}$. We claim that $I = \langle d \rangle$ 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 $I = \langle d \rangle$. Since $d \in I$

and I is an ideal, it is clear that $\langle d\rangle \subseteq I$. Conversely suppose that $n \in I$. Using division with remainder (see Fact 1.3.5), we may write $n = q\cdot d + r$ for integers q and r such that $0 \leq r \leq d-1$. Since $r = n - q\cdot d$ 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. This implies that $r = 0$, meaning that $n = q\cdot d$, is an element of $\langle d\rangle$. We have shown that $I \subseteq \langle d\rangle$. All in all we may conclude that $I = \langle d\rangle$. ■

Example 8.2.2 This is a continuation of Example 8.3.3. From the above proposition we see that any ideal I of $\mathbb{Z}$ is of the form $n\mathbb{Z}$. Moreover, if $a + n\mathbb{Z}$ is a coset of such an ideal, we may assume that $a \in \mathbb{Z}_n$, since $a + n\mathbb{Z} = (a \bmod n) + n\mathbb{Z}$ (see Theorem 1.3.7). If $a \in \mathbb{Z}_n$, we say that a is the standard representative of $a + n\mathbb{Z}$ and that $a + n\mathbb{Z}$ is given in standard form.

Given a concrete ideal, Proposition 8.2.1 does not explain how to find its generator. In a concrete situation, typically one considers an ideal generated by finitely many integers, say $I = \langle d_1, d_2, \ldots, d_\ell\rangle$. If the ideal is generated by two elements, say $I = \langle d_1, d_2\rangle$, Lemma 8.1.15 applies and the generator of I can be chosen as $\gcd(d_1, d_2)$. Now consider the ideal $I = \langle d_1, d_2, d_3\rangle \subseteq \mathbb{Z}$, generated by three integers $d_1, d_2, d_3 \in \mathbb{Z}$. Lemma 8.1.15 then implies that

$$I = \langle d_1, d_2, d_3\rangle = \langle \gcd(d_1, d_2), d_3\rangle = \langle \gcd(\gcd(d_1, d_2), d_3)\rangle = \langle \gcd(d_1, d_2, d_3)\rangle.$$

Inductively, one obtains that $\langle d_1, \ldots, d_\ell\rangle = \langle \gcd(d_1, \ldots, d_\ell)\rangle$.

A second standard example of a principal ideal domain will be studied now:

Proposition 8.2.3 *Let $(\mathbb{F}, +, \cdot)$ be a field. Then the polynomial ring $(\mathbb{F}[X], +, \cdot)$ is a PID.*

Proof. Let I be an ideal of $\mathbb{F}[X]$. As for the ring of integers, we claim that $I = \langle d(X)\rangle$ 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 $I = \langle d(X)\rangle$. The inclusion $\langle d(X)\rangle \subseteq I$ 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 $n(X) = q(X)\cdot d(X) + r(X)$ for polynomials $r(X)$ and $q(X)$ such that $\deg r(X) < \deg d(X)$. Since $r(X) = n(X) - q(X)\cdot d(X)$, 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 $n(X) \in \langle d(X)\rangle$, implying that $I \subseteq \langle d(X)\rangle$. Combining the above, we conclude that $I = \langle d(X)\rangle$. ■

Again the proof of this proposition does not tell one how to find the generator of an ideal, but a very similar procedure as in the case of the ring of integers works. As for integers, one can define a divisor and a greatest common divisor for polynomials with coefficients in a field $(\mathbb{F}, +, \cdot)$. A polynomial $d(X)$ is a *divisor* of a polynomial $p(X)$, if there exists a polynomial $q(X)$ such that $p(X) = q(X)\cdot d(X)$. Given two polynomials $p_1(X)$ and $p_2(X)$, we would like to define a greatest common divisor of $p_1(X)$ and $p_2(X)$. In this context "greatest" should be interpreted as "having largest degree". For example a greatest common divisor of the polynomials $X - 1$ and $(X-1)(X+1)$ is $X-1$. We wrote **a** greatest common divisor on purpose. A polynomial of the form $aX - a$ for any nonzero $a \in \mathbb{F}$ is also a choice for a greatest common divisor of $X - 1$ and $(X - 1)(X + 1)$, since $aX - a$ has the same degree as $X - 1$ and divides both $X - 1$ and $(X-1)(X+1)$ (we have $X - 1 = a^{-1}\cdot(aX - a)$ and $(X-1)(X+1) = a^{-1}(X+1)\cdot(aX - a)$). In other words: given a greatest common divisor $d(x)$ of two polynomials $p_1(X)$ and $p_2(X)$,

any nonzero scalar multiple of it, that is $a \cdot d(x)$ with $a \in \mathbb{F}^*$, is an equally valid choice as a greatest common divisor. Hence a greatest common divisor is only unique up to multiplication with a nonzero constant from the field $\mathbb{F}$. A commonly used way to get around this, is to say that a greatest common divisor of two polynomials has to be monic (leading coefficient equal to one). This convention makes greatest common divisors of polynomials uniquely determined. For example, one obtains that $\gcd((X-1)(X+1), X-1) = X-1$.

Apart from this minor detail, the theory of divisors and greatest common divisors of polynomials is very similar to the corresponding theory in case of integers. In particular, the extended Euclidean algorithm also works for polynomials with coefficients in a field. Also Lemma 8.1.15 has an analogue:

Lemma 8.2.4 *Let $f(x), g(x) \in \mathbb{F}[X]$ be two polynomials with coefficients in a field $\mathbb{F}$. Then $\langle f(x), g(X)\rangle = \langle \gcd(f(X), g(X))\rangle$.*

Proof. The proof is left as an exercise. See Exercise 14 ■

Aside 8.2.5 *If a ring $(S, +, \cdot)$ contains a field $\mathbb{F}$ as a subring, we can for any element $a \in S$ construct a ring homomorphism $\psi : \mathbb{F}[X] \to S$ as $\psi(p(X)) = p(a)$, see Proposition 8.1.2. If the kernel of ψ is not the zero ideal, the monic generator of the ideal* $\ker \psi$ *is called the* minimal polynomial *of a over $\mathbb{F}$. Consider for example $S = M_{n\times n}(\mathbb{C})$, the ring of n by n matrices with coefficients from the field of complex numbers. Then $\mathbb{C}$ can be identified with the subring of S consisting of matrices of the form $a \cdot I$, where I is the n by n identity matrix and $a \in \mathbb{C}$. Hence we can say in this way that $M_{n\times n}(\mathbb{C})$ contains $\mathbb{C}$ as a subring, or more properly a subring isomorphic to $(\mathbb{C}, +, \cdot)$. This allows us to so that we can define the minimal polynomial of a matrix, in this case over $\mathbb{C}$. One can for example show that matrix $A := \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$ has minimal polynomial $(X-1)^2$. It turns out that the minimal polynomial of a matrix contains a lot of information about the matrix. For example, the eigenvalues of a matrix over $\mathbb{C}$ are precisely the roots of its minimal polynomial. Moreover, a matrix is diagonalizable over $\mathbb{C}$ if and only if its minimal polynomial has no multiple roots (that is to say, roots with multiplicity 2 or more). In particular, the matrix A given above is not diagonalizable.*

It will be useful when considering quotient rings, to study the cosets of an ideal $\langle f(X)\rangle$ in $\mathbb{F}[X]$. We do this in the following lemma.

Lemma 8.2.6 *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 $I := \langle f(X)\rangle$ can uniquely be described in the form $r(X) + I$, with $\deg r(X) < \deg f(X)$.*

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 $r_1(X) + I = r_2(X) + I$ and that for both $i = 1$ and $i = 2$ it holds that $\deg r_i(X) < \deg f(X)$. Then $\deg(r_1(X) - r_2(X)) < \deg f(X)$. On the

other hand $r_1(X) + I = r_2(X) + I$ implies that $r_1(X) - r_2(X) \in I$, 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 either $q(X) = 0$ or $\deg(r_1(X) - r_2(X)) \geq \deg f(X)$, which would give a contradiction. Apparently $q(X) = 0$ and therefore $r_1(X) = r_2(X)$. ■

Given a coset of I, we say that the polynomial $r(X)$ from the above lemma is the standard representative of the coset and that $r(x) + I$ is its standard form.

Example 8.2.7 Let $R = \mathbb{R}[X]$ and $I := \langle X^2 + 1 \rangle$. Any coset of this particular ideal can be described as $a + bX + I$, with $a, b \in \mathbb{R}$. For example, consider the coset $X^3 + 2X^2 + 1 + \langle X^2 + 1 \rangle$. It is not in standard form, but we have $X^3 + 2X^2 + 1 = (X + 2)(X^2 + 1) + (-X - 1)$. Hence $X^3 + 2X^2 + 1 + \langle X^2 + 1 \rangle = -X - 1 + \langle X^2 + 1 \rangle$, which gives the coset in standard form. The polynomial $-X - 1$ is the standard representative of the coset $X^3 + 2X^2 + 1 + \langle X^2 + 1 \rangle$.

8.3 Quotient rings

Keywords: quotient rings

We will now use ideals to construct the ring-theoretical analogue of quotient groups: quotient rings. The situation is analogous to groups, where one uses normal subgroups to construct quotient groups, but now ideals are used instead. First of all, since an ideal I of the ring $(R, +, \cdot)$ is a subgroup of $(R, +)$, we can define cosets of I in R by r as:

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

Recall that any element r_1 in the coset $r + I$ is called a representative of that coset. Then we have $r_1 + I = r + I$, since cosets are either disjoint or identical and r_1 lies in both coset $r + I$ and $r_1 + I$. Also recall that $r + I = s + I$ if and only if $r - s \in I$. All this follows from Theorem 4.3.4 applied to the subgroup I of the group $(R, +)$. 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 cosets of an ideal I also can be multiplied in a meaningful way:

Lemma 8.3.1 *Let $I \subseteq R$ 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 (respectively s_1) is an arbitrary representative of the coset $r + I$ (respectively $s + I$), then

$$r \cdot s + I = r_1 \cdot s_1 + I.$$

Equivalently, we need to show that $(r \cdot s) - (r_1 \cdot s_1) \in I$. However, we have

$$(r \cdot s) - (r_1 \cdot s_1) = (r \cdot s) - (r \cdot s_1) + (r \cdot s_1) - (r_1 \cdot s_1) = r \cdot (s - s_1) + (r - r_1) \cdot s_1. \qquad (8.1)$$

Since $s + I = s_1 + I$, we have $s - s_1 \in I$ and since I is an ideal, we can deduce that $r \cdot (s - s_1) \in I$. Similarly, one obtains that $(r - r_1) \cdot s_1 = s_1 \cdot (r - r_1) \in I$. Therefore, equation (8.1) implies that $(r \cdot s) - (r_1 \cdot s_1) \in I$, which was what we wanted to show. ■

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

Theorem 8.3.2 *Let $I \subseteq R$ 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 one element $1 + I$.

Proof. We check that for example the first 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 $(R/I, +, \cdot)$, as well as commutativity, follow from the fact that they are satisfied in R. ■

Example 8.3.3 Consider the ring $(\mathbb{Z}, +, \cdot)$ and choose $I = \langle n \rangle = n\mathbb{Z}$, 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 $a + n\mathbb{Z}$ in standard form with $a \in \{0, 1, \ldots, n - 1\}$.

The ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ is in fact isomorphic to the ring $(\mathbb{Z}_n, +_n, \cdot_n)$. A ring isomorphism $\theta : \mathbb{Z}/n\mathbb{Z} \to \mathbb{Z}_n$ is obtained by mapping $a + n\mathbb{Z}$ to $(a \mod n) \in \mathbb{Z}_n$. We have already seen in Example 6.3.3 that this map defines an isomorphism of the groups $(\mathbb{Z}/n\mathbb{Z}, +)$ and $(\mathbb{Z}_n, +_n)$. Further $\theta(1 + n\mathbb{Z}) = 1$ and

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

Hence θ is an isomorphism of the rings $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ and $(\mathbb{Z}_n, +_n \cdot_n)$.

Example 8.3.4 The multiplication in the quotient ring $\mathbb{R}[X]/\langle X^2 + 1 \rangle$ can be described explicitly using the standard form of the cosets of $\langle X^2 + 1 \rangle$. For convenience, we write $I = \langle X^2 + 1 \rangle$ in this example. 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 other. We see that the quotient ring $\mathbb{R}[X]/\langle X^2 + 1\rangle$ is isomorphic to the field of complex numbers $\mathbb{C}$ by identifying $a + bX + \langle X^2 + 1\rangle$ and $a + bi$.

This means that the element $X + \langle X^2 + 1\rangle$ of the quotient ring $(\mathbb{R}[X]/\langle X^2 + 1\rangle, +, \cdot)$ should behave as a square root of minus one. Indeed, we have:

$$(X + \langle X^2 + 1\rangle)^2 = X^2 + \langle X^2 + 1\rangle = -1 + \langle X^2 + 1\rangle,$$

where the last equality holds, since $X^2 - (-1) \in \langle X^2 + 1\rangle$.

The ring isomorphisms found in the previous examples are special cases of a general result: the isomorphism theorem for rings. We state and prove it now:

Theorem 8.3.5 *Let $\psi : R \to S$ be a ring homomorphism between the rings $(R, +, \cdot)$ and $(S, +, \cdot)$. Then the map $\overline{\psi} : R/\ker(\psi) \to \mathrm{im}(\psi)$ defined by $\overline{\psi}(r + \ker(\psi)) = \psi(r)$ is a ring isomorphism.*

Proof. First let us check that $(\mathrm{im}(\psi), +, \cdot)$ is in fact a ring. What might go wrong is that $\mathrm{im}(\psi)$ is not closed under the ring operators $+$ and $\cdot$ or that the elements 0_S and 1_S are not contained in $\mathrm{im}(\psi)$. However, we already know that $\mathrm{im}(\psi)$ is a subgroup of $(S, +)$, hence we only need to check that $\mathrm{im}(\psi)$ contains 1_S and is closed under the multiplication $\cdot$. First of all $1_S \in \mathrm{im}(\psi)$, since $\psi(1_R) = 1_S$. Second, if $s_1, s_2 \in \mathrm{im}(\psi)$, then there exist $r_1, r_2 \in R$ such that $\psi(r_1) = s_1$ and $\psi(r_2) = s_2$. Then

$$s_1 \cdot s_2 = \psi(r_1) \cdot \psi(r_2) = \psi(r_1 \cdot r_2),$$

which implies that $s_1 \cdot s_2$ is an element of $\mathrm{im}(\psi)$. Hence $(\mathrm{im}(\psi), +, \cdot)$ is a ring.

The map $\overline{\psi} : R/\ker(\psi) \to \mathrm{im}(\psi)$ defined by $\overline{\psi}(r + \ker(\psi)) = \psi(r)$ is an isomorphism of the groups $(R/\ker(\psi), +)$ and $(\mathrm{im}(\psi), +)$. To show that it is a ring isomorphism, we only need to check items 2 and 3 from Definition 8.1.1. However, we have

$$\overline{\psi}(1_R + \ker(\psi) = \psi(1_R) = 1_S$$

and

$$\overline{\psi}((r + \ker(\psi)) \cdot (s + \ker(\psi))) = \overline{\psi}(r \cdot s + \ker(\psi)) = \psi(r \cdot s) = \psi(r) \cdot \psi(s) = \overline{\psi}(r + \ker(\psi)) \cdot \overline{\psi}(s + \ker(\psi)).$$

Hence $\overline{\psi}$ is in fact a ring isomorphism. ■

Example 8.3.6 This is a continuation of Example 8.3.3. Consider the ring homomorphism $\psi : \mathbb{Z} \to \mathbb{Z}_n$ defined by $\psi(a) = a \bmod n$. Then ψ is surjective, while $\ker(\psi) = n\mathbb{Z}$. Hence the isomorphism theorem of rings implies that the rings $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ and $(\mathbb{Z}_n, +_n, \cdot_n)$ are isomorphic. This explains the isomorphism found in Example 8.3.3.

If $n = p$, a prime number, then we have seen in Example 7.2.6 that $(\mathbb{Z}_p, +_p, \cdot_p)$ is a finite field with p elements, usually denoted by $(\mathbb{F}_p, +, \cdot)$. Since the rings $(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$ and $(\mathbb{Z}_p, +_p, \cdot_p)$ are isomorphic, this field can also be thought of as the quotient ring $(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$. One can show that any finite field with p elements is isomorphic to $(\mathbb{Z}_p, +_p, \cdot_p)$. Therefore one usually speaks about *the* finite field with p elements and, as we already mentioned, denotes it by $(\mathbb{F}_p, +, \cdot)$.

Example 8.3.7 This is a continuation of Example 8.3.4. Consider the ring $(\mathbb{R}[X], +, \cdot)$ of polynomials with coefficients in the real numbers and denote by $(\mathbb{C}, +, \cdot)$ the field of complex numbers. In particular, let $i \in \mathbb{C}$ denote the imaginary number with imaginary part 1. The map $\psi : \mathbb{R}[X] \to \mathbb{C}$ defined by $\psi(p(X)) = p(i)$ is a ring homomorphism by Proposition 8.1.2.

The image of ψ is all of $\mathbb{C}$, since for any $a, b \in \mathbb{R}$, we have $\psi(a+bX) = a+bi$. The kernel of ψ consists of all polynomials $p(X)$ with real coefficients having i as a root. But then $-i$ is a root of $p(X)$ as well, which implies that $X^2 + 1$ divides $p(X)$. Conversely, if $X^2 + 1$ divides $p(X)$, then $p(i) = 0$. Hence $\ker(\psi) = \langle X^2 + 1 \rangle$, the ideal of $\mathbb{R}[X]$ generated by $X^2 + 1$. The isomorphism theorem for rings implies that the ring $(\mathbb{R}[X]/\langle X^2 + 1 \rangle, +, \cdot)$ is isomorphic to the field $(\mathbb{C}, +, \cdot)$. The isomorphism is given by $\overline{\psi}(a + bX + \langle X^2 + 1 \rangle) = \psi(a + bX) = a + bi$. This is exactly the identification between the elements of $\mathbb{R}[X]/\langle X^2 + 1 \rangle$ and $\mathbb{C}$ that we found in Example 8.3.4.

8.4 The Euclidean algorithms for polynomials with coefficients in a field

It is assumed in these notes that the reader is familiar with the extended Euclidean algorithm for integers and has seen a variant for polynomials with coefficients in the field of rational numbers $(\mathbb{Q}, +, \cdot)$. However, for those readers that are not or that have forgotten the details, this section can be used as a crash course or as a brush up. For polynomials with coefficients in an arbitrary field, the algorithm works in exactly the same way.

Euclid's (extended) algorithm can also be formulated for polynomials. As for integers, the algorithm starts with a suitable 2×3 matrix and then performs certain row operations on this matrix. In case of integers, there were two possible operations: interchanging the rows or subtracting row two from row one. In the polynomial case we allow interchanging the rows and subtracting a *multiple* of row two from row one. In fact, we could have allowed this in the case of integers as well.

Let us as for example compute a greatest common divisor of the polynomials $p_1(X) = 2X^3 + 6X^2 - 8$ and $p_2(X) = X^2 + 2X$, as well as polynomials $r(X)$ and $s(X)$ such that $r(X)p_1(X) + s(X)p_2(X)$ equals this greatest common divisor. In this example, we work over the field $(\mathbb{Q}, +, \cdot)$ of rational numbers.

$$\begin{bmatrix} 2X^3 + 6X^2 - 8 & 1 & 0 \\ X^2 + 2X & 0 & 1 \end{bmatrix} \underset{R_1 \text{ - } 2X \cdot R_2}{\longrightarrow} \begin{bmatrix} 2X^2 - 8 & 1 & -2X \\ X^2 + 2X & 0 & 1 \end{bmatrix} \underset{R_1 \text{ - } 2 \cdot R_2}{\longrightarrow}$$

$$\left[\begin{array}{ccc} -4X-8 & 1 & -2X-2 \\ X^2+2X & 0 & 1 \end{array}\right] \begin{array}{c} \longrightarrow \\ R_2 \leftrightarrows R_1 \end{array} \left[\begin{array}{ccc} X^2+2X & 0 & 1 \\ -4X-8 & 1 & -2X-2 \end{array}\right] \begin{array}{c} \longrightarrow \\ R_1 + \frac{1}{4}X \cdot R_2 \end{array}$$

$$\left[\begin{array}{ccc} 0 & \frac{1}{4}X & -\frac{1}{2}X^2-\frac{1}{2}X+1 \\ -4X-8 & 1 & -2X-2 \end{array}\right] \begin{array}{c} \longrightarrow \\ R_2 \leftrightarrows R_1 \end{array} \left[\begin{array}{ccc} -4X-8 & 1 & -2X-2 \\ 0 & \frac{1}{4}X & -\frac{1}{2}X^2-\frac{1}{2}X+1 \end{array}\right]$$

We see that $-4X-8$ is a valid choice for a greatest common divisor of $2X^3+6X^2-8$ and X^2+2X, but using the convention that greatest common divisors have to be monic polynomials, we obtain that $\gcd(2X^3+6X^2-8, X^2+2X) = X+2$. Note that

$$1 \cdot (2X^3+6X^2-8) + (-2X-2) \cdot (X^2+2X) = -4X-8,$$

and hence

$$\frac{-1}{4} \cdot (2X^3+6X^2-8) + \frac{X+1}{2} \cdot (X^2+2X) = X+2.$$

Hence the sought polynomials $r(X)$ and $s(X)$ are $r(X) = -1/4$ and $s(X) = (X+1)/2$.

Remark 8.4.1 *The point of allowing the more general row operation, which allowed us to subtract a multiple of row two from row one, is that this multiple can be chosen such that each time the degree of the polynomial in the first coordinate of the first row drops. Moreover, the allowed multiples of row two, are always of the form aX^m times row two. This means that the number of operations of this row operation is of order of magnitude n, the maximum of the degrees of the input polynomials. This means that the total complexity of this version of the extended Euclidean algorithm in terms of number of operations in the field $(\mathbb{F}, +, \cdot)$, is $\mathcal{O}(n^2)$.*

8.5 Extra: the Chinese remainder theorem for rings.

Just as for groups, it is straightforward to define the direct product of two rings. If $(R, +_R, \cdot_R)$ and $(S, +_S, \cdot_S)$ are two rings, then one can define an addition $+_R \times +_S$ and multiplication $\cdot_R \times \cdot_S$ on $R \times S$ in a straightforward way: $(r_1, s_1) +_R \times +_S (r_2, s_2) = (r_1 +_R r_2, s_1 +_S s_2)$ and $(r_1, s_1) \cdot_R \times \cdot_S (r_2, s_2) = (r_1 \cdot_R r_2, s_1 \cdot_S s_2)$. This gives $R \times S$ the structure of a ring called the direct product ring of R and S. In Theorem 1.6.3, we have seen that if n_1 and n_2 are positive integers such that $\gcd(n_1, n_2) = 1$ and $n = n_1 \cdot n_2$, then the map $\varphi : \mathbb{Z}_n \to \mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$ defined by $\varphi(a) := (a \bmod n_1, a \bmod n_2)$ has an inverse. Actually the map φ is a ring homomorphism if we interpret $\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}$ as the direct product ring of $\mathbb{Z}_{n_1}$ and $\mathbb{Z}_{n_2}$. Theorem 1.6.3 can therefore we reformulated by stating that the rings $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}_{n_1} \times \mathbb{Z}_{n_2}, +_{n_1} \times +_{n_2}, \cdot_{n_1} \times_{n_2})$ are isomorphic. Since for any positive integer n the rings $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}/\langle n \rangle, +, \cdot)$ are isomorphic, we can also say that $\mathbb{Z}/\langle n \rangle$ and the direct product of $\mathbb{Z}/\langle n_1 \rangle$ and $\mathbb{Z}/\langle n_2 \rangle$ are isomorphic.

This statement has a very elegant generalization in ring theory called the Chinese remainder theory for rings. It is easiest to formulate it using the notion of coprime ideals: two ideals I and J of a ring $(R, +, \cdot)$ are called coprime is there exist $i \in I$ and $j \in J$ such that $i + j = 1$.

Theorem 8.5.1 *Let $(R, +, \cdot)$ be a commutative ring and suppose that I and J are two coprime ideals of R. Then the map*

$$\varphi : R/(I \cap J) \to R/I \times R/J$$
$$r + (I \cap J) \mapsto (r + I, r + J)$$

is an isomorphism of rings.

Proof. We start by considering the ring homomorphism $\psi : R \to R/I \times R/J$ defined by $\psi(r) = (r + I, r + J)$. Checking that this is a ring homomorphism can be done by writing out the definitions and is left to the reader. First of all, we claim that ψ is surjective. Indeed, let two cosets $r_1 + I$ and $r_2 + J$ be given. Using that there exist $i \in I$ and $j \in J$ such that $i + j = 1$, we obtain that

$$r_1 j + r_2 i + I = r_1 j + I = r_1(1 - i) + I = r_1 - r_1 i + I = r_1 + I$$

and

$$r_1 j + r_2 i + J = r_2 i + J = r_2(1 - j) + J = r_2 - r_2 j + I = r_2 + J.$$

Therefore $\psi(r_1 j + r_2 i) = (r_1 + I, r_2 + J)$, showing that ψ is surjective.

The kernel of ψ is given by

$$\ker \psi = \{r \in R \mid r + I = 0 + I \text{ and } r + J = 0 + J\} = \{r \in R \mid r \in I \text{ and } r \in J\} = I \cap J.$$

The result now follows from the isomorphism theorem for rings, see Theorem 8.3.5. ■

There exist a version of the Chinese remainder theorem where not just two ideals, but several ideals are considered. One then considers ideals $I_1, \ldots, I_n$ such that any two of them are coprime. In this case, one can show that the map

$$\varphi : R/(\bigcap_{i=1}^{n} I_i) \to R/I_1 \times R/I_n$$
$$r + \bigcap_{i=1}^{n} I_i \mapsto (r + I_1, \ldots, r + I_n)$$

is an isomorphism of rings.

We have previously seen in Proposition 7.4.3 that for any element $a \in \mathbb{F}$, where $(\mathbb{F}, +, \cdot)$ is a field, and any polynomial $p(X) \in \mathbb{F}[X]$, there exists a polynomial $q(X) \in \mathbb{F}[X]$ such that $p(X) = q(X)(X - a) + p(a)$. In particular this implies that $p(X) + \langle X - a \rangle = p(a) + \langle X - a \rangle$. In particular, $p(a)$ is the standard representative of the coset $p(X) + \langle X - a \rangle$. In this sense, the "evaluation in a map" $\psi : \mathbb{F}[X] \to \mathbb{F}$ defined by $\psi(p(X)) := p(a)$, is essentially the same as the map $\phi : \mathbb{F}[X] \to \mathbb{F}[X]/\langle X - a \rangle$ sending $p(X)$ to $p(X) + \langle X - a \rangle$, since $p(a)$ and $p(X)$ give rise to the same coset.

In some settings, it is needed to evaluate a given polynomial $p(X) \in \mathbb{F}[X]$ in n distinct values of $\mathbb{F}$, say in distinct $a_1, \ldots, a_n \in \mathbb{F}$. In other words, we are interested in the map $\phi^{(n)} : \mathbb{F}[X] \to \mathbb{F}^n$, defined by $\phi^{(n)}(p(X)) := (p(a_1), \ldots, p(a_n))$. The problem of computing this map is sometimes called *multipoint evaluation*. Like for evaluation in one point, we can alternatively think about

multipoint evaluation in $a_1, \ldots, a_n$ using the map $\psi^{(n)} : \mathbb{F}[X] \to \mathbb{F}[X]/\langle X - a_1\rangle \times \cdots \times \mathbb{F}[X]/\langle X - a_n\rangle$ sending a polynomial $p(X)$ to $(p(X) + \langle X - a_1\rangle, \ldots, p(X) + \langle X - a_n\rangle)$.

The kernel of the ring homomorphism $\psi^{(n)}$ consists exactly of those polynomials satisfying $p(a_1) = \cdots = p(a_n) = 0$. Since all the a_i are assumed to be distinct, this implies that any polynomial in the kernel is a multiple of $\prod_{i=1}^{n}(X - a_i)$. In other words: $\ker(\psi^{(n)}) = \langle \prod_{i=1}^{n}(X - a_i)\rangle$. It is not so hard to see that $\psi^{(n)}$ is surjective. Any coset of $\langle X - a_i\rangle$ in standard form is of the form $r_i + \langle X - a_i\rangle$ for some $r_i \in \mathbb{F}$. Now given any $r_1, \ldots, r_n \in \mathbb{F}$, the polynomial

$$\sum_{i=0}^{n} r_i \prod_{\substack{j=1 \\ j\neq i}}^{n} \frac{X - a_j}{a_i - a_j}$$

is mapped to $(r_1 + \langle X - a_1\rangle, \ldots, r_n + \langle X - a_n\rangle)$ under $\psi^{(n)}$. The isomorphism theorem for rings then implies that the rings $\mathbb{F}[X]/\langle \prod_{i=1}^{n}(X - a_i)\rangle$ and $\mathbb{F}[X]/\langle X - a_1\rangle \times \cdots \times \mathbb{F}[X]/\langle X - a_n\rangle$ are isomorphic. Since $\bigcap_{i=1}^{n}\langle X - a_i\rangle = \langle \prod_{i=1}^{n}(X - a_i)\rangle$, this is simply a special case of the Chinese remainder theorem.

Multipoint evaluation is a computationally very interesting problem. Evaluating a polynomial $p(X)$ of degree up to $n - 1$ in n values $a_1, \ldots, a_n$ can of course be done by evaluating $p(X)$ in one value a_i at the time. This would result in an algorithm needing $\mathcal{O}(n^2)$ operations in the field $\mathbb{F}$. Using fast algorithms to carry out division with remainder in combination with a clever use of the Chinese remainder theorem, this can be improved to $\mathcal{O}(n \log^2(n) \log\log(n))$. The well-known interpolation problem is to compute the inverse of this isomorphism: given $r_1, \ldots, r_n \in \mathbb{F}$, compute a polynomial $p(X) \in \mathbb{F}[X]$ of degree at most $n - 1$ such that for all i, $p(a_i) = r_i$. Also interpolation can be done using $\mathcal{O}(n \log^2(n) \log\log(n))$ operations in $\mathbb{F}$.

To illustrate how fast multipoint evaluation and the Chinese remainder theory are connected, let us look at a small example. As field, let us choose $\mathbb{F}_{11}$ and let us choose $a_i = i$ for $i = 1, \ldots, 8$. Now suppose that we wish to evaluate $p(X) = X^7 + X + 1$ in $a_1, \ldots, a_8$. The idea is to not compute the remainder of $p(X)$ modulo the $X - a_i$ individually right away, but to bundle these polynomials together in groups. More precisely, we define

$$m_1(X) := (X - 1)(X - 2)(X - 3)(X - 4) = X^4 + X^3 + 2X^2 + 5X + 2$$

and

$$m_2(X) := (X - 5)(X - 6)(X - 7)(X - 8) = X^4 + 7X^3 + 9X^2 + X + 8.$$

Then the remainders $r_1(X)$ and $r_2(X)$ when dividing $p(X)$ by $m_1(X)$ and $m_2(X)$ are

$$r_1(X) = 7X^3 + 2X + 5 \text{ and } r_2(X) = 10X^3 + 9X^2 + 9X + 7.$$

The point of this is that $r_1(X)$ is a first step on the way towards computing the remainder of $p(X)$ when dividing by $X - a_i$ for $i = 1, \ldots, 4$, while similarly $r_2(X)$ is that for $i = 4, \ldots, 8$. Now for the next step, we define

$$m_3(X) := (X - 1)(X - 2) = X^2 + 8X + 2, \quad m_4(X) := (X - 3)(X - 4) = X^2 + 4X + 1,$$

$$m_5(X) := (X - 5)(X - 6) = X^2 + 8, \quad \text{and} \quad m_6(X) := (X - 7)(X - 8) = X^2 + 7X + 1.$$

and compute the remainders $r_3(X), r_4(X), r_5(X)$, and $r_6(X)$ of $p(X)$ with respect to these polynomials. Note though that since $m_3(X)$ and $m_4(X)$ divide $m_1(X)$, we do not have to start from

scratch, but simply can compute the remainder of $r_1(X)$ modulo $m_3(X)$ and $m_4(X)$. Similarly, to compute the remainder of $p(X)$ modulo $m_5(X)$ and $m_6(X)$, we simply compute the remainder of $r_2(X)$ modulo $m_5(X)$ and $m_6(X)$. This observation is the key to why this approach is computationally fast. The result is:

$$r_3(X) = 7X + 7, \quad r_4(X) = 8X, \quad r_5(X) = 6X + 1, \quad \text{and} \quad r_6(X) = 8X + 2.$$

In the last step we compute the remainder $r_7(X), \ldots, r_{14}(X)$ of $r_3(X)$ modulo $X-1$ and $X-2$, of $r_4(X)$ modulo $X-3$ and $X-4$, of $r_5(X)$ modulo $X-5$ and $X-6$, and of $r_6(X)$ modulo $X-7$ and $X-8$. There remainders are precisely the evaluations of $p(X)$ in $a_1, \ldots, a_8$.

$$r_7(X) = 3, \quad r_8(X) = 10, \quad r_9(X) = 2, \quad r_{10}(X) = 10,$$

$$r_{11}(X) = 9, \quad r_{12}(X) = 4, \quad r_{13}(X) = 3 \quad \text{and} \quad r_{14}(X) = 0.$$

Ring theoretically, the situation is quite simple. Rather than computing the map

$$\psi^{(n)} : \mathbb{F}_{11}[X] \to \mathbb{F}_{11}[X]/\langle X-1\rangle \times \cdots \times \mathbb{F}[X]/\langle X-8\rangle$$

directly one coordinate at the time, we have split $\phi^{(n)}$ up into three maps:

$$\begin{aligned} \mathbb{F}_{11}[X] &\to \mathbb{F}_{11}[X]/\langle m_1(X)\rangle \times \mathbb{F}_{11}[X]/\langle m_2(X)\rangle \\ &\to \mathbb{F}_{11}[X]/\langle m_3(X)\rangle \times \mathbb{F}_{11}[X]/\langle m_4(X)\rangle \times \mathbb{F}_{11}[X]/\langle m_5(X)\rangle \times \mathbb{F}_{11}[X]/\langle m_6(X)\rangle \\ &\to \mathbb{F}_{11}[X]/\langle X-1\rangle \times \cdots \times \mathbb{F}[X]/\langle X-8\rangle \end{aligned}$$

The Chinese remainder theorem implies that this is mathematically sound.

8.6 Exercises

Multiple choice exercises

1. Let $\phi : R \to S$ be a ring homomorphism. Assume that $(S, +_S, \cdot_S)$ is not the zero ring. Then

   A. $\phi(0_R) = 1_S$
   TRUE ☐ FALSE ☐

   B. $\phi(0_R) = 0_S$
   TRUE ☐ FALSE ☐

   C. ϕ is not necessarily a group homomorphism between $(R, +_R)$ and $(S, +_S)$
   TRUE ☐ FALSE ☐

   D. $\phi(1_R) \in S^*$
   TRUE ☐ FALSE ☐

   E. for all $r, s \in R$ it holds $\phi(r \cdot_R s) \neq \phi(r) \cdot_S \phi(s)$
   TRUE ☐ FALSE ☐

   F. there cannot exist $r \in R$ such that $\phi(r) \cdot_S \phi(1_R) \neq \phi(r)$
   TRUE ☐ FALSE ☐

G. $\ker(\phi)$ is the set of elements $r \in R$ such that $\phi(r) = 1_S$
TRUE ☐ FALSE ☐

H. $\ker(\phi)$ is an ideal of R
TRUE ☐ FALSE ☐

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

A. ☐ $r +_R x \in I$ for all $r \in R$ and $x \in I$

B. ☐ $(I, \cdot_R)$ is a group

C. ☐ $(I, +_R)$ is an additive group and $r \cdot_R x \in I$ for all $r \in R$ and $x \in I$

3. Let $(R, +_R, \cdot_R)$ be a commutative ring and I an ideal

A. it can happen that $I = R$ or $I = \{0_R\}$
TRUE ☐ FALSE ☐

B. $1_R \in I$ if and only if $I = R$
TRUE ☐ FALSE ☐

C. if $I = \langle x \rangle$ for some $x \in R$ it means that I consists exactly of all multiples $x \cdot_R r$ with $r \in R$
TRUE ☐ FALSE ☐

D. if $n > 1$ and $I = \langle x_1, \ldots, x_n \rangle$ for some $x_1, \ldots, x_n \in R$ then I consists exactly of all multiples $x_i \cdot_R r$ with $r \in R$ and $i = 1, \ldots, n$
TRUE ☐ FALSE ☐

E. if $I = \langle 71, 11 \rangle$ and $R = \mathbb{Z}$ then $I = \mathbb{Z}$
TRUE ☐ FALSE ☐

4. A *principal ideal ring* is

A. ☐ a commutative ring in which every ideal can be generated by exactly one element

B. ☐ a commutative ring in which every ideal can be generated by one or more element

C. ☐ a commutative ring admitting just the trivial ideals (i.e. $\{0_R\}$ and R)

5. You are given a commutative ring $(R, +_R, \cdot_R)$ and two polynomials $f(X), g(X) \in R[X]$ where $f(X)$ is non-constant of degree d. To compute the *standard form* of $g(X) + \langle f(X) \rangle \in R[X]/\langle f(X) \rangle$ one can

A. ☐ compute the division with remainder $g(X) = q(X)f(X) + r(X)$ where either $r(X) = 0$ or $\deg r(X) < d$. The standard form is then $q(X) + \langle f(X) \rangle$

B. ☐ compute the division with remainder $g(X) = q(X)f(X) + r(X)$ where either $r(X) = 0$ or $\deg r(X) < d$. The standard form is then $r(X) + \langle f(X) \rangle$

C. ☐ compute the division with remainder $f(X) = q(X)g(X) + r(X)$ where either $r(X) = 0$ or $\deg r(X) < \deg g(X)$. The standard form is then $r(X) + \langle f(X) \rangle$

Exercises to get to know the material better

6. Is $\{0,1,2,4\} \subset \mathbb{Z}_8$ an ideal of the ring $(\mathbb{Z}_8, +_8, \cdot_8)$? Compute all the ideals of $(\mathbb{Z}_8, +_8, \cdot_8)$.

7. Let R be the set of upper triangular matrices

$$R = \left\{ \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} \mid a, b, c \in \mathbb{Z} \right\}.$$

Show that R is a ring with usual addition and multiplication of matrices. Prove also that the map $\phi : R \to \mathbb{Z} \times \mathbb{Z}$ defined by

$$\phi \begin{pmatrix} a & b \\ 0 & c \end{pmatrix} = (a, c),$$

is a surjective homomorphism of rings and describe its kernel.

8. Let $X^4 - 16 \in \mathbb{Z}[X]$ and let R be the quotient ring $R = \mathbb{Z}[X]/\langle X^4 - 16 \rangle$.

   - Compute the division with remainder of the polynomials $7X^{13} - 11X^9 + 5X^5 - 2X^3 + 3$ and $X^4 - 16$
   - use the previous computation to obtain the standard form of $7X^{13} - 11X^9 + 5X^5 - 2X^3 + 3 + \langle X^4 - 16 \rangle \in R$
   - prove that $X + 2 + \langle X^4 - 16 \rangle$ and $X - 2 + \langle X^4 - 16 \rangle$ are zero-divisors in R.
     [**Hint:** prove that $X^{14} - 16$ is a multiple of both $X + 2$ and $X - 2$ and write $X^{14} - 6 = (X + 2) \cdot p_1(X)$ and $X^{14} - 6 = (X - 2) \cdot p_2(X)$. Recall that $X^{14} - 16 + \langle X^{14} - 16 \rangle = 0_R$]

   Remark: the reason why $X + 2 + \langle X^4 - 16 \rangle$ and $X - 2 + \langle X^4 - 16 \rangle$ are zero-divisors in $\mathbb{Z}[X]/\langle X^4 - 16 \rangle$ will be formalized in the next chapter (see Proposition 9.1.2).

9. Compute the standard form of the coset

$$3X^3 + 5X^2 + 4X + 2 + \langle 3X^2 + 1 \rangle$$

in $\mathbb{Z}_7[X]$.

10. Let I and J be two ideals of a commutative ring $(R, +_R, \cdot_R)$. Show that the intersection $I \cap J$ also is an ideal of $(R, +_R, \cdot_R)$. Also, find an example of a ring R and ideals I and J such that $I \cup J$ is not an ideal of $(R, +_R, \cdot_R)$.

    [**Hint:** think about which property of the ones required to be an ideal can fail in $I \cup J$]

11. Let I be an ideal of a ring $(R, +_R, \cdot_R)$ and let $[I : R] = \{x \in R \mid r \cdot_R x \in I \text{ for all } r \in R\}$. Prove that $[I : R]$ is an ideal of $(R, +_R, \cdot_R)$ containing I.

12. If I, J are ideals of a ring $(R, +_R, \cdot_R)$, show that also $I + J := \{u +_R v \mid u \in I,\ v \in J\}$ is an ideal of $(R, +_R, \cdot_R)$.

13. Using Lemmas 8.1.15 and 8.2.4, find a generator for the following (principal) ideals:

    - $\langle 12, 20 \rangle$ in $\mathbb{Z}$
    - $\langle X^2 - 1, X^3 - 1 \rangle$ in $\mathbb{Q}[X]$. As usual, $\mathbb{Q}$ denotes the set of rational numbers
    - $\langle X^2 + 1, X^5 + X + 1 \rangle$ in $\mathbb{C}[X]$. As usual, $\mathbb{C}$ denotes the set of complex numbers.

Exercises to get around in the theory

14. In Exercise 13 you have seen how to use Lemmas 8.1.15 and 8.2.4 in practise. In this exercise you are asked to convince yourself that this method actually works: prove Lemma 8.2.4.

    [**Hint:** Try to adapt the proof of Lemma 8.1.15].

15. Let $(\mathbb{F}, +_{\mathbb{F}}, \cdot_{\mathbb{F}})$ be a field. Show that there are only two possible ideals $I \subset \mathbb{F}$.

16. Prove that for any homomorphism $\phi : \mathbb{F} \to S$ of a field $(\mathbb{F}, +_{\mathbb{F}}, \cdot_{\mathbb{F}})$ to a commutative ring $(S, +_S, \cdot_S)$ either $\ker \phi = \{0_{\mathbb{F}}\}$ or $\mathrm{im}\phi = \{0_S\}$.

    [**Hint:** Use the previous exercise].

17. Let $(R, +_R, \cdot_R)$ be a commutative ring.

    - Let n be a positive integer. Show that if I is an ideal of R and $r_1, \ldots, r_n \in I$ then $\langle r_1, \ldots, r_n \rangle \subseteq I$
    - for $a \in R$ and $u \in R^*$, show that $\langle u \cdot_R a \rangle = \langle a \rangle$
    - for $a, b \in R$, show that $\langle a, a \cdot_R b \rangle = \langle a \rangle$
    - if $(R, +_R, \cdot_R)$ is an integral domain and $a, b \in R$ show that $\langle a \rangle = \langle b \rangle$ if and only if $a = u \cdot_R b$ for some unit $u \in R^*$.

18. Let $\phi : R \to S$ be a ring homomorphism. Prove that

    - If $(R, +_R, \cdot_R)$ is commutative then $(\mathrm{im}\phi, +_S, \cdot_S)$ is commutative
    - if $(R, +_R, \cdot_R)$ is a field then $(\mathrm{im}\phi, +_S, \cdot_S)$ is either a field or the zero ring.

19. Let $(\mathbb{F}_2, +, \cdot)$ be the finite field with two elements.

    - Compute the standard form of the coset $X^3 + X^2 + \langle X^3 + X + 1 \rangle$ in $\mathbb{F}_2[X]$
    - Compute a greatest common divisor of the polynomials $X^3+X+1, X^2+X+1 \in \mathbb{F}_2[X]$ as well as polynomials $r(X), s(X) \in \mathbb{F}_2[X]$ such that

      $$r(X) \cdot (X^3 + X + 1) + s(X) \cdot (X^2 + X + 1) = \gcd(X^3 + X + 1, X^2 + X + 1)$$

      using the extended Euclidean algorithm for polynomials.

20. Suppose that $R = \{0, 1, a, b\}$ consists of four distinct elements. The goal of this exercise is to show that it is possible to find an addition $+$ and multiplication $\cdot$ such that $(R, +, \cdot)$ is a finite field with four elements.

    - Show that if $(R, +, \cdot)$ is a finite field with four elements, then $1 + 1 = 0$

      [**Hint:** first determine the possible orders of the element 1 in the group $(R, +)$ and show that the order cannot be 4 if R is a field]
    - Show that if $(R, +, \cdot)$ is a finite field with four elements, then $a + 1 = b$

      [**Hint:** if $a + 1$ is not b then it has to be equal to 0, 1 or a. Show that this is not possible]
    - Show that if $(R, +, \cdot)$ is a finite field with four elements, then $a^3 = 1$ and $a^2 = b = a+1$

      [**Hint:** Observe that the group $(R^*, \cdot)$ has order 3]

- Now show that there exists a finite field with four elements and identify it with the quotient ring $(\mathbb{F}_2[X]/\langle X^2 + X + 1\rangle, +, \cdot)$.

21. Decide which of the following are ideals of the ring $\mathbb{Z}[X]$:

    - the set of all polynomials whose constant term is a multiple of 5
    - the set of all polynomials whose coefficient of X^2 is a multiple of 5
    - the set of all polynomials whose constant term, coefficient of X and coefficient of X^2 are zero
    - $\mathbb{Z}[X^2]$, that is, the set of polynomials in which only even powers of X appear
    - the set of polynomials $p(X)$ such that the usual derivative $p'(X)$ of $p(X)$ satisfies $p'(0) = 0$.

22. The Gaussian integers $\mathbb{Z}[i]$ are defined as the set $\mathbb{Z}[i] := \{a + bi \mid a, b \in \mathbb{Z}\}$. Here as usual, i denotes the imaginary complex number with real part 0 and imaginary part 1. In this exercise we consider the ring of Gaussian integers $(\mathbb{Z}[i], +, \cdot)$ and one of its quotient rings which turns out to be a finite field with nine elements.

    - Show that the ideal $\langle 3\rangle$ has exactly nine distinct cosets in $\mathbb{Z}[i]$, namely $\langle 3\rangle, \pm 1 + \langle 3\rangle, \pm i + \langle 3\rangle$ and $\pm 1 \pm i + \langle 3\rangle$. Conclude that the quotient ring $(\mathbb{Z}[i]/\langle 3\rangle, +, \cdot)$ contains exactly nine elements
    - Prove that in fact $(\mathbb{Z}[i]/\langle 3\rangle, +, \cdot)$ is a finite field with 9 elements, by showing that each of the elements $\pm 1 + \langle 3\rangle, \pm i + \langle 3\rangle$ and $\pm 1 \pm i + \langle 3\rangle$ has a multiplicative inverse.

23. This exercise can be seen as a continuation of the previous one. Again we consider the ring of Gaussian integers $(\mathbb{Z}[i], +, \cdot)$ and some of its quotient rings.

    - Consider the quotient ring $(\mathbb{Z}[i]/\langle 5\rangle, +, \cdot)$. Show that this ring has zero-divisors and hence it is not a field

      [**Hint:** can one factor the integer 5 in the ring of Gaussian integers?]
    - Show that the quotient ring $(\mathbb{Z}[i]/\langle 1 + 2i\rangle, +, \cdot)$ has precisely five elements, namely $\langle 1 + 2i\rangle, \pm 1 + \langle 1 + 2i\rangle$ and $\pm i + \langle 1 + 2i\rangle$
    - Prove that the map $\psi : \mathbb{Z}[i] \to \mathbb{Z}_5$, defined by $\psi(a + ib) := (a + 2b) \pmod 5$, is a ring homomorphism
    - Prove that ψ is surjective and that $\ker \psi = \langle 1 + 2i\rangle$. Then use the isomorphism theorem for rings to conclude that the ring $(\mathbb{Z}[i]/\langle 1 + 2i\rangle, +, \cdot)$ is a finite field with 5 elements.

      [**Hint:** When proving the part about the kernel, the identities $a+bi = a+2b+bi(1+2i)$ and $5 = (1 + 2i)(1 - 2i)$ may come in handy].

24. Let $\phi : R \to S$ be a ring homomorphism.

    - Prove that if J is an ideal of S then $\phi^{-1}(J) = \{r \in R \mid \phi(r) \in J\}$ is an ideal of R
    - Prove that if ϕ is surjective and I is an ideal of R then $\phi(I)$ is an ideal of S. Give an example where this fails if ϕ is not surjective.

25. Let $(D, +, \cdot)$ be an integral domain and define the relation $\sim$ on $D \times D$ such that $b, d \neq 0$ and $(a, b) \sim (c, d)$ if and only if $a \cdot d = b \cdot c$. Let F be the set of equivalence classes $[(a, b)]$,

with $(a, b) \in D \times D, b \neq 0$. The equivalence class $[(a \cdot x, x)]$ for some $x \neq 0$ is denoted by $[(a, 1)]$. The addition and multiplication in F are defined respectively by

$$[(a, b)] + [(c, d)] = [(a \cdot d + b \cdot c, b \cdot d)],$$

$$[(a, b)] \cdot [(c, d)] = [(a \cdot c, b \cdot d)].$$

Show that:

- (The multiplication operation is well-defined) Prove that if $[(a, b)] = [(a', b')]$ and $[(c, d)] = [(c', d')]$ then $[(a, b)] \cdot [(c, d)] = [(a', b')] \cdot [(c', d')]$.
- (The addition operation is well-defined) Prove that if $[(a, b)] = [(a', b')]$ and $[(c, d)] = [(c', d')]$ then $[(a, b)] + [(c, d)] = [(a', b')] + [(c', d')]$
- Prove the distributive law in F
- Prove that the map $\phi : D \to F$ defined by $\phi(a) = [(a, 1)]$ is a homomorphism of D to F and $\ker \phi = \{0\}$.

Chapter 9

Finite fields

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

9.1 Quotients of polynomial rings with coefficients in a field

We have seen in Example 8.3.6 that the rings $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ are isomorphic. Further Proposition 7.1.15 stated that a nonzero element in $\mathbb{Z}_n$ is a unit precisely if $\gcd(a, n) = 1$ and a zero-divisor otherwise. Using the isomorphism between $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$, we obtain that: if $a + n\mathbb{Z}$ is a nonzero element in $\mathbb{Z}/n\mathbb{Z}$ (that is to say: if $a \notin n\mathbb{Z}$), then $a + n\mathbb{Z}$ is a unit in $\mathbb{Z}/n\mathbb{Z}$ if and only if $\gcd(a, n) = 1$ and a zero-divisor otherwise. This leads to the following reformulation of Proposition 7.1.15:

Proposition 9.1.1 *Let n be a positive integer and $a \in \mathbb{Z}$ an arbitrary integer. Then exactly one of the following three cases holds:*

1. *$a + n\mathbb{Z}$ is the zero element in $\mathbb{Z}/n\mathbb{Z}$, equivalently:* $\gcd(a, n) = n$.
2. *$a + n\mathbb{Z}$ is a zero-divisor in $\mathbb{Z}/n\mathbb{Z}$, equivalently:* $1 < \gcd(a, n) < n$.
3. *$a + n\mathbb{Z}$ is a unit in $\mathbb{Z}/n\mathbb{Z}$, equivalently:* $\gcd(a, n) = 1$.

Proof. The key ingredients of the proof are all contained in the discussion just before this proposition. We have $a + n\mathbb{Z} = 0 + n\mathbb{Z}$ if and only if $a - 0 \in n\mathbb{Z}$, which is true if and only if $\gcd(a, n) = n$. Also note that $\gcd(a, n) = \gcd(a \bmod n, n)$.

Using the isomorphism between the rings $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$, we see that $a + n\mathbb{Z}$ is a unit (respectively zero-divisor) in $\mathbb{Z}/n\mathbb{Z}$ if and only if $(a \bmod n)$ is a unit (respectively zero-divisor) in $\mathbb{Z}_n$. By Proposition 7.1.15, $(a \bmod n)$ is a unit if and only if $\gcd(a, n) = \gcd(a \bmod n, n) = 1$. This leaves the case where $\gcd(a, n)$ is a number between 2 and $n - 1$. In this case $a \bmod n \neq 0$ and hence Proposition 7.1.15 implies that $1 < \gcd(a, n) < n$ if and only if $a \bmod n$ is a zero-divisor in $\mathbb{Z}_n$. Again using the isomorphism between $(\mathbb{Z}_n, +_n, \cdot_n)$ and $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$, this is equivalent to $a + n\mathbb{Z}$ being a zero-divisor in $\mathbb{Z}/n\mathbb{Z}$. ■

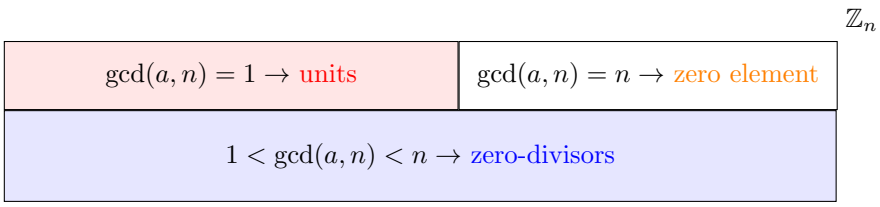


Figure 9.1: The elements in $\mathbb{Z}_n$ collected according to their gcd with n

A similar proposition holds for quotient of polynomial rings with coefficients in a field:

Proposition 9.1.2 *Let $(\mathbb{F}, +, \cdot)$ be a field and suppose that $f(X) \in \mathbb{F}[X]$ is a non-constant polynomial of degree d. Further let $g(X) \in \mathbb{F}[X]$ be an arbitrary polynomial. Then exactly one of the following three cases holds:*

1. *$g(X) + \langle f(X) \rangle = 0 + \langle f(X) \rangle$, the zero element in $\mathbb{F}[X]/\langle f(X) \rangle$. This occurs if and only if $\deg(\gcd(g(X), f(X))) = d$,*
2. *$g(X) + \langle f(X) \rangle$ is a zero-divisor in $\mathbb{F}[X]/\langle f(X) \rangle$. Equivalently: $0 < \deg(\gcd(g(X), f(X))) < d$,*
3. *$g(X) + \langle f(X) \rangle$ is a unit in $\mathbb{F}[X]/\langle f(X) \rangle$. Equivalently: $\deg(\gcd(g(X), f(X))) = 0$.*

Proof. We prove the three items one at the time.

1. First of all, since $d = \deg(f(X))$, we have that $\deg(g(X), f(X)) = d$ if and only if $f(X)$ divides $g(X)$. By definition, the ideal $\langle f(X) \rangle$ consists of all multiples of $f(X)$. Therefore $\deg(g(X), f(X)) = d$ if and only if $g(X) \in \langle f(X) \rangle$, which in turn is equivalent to $g(X) + \langle f(X) \rangle = 0 + \langle f(X) \rangle$.
2. Suppose first that $g(X) + \langle f(X) \rangle$ is a zero-divisor, then there exists a nonzero element $i(X) + \langle f(X) \rangle$ such that

$$(g(X) + \langle f(X) \rangle)(i(X) + \langle f(X) \rangle = 0 + \langle f(X) \rangle.$$

   This implies that $g(X)i(X) \in \langle f(X) \rangle$ and hence that $f(X)$ divides $g(X)i(X)$ Since both $g(X) + \langle f(X) \rangle$ and $i(X) + \langle f(X) \rangle$ are not the zero element $0 + \langle f(X) \rangle$, we see that $f(X)$

does not divide $g(X)$ nor $i(X)$. But then $f(X)$ and $g(X)$ have a common divisor whose degree is positive, but not equal to d.

Now suppose that $0 < \deg(\gcd(g(X), f(X))) < d$ and write $h(X)$ for this greatest common divisor. Further write $f(X) = h(X)i(X)$ and $g(X) = h(X)j(X)$. Since $0 < \deg(h(X))$, the polynomial $h(X)$ is not a constant. Since $\deg(h(X)) < d = \deg(f(X))$, the polynomial $i(X)$ has degree strictly less than d. Indeed, if $\deg(i(X)) = d$, then Corollary 7.3.7 would imply that $\deg(h(X)) = 0$. Then we obtain:

$$(g(X) + \langle f(X)\rangle)(i(X) + \langle f(X)\rangle) = g(X)i(X) + \langle f(X)\rangle = \\ h(X)j(X)i(X) + \langle f(X)\rangle = f(X)j(X) + \langle f(X)\rangle = 0 + \langle f(X)\rangle.$$

Hence $g(X) + \langle f(X)\rangle$ is a zero-divisor.

3. Suppose that $g(X) + \langle f(X)\rangle$ is a unit. We know that a unit is not the zero element, nor a zero-divisor. Hence the previous two parts imply that $\deg(\gcd(g(X), f(X))) = 0$, since this is the only possibility left for the degree of $\gcd(g(X), f(X))$.

   Conversely, assume that $\deg(\gcd(g(X), f(X))) = 0$. This means that $\gcd(f(X), g(X))$ is a nonzero constant. Since we use the convention that greatest common divisors of polynomials are monic, we see that $\gcd(g(X), f(X)) = 1$. By Lemma 8.2.4, or alternatively the extended Euclidean algorithm, we can find polynomials $r(X)$ and $s(X)$ such that $r(X)f(X) + s(X)g(X) = 1$. Then

$$(g(X) + \langle f(X)\rangle) \cdot (s(X) + \langle f(X)\rangle) = g(x) \cdot s(X) + \langle f(X)\rangle = \\ (1 - r(X) \cdot f(X)) + \langle f(X)\rangle = 1 + \langle f(X)\rangle.$$

   The last equality follows, since $r(X)f(X) \in \langle f(X)\rangle$. We see that $g(X) + \langle f(X)\rangle$ is a unit, since it has a multiplicative inverse in the quotient ring $(\mathbb{F}[X]/\langle f(X)\rangle, +, \cdot)$.

■

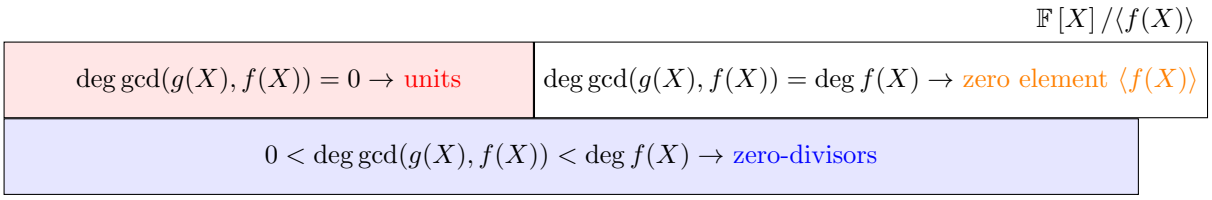


Figure 9.2: The partition of $\mathbb{F}[X]/\langle f(X)\rangle$ according to the gcd of its elements $g(X) + \langle f(X)\rangle$

9.2 Construction of fields using irreducible polynomials

Keywords: irreducible polynomials, construction of new fields

We have seen in Example 8.3.6 that the quotient ring $(\mathbb{Z}/p\mathbb{Z}, +, \cdot)$ is a finite field with p elements if p is a prime number. We have seen this field before in another description, namely

as $(\mathbb{Z}_p, +_p, \cdot_p)$ and usually denote this field by $(\mathbb{F}_p, +, \cdot)$. Another procedure to produce fields is by considering quotient rings of $\mathbb{F}[X]$, the polynomial ring with coefficients in a field $\mathbb{F}$. As we saw in Examples 8.3.4 and 8.3.7, the complex numbers $\mathbb{C}$ are obtained in this way. As we will see, what we will need is the analogue of a prime number in a polynomial ring: an irreducible polynomial.

Definition 9.2.1 *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 nonzero constants, this means that $g(X)$ or $h(X)$ is a nonzero constant.

Example 9.2.2 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 9.2.3 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).

It is in fact easy in general to check whether a polynomial of degree at most 3 is irreducible. We provide an analysis of such polynomials in the following two lemmas.

Lemma 9.2.4 *Let $(\mathbb{F}, +, \cdot)$ be a field. If $f(X) \in \mathbb{F}[X]$ is a polynomial of degree 1, then $f(X)$ is irreducible.*

Proof. To check irreducibility we write $f(X) = g(X) \cdot h(X)$ for some $g(X), h(X) \in \mathbb{F}[X]$ with $\deg g(X) \geq 0$ and $\deg h(X) \geq 0$. Considering the corresponding degrees, one has $1 = \deg f(X) = \deg g(X) + \deg h(X)$. Since $\deg g(X) \geq 0$ and $\deg h(X) \geq 0$ the only possibility is that either $\deg g(X) = 1$ (and $\deg h(X) = 0$) or $\deg h(X) = 1$ (and $\deg g(X) = 0$). Hence either $\deg g(X) = \deg f(X)$ or $\deg h(X) = \deg f(X)$ and $f(X)$ is irreducible. ■

Lemma 9.2.5 *Let $(\mathbb{F}, +, \cdot)$ be a field and $f(X) \in \mathbb{F}[X]$ be a polynomial of degree 2 or 3. Then $f(X)$ is irreducible if and only if it has no roots in $\mathbb{F}$.*

Proof. We will prove the logically equivalent statement: $f(X)$ is reducible if and only if it has a root in $\mathbb{F}$. Recall that $f(X)$ has a root $a \in \mathbb{F}$ if and only if $f(X) = (X - a) \cdot q(X)$ for some $q(X) \in \mathbb{F}[X]$ with $\deg q(X) = \deg f(X) - 1$ from Proposition 7.4.3. This implies immediately that whenever $f(X)$ has a root then it is reducible. We want now to show the other implication, that is, $f(X)$ reducible implies that $f(X)$ has a root in $\mathbb{F}$. Assume that $f(X)$ is reducible, and write $f(X) = g(X) \cdot h(X)$ for some $g(X), h(X) \in \mathbb{F}[X]$ where $\deg g(X) > 0$ and $\deg h(X) > 0$. One has that either $2 = \deg f(X) = \deg g(X) + \deg h(X)$ or $3 = \deg f(X) = \deg g(X) + \deg h(X)$. Since $\deg g(X) > 0$ and $\deg h(X) > 0$, it necessarily holds that either $\deg g(X) = 1$ or $\deg h(X) = 1$. This proves that if $f(X)$ is reducible then it is divisible by a polynomial of degree 1, that is, it is

divisible by $x - a$ for some $a \in \mathbb{F}$ (up to dividing by the leading coefficient). The fact that $f(X)$ has a root follows at this point again from Proposition 7.4.3. ■

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 $X^4 + 2X^2 + 1 = (X^2 + 1)^2$.

Before showing how irreducible polynomials can be used to produce fields, let us analyze some additonal interesting properties that they have in common with prime numbers. We have in fact already mentioned that irreducible polynomials are an analogue of prime numbers in the setting of polynomials over fields, because of their being only divisible by a constant times themselves and constants. One of the main reasons of importance of prime numbers is however also that they are the building blocks of the positive integers under multiplication, meaning that every integer $n > 1$ can be written uniquely as product of prime numbers. The fact that every integer $n > 1$ can be written as a product of primes has been analyzed in Exercise 26 from Chapter 2.

Surprising enough, also from this point of view irreducible polynomials behave like prime numbers. In fact, a similar property to the one just mentioned for integers holds also for polynomials. Namely every polynomial of degree at least 1 can be written uniquely as a product of irreducible polynomials (where uniqueness here is only up to multiplication by constant polynomials).

Hence our next aim exactly is to prove this remarkable property. Before doing so, let us first prove the uniqueness of the factorization in $\mathbb{Z}$ into prime numbers. We will see that the proof in the polynomial setting is essentially the same. A first technical lemma is needed.

Lemma 9.2.6 *If p is a prime number and p divides a product $a_1 \dots a_k$ for some integers a_i and $k \geq 1$, then p divides a_i for some $i = 1, \dots, k$.*

Proof. The case $k = 1$ is trivial. We consider the case $k = 2$. Suppose that p does not divide a_1. Then $\gcd(p, a_1) = 1$ as p is a prime. Using the Euclidean algorithm (Bezout's identity) we can write $1 = pn + a_1 m$ for some integers m and n. Multiplying by a_2 we get that $a_2 = pa_2 n + a_1 a_2 m$. Since p divides both $pa_2 n$ and $a_1 a_2 m$ we get that p must divide a_2. The proof proceeds by induction on k, using $k = 2$ as a base case. Suppose the statement is true for the product of k integers (induction hypothesis), and consider the case of $k + 1$ integers. If p divides $a_1 \dots a_{k+1}$, then in particular p divides the product of the two integers $a_1 \dots a_k$ and a_{k+1}. From the analysis of the case $k = 2$ we get that either p divides $a_1 \dots a_k$ or p divides a_{k+1}. If p divides $a_1 \dots a_k$ then p divides a_i for some $i = 1, \dots, k$ from the induction hypothesis. Including with this the option that p divides a_{k+1} we get that in any case p divides a_i for some $i = 1, \dots, k + 1$. ■

Theorem 9.2.7 *Every integer $n > 1$ can be written uniquely as a product of prime numbers, that is, if $p_1 \dots p_r = q_1 \dots q_s$ where the $p_i's$ and the $q_i's$ are primes, then $r = s$ and after relabeling the factors we have $p_i = q_i$ for all i.*

Proof. The first key step is to prove that if $p_1 \dots p_r = q_1 \dots q_s$ then $p_1 = q_j$ for some j. This is because $p_1 \dots p_r = q_1 \dots q_s$ implies that p_1 divides the product $q_1 \dots q_s$. Hence from Lemma

9.2.6 $p_1 = q_j$ for some $j = 1, \ldots, s$ (a prime has no factor greater than 1 other than itself). To prove the theorem, we will proceed by induction on the total number of prime factors involved in the two prime factorizations, which is $r + s$, which allows repeated primes. The base case is $r+s = 2$ when the two equal prime factorizations turn into $p_1 = q_1$ and the conclusion is obvious. Suppose next that $r + s > 2$ and the theorem is true for all pairs of equal prime factorizations where the total number of prime factors is less than $r + s$. If we have $p_1 \ldots p_r = q_1 \ldots q_s$ then $r > 1$ and $s > 1$. Indeed is $r = 1$ or $s = 1$ then one side is a prime number and therefore the other side has to be a prime number too, giving $r = s = 1$, which is not possible as $r + s > 2$. From $p_1 \ldots p_r = q_1 \ldots q_s$ we already know from the first step that $p_1 = q_j$ for some j. Up to relabeling we can assume $p_1 = q_1$. Our equal factorizations become $p_1 \ldots p_r = p_1 \ldots q_s$, implying that

$$p_2 \ldots p_r = q_2 \ldots q_s.$$

In this equation of equal prime factorizations the total number of prime factors appearing is $(r-1)+(s-1) = r+s-2 < r+s$. By our inductive hypothesis we conclude that $r-1 = s-1$, and hence $r = s$, and after relabeling $p_i = q_i$ for all $i \geq 2$. Combining this with $p_1 = q_1$ completes the proof. ■

We turn next to unique factorization of polynomials, and we will see that almost everything we have done for integers will carry over to the polynomial setting except that we have to keep track of constant multiples of irreducible factors. We start by proving that irreducible factorizations exist.

Theorem 9.2.8 *Let $(\mathbb{F}, +, \cdot)$ be a field and $f(X) \in \mathbb{F}[X]$ be a polynomial of positive degree d. The $f(X)$ can be written as a product of irreducible polynomials.*

Proof. We will prove the statement by induction on d. The base case is $d = 1$. As proven in Lemma 9.2.4, $f(X)$ is irreducible itself, so there is nothing to prove. Now assume $d > 1$ and that every polynomial in $\mathbb{F}[X]$ of degree less than d admits an irreducible factorization. Two cases can occur for our polynomial $f(X)$ of degree d: either $f(X)$ is irreducible, or it is reducible. If $f(X)$ is irreducible that $f(X)$ is a one-term product of irreducibles, and there is nothing to prove. If $f(X)$ is reducible then $f(X) = g(X)h(X)$ for some $g(X), h(X) \in \mathbb{F}[X]$ both of degree larger than zero and less than d. By the induction hypothesis both $g(X)$ and $h(X)$ can be written as a product of irreducible polynomials as $g(X) = p_1(X) \ldots p_s(X)$ and $h(x) = q_1(X) \ldots q_r(X)$. Hence $f(X)$ can be written as a product of irreducible polynomials as $p_1(X) \ldots p_s(X) q_1(X) \ldots q_r(X)$. ■

To prove the uniqueness of the irreducible factorization we need the following analogue of Lemma 9.2.6.

Lemma 9.2.9 *Let $(\mathbb{F}, +, \cdot)$ be a field. If $p(X) \in \mathbb{F}[X]$ is an irreducible polynomial and $p(X)$ divides a product $a_1(X) \ldots a_k(X)$ for some polynomials $a_i(X) \in \mathbb{F}[X]$ and $k \geq 1$, then $p(X)$ divides $a_i(X)$ for some $i = 1, \ldots, k$.*

Proof. The proof is identical to that of Lemma 9.2.6, as also in this case we can use the (extended) Euclidean algorithm. The case $k = 1$ is trivial. We consider the case $k = 2$. Suppose that $p(X)$ does not divide $a_1(X)$. Then $\gcd(p(X), a_1(X)) = 1$ as $p(X)$ is irreducible. Using the extended Euclidean algorithm (Bezout's identity) we can write $1 = p(X)n(X) + a_1(X)m(X)$ for

some polynomials $m(X)$ and $n(X)$. Multiplying by $a_2(X)$ we get that $a_2(X) = p(X)a_2(X)n(X)+ a_1(X)a_2(X)m(X)$. Since $p(X)$ divides both $p(X)a_2(X)n(X)$ and $a_1(X)a_2(X)m(X)$ we get that $p(X)$ must divide $a_2(X)$. The proof proceeds by induction on k, using $k = 2$ as a base case. Suppose the statement is true for the product of k polynomials (induction hypothesis), and consider the case of $k + 1$ polynomials. If $p(X)$ divides $a_1(X) \dots a_{k+1}(X)$, then in particular $p(X)$ divides the product of the two polynomials $a_1(X) \dots a_k(X)$ and $a_{k+1}(X)$. From the analysis of the case $k = 2$ we get that either $p(X)$ divides either $a_1(X) \dots a_k(X)$ or $p(X)$ divides $a_{k+1}(X)$. If $p(X)$ divides $a_1(X) \dots a_k(X)$ then $p(X)$ divides $a_i(X)$ for some $i = 1, \dots, k$ from the induction hypothesis. Including with this the option that $p(X)$ divides $a_{k+1}(X)$ we get that in any case $p(X)$ divides $a_i(X)$ for some $i = 1, \dots, k + 1$. ■

With this lemma, we are ready to prove the uniqueness of the factorization into irreducible polynomials in $\mathbb{F}[X]$. Note that uniqueness in this case is only up to constant factors, as the following theorem explains.

Theorem 9.2.10 *Let $(\mathbb{F}, +\cdot)$ be a field. Every polynomial $f(X) \in \mathbb{F}[X]$ of degree at least 1 can be written uniquely as a product of irreducible polynomials, that is, if $p_1(X) \dots p_r(X) = q_1(X) \dots q_s(X)$ where the $p_i(X)'s$ and the $q_i(X)'s$ are irreducible polynomials, then $r = s$ and after relabeling the factors we have $p_i(X) = c_i q_i(X)$ for all i, where the $c_i \in \mathbb{F}$ are nonzero.*

Proof. The proof is essentially the same as that of Theorem 9.2.7, apart from some extra needed bookkeeping to account for constant multiples. The first key step is to prove that if $p_1(X) \dots p_r(X) = q_1(X) \dots q_s(X)$ then $p_1(X) = c_j q_j(X)$ for some j and some $c_j \in \mathbb{F}$. This is because $p_1(X) \dots p_r(X) = q_1(X) \dots q_s(X)$ implies that $p_1(X)$ divides the product $q_1(X) \dots q_s(X)$. Hence from Lemma 9.2.6 $p_1(X)$ divides $q_j(X)$ for some $j = 1, \dots, s$. Up to relabeling we can assume that $j = 1$, that is, $p_1(X)$ divides $q_1(X)$. This implies that $p_1(X)$ is a constant multiple of $q_1(X)$ since the only nonconstant factors of an irreducible polynomial in $\mathbb{F}[X]$ are constant multiples of itself. Therefore $p_1(x) = c_1 q_1(x)$ for some constant $c_1 \in \mathbb{F}$. Our theorem will be proved by induction on the total number of irreducible factors involved in the two equal irreducible factorizations, which is $r + s$. The base case is $r + s = 2$ when the two equal irreducible factorizations turn into $p_1(X) = q_1(X)$ and the conclusion is obvious. Suppose next that $r+s > 2$ and the theorem is true for all pairs of equal irreducible factorizations where the total number of irreducible factors is less than $r + s$. If $p_1(X) \dots p_r(X) = q_1(X) \dots q_s(X)$ then $r > 1$ and $s > 1$, by the same proof as for the integer case.

From $p_1(X) \dots p_r(X) = q_1(X) \dots q_s(X)$ we already know, up to relabeling, that $p_1(X) = c_1 q_1(X)$ for some $c_1 \in \mathbb{F}$. Note that $c_1 \neq 0$ and hence it is a unit in $\mathbb{F}$. Hence $p_1(X) \dots p_r(X) = c_1^{-1} p_1(X) \dots q_s(X)$, implying that

$$p_2(X) \dots p_r(X) = c_1^{-1} q_2(X) \dots q_s(X).$$

We need to be careful here: the factor c_1^{-1} is not irreducible, it is just a constant. The polynomial $c_1^{-1} q_2(X)$ is though irreducible. Hence in the equation of equal irreducible factorizations written above, the total number of irreducible factors appearing is $(r - 1) + (s - 1) = r + s - 2 < r + s$. By our inductive hypothesis we conclude that $r - 1 = s - 1$, and hence $r = s$, and after relabeling $p_i(X) = c_i q_i(X)$ for all $i \geq 3$, while $p_2(X) = c c_1^{-1} q_2(X)$. Setting $c_2 = c c_1^{-1}$ and combining this with $p_1(X) = c_1 q_1(X)$ gives $p_i(X) = c_i q_1(X)$, which completes the proof. ■

Now we can formalize what we anticipated at the beginning of the section, namely that

irreducible polynomials are useful, as it turns out that they can be used to construct fields. One such field we have already encountered in Examples 8.3.4 and 8.3.7. 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 construct finite fields that are very useful in several areas of discrete mathematics, but for now we give a different example:

Example 9.2.11 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. It turns out that the quotient ring $(\mathbb{F}_2[X]/\langle X^3 + X + 1\rangle, +, \cdot)$ is a field from a general result that we will prove now.

Theorem 9.2.12 *Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X) \in \mathbb{F}[X]$ be a non-constant polynomial. Then the quotient ring $(\mathbb{F}[X]/\langle f(X)\rangle, +, \cdot)$ is a field if and only if $f(X)$ is irreducible.*

Proof. First of all, assume that $f(X)$ is reducible, say $f(X) = g(X) \cdot h(X)$ for non-constant polynomials $g(X)$ and $h(X)$. Since both $g(X)$ and $h(X)$ have degree strictly less than $\deg(f(X))$, the cosets $g(X) + \langle f(X)\rangle$ and $h(X) + \langle f(X)\rangle$ are in standard form. In particular neither of these cosets is equal to $0 + \langle f(X)\rangle$. However,

$$(g(X) + \langle f(X)\rangle)(h(X) + \langle f(X)\rangle) = f(X) + \langle f(X)\rangle) = 0 + \langle f(X)\rangle).$$

Therefore, the quotient ring $(\mathbb{F}[X]/\langle f(X)\rangle, +, \cdot)$ has zero-divisors. As a consequence it is not a domain. Since any field is a domain, $(\mathbb{F}[X]/\langle f(X)\rangle, +, \cdot)$ is not a field either.

Now assume that $f(X)$ is irreducible and let $g(X) + \langle f(X)\rangle$ be an arbitrary coset in standard form not equal to $0 + \langle f(X)\rangle$. In particular, this implies that $f(X)$ does not divide $g(X)$. Since $f(X)$ is irreducible, we conclude that $\gcd(g(X), f(X)) = 1$. Then Proposition 9.1.2 implies that $g(X) + \langle f(X)\rangle$ is a unit in $\mathbb{F}[X]/\langle f(X)\rangle$. Since the coset $g(X) + \langle f(X)\rangle$ was chosen arbitrarily (but different from $0 + \langle f(X)\rangle$), we conclude that any nonzero element of $\mathbb{F}[X]/\langle f(X)\rangle$ is a unit. Hence $(\mathbb{F}[X]/\langle f(X)\rangle, +, \cdot)$ is a field. ■

If $f(X) \in \mathbb{F}[X]$ is any non-constant polynomial, we can see $\mathbb{F}$ as a subset of $\mathbb{E} := \mathbb{F}[X]/\langle f(X)\rangle$ by identifying $a \in \mathbb{F}$ with $a + \langle f(X)\rangle$. To say the same thing in a different way: the map $\psi : \mathbb{F} \to \mathbb{E}$ defined by $\psi(a) = a + \langle f(X)\rangle$ is an injective ring homomorphism. If $f(X)$ is irreducible as well, one says that $\mathbb{F}$ is a *subfield* of $\mathbb{E}$ and conversely that $\mathbb{E}$ is an *extension field* of $\mathbb{F}$. This means that we can interpret a polynomial in $\mathbb{F}[X]$ as a polynomial with coefficients in $\mathbb{E}$. It turns out that $f(X)$ interpreted as a polynomial with coefficients in $\mathbb{E}$ has a root in $\mathbb{E}$, when seen as a polynomial with coefficients in $\mathbb{E}$.

Lemma 9.2.13 *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 $f(T)$ when seen as an element of the set of polynomials $\mathbb{E}[T]$.*

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 $a + \langle f(X)\rangle$. Therefore we can interpret polynomials with coefficients in $\mathbb{F}$ as polynomials

with coefficients in $\mathbb{E}$. To avoid confusing elements from $\mathbb{E}$ with polynomials with coefficients in $\mathbb{F}$, we use the variable T when talking about polynomials with coefficients in $\mathbb{E}$. Concretely, this means that we identity the polynomial $a_0 + a_1X + \cdots a_dX^d \in \mathbb{F}[X]$ with the polynomial $(a_0 + \langle f(X)\rangle) + (a_1 + \langle f(X)\rangle)T + \cdots (a_1 + \langle f(X)\rangle)T^d \in \mathbb{E}[T]$. In particular, the polynomial

$$f(T) = \sum_{i=0}^{d} a_i X^i \in \mathbb{F}[X],$$

is identified with the polynomial

$$f(T) = \sum_{i=0}^{d} (a_i + \langle f(X)\rangle) T^i \in \mathbb{E}[T].$$

To show the lemma, we now need to show that $X + \langle f(X)\rangle$ is a root of $f(T)$. 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. ■

If $f(X) = X^2 + 1$ and $\mathbb{F} = \mathbb{R}$, then we have seen in Example 8.3.4 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 indeed has a complex root in $\mathbb{C}$, namely i (of course $-i$ is a root as well). The complex number i was identified with $X + \langle X^2 + 1\rangle \in \mathbb{R}[X]/\langle X^2 + 1\rangle$, which is exactly the root that the previous lemma indicated. Briefly put, Lemma 9.2.13 states that for any irreducible polynomial $f(X)$ with coefficients in a field $\mathbb{F}$, one can construct an extension field $\mathbb{E}$ of $\mathbb{F}$ in which $f(X)$ has a root.

9.3 Construction of finite fields

Keywords: finite fields

If p is a prime number, we have already encountered finite fields with p elements: $(\mathbb{F}_p, +, \cdot)$, also see Examples 7.2.6 and 8.3.6. We will now introduce more finite fields using the construction from the previous section. In fact we have already seen one such a field in Example 9.2.11. 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)). We start with an observation on the number of elements in a quotient ring of the form $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$.

Lemma 9.3.1 *Let $f(X) \in \mathbb{F}_p[X]$ be a nonzero polynomial of degree d. Then the quotient ring $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$ is a finite ring with p^d elements.*

Proof. By Lemma 8.2.6, any element in $\mathbb{F}_p[X]/\langle f(x)\rangle$ can be described uniquely in the standard form $r(X)+\langle f(X)\rangle$ where $r(X) = a_0+a_1X+\cdots a_{d-1}X^{d-1}$ 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 $\mathbb{F}_p[X]/\langle f(X)\rangle$. This concludes the proof. ■

With this lemma in place, we explain how to construct a whole class of finite fields:

Theorem 9.3.2 *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. The previous lemma implies that $|\mathbb{F}_p[X]/\langle f(x)\rangle| = p^d$, while by Theorem 9.2.12, the quotient ring $(\mathbb{F}_p[X]/\langle f(x)\rangle, +, \cdot)$ is a field. ■

We will see later that for any prime p and positive integer d there exists at least one irreducible polynomial in $\mathbb{F}_p[X]$ of degree d. Hence, using the previous theorem, we see that we can construct a finite field with p^d elements for any prime p and any positive integer d. To get some impression on how many irreducible polynomials there are, a small table of irreducible polynomials of degree 2 in $\mathbb{F}_p[X]$ for small values of p is given below:

p	d	monic, irreducible polynomials of degree d
2	2	X^2+X+1
	3	X^3+X+1, X^3+X^2+1
	4	$X^4+X+1, X^4+X^3+1, X^4+X^3+X^2+X+1$
	5	$X^5+X^4+X^3+X+1, X^5+X^4+X^2+X+1, X^5+X^3+X^2+X+1,$
		$X^5+X^3+1, X^5+X^4+X^3+X^2+1, X^5+X^2+1$
3	2	X^2+1, X^2+X+2, X^2+2X+2
	3	$X^3+2X^2+X+1, X^3+2X^2+2X+2, X^3+2X^2+1, X^3+2X+2,$
		$X^3+X^2+2X+1, X^3+X^2+X+2, X^3+2X+1, X^3+X^2+2$
5	2	$X^2+4X+2, X^2+2X+3, X^2+2X+4, X^2+4X+1, X^2+3$
		$X^2+X+2, X^2+3X+3, X^2+X+1, X^2+2, X^2+3X+4$

As the table illustrates, there are usually many different ways to construct a finite field with p^d elements of the form $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$, since there are many choices for the irreducible polynomial $f(X)$. This gives rise to the question if there are are distinct, or more precisely, non-isomorphic finite fields with the same number $q = p^d$ of elements. We will see later that up to isomorphism, there is exactly one finite field with $q = p^d$ elements. This motivates and explains the notation $(\mathbb{F}_q, +, \cdot)$ or $(GF(q), +, \cdot)$ for these fields. We will see in the exercises that if q is not a prime power, a finite field with q elements does not exist. This means that we have now constructed all possible finite fields!

Example 9.3.3 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 9.2.11 is a finite field with 8 elements.

9.4 Primitive elements in finite fields

Finite fields have a very rich structure. One that is important for applications concerns the multiplicative structure of a finite field. We have already seen that in general the units of a ring give rise to a group $(R^*, \cdot)$. The precise structure of this group depends very much on the ring $(R, +, \cdot)$. In the case of a finite field $(\mathbb{F}_q, +, \cdot)$, where $q = p^d$ for some prime number p, the structure of the group $(\mathbb{F}_q^*, \cdot)$ is quite simple:

Theorem 9.4.1 *Let $(\mathbb{F}_q, +, \cdot)$ be a finite field with q elements, where $q = p^d$ for some prime number p and d is a positive integer. 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 4.4.4 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 7.4.4 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 have order m (where $\phi(m)$ is the Euler totient function from Corollary 4.4.6). 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

$$q - 1 \leq \sum_{m \mid q-1} \phi(m). \tag{9.1}$$

On the other hand, from Corollary 3.2.7, we know that $\sum_{m \mid q-1} \phi(m) = q - 1$. Apparently, equality holds in equation (9.1), which implies that for any m dividing $q - 1$ there exist exactly $\phi(m)$ elements in $\mathbb{F}_q^*$ of order m. In particular, there exists an element $c \in \mathbb{F}_q^*$ having order $q - 1$. But then $(\mathbb{F}_q^*, \cdot)$ is a cyclic group. ■

Since $(\mathbb{F}_q^*, \cdot)$ is a cyclic group, Corollary 3.2.6 directly applies and we obtain the following.

Corollary 9.4.2 *Let $(\mathbb{F}_q, +, \cdot)$ be a finite field with q elements, where $q = p^d$ for some prime number p and d is a positive integer. Further let m be a divisor of $q - 1$. Then there exist exactly $\phi(m)$ elements in $\mathbb{F}_q^*$ with multiplicative order m.*

An element $c \in \mathbb{F}_q^*$ with multiplicative order $q - 1$ is called a *primitive element* of $\mathbb{F}_q$. The previous corollary implies that there exist exactly $\phi(q - 1)$ primitive elements in $\mathbb{F}_q$. This means in practice that $\mathbb{F}_q$ contains many primitive elements. Hence it is not too hard to find one by

trial and error. The following table gives for some values of q the number of primitive elements of $\mathbb{F}_q$ and the relative number of primitive elements $\phi(q)/(q-1)$.

q	$\phi(q-1)$	$\frac{\phi(q-1)}{q-1}$	q	$\phi(q-1)$	$\frac{\phi(q-1)}{q-1}$	q	$\phi(q-1)$	$\frac{\phi(q-1)}{q-1}$
2	1	1.00	3	1	0.50	5	2	0.50
2^2	2	0.66	3^2	4	0.50	5^2	8	0.33
2^3	6	0.85	3^3	12	0.46	5^3	60	0.48
2^4	8	0.53	3^4	32	0.40	5^4	192	0.30
2^5	30	0.96	3^5	110	0.45	5^5	1400	0.44
2^6	36	0.57	3^6	288	0.39	5^6	4320	0.27
2^7	126	0.99	3^7	1092	0.49	5^7	39060	0.49
2^8	128	0.50	3^8	2560	0.39	5^8	119808	0.30
2^9	432	0.84	3^9	9072	0.46	5^9	894240	0.45
2^{10}	600	0.58	3^{10}	26400	0.44	5^{10}	2912000	0.29

It is also worth noting that any element in $\mathbb{F}_q$ is a root of the polynomial $X^q - X \in \mathbb{F}_p[X]$. Indeed, since $(F_q^*, \cdot)$ is a cyclic group of order $q-1$, we already see that $\alpha^{q-1} = 1$ for any nonzero $\alpha \in \mathbb{F}_q$. Hence $\alpha^q = \alpha$ for any nonzero $\alpha \in \mathbb{F}_q$, but for $\alpha = 0$, the equality $\alpha^q = \alpha$ clearly holds as well.

Example 9.4.3 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 $c^6 = c^3 \cdot c^3 = 1$.

Example 9.4.4 Consider the finite field with $p = 521$ elements. Note that 521 is a prime number and that $p - 1 = 520 = 2^3 \cdot 5 \cdot 13$. We claim that 3 is a primitive element of $\mathbb{F}_{521}$. Since ord(3) divides 520, the only way it cannot be equal to 520 is if it divides $520/2 = 260$, $520/5 = 104$, or $520/13 = 40$. However, using for example a computer, one sees that $3^{260} \bmod 521 = 520$, so ord(3) does not divide 260 and similarly $3^{104} \bmod 521 = 25$ and $3^{40} \bmod 521 = 422$. Hence ord(3) = 520, which means that 3 is a primitive element of $\mathbb{F}_{521}$.

Given a primitive element $c \in \mathbb{F}_q^*$ and an arbitrarily chosen 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, such as the Diffie-Hellman key-exchange protocol. Also in other areas of discrete mathematics, finite fields are used with great success. For example the well-known Reed–Solomon codes, used among others to achieve reliable communication over a noisy channel, make extensive use of finite fields.

9.5 Extra: uniqueness and existence of a finite field with $q = p^d$ elements

In this section, we will show that the notation $\mathbb{F}_{p^d}$ is justified. That is to say, we will show that such fields always exist and that up to isomorphism there is only one finite field with p^d elements. We will need various properties of irreducible polynomials in $\mathbb{F}_p[X]$ for this. We start with a lemma.

Lemma 9.5.1 *Let p be a prime number and $g(X) \in \mathbb{F}_p[X]$ an irreducible polynomial of degree $d > 0$, then $g(X)$ divides $X^{p^d} - X$.*

Proof. Consider the ring homomorphism $\phi : \mathbb{F}_p[X] \to \mathbb{F}_p[X]/\langle g(X)\rangle$. It is clear that ϕ is surjective and has kernel $\ker\phi = \langle g(X)\rangle$. On the other hand, since $g(X)$ is irreducible, $(\mathbb{F}_p[X]/\langle g(X)\rangle, +, \cdot)$ is a finite field with q elements and hence any of its elements is a root of the polynomial $X^q - X$. This means that $X^q - X \in \ker\phi$. But since $\ker\phi$ is a principal ideal generated by $g(X)$, this implies that $g(X)$ divides $X^q - X$ in $\mathbb{F}_p[X]$. ■

This lemma is already enough to show that up to isomorphism, finite fields with q elements are unique.

Theorem 9.5.2 *Let $(\mathbb{F}_q^{(1)}, +, \cdot)$ and $(\mathbb{F}_q^{(2)}, +, \cdot)$ be two finite fields with $q = p^d$ elements, where $d \geq 1$ is an integer and p is a prime number. Then $(\mathbb{F}_q^{(1)}, +, \cdot)$ and $(\mathbb{F}_q^{(2)}, +, \cdot)$ are isomorphic fields.*

Proof. Let α be a primitive element of for $\mathbb{F}_q^{(1)}$ and consider the ring homomorphism $\phi : \mathbb{F}_p[X] \to \mathbb{F}_q^{(1)}$ which sends X to α. That is to say, if $f(X) \in \mathbb{F}_p[X]$, then $\phi(f(X)) = f(\alpha)$. Since α is a primitive element, it is clear that ϕ is surjective. Indeed, $\phi(0) = 0$ and $\phi(X^i) = \alpha^i$ for $0 \leq i \leq q-2$, but we know that $\mathbb{F}_q^{(1)} = \{0, 1, \alpha, \ldots, \alpha^{q-2}\}$. Since $(\mathbb{F}_p[X], +, \cdot)$ is a PID, we see that $\ker\phi$ is a principal ideal, say generated by a monic polynomial $g^{(1)}(X)$. Using the isomorphism theorem for rings, we see that $(\mathbb{F}_p[X]/\langle g^{(1)}(X)\rangle, +, \cdot)$ is isomorphic to $(\mathbb{F}_q^{(1)}, +, \cdot)$. Since $(\mathbb{F}_q^{(1)}, +, \cdot)$ is a field with $q = p^d$ elements, we may conclude that $g^{(1)}(X)$ is an irreducible polynomial of degree d. A similar argument shows that there exists a monic, irreducible polynomial $g^{(2)}(X) \in \mathbb{F}_p[X]$ such that $(\mathbb{F}_p[X]/\langle g^{(2)}(X)\rangle, +, \cdot)$ is isomorphic to $(\mathbb{F}_q^{(2)}, +, \cdot)$.

To prove the theorem, it is sufficient to prove that $(\mathbb{F}_p[X]/\langle g^{(1)}(X)\rangle, +, \cdot)$ and $(\mathbb{F}_p[X]/\langle g^{(2)}(X)\rangle, +, \cdot)$ are isomorphic. If $g^{(1)}(X) = g^{(2)}(X)$ this is obvious, so we assume from now on that $g^{(1)}(X) \neq g^{(2)}(X)$. With slight abuse of notation, we write $\mathbb{F}_q^{(1)} := \mathbb{F}_p[X]/\langle g^{(1)}(X)\rangle$ and $\mathbb{F}_q^{(2)} := \mathbb{F}_p[X]/\langle g^{(2)}(X)\rangle$.

We know from Lemma 9.5.1 that both $g^{(1)}(X)$ and $g^{(2)}(X)$ divide $X^q - X$ viewed as elements in $\mathbb{F}_p[X]$. We also know that all elements $\mathbb{F}_q^{(1)}$ are roots of the polynomials $X^q - X$. Hence, when viewed as an element of $\mathbb{F}_q^{(1)}[X]$, we have the factorization $X^q - X = \prod_{\alpha \in \mathbb{F}_q^{(1)}} (X - \alpha)$. Since $g^{(2)}(X)$ divides $X^q - X$, this implies that also $g^{(2)}(X)$, viewed as element of $\mathbb{F}_q^{(1)}[X]$, is the product of certain degree one polynomials and in particular that $g^{(2)}(X)$ has a root in $\mathbb{F}_q^{(1)}$. Let

us denote one such a root by β and consider the ring homomorphism $\psi : \mathbb{F}_p[X] \to \mathbb{F}_q^{(1)}$ sending $f(X)$ to $f(\beta)$. Then $g^{(2)}(X) \in \ker \psi$, since β was chosen as a root of $g^{(2)}(X)$. In particular, $\langle g^{(2)}(X) \rangle \subseteq \ker \psi$. We claim that $\langle g^{(2)}(X) \rangle = \ker \psi$. Since $\mathbb{F}_p[X]$ is a PID, we know that $\ker \psi = \langle d(X) \rangle$ for some monic polynomial $d(X) \in \mathbb{F}_p[X]$. The inclusion $\langle g^{(2)}(X) \rangle \subseteq \ker \psi$ implies that $d(X)$ divides $g^{(2)}(X)$. Since $g^{(2)}(X)$ is irreducible, this implies that either $d(X) = 1$ or $d(X) = g^{(2)}(X)$. However, $d(X) = 1$ would imply that $\ker \psi = \mathbb{F}_p[X]$, which is not true, since $\psi(X) = \beta \neq 0$.

Now that we know that $\ker \psi = \langle g^{(2)}(X) \rangle$, we can conclude that ψ gives rise to an injective ring homomorphism $\tilde{\psi} : \mathbb{F}_q^{(2)} \to \mathbb{F}_q^{(1)}$. Since $|\mathbb{F}_q^{(2)}| = |\mathbb{F}_q^{(1)}|$, this implies that $\tilde{\psi}$ is surjective as well and hence is an isomorphism. ■

What is left to show is the existence of a finite field with p^d elements for all prime numbers p and positive integers d. This amounts to showing that there exists at least one irreducible polynomial of degree d in $\mathbb{F}_p[X]$. We will do better than that and give a formula for the number of monic, irreducible polynomials of degree d. Again Lemma 9.5.1 is crucial and we will start out by strengthening it. In order to do this we need the following elegant result.

Lemma 9.5.3 *Let p be a prime number and d, e positive integers. Then the greatest common divisor of the polynomials $X^{p^e} - X, X^{p^d} - X \in \mathbb{F}_p[X]$ equals $X^{p^{\gcd(d,e)}} - X$. In particular, if e divides d, then $X^{p^e} - X$ divides $X^{p^d} - X$.*

Proof. We may assume that $e \leq d$, interchanging d and e if needed. We will show by strong induction on d that the lemma is true. If $d = 1$, then necessarily $e = 1$, since $e \leq d$ and d, e are positive integers. In this case, we have $\gcd(d, e) = 1$ and $\gcd(X^p - X, X^p - X) = X^p - X$, so the lemma is true. Now assume that the lemma holds for all integers $d - 1, f$ satisfying $f \leq d - 1$ for some $d \geq 2$. Assume a pair of positive integers e, d is given, where $e \leq d$. Since $(X^{p^d} - X) - (X^{p^e} - X)^{p^{d-e}} = X^{p^{d-e}} - X$, we see that $\gcd(X^{p^d} - X, X^{p^e} - X) = \gcd(X^{p^{d-e}} - X, X^{p^e} - X)$. Since $\max\{d - e, e\} < d$, the induction hypothesis implies that $\gcd(X^{p^d} - X, X^{p^e} - X) = X^{p^{\gcd(d-e,e)}} - X = X^{p^{\gcd(d,e)}} - X$, where in the last equality we used that $\gcd(d-e, e) = \gcd(d, e)$.

If e divides d, then $\gcd(d, e) = e$. Hence in this case $X^{p^e} - X$ divides $X^{p^d} - X$. ■

Now we strengthen Lemma 9.5.1.

Lemma 9.5.4 *Let p be prime number and d a positive integer. Let $g(X) \in \mathbb{F}_p[X]$ an irreducible polynomial of degree e dividing d, then $g(X)$ divides $X^{p^d} - X$. Moreover, if $g(X)$ divides $X^{p^f} - X$ for some positive integer f, then $f \geq e$.*

Proof. The first half of the lemma follows directly from Lemma 9.5.1, since $X^{p^e} - X$ divides $X^{p^d} - X$ by Lemma 9.5.3. Note that we can use Lemma 9.5.3, since we assume that e divides d.

Now suppose that $g(X)$ divides $X^{p^f} - X$. Define $\mathbb{F} := \mathbb{F}_p[X]/\langle g(X) \rangle$. Since $g(X) \in \mathbb{F}_p[X]$ is an irreducible polynomial of degree e, we know that $\mathbb{F}$ is a finite field with p^e elements. We know from Lemma 9.2.13 that the element $\alpha := X + \langle g(X) \rangle \in \mathbb{F}_p[X]/\langle g(X) \rangle$ is a root of

$g(X)$ when viewed as polynomial in $\mathbb{F}[X]$. Since $g(X)$ divides $X^{p^f} - X$, we can conclude that α is also a root of $X^{p^f} - X$, that is to say, we have $\alpha^{p^f} = \alpha$. We claim that in fact any element of $\mathbb{F}$ is a root of $X^{p^f} - X$. Indeed, any element of $\mathbb{F}$ can be written in standard form: $a_0 + a_1 X + \cdots + a_{e-1}X^{e-1} + \langle g(X)\rangle = a_0 + a_1\alpha + \cdots + a_{e-1}\alpha^{e-1}$, with $a_0, \ldots, a_{e-1} \in \mathbb{F}_p$. Using the Freshman's dream, see Exercise 18, we see that

$$(a_0 + a_1\alpha + \cdots + a_{e-1}\alpha^{e-1})^p = a_0^p + a_1^p\alpha^p + \cdots + a_{e-1}^p\alpha^{(e-1)p} = a_0 + a_1\alpha^p + \cdots + a_{e-1}\alpha^{(e-1)p},$$

where we used that $a^p = a$ for any $a \in \mathbb{F}_p$. Taking the p-th power f times and using that $\alpha^{p^f} = \alpha$, this implies that

$$(a_0 + a_1\alpha + \cdots + a_{e-1}\alpha^{e-1})^{p^f} = a_0 + a_1\alpha + \cdots + a_{e-1}\alpha^{e-1},$$

implying that indeed any of the p^e elements in $\mathbb{F}$ is a root of $X^{p^f} - X$. Since a polynomial of degree D with coefficients in a field can have at most D roots in that field, we can conclude that $p^e \leq p^f$ and hence that $f \leq e$. ■

A surprising result is that the converse is true as well.

Theorem 9.5.5 *Let p be a prime number and d a positive integer, and $g(X) \in \mathbb{F}_p[X]$ an irreducible polynomial. Then $g(X)$ divides $X^{p^d} - X$ if and only if $\deg g(X)$ divides d.*

Proof. Lemma 9.5.4 proves the implication from right to left. Now let $g(X) \in \mathbb{F}_p[X]$ be an irreducible polynomial of degree e dividing $X^{p^d} - X$. The second half of Lemma 9.5.4 implies that $e \leq d$. We will show using strong induction on d that e divides d. If $d = 1$, then $e = 1$ and we are done. Now suppose the result holds for up to $d-1$ for some $d > 1$ and that $g(X) \in \mathbb{F}_p[X]$ is an irreducible polynomial of degree e dividing $X^{p^d} - X$. If e divides d, there is nothing to show. Hence assume that e does not divide d. From Lemma 9.5.1, we see that $g(X)$ divides $X^{p^e} - X$. Hence $g(X)$ divides $\gcd(X^{p^e} - X, X^{p^d} - X)$. Hence Lemma 9.5.3 implies that $g(X)$ divides $X^{p^{\gcd(d,e)}} - X$. Since $e \leq d$, we either have $e = d$, or $\gcd(d, e) \leq e < d$. In the first case, clearly e divides d, in the second case the induction hypothesis applies and we can conclude that e divides $\gcd(d, e)$, which is only possible if e divides d. ■

Definition 9.5.6 *Let p be a prime number and d a positive integer. Then we define $I_d(p)$ to be the number of monic, irreducible polynomials in $\mathbb{F}_p[X]$ of degree d.*

Theorem 9.5.7 *Let p be a prime number and d a positive integer. Then $p^d = \sum_{e|d} eI_e(p)$, where the sum is over all positive integers e dividing d. Moreover, $I_1(p) = p$.*

Proof. By Theorem 9.5.5, the polynomial $X^{p^d} - X$ is the product of all monic, irreducible polynomials in $\mathbb{F}_p[X]$ of degree dividing d. Comparing degrees, we see that $\deg(X^{p^d} - X) = \sum_{e|d} eI_e(p)$. The last part of the theorem follows because the monic, irreducible polynomials in $\mathbb{F}_p[X]$ of degree 1 are just the polynomials of the form $X + a$, with $a \in \mathbb{F}_p$. ■

This theorem gives a recursive formula for the numbers $I_d(p)$. For example for $d = 2$, we find that $p^2 = 2I_2(p) + I_1(p) = 2I_2(p) + p$ and hence

$$I_2(p) = \frac{p^2 - p}{2}. \tag{9.2}$$

Similarly, one can for example obtain

$$I_3(p) = \frac{p^3 - p}{3} \quad I_4(p) = \frac{p^4 - p^2}{4} \quad I_5(p) = \frac{p^5 - p}{5} \text{ and } I_6(p) = \frac{p^6 - p^3 - p^2 + p}{6}. \tag{9.3}$$

It is now not so hard anymore to show that for any prime number p and any positive integer d, there exists an irreducible polynomials in $\mathbb{F}_p[X]$ of degree d, or equivalently that $I_d(p) > 0$.

Corollary 9.5.8 *Let p be a prime number and d a positive integer, then $I_d(p) > 0$. In particular, there exists a finite field with p^d elements.*

Proof. The second part follows directly from the first part, since once we have an irreducible polynomial $g(X) \in \mathbb{F}_p[X]$ of degree d, the quotient ring $(\mathbb{F}_p[X]/\langle g(X)\rangle, +, \cdot)$ will be a finite field with p^d elements.

To show $I_d(p) > 0$, first observe that Theorem 9.5.7 implies that $dI_d(p) \leq p^d$ for all positive integers d. Using this in combination with Theorem 9.5.7, we find that

$$p^d \leq dI_d(p) + \sum_{e=1}^{d-1} eI_e(p) \leq dI_d(p) + \sum_{e=1}^{d-1} p^e = dI_d(p) + \frac{p^d - p}{p-1}.$$

Hence

$$dI_d(p) \geq p^d - \frac{p^d - p}{p-1} = \frac{p^d(p-2) + p}{p-1} > 0.$$

■

Remark 9.5.9 *The above argument can easily be modified to count the number of polynomials with coefficients in a finite field $\mathbb{F}_q$. If one defines $I_q(d)$ to be the number of monic, irreducible polynomials in $\mathbb{F}_q[X]$ of degree d, then $I_q(1) = q$ and $q^d = \sum_{e|d} eI_e(q)$ just as in Theorem 9.5.7.*

Aside 9.5.10 *This corollary together with Theorem 9.5.2 shows that finite fields with $q = p^d$ elements exists and are unique up to isomorphism. For this reason many authors talk about* the *finite field with q elements. However, in some areas of mathematics, in particular in category theory, a more strict notion of uniqueness is used. There an object with certain defining properties is called universal if any two objects having the defining properties are isomorphic up to unique isomorphism. In other words, not only should the objects be isomorphic, there should be only one possible choice for the isomorphism.*

If we apply this stricter uniqueness criterion, finite fields with p^d elements are not universal unless $d = 1$. This stems from the fact that if $d > 1$ there exist non-trivial isomorphisms from a finite field with p^d elements to itself (such isomorphisms are called automorphisms). More precisely, the Frobenius map mentioned in Exercise 19, is such an automorphism. In fact one can show that powers of the Frobenius map are all possible automorphisms of a finite field and therefore that a finite field with p^d elements has exactly d automorphisms, including the identity.

The characteristic $\mathrm{char}\mathbb{F}$ of a field $\mathbb{F}$ is the additive order of its 1-element if this order is finite. Otherwise one defines $\mathrm{char}\mathbb{F} = 0$. Since a finite field only has finitely many elements, its

characteristic is positive. If char$\mathbb{F} = n > 0$, then $\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{n \text{ times}} = 0_{\mathbb{F}}$, while $\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{k \text{ times}} \neq 0_{\mathbb{F}}$ for all k between 1 and $n-1$. Now suppose that $n = k \cdot \ell$ with $k > 1$ and $\ell > 1$. Then using the distributive laws:

$$(\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{k \text{ times}}) \cdot (\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{\ell \text{ times}}) = \underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{k \text{ times}} = 0_{\mathbb{F}},$$

while $\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{k \text{ times}} \neq 0_{\mathbb{F}}$ and $\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{\ell \text{ times}} \neq 0_{\mathbb{F}}$. This would imply that the field contains zero-divisors. Hence, if the characteristic of a field is positive, it needs to be a prime number.

This has an interesting consequence: a finite field $\mathbb{F}$ always contains a field $\mathbb{F}_p$ as a *subfield*, where p is the characteristic, where a subfield of a field $(\mathbb{F}, +, \cdot)$ is a subring that is a field as well. The elements in this subfield are simply the elements in the set $\{\underbrace{1_{\mathbb{F}} + \ldots + 1_{\mathbb{F}}}_{k \text{ times}} \mid 0 \leq k \leq p-1\}$. But this means that we can view a finite field with characteristic p as a vector space over $\mathbb{F}_p$. Choosing a basis $\{b_1, \ldots, b_d\}$ of $\mathbb{F}$ when viewed as vector space over $\mathbb{F}_p$, we can then conclude that $\mathbb{F} = \{a_1 b_1 + \cdot + a_d b_d \mid a_1, \ldots, a_d \in \mathbb{F}_p\}$. Hence $\mathbb{F}$ contains exactly p^d elements. This shows that the cardinality of an arbitrary finite field is of the form p^d, where p is a prime number. Hence we have constructed all possible finite fields in the previous!

As a last remark on finite fields, we will discuss containment among finite fields. We have just seen, and indeed used implicitly before, that a finite field $\mathbb{F}_{p^d}$ contains $\mathbb{F}_p$ as a subfield. It is natural to ask though if $\mathbb{F}_{p^d}$ can contain other finite fields as subfields. The following theorem answers this question.

Theorem 9.5.11 *Let p be a prime number and d a positive integer. Any subfield of $(\mathbb{F}_{p^d}, +, \cdot)$ is isomorphic to a field $(\mathbb{F}_{p^e}, +, \cdot)$, where e divides d. Conversely, for any e dividing d, there exists precisely one subfield of $(\mathbb{F}_{p^d}, +, \cdot)$ isomorphic to $(\mathbb{F}_{p^e}, +, \cdot)$.*

Proof. Clearly, a subfield $(\mathbb{F}, +, \cdot)$ of $(\mathbb{F}_{p^d}, +, \cdot)$ is a finite field. Moreover, the subfield will contain $(\mathbb{F}_p, +, \cdot)$, since 1 must be in the subfield $\mathbb{F}$. This shows that $|\mathbb{F}| = p^e$ for some $e \leq d$. We have seen that any element of $\mathbb{F}_{p^d}$ is a root of the polynomial $X^{p^d} - X$. Similarly any element of $\mathbb{F}$ is a root of the polynomial $X^{p^e} - X$. Since $\mathbb{F} \subseteq \mathbb{F}_{p^d}$, this implies that $X^{p^e} - X$ divides $X^{p^d} - X$. Lemma 9.5.3 then implies that e divides d. It also shows that $\mathbb{F}_{p^d}$ can contain only one subfield with p^e elements, since the elements of such a subfield need to be the roots of the polynomial $X^{p^e} - X$.

Now assume that e divides d. Then $X^{p^e} - X$ divides $X^{p^d} - X$, which implies that $X^{p^e} - X$ has precisely p^e roots in $\mathbb{F}_{p^d}$. Denote by $\mathbb{F}$ the set of these roots. We claim that $(\mathbb{F}, +, \cdot)$ is a field. Indeed, using the Freshman's dream, we see that if $a, b \in \mathbb{F}$, then $(a-b)^{p^e} = a^{p^e} - b^{p^e} = a - b$, implying that $(\mathbb{F}, +)$ is a subgroup of $(\mathbb{F}_{p^d}, +)$. Further, if $a, b \in \mathbb{F} \setminus \{0\}$, then $(ab^{-1})^{p^e} = a^{p^e}(b^{p^e})^{-1} = ab^{-1}$, implying that $(\mathbb{F} \setminus \{0\}, \cdot)$ is a subgroup of $(\mathbb{F}_{p^d} \setminus \{0\}, \cdot)$. This is enough to conclude that $(\mathbb{F}, +, \cdot)$ is a subfield of $(\mathbb{F}_{p^d}, +, \cdot)$ with p^e elements. ■

For example $\mathbb{F}_{p^2}$ is not a subfield of $\mathbb{F}_{p^3}$, but $\mathbb{F}_{p^2}$ is a subfield of $\mathbb{F}_{p^4}$.

9.6 Exercises

Multiple choice exercises

1. Let n be a positive integer and $a \in \mathbb{Z}$. Then
   A. $a + n\mathbb{Z}$ is a zero-divisor of the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ if and only if $\gcd(a, n) = 1$
   TRUE ☐ FALSE ☐
   B. $a + n\mathbb{Z}$ is a unit of the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ if and only if $\gcd(a, n) = 1$
   TRUE ☐ FALSE ☐
   C. $a + n\mathbb{Z}$ is the zero-element of the ring $(\mathbb{Z}/n\mathbb{Z}, +, \cdot)$ if and only if $\gcd(a, n) = n$
   TRUE ☐ FALSE ☐

2. Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X), g(X) \in \mathbb{F}[X]$ with $f(X)$ non-constant of degree d. You are asked to decide whether $g(X) + \langle f(X)\rangle \in \mathbb{F}[X]/\langle f(X)\rangle$ is the zero element, or a zero-divisor, or a unit. Then you would
   A. ☐ compute $m := \gcd(f(X), g(X))$ using the Euclidean algorithm and see whether $\deg(m) = d$, or $0 < \deg(m) < d$ or $\deg(m) = 0$
   B. ☐ try to find $h(X) \in \mathbb{F}[X]$ such that $(g(X) + \langle f(X)\rangle) \cdot (h(X) + \langle f(X)\rangle) = 0_{\mathbb{F}} + \langle f(X)\rangle$
   C. ☐ try to find $h(X) \in \mathbb{F}[X]$ such that $(g(X) + \langle f(X)\rangle) \cdot (h(X) + \langle f(X)\rangle) = 1_{\mathbb{F}} + \langle f(X)\rangle$

3. Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X), g(X) \in \mathbb{F}[X]$ where $f(X)$ is non-constant of degree d. Let $\gcd(f(X), g(X)) = p_1(X) \cdot f(X) + p_2(X) \cdot g(X)$ with $p_1(X), p_2(X) \in \mathbb{F}[X]$, $f(X) = p_3(X) \cdot \gcd(f(X), g(X))$ and $m = \gcd(f(X), g(X))$. Then
   A. when $m = 1$, $g(X) + \langle f(X)\rangle \in \mathbb{F}[X]/\langle f(X)\rangle$ is a unit and its multiplicative inverse is $p_1(X) + \langle f(X)\rangle$
   TRUE ☐ FALSE ☐
   B. when $m = 1$ the inverse of $g(X) + \langle f(X)\rangle \in \mathbb{F}[X]/\langle f(X)\rangle$ is $p_2(X) + \langle f(X)\rangle$
   TRUE ☐ FALSE ☐
   C. when $0 < \deg(m) < d$, $(g(X) + \langle f(X)\rangle) \cdot (p_3(X) + \langle f(X)\rangle) = 0_{\mathbb{F}[X]/\langle f(X)\rangle}$
   TRUE ☐ FALSE ☐

4. Let $(\mathbb{F}, +, \cdot)$ be a field and let $f(X) \in \mathbb{F}[X]$. Then $f(X)$ is *irreducible* if
   A. ☐ $f(X)$ cannot be written as the product of two polynomials. Equivalently $\mathbb{F}[X]/\langle f(X)\rangle$ is a PID
   B. ☐ $f(X)$ cannot be written as the product of two polynomials unless one of the two is constant. Equivalently $\mathbb{F}[X]/\langle f(X)\rangle$ is a field
   C. ☐ $f(X)$ has no roots in $\mathbb{F}$

5. Let $f(X) \in \mathbb{F}_p[X]$ where p is a prime number and $d = \deg(f(X))$. Then
   A. the quotient ring $\mathbb{F}_p[X]/\langle f(X)\rangle$ has p^d elements
   TRUE ☐ FALSE ☐
   B. if $f(X)$ is irreducible, then $\mathbb{F}_p[X]/\langle f(X)\rangle$ is a finite field with p^{d-1} elements
   TRUE ☐ FALSE ☐

C. when $d = 2, 3$, it is enough to understand whether $f(X)$ has roots in $\mathbb{F}_p$ to check whether $f(X)$ is irreducible
TRUE ☐ FALSE ☐

6. You are given a finite field $(\mathbb{F}_q, +, \cdot)$ and an element $a \in \mathbb{F}_q$. You need to decide whether a is a *primitive element* in $\mathbb{F}_q$. Then
   A. ☐ you check that $a \neq 0$ and $a^{q-1} = 1$
   B. ☐ you check that $a \neq 0$ and $a^m \neq 1$ for all $m \neq q-1$ divisor of $q-1$
   C. ☐ you check that a has order dividing $q-1$

Exercises to get to know the material better

7. Let $(\mathbb{F}, +, \cdot)$ be a field and let $a \in \mathbb{F}$ be a given element.
   - Compute the kernel and the image of the ring homomorphism $\psi : \mathbb{F}[X] \to \mathbb{F}$ defined by $\psi(p(X)) := p(a)$
   - show that the quotient ring $(\mathbb{F}[X]/\langle X - a\rangle, +, \cdot)$ is isomorphic to $(\mathbb{F}, +, \cdot)$.

8. Let $\mathbb{F}_5 = \mathbb{Z}_5$. Then $(\mathbb{F}_5, +, \cdot)$ is a finite field with 5 elements. In this exercise several elements from the quotient ring $(\mathbb{F}_5[X]/\langle X^2 + 3X + 2\rangle, +, \cdot)$ are investigated.
   - Is the element $X^3 + 2X^2 + 4X + 3 + \langle X^2 + 3X + 2\rangle$ the zero-element, a zero-divisor or a unit of the quotient ring $(\mathbb{F}_5[X]/\langle X^2 + 3X + 2\rangle, +, \cdot)$?
     [**Hint:** write the element in standard form]
   - Show that the element $r := X^2 + X + \langle X^2 + 3X + 2\rangle$ is a zero-divisor and find $s \in \mathbb{F}_5[X]/\langle X^2 + 3X + 2\rangle$ such that $r \cdot s = 0 + \langle X^2 + 3X + 2\rangle$ but $s \neq 0 + \langle X^2 + 3X + 2\rangle$
   - Show that the element $u := X + \langle X^2 + 3X + 2\rangle$ is a unit and find $v \in \mathbb{F}_5[X]/\langle X^2 + 3X + 2\rangle$ such that $u \cdot v = 1 + \langle X^2 + 3X + 2\rangle$.

9. Is $(\mathbb{F}_3[X]/\langle X^3 + X + 1\rangle, +, \cdot)$ a field?

10. Write the following polynomials as a product of irreducible polynomials
    - $X^2 + 1 \in \mathbb{F}_2[X]$
    - $X^2 + 1 \in \mathbb{F}_3[X]$
    - $X^2 + 1 \in \mathbb{F}_5[X]$

Exercises to get around in the theory

11. - Find irreducible polynomials $f(X) \in \mathbb{F}_2[X]$ of degrees $2, 3$ and 4
      [**Hint:** to find an irreducible polynomial of degree 4 with coefficients in $\mathbb{F}_2$, first find all irreducible polynomials of degree 2 with coefficients in $\mathbb{F}_2$]
    - Find an irreducible polynomial $f(X) \in \mathbb{F}_3[X]$ of degree 2. Show using the division algorithm that $f(X)$ divides the polynomial $X^9 - X \in \mathbb{F}_3[X]$.
      Remark: more generally for any prime p and any irreducible polynomial $f(X) \in \mathbb{F}_p[X]$ of degree d, it can be shown that $f(X)$ divides the polynomial $X^{p^d} - X \in \mathbb{F}_p[X]$.

12. Let $(\mathbb{F}_2, +, \cdot)$ be the finite field with two elements and let us write $\mathbb{F}_2 = \{0, 1\}$. We define $R := \mathbb{F}_2[X]/\langle X^3 \rangle$.

    - How many elements does R have? Why?
    - Find a zero-divisor in the ring $(R, +, \cdot)$
    - Find the multiplicative inverse of $X + 1 + \langle X^3 \rangle \in R$
    - How many elements does R^* contain?

    [**Hint:** Use the characterization in Proposition 9.1.2]

13. Let $(\mathbb{F}_q, +, \cdot)$ be a finite field of odd order q. An element $a \in \mathbb{F}_q^*$ is a *quadratic residue* in $\mathbb{F}_q$ if the binomial $X^2 - a \in \mathbb{F}_q[X]$ has roots in $\mathbb{F}_q$. Prove that

    - the number of quadratic residues is equal to $(q-1)/2$
    - a is a quadratic residue if $a^{(q-1)/2} = 1$, and it is not a quadratic residue if $a^{(q-1)/2} = -1$.

14. Check that $(\mathbb{F}_2[X]/\langle X^4 + X + 1 \rangle, +, \cdot)$ is a field and find a primitive element of it.

    [**Hint:** use that in Exercise 11, you found that X^2+X+1 is the only irreducible polynomial of degree 2 in $\mathbb{F}_2[X]$]

15. If $f(X) \in \mathbb{Z}_3[X]$ is an irreducible polynomial of degree 3, prove that either $X + \langle f(X) \rangle$ or $2X + \langle f(X) \rangle$ is a primitive element in $\mathbb{Z}_3[X]/\langle f(X) \rangle$.

    [**Hint:** can $X + \langle f(X) \rangle$ have order 2? If not, which order can it have?]

16. 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?

    [**Hint:** as in Exercise 14, use that $X^2 + X + 1$ is the only irreducible polynomial of degree 2 in $\mathbb{F}_2[X]$]

17. Let p be a prime number. The aim of this exercise is to compute the characteristic of a finite field (see Remark 2 in Exercise 16 from Chapter 7 for the definition of the characteristic of a ring).

    - Compute the characteristic of $\mathbb{F}_q := \mathbb{Z}_p[X]/\langle f(X) \rangle$ where $f(X) \in \mathbb{Z}_p[X]$ is irreducible of degree $n > 0$
    - Show that the sum of elements in a finite field with q elements is one if $q = 2$ and zero otherwise, while the product of all non-zero elements is in both cases equal to -1.

    [**Hint (for the second item):** recall that for a natural number n: $\sum_{i=0}^{n} w^i = (w^{n+1} - 1)/(w-1)$ while $\sum_{i=1}^{n} i = n(n+1)/2$. For the statement about the product, analyze the cases q even and q odd separately. Use the first item of the exercise when q is even, while for q odd show that $a^{(q-1)/2} = -1$ with a primitive element of $\mathbb{F}_q$]

    Remark: When $\mathbb{F}_q = \mathbb{Z}_p$, the second statement in this exercise, i.e. that the product of all non-zero elements in a finite field is equal to -1, is a famous result called *Wilson's theorem*. The most common formulation of Wilson's Theorem states that $(p-1)!$ is congruent to $-1 \pmod{p}$.

18. Let p be a prime number, $(\mathbb{F}_p[X], +, \cdot)$ be the ring of polynomials with coefficients in $\mathbb{F}_p$ and let $f(X) \in \mathbb{F}_p[X]$ be a polynomial of degree $d \geq 1$. Show that for any $a, b \in \mathbb{F}_p[X]/\langle f(X)\rangle$ it holds that $(a+b)^p = a^p + b^p$. You may assume the fact that p divides the binomial coefficients $\binom{p}{i}$ for $1 \leq i \leq p-1$.

    Remark: the result $(a+b)^p = a^p + b^p$ for $a, b \in \mathbb{F}_p[X]/\langle f(X)\rangle$ is known as the *freshman's dream*. This funny name comes from the erroneous equation $(x+y)^k = x^k + y^k$, where $k > 1$ is an integer. Beginning calculus students commonly make this error when computing the power of a sum of real numbers, falsely assuming powers distribute over sums. Of course even in $\mathbb{F}_p[X]/\langle f(X)\rangle$, the element $(a+b)^k$ does not need to be equal to $a^k + b^k$ for all k, but as you proved above, for $k = p$ it is. This means that if $k = p$, we really have the realization of the freshman's dream!

19. Let p be a prime, $f(X) \in \mathbb{F}_p[X]$ a polynomial of degree $d \geq 1$ and $(\mathbb{F}_p[X]/\langle f(X)\rangle, +, \cdot)$ a finite field with $q := p^d$ elements. As is customary, we write $\mathbb{F}_q = \mathbb{F}_p[X]/\langle f(X)\rangle$. Show that the map $\phi : \mathbb{F}_q \to \mathbb{F}_q$, with $\phi(a) = a^p$ is an isomorphism of rings.

    [**Hint:** prove first that ϕ is injective and then use Lemma 2.1.6.]

    Remark: the map ϕ is called the *Frobenius map*.

20. Let $(\mathbb{F}, +, \cdot)$ be a field whose characteristic is not 2. We know already that if $(\mathbb{F}, +, \cdot)$ is a finite field then $(\mathbb{F}^*, \cdot)$ is a cyclic group (see Theorem 9.4.1). In this exercise we show that the inverse implication also holds.

    - Prove that if $\mathbb{F}^*$ is cyclic then $\mathbb{F}$ is a finite field.

    [**Hint:** recall that since the characteristic of $\mathbb{F}$ is not 2, then $1_{\mathbb{F}} \neq -1_{\mathbb{F}}$]

Chapter 10

Solutions to selected exercises

10.1 Chapter 1

- *Ch. 1, Ex. 1* A. TRUE, B. TRUE, C. TRUE, D. FALSE, E. FALSE, F. TRUE, G. FALSE, H. FALSE, I. TRUE.
- *Ch. 1, Ex. 2* B.
- *Ch. 1, Ex. 3* C.
- *Ch. 1, Ex. 4* A. TRUE, B. FALSE, C. TRUE, D. TRUE, E. TRUE.
- *Ch. 1, Ex. 5* A, C, D.
- *Ch. 1, Ex. 6* A, B.
- *Ch. 1, Ex. 7* A. TRUE, B. TRUE, C. FALSE, D. FALSE.
- *Ch. 1, Ex. 8* A. TRUE, B. FALSE, C. TRUE.
- *Ch. 1, Ex. 9* (a) True, since $-2 - 97 = -99 = (-9) \cdot 11$.

  (b) True.

  (c) False, since $10 - 3 = 7$ is not a multiple of 13.

  (d) True: any number is a multiple of 1.

  (e) False.
- *Ch. 1, Ex. 10* Quotient 6, remainder 11.
- *Ch. 1, Ex. 11* The congruence class we are looking for is $11 + 7\mathbb{Z}$. It contains 11 itself and the five element of $11+7\mathbb{Z}$ of smallest absolute value can be found by some trial and error by adding small multiples of 7 to it. We obtain that $11 + 7\mathbb{Z} = \{\ldots, -17, -10, -3, 4, 11 \ldots\}$. The elements written are the five representatives of $11 + 7\mathbb{Z}$ of smallest absolute value.

- *Ch. 1, Ex. 12* The integer 4 is a representative of the congruence class $4 + 11\mathbb{Z}$. On the other hand, since $77 \equiv 0 \pmod{11}$, the integer 77 is a representative of the congruence class $0 + 11\mathbb{Z}$. According to Theorem 1.3.7, the congruence classes $4 + 11\mathbb{Z}$ and $0 + 11\mathbb{Z}$ are distinct, so the answer is no.

- *Ch. 1, Ex. 13* Running the extended Euclidean algorithm, one obtains that $1 = 11 \cdot 101 - 30 \cdot 37$. This implies that $1 \equiv -30 \cdot 37 \equiv 71 \cdot 37 \pmod{101}$. Therefore 71 is the inverse of 37 modulo 101. It is also common and in principle equally correct, to just say that -30 is the inverse of 37 modulo 101.

- *Ch. 1, Ex. 14* One can substitue all values between 0 and 7 in the equation and see what happens. Doing so one finds the correct answer $x = 4$.

- *Ch. 1, Ex. 15* The last digit is 1.

- *Ch. 1, Ex. 16* Zero.

- *Ch. 1, Ex. 17* (a) 163.

  (b) 1.

  (c) 1075. The inverse of 35 is 1093. The equality is false.

- *Ch. 1, Ex. 18*

  (a) For $a \in A \setminus \{0\}$ we have $[a]_\sim = \{a, -a\}$ while $[0]_\sim = \{0\}$.

  (b) For a given $a \in A$ the congruence class $[a]_\sim$ is the set of real numbers b such that $a^2 + a = b^2 + b$, that is, $b^2 + b - a^2 - a = (b - a)(b + a + 1) = 0$. This means that $[a]_\sim = \{a, -a - 1\}$ for all $a \in \mathbb{R}$.

  (c) The congruence class of a point P of the plane $\mathbb{R}^2$ is the set of all points of the plane having the same distance as P from the origin, say d. Geometrically this means that the congruence class of P is nothing else but the points of the circle with center the origin and radius d.

  (d) Let a be a fixed natural number. If a is a square then $[a]_\sim = \{b^2 \mid b \in \mathbb{N}_0\}$. If a is not a square then $[a]_\sim = \{ab^2 \mid b \in \mathbb{N}_0\}$.

  (e) Assume that $(a, b) \in \mathbb{R} \times \mathbb{R}$ is fixed. We want to compute $[(a, b)]_\sim$. This consists of all points (x, y) of the conic $x^2 + y^2 = a^2 + b^2$.

- *Ch. 1, Ex. 19* The statement $a \equiv b \pmod{1}$ is true if and only if $a - b$ is a multiple of 1. Since the identity $a - b = (a - b) \cdot 1$ holds for any integers a and b, we may conclude that $a \equiv b \pmod{1}$ is true for any two integers a and b.

- *Ch. 1, Ex. 20* The fact that $a \equiv b \pmod{n}$ means that there exists an integer k such that $a - b = kn$. Let d be a fixed common divisor of a and b. Then we can also write $a = da_1$ and $b = db_1$ for some a_1 and b_1 integers. This means that

  $$a - b = da_1 - db_1 = d(a_1 - b_1) = kn,$$

  and hence

  $$a_1 - b_1 = \frac{kn}{d}.$$

  Since $\gcd(n, d)$ trivially divides both n and d, we can finally write $n = n_1 \gcd(n, d)$ and $d = d_1 \gcd(n, d)$ where $n_1 = n/\gcd(n, d)$ and $d_1 = d/\gcd(n, d)$ are integers. Then

$$a_1 - b_1 = \frac{kn}{d} = \frac{kn}{d_1 \gcd(n,d)} = \frac{k}{d_1} \cdot \frac{n}{\gcd(n,d)}.$$

Since this shows that $a_1 - b_1$ is a multiple of $n/\gcd(n,d)$ the result follows.

- *Ch. 1, Ex. 21* (a) $10/3 = 3.3333...$ Hence $\lfloor 10/3 \rfloor = 3$. On the other hand, $10 = 3 \cdot 3 + 1$, so that 10 quot 3 = 3. Similarly $-11/5 = -2.2$ and hence $\lfloor -11/5 \rfloor = -3$, while $-11 = -3 \cdot 5 + 4$ so that -11 quot $5 = -3$.

  (b) Directly from the definition of the floor function, we can deduce that $\lfloor x \rfloor \leq x$. If $\lfloor x \rfloor \leq x - 1$, then $1 + \lfloor x \rfloor \leq x$. Since $\lfloor x \rfloor$ by definition is the largest integer less than or equal to x, this is impossible. Hence $\lfloor x \rfloor > x - 1$, which is equivalent to $x - 1 < \lfloor x \rfloor$. This shows that $\lfloor x \rfloor$ is an integer in the interval $]x-1, x]$. Since the distance between $x-1$ and x is 1, it is intuitively clear that there can not be two distinct integers in $]x-1, x]$. To make this precise, assume that n is another integer in $]x-1, x]$. Then $n \leq x$, implying that $n < \lfloor x \rfloor$, since $\lfloor x \rfloor$ is the largest integer less than or equal to x. This would imply that $x - 1 < n < \lfloor x \rfloor \leq x$ and hence that $1 = x - (x-1) > \lfloor x \rfloor - n \geq 1$. In the last inequality, we used that both $\lfloor x \rfloor$ and n are integers. Hence $1 > 1$, which is absurd. This shows that $\lfloor x \rfloor$ is the only integer in the interval $]x-1, x]$.

  (c) Let $a \in \mathbb{Z}$ and $n \in \mathbb{Z}_{\geq 1}$ be given. Applying part b) to $x = a/n$, we see that $a/n - 1 < \lfloor a/n \rfloor \leq a/n$. Multiplying by n, we deduce that $a - n < \lfloor a/n \rfloor \cdot n \leq a$, which is equivalent to $0 \leq a - \lfloor a/n \rfloor \cdot n < n$. The strict inequality $a - \lfloor a/n \rfloor \cdot n < n$ implies $a - \lfloor a/n \rfloor \cdot n \leq n - 1$, since $a - \lfloor a/n \rfloor \cdot n$ is an integer. This shows that the pair $(q, r) := (\lfloor a/n \rfloor, a - \lfloor a/n \rfloor \cdot n) \in \mathbb{Z}^2$ satisfies the requirements from Fact 1.3.5. Hence $(\lfloor a/n \rfloor, a - \lfloor a/n \rfloor \cdot n) = (a \text{ quot } n, a \bmod n)$.

  We know that there exist unique integers $q = a$ quot n and $r = a \bmod n$ such that $a = q \cdot n + r$ and $0 \leq r < n$. Dividing by n, we obtain that

$$\frac{a}{n} = q + \frac{r}{n} \quad \text{and} \quad 0 \leq \frac{r}{n} < 1.$$

  But then we may conclude that a quot $n = q = \lfloor a/n \rfloor$.

- *Ch. 1, Ex. 22* We check the transitivity in full detail, but leave reflexivity and symmetry to the reader. Suppose that $a \equiv b \pmod{n}$ and $b \equiv c \pmod{n}$. Then there exist integers $k \in \mathbb{Z}$ and $\ell \in \mathbb{Z}$ such that $a - b = k \cdot n$ and $b - c = \ell \cdot n$. This implies that

$$a - c = (a - b) + (b - c) = k \cdot n + \ell \cdot n = (k + \ell) \cdot n.$$

  Hence $a - c$ is a multiple of n, so that we may conclude that $a \equiv c \pmod{n}$.

- *Ch. 1, Ex. 23* We need to show that $(a \cdot_n b) \cdot_n c = a \cdot_n (b \cdot_n c)$.

  By the definition of $\cdot_n$, we know that $a \cdot_n b \equiv a \cdot b \pmod{n}$ and therefore we may conclude that

$$(a \cdot_n b) \cdot c \equiv (a \cdot b) \cdot c \pmod{n}.$$

  Similarly, by definition of $\cdot_n$, we have $(a \cdot_n b) \cdot_n c \equiv (a \cdot_n b) \cdot c \pmod{n}$. Combining the above (using transitivity of the relation $\equiv \pmod{n}$), we see

$$(a \cdot_n b) \cdot_n c \equiv (a \cdot b) \cdot c \pmod{n}.$$

Similarly, one can obtain that

$$a \cdot_n (b \cdot c) \equiv a \cdot (b \cdot c) \pmod{n}.$$

Since for integers, it holds that $a \cdot (b \cdot c) = (a \cdot b) \cdot c$, we derive from the above the identity:

$$(a \cdot_n b) \cdot_n c \equiv a \cdot_n (b \cdot_n c) \pmod{n}, \text{ or equivalently } (a \cdot_n b) \cdot_n c + n\mathbb{Z} = a \cdot_n (b \cdot_n c) + n\mathbb{Z}.$$

Since $(a \cdot_n b) \cdot_n c$ and $a \cdot_n (b \cdot_n c)$ are both integers between 0 and $n-1$, they are equal (using for example Theorem 1.3.7).

- *Ch. 1, Ex. 24* A direct computation shows that the relation is reflexive, symmetric and transitive.

- *Ch. 1, Ex. 25* (a) The relation can be simplified. Indeed we know that the product of two integers is equal to zero if and only if one of the two is equal to zero. Given $(a, b) \sim (c, d)$ means that there must exist an integer $e \neq 0$ that multiplied with the other integer $ad - bc$ is equal to zero. The only way in which this can happen is if $ad - bc = 0$. Hence the equivalence relation reads as $(a, b) \sim (c, d)$ if and only if $ad - bc = 0$. Reflexivity and symmetry can now be simply checked by hands. For the transitivity suppose that $(a, b) \sim (c, d) \sim (f, g)$, where $b, d, g \neq 0$. Then $ad - bc = 0$ and $cg - df = 0$, while we want to show $ag - bf = 0$. The first condition reads $ad = bc$ and the second one $cg = df$. Multiplying the first condition by f gives $adf = bcf$ while multiplying the second by a gives $cga = dfa$. Hence $cga = dfa = bcf$, that is, $c(ga - bf) = 0$. This means that either $ga - bf = 0$ (what we want) or $c = 0$. Suppose that $c = 0$. Then $ad - bc = 0$ and $cg - df = 0$ give $ad = 0$ and $df = 0$. Since $d \neq 0$ we get $a = f = 0$ and as we want $ga - bf = 0$.

  (b) In (a) we showed that the equivalence class of (a, b) with $b \neq 0$ is given by all the (c, d) where $d \neq 0$ and $ad - bc = 0$. Dividing by bd (which is allowed as they are both not zero) gives $a/b = c/d$.

  (c) Existence: consider the equivalence class of (a, b) with $b \neq 0$. If $\gcd(a, b) = 1$ then (a, b) is itself the right representative. Suppose so $\gcd(a, b) = d \neq 1$ and write $a = d \cdot a_1$ and $b = d \cdot b_1$ for some integers a_1, b_1. Note that a_1 and b_1 are coprime. In (b) we showed that the equivalence class of (a, b) is given by all (c, d) with $d \neq 0$ such that $a/b = c/d$. Considering (a_1, b_1) we see that $a/b = a_1/b_1$ and (a_1, b_1) is the right representative.

  Uniqueness: suppose by contradiction there are two such representative (a_1, b_1) and (a_2, b_2) where $\gcd(a_1, b_1) = \gcd(a_2, b_2) = 1$. Then $a_1 b_2 = b_1 a_2$ so that both a_1 and b_2 divides $b_1 a_2$. Since $\gcd(a_1, b_1) = 1 = \gcd(a_2, b_2)$ we deduce that a_1 divides a_2 and b_2 divides b_1. However the same equality $a_1 b_2 = b_1 a_2$ gives also that both b_1 and a_2 divide $a_2 b_2$. For the same reason we get that b_1 divides b_2 and a_2 divides a_1. Since two integers dividing eachother coincide, the proof is complete.

- *Ch. 1, Ex. 27* One can check the axioms of an equivalence relation by hands. The classes are exactly the sets A_i.

- *Ch. 1, Ex. 28* Let (x, y) taken from the putative subset. Then $x \equiv y \pmod 6$, that is, $x - y = 6k$ for some integer k. In particular $x - y$ is a multiple of 3 and hence $x \equiv y \pmod 3$.

10.2 Chapter 2

- *Ch. 2, Ex. 1* A. FALSE, B. FALSE, C. TRUE, D. TRUE.
- *Ch. 2, Ex. 2* B.
- *Ch. 2, Ex. 3* B.
- *Ch. 2, Ex. 4* A. TRUE, B. FALSE, C. FALSE, D. FALSE, E. TRUE, F. TRUE.
- *Ch. 2, Ex. 5* B.
- *Ch. 2, Ex. 6* A. FALSE, B. FALSE.
- *Ch. 2, Ex. 7* A. TRUE, B. TRUE, C. TRUE, D. TRUE.
- *Ch. 2, Ex. 8* A, C, D, E.
- *Ch. 2, Ex. 9* The correct answer is $(1\,2\,3\,7\,6)(4\,5)$. Note that one does not write commas between the numbers in a cycle. So one does not write $(1,2,3,7,6)(4,5)$. Also note that a cycle can be written in several ways, so some different looking answers can be correct nonetheless. The following cycles are in fact all the same: $(1\,2\,3\,7\,6)$, $(6\,1\,2\,3\,7)$, $(7\,6\,1\,2\,3)$, $(3\,7\,6\,1\,2)$, and $(2\,3\,7\,6\,1)$.
- *Ch. 2, Ex. 10* $f \circ g = (1\,6)(2\,4\,5)$ and $g \circ f = (1\,5\,6)(3\,4)$.
- *Ch. 2, Ex. 11* Let $g = (a\,b) \in S_n$ and consider the matrix M_f. Developing after all rows except rows a and b, one ends up showing that $\det(M_f)$ is the same as the determinant of the matrix $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$, which of course is -1.

  Another proof sketch: First show that $\det(M_f) = -1$ if $f = (1\,2)$. If $f = (a\,b)$, an arbitrary 2-cycle of S_n, choose N to be the matrix obtained from the identity matrix by interchanging row 1 and a as well as row 2 and row b. Then $N \cdot M_f \cdot N$ is simply the matrix $M_{(1\,2)}$, while N^2 is the identity matrix. Hence $\det M_f = \det M_{(1\,2)} \det N^2 = \det M_{(1\,2)} = -1$.
- *Ch. 2, Ex. 12* (a) Surjective (because $f[i] = i$ for all $i = 0, 1, 2$) but not injective (indeed $f[0] = f[3] = 0$.

  (b) Not injective (because $f[0] = f[2] = 0$) and not surjective (because $\mathrm{im} f = \{0, 1\}$).

  (c) Bijective and hence a permutation.

  (d) Bijective and hence a permutation.

  (e) Not injective (as $f(1) = f(-1) = 1$) and not surjective (as $\mathrm{im} f = \mathbb{N}_0$).
- *Ch. 2, Ex. 13* (a) Permutation (as it is bijective from $\mathbb{R}$ to itself).

  (b) Suppose that the matrix A is of type

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

  Then $\det A = ad - bc = 0$ as A has rank 1.The map is not injective. Indeed using $ad - bc = 0$ one can see that $f(0,0) = f(d,-c) = (0,0)$. The map is not surjective either because A has rank 1 (and hence the image is properly contained in $\mathbb{R}^{\not=}$).

- *Ch. 2, Ex. 14* (a) (1532).

  (b) (184567).

  (c) (1234).

  (d) (1352).

  (e) (123456).

- *Ch. 2, Ex. 15* (a) It is enough to consider the product of disjoint 2-cycle and 3-cycle. An example is: (12)(345).

  (b) Assume by contradiction that such an element of order 8 exists, say f. We write f as a composition of disjoint cycles. There are few possibilities:

  - A 5-cyclic appears in the decomposition. Then f is a 5-cycle and hence its order is 5, a contradiction.
  - A 4-cycle appears in the decomposition. Then f is a 4-cycle (as only one element is left, and so it is fixed) and its order is 4, a contradiction.
  - A 3-cycle appears in the decomposition. Then either f is a 3-cycle, or the composition of a 3-cycle and a 2-cycle. In the first case the order of f is 3 and in the second case it is 6, a contradiction.
  - A 2-cycle appears in the decomposition (but not a 3-cycle). Then f is the product of 2-cycles and hence its order is 2, a contradiction.
  - None of the last cases occurs. Then f is the identity permutation and hence its order is 1, a contradiction.

  (c) The answer is already given in (b). The possible orders are $1, 2, 3, 4, 5, 6$.

- *Ch. 2, Ex. 16* One can simply check whether $\sum_{i=1,\ldots,n} it_i = n$. (a) No.

  (b) No.

  (c) Yes.

  (d) Yes.

- *Ch. 2, Ex. 17* Once we have the disjoint cycle description computing the order is just computing the least common multiple of the orders of the cycles. (a) 4.

  (b) 6.

  (c) 4.

  (d) 4.

  (e) 6.

- *Ch. 2, Ex. 18* One can use Exercise 15. The largest order is 6.

- *Ch. 2, Ex. 19* (a) First of all, one should show that $f \circ g$ is a function from A to A: Take $a \in A$, then $g[a] \in A$, since g is a function from A to A. This means that $f[g[a]] \in A$, since also f is a function from A to A.

  Next, one should show that $f \circ g$ is a bijection. There are several ways to do this. One is to show that $f \circ g$ is an injective function as well as a surjective function using the definition of injective and surjective. Here is another way: One may use that a function $h : A \to B$ is a bijection if and only if it has an inverse function $h^{-1} : B \to A$. This inverse function has

to satisfy that $h \circ h^{-1} = \mathrm{id}_B$ (the identity function on B) and $h^{-1} \circ h = \mathrm{id}_A$ (the identity function on A). The claim is that $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$. Indeed, if $a \in A$, then

$$\begin{aligned}(f \circ g) \circ (g^{-1} \circ f^{-1}) =& f \circ (g \circ (g^{-1} \circ f^{-1})) = f \circ ((g \circ g^{-1}) \circ f^{-1}) \\ =& f \circ (\mathrm{id} \circ f^{-1}) = f \circ f^{-1} = \mathrm{id}.\end{aligned}$$

Note the we have used associativity of the operator $\circ$ several times. Also we have used that $\mathrm{id} \circ f^{-1}$, which follows from the previous part of this exercise. Similarly one can show that $(g^{-1} \circ f^{-1}) \circ (f \circ g) = \mathrm{id}$, but that part of the solution is left out here.

(b) For any $a \in A$, we have $(\mathrm{id} \circ f)[a] = \mathrm{id}[f[a]] = f[a]$, since the identity function fixes every element of A, hence also $f[a]$. This means that the permutations $\mathrm{id} \circ f$ and f have the same effect on the elements of A. Hence $\mathrm{id} \circ f = f$. Similarly, one shows $f \circ \mathrm{id} = f$, but that solution is left out here.

- *Ch. 2, Ex. 20* There are 12 rotational symmetries. The rotation axes either connect midpoints of opposite edges or connect a vertex with the midpoint of the opposing face. Hence not all $4! = 24$ elements in S_4 are obtained. The set of permutations that are obtained, is

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

- *Ch. 2, Ex. 21* (a) $(a\,b\,c) = (a\,b)(b\,c)$. Another possibility is $(a\,b\,c) = (a\,c)(a\,b)$.

  (b) $(a_1\,a_2 \ldots a_m) = (a_1\,a_2)(a_2\,a_3) \ldots (a_{m-1}\,a_m)$. A proof can be obtained by either arguing that these two permutation have the same effect on any element of A, or using induction on m. In the induction step one would need to prove that $(a_1\,a_2 \ldots a_m) = (a_1\,a_2 \ldots a_{m-1}) \circ (a_{m-1}\,a_m)$. Another correct answer is $(a_1\,a_2 \ldots a_m) = (a_1\,a_m) \ldots (a_1\,a_3)(a_1\,a_2)$.

  (c) Since any permutation can be written as the composite of (mutually disjoint) cycles, the follows from the previous.

  (d) A mirror symmetry of the tetrahedron gives rise to a 2-cycle by considering its effect/action on the 4 vertices of the tetrahedron. Moreover, all six 2-cycles in S_4 are obtained in this way. By the previous, any permutation in S_4 can be written as the composite of 2-cycles. Hence we can conclude that any rotational symmetry of the tetrahedron can be written as a composite of mirror symmetries of the tetrahedron.

- *Ch. 2, Ex. 22* Let us write $f = (a_1\,a_2 \ldots a_m)$ and $g = (b_1\,b_2 \ldots b_n)$. One can divide up that set A into three disjoint subsets: $A_1 = \{a_1, \ldots, a_m\}$, $A_2 = \{b_1, \ldots, b_n\}$ and $A_3 = A \backslash (A_1 \cup A_2)$. Note that since f and g are disjoint cycles, we have $A_1 \cap A_2 = \emptyset$. Now prove that for any $a \in A$ one has $(f \circ g)[a] = (g \circ f)[a]$ by first considering $a \in A_1$, then $a \in A_2$ and lastly $a \in A_3$.

- *Ch. 2, Ex. 23* The matrices M_f and M_g contain exactly one 1 per row or column and otherwise zeroes. This makes the formula $(A \cdot B)_{ij} = A_{i1}B_{1j} + A_{i2}B_{2j} + \cdots + A_{in}B_{nj}$ useful. Since the only nonzero entry in the i-th column of M_f is $(M_f)_{if[i]}$, we find

  $$(M_f \cdot M_g)_{ij} = (M_f)_{i1}(M_g)_{1j} + (M_f)_{i2}(M_g)_{2j} + \cdots + (M_f)_{in}(M_g)_{nj} = (M_f)_{if[i]}(M_g)_{f[i]j}.$$

  This could still be zero, since $(M_g)_{f[i]j}$ could be zero. However, $(M_g)_{f[i]j} = 1$ if and only if $g[f[i]] = j$. Therefore $(M_f \cdot M_g)_{ij} = 1$ if $g[f[i]] = j$ and zero otherwise. However, $g[f[i]] = (g \circ f)[i]$, so apparently $(M_f \cdot M_g)_{ij} = (M_{g \circ f})_{ij}$. Since i and j were arbitrary, we have shown that $M_f \cdot M_g = M_{g \circ f}$.

- *Ch. 2, Ex. 25* The sign of a 2-cycle is -1. Since $\text{sign}(f \circ g) = \text{sign}(f)\text{sign}(g)$, this implies that if a permutation is the composition of ℓ 2-cycles, its sign is $(-1)^\ell$. Since $\text{sign}(\,(123)\,) = 1$ it can therefore not be written as the composition of an odd number of 2-cycles.

- *Ch. 2, Ex. 26* The proof can be given by induction. It is enough to prove the result for all positive integers as if $n > 0$ is the product of primes, so is $-n$. We start with the base case $n = 2$. Then n is a prime and there is nothing to prove (a prime is trivially a product of primes!).

  Suppose that $2 < n$ and that we have proved the result for all numbers m such that $2 \leq m < n$. We wish to show that n is a product of primes. If n is a prime, there is nothing to do. If n is not a prime, then we can write $n = m \cdot t$, where $2 \leq m, t < n$. By the induction step both m and t are products of primes and thus so is n.

- *Ch. 2, Ex. 27* (a) An example can be constructed as follows. Let $\alpha = (1\,2)(3\,4)(5\,6)$, which is in S_n as $n \geq 20$. Then α has order 2 (the least common multiple of the orders of the disjoint cycles). Check that $\alpha^3 = \alpha$. Since α itself is the product of 3 non-trivial disjoint cycles, so is α^3.

  (b) If $\alpha = (123\ldots19\,20)$ then α is in S_n as $n \geq 20$. One has

  $$\alpha^{12} = (1\,13\,5\,17\,9)(2\,14\,6\,18\,10)(3\,15\,7\,19\,11)(4\,16\,8\,20\,12).$$

- *Ch. 2, Ex. 28* If the two permutations have the same cycle type, then they can be written structurally in the same way as product of disjoint cycles. The order of a cycle is nothing else but its length, and we know that the order of the product of disjoint cycles is just the least common multiple of their orders. Since these quantities coincide for two permutations with the same cycle type, they must have the same order.

- *Ch. 2, Ex. 29* We can write our cycle as $f = (a_0\, a_1\, \ldots\, a_{m-1})$ where $m = 2k + 1$ and $k \geq 1$. Why is f^2 an m-cycle as well? We can start with a_0 trying to write the disjoint cycle decomposition of f^2. Doing so we obtain $f^2 = (a_0\, a_2\, \ldots, a_{2k}\, a_1 \ldots a_3 \ldots a_{2k+1})$. Since m is odd this cycle has length m, and hence its order is equal to the one of f.

  This is not true if m is even. What can go wrong? An easy example is given by taking a 2-cycle: $f = (1\,2)$, so that $f^2 = \text{id}$. If we take $f = (1\,2\,3\,4)$ then $f^2 = (1\,3)(2\,4)$ and the orders are different again. The general reason is that if m is even then f^2 has order $m/2$.

- *Ch. 2, Ex. 30*

  1. If f has a left inverse g and $f[a_1] = f[a_2]$, then $a_1 = (g \circ f)[a_1] = g[f[a_1]] = g[f[a_2]] = (g \circ f)[a_2] = a_2$. Hence f is injective. Conversely, if f is injective, we can define a function $h : \text{im}(f) \to A$ by $f[a] \mapsto a$. The function h is well defined, since the preimage of $f[a]$ under f only contains a (using the injectivity of f). Now we can extend the function $h : \text{im}(f) \to A$ to a function $g : B \to A$, for example by mapping $b \in B \setminus \text{im}(f)$ to a fixed chosen $a_0 \in A$. Then g is a left inverse of f, since for any $a \in A$, we have $(g \circ f)[a] = g[f[a]] = a$.
  2. If f has a right inverse g, and b is an arbitrarily chosen element of B, then $b = (f \circ g)[b] = f[g[b]]$. Hence f maps $g[b]$ to b. Since $b \in B$ was arbitrary, this shows that f is surjective. Conversely, assume that f is surjective and let $g : B \to A$ be a function that maps $b \in B$ to a preimage of b under f. Since f is surjective, at least one such a preimage always exist. Then for any $b \in B$, we have $(f\ circg)[b] = f[g[b]] = b$,

where the last equality follows since $g[b]$ lies in the preimage of b under f. Hence f has a right inverse.

10.3 Chapter 3

- *Ch. 3, Ex. 1* A. TRUE, B. FALSE, C. FALSE, D. TRUE.
- *Ch. 3, Ex. 2* B,C.
- *Ch. 3, Ex. 3* A. TRUE, B. TRUE, C. FALSE, D. TRUE.
- *Ch. 3, Ex. 4* A, C.
- *Ch. 3, Ex. 5* A. FALSE, B. TRUE, C. TRUE, D. FALSE.
- *Ch. 3, Ex. 6* A. TRUE, B. FALSE, C. TRUE, D. TRUE.
- *Ch. 3, Ex. 7* C.
- *Ch. 3, Ex. 8* No, it is not. There are no negative numbers in $\mathbb{N}$ and therefore not always inverse elements.
- *Ch. 3, Ex. 9* It is. Multiplication of real numbers is an associative operation (you do not have to show this, but may assume this to be true). Moreover, the product of positive real numbers is a positive real number again, so the operation $\cdot$ gives a group operation which is associative. The identity element is 1. Finally, the inverse of a positive real number r is $r^{-1} = 1/r$. Note that if $r > 0$, then $1/r > 0$ as well.
- *Ch. 3, Ex. 10* No, indeed $3 \cdot_6 2 = 0$ and 0 is not an element of the given set.
- *Ch. 3, Ex. 11* (a) Denote with n the order of f. Note that, from the definition of order and the associative property,

$$(f^{-1})^n = e \cdot (f^{-1})^n = f^n \cdot (f^{-1})^n = (f \cdot \ldots \cdot (f \cdot (f \cdot f^{-1}) \cdot f^{-1}) \cdot \ldots f^{-1}) = e.$$

Hence the order of f^{-1} divides n. Suppose by contradiction that it is smaller than n, say m. Using a similar trick as before

$$f^m = f^m \cdot (f^{-1})^m = (f \cdot \ldots (f \cdot (f \cdot f^{-1}) \cdot f^{-1}) \ldots \cdot f^{-1}) = e.$$

This is not possible as $m < n$ and n is the order of f.

(b) Let n be the order of $f \cdot g$. Then since $(f\dot{g})^n = e$ we have $(f \cdot g)^{n-1} = (f \cdot g)^{-1} = g^{-1} \cdot f^{-1}$ and

$$(g \cdot f)^n = g \cdot (f \cdot g) \cdot \ldots \cdot (f \cdot g) \cdot f = g \cdot (f \cdot g)^{n-1} \cdot f = g \cdot (g^{-1} \cdot f^{-1}) \cdot f = e.$$

Hence the order of $g \cdot f$ divides n. Suppose by contradiction that it is smaller than n, say m. Using a similar trick as before we get

$$(f \cdot g)^m = f \cdot (g \cdot f) \cdot \ldots \cdot (g \cdot f) \cdot g = f \cdot (g \cdot f)^{m-1} \cdot g = f \cdot (f^{-1} \cdot g^{-1}) \cdot g = e.$$

This is not possible as $m < n$ and n is the order of $f \cdot g$.

For the second statement, one can simply use what just proved for $f_1 \cdot g_1$ and $g_1 \cdot f_1$ where $f_1 = f$ and $g_1 = g \cdot h$.

(d) Take for example $f = (123)$, $g = (12)$ and $h = (13)$. Then

$$f \cdot g \cdot h = (123)(12)(13) = Id$$

while

$$g \cdot f \cdot h = (12)(123)(13) = (123).$$

Hence $f \cdot g \cdot h$ has order 1 while $g \cdot f \cdot h$ has order 3.

- *Ch. 3, Ex. 12*

  (a) Since $r^{91} = e$, we have $r^{98} = r^7$. Using that 7 divides 91, we conclude from Theorem 3.2.5 that the order of r^7 is $91/7 = 13$. Hence r^{98} is not a generator of C_{91}, since a generator needs to have order 91.

  (b) Since $sr^i = r^{-i}s$ for all $i \in \mathbb{Z}$, $r^{91} = e$, and $s^2 = e$, we obtain

  $$r^{98}sr^{-17}s^3 = r^7r^{17}ss^3 = r^{24}s^0.$$

  Hence we obtain $i = 24$ and $j = 0$.

- *Ch. 3, Ex. 13* We can choose $a_1 = e$. If $n = 1$ then we are done as $c = a_1 = e$. So we assume $n \geq 2$. We distinguish two cases: (a) There is not $i = 2, \ldots, n$ such that $a_i^{-1} = a_i$ and (b) there is some $i = 2, \ldots, n$ such that $a_i^{-1} = a_i$. Suppose that (a) occurs. Since for each i there exists a unique $j \neq i$ such that $a_j = a_i^{-1}$ we see every element and its inverse appears in the product $a_1 \cdot a_2 \cdot \ldots \cdot a_n$, and hence this product c must be the identity element e (recall that the group is abelian!).

  Suppose now that (b) occurs. The difference with respect to case (a) is that for all $i = 2, \ldots, n$ such that $a_i = a_i^{-1}$, a_i and a_i^{-1} do not appear at the same time in c (as they coincide and each element in G appears only once in c). Every other element in G for which this is not the case cancels exactly as in case (a). This means that c is nothing else but the product of all $i = 2, \ldots, n$ such that $a_i = a_i^{-1}$ and hence if we multiply c by itself we obtain the identity (recall that G is abelian!).

- *Ch. 3, Ex. 14* (a) One can simply check all the axioms where the identity element is given by $(0, 1)$, since $(x, s) \cdot (0, 1) = (x, s) = (0, 1) \cdot (x, s)$.

  (b) To show that an element of type $(x, -1)$ with $x \neq 0$ has order 2, it is sufficient to note that it does not have order 1 (as the identity element, only element of order 1, is $(0, 1)$) and that $(x, -1) \cdot (x, -1) = (0, 1)$. On the other hand $(x, 1) \cdot (x, 1) = (2x, 1)$ and a simple induction shows $(x, 1)^n = (nx, 1)$ for all $n \in \mathbb{N}$. Since $nx \neq 0$ for $n \neq 0$ we get that $(x, 1)$ has infinite order.

- *Ch. 3, Ex. 15* Let $(G_1, \cdot_1)$ and $(G_2, \cdot_2)$ be two abelian group. We want to show that for all $(f_1, f_2), (g_1, g_2) \in G_1 \times G_2$ it holds that $(f_1, f_2) \cdot (g_1, g_2) = (g_1, g_2) \cdot (f_1, f_2)$. However by definition of the operation in the direct product and G_1 and G_2 abelian we get:

  $$(f_1, f_2) \cdot (g_1, g_2) = (f_1 \cdot_1 g_1, f_2 \cdot_2 g_2) = (g_1 \cdot_1 f_1, g_2 \cdot_2 f_2) = (g_1, g_2) \cdot (f_1, f_2).$$

- *Ch. 3, Ex. 16* (a) The identity element in $\mathbb{Z}_4 \times \mathbb{Z}_{12}$ is $(0, 0)$, which is not the given element. Hence the order of $(2, 6)$ is at least 2. Actually,

  $$(2, 6) + (2, 6) = (2 +_4 2, 6 +_{12} 6) = (0, 0),$$

and hence the order of $(2,6)$ is exactly 2.

(b) Note that 8 has order 3 in $(\mathbb{Z}_{12}, +_{12})$ as $8 \neq 0$ and $8 +_{12} 8 \neq 0$, while $8 +_{12} 8 +_{12} 8 = 0$. Analogously we see that the order of 10 in $(\mathbb{Z}_{18}, +_{18})$ is 9. This implies that $(8,10)$ has order 9 (one should spend two words to explain why it is not smaller also!).

- *Ch. 3, Ex. 17* First we show that each element in G appears at most once in each row and column. Suppose indeed that some $a \in G$ appears at least twice in some column of the multiplication table (the case of a row is completely identical!). Then for some $b \in G$, the equation $x \cdot b = a$ has at least two distinct solutions, say c and d, with $c \neq d$. So $cb = a = db$. Since there is some $f \in G$ such that $bf = e$ (that is b^{-1}), we get $c = ce = cbf = dbf = de = d$, a contradiction.

  We are left with showing that each element appears at least once in each row and column, but this intuitive enough. We show it here for a row but for a column it is exactly the same strategy. In general, a row is nothing else but all (left-)multiplications by a given element $b \in G$ with any other element G, that is all $b \cdot x$ for $x \in G$. So if we fix $a \in G$, then this element is contained in the row with all (left-)multiplications by b as we can choose $x = b^{-1} \cdot a$ obtaining $b \cdot x = b \cdot b^{-1} \cdot a = a$.

- *Ch. 3, Ex. 18* From the previous exercise we see that the first table is not a multiplication table, as a is repeated twice in the first row. The same holds for the last table as b is repeated twice in the second column. The other two tables are multiplication tables (identity element a, every element has an inverse as a appears exactly once in each row and column, associativity can be checked by hand).

- *Ch. 3, Ex. 19* It is enough to find two matrices A and B such that $A \cdot B \neq B \cdot A$. One possibility if to take

$$A := \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \text{ and} A := \begin{pmatrix} 2 & 1 \\ 1 & 3 \end{pmatrix}.$$

- *Ch. 3, Ex. 20* We want to show that $f \cdot g = g \cdot f$ for all $f, g \in G$. To do that, we first see that

$$f \cdot f \cdot g \cdot g = f^2 \cdot g^2 = (f \cdot g)^2 = f \cdot g \cdot f \cdot g.$$

  Hence $f \cdot f \cdot g \cdot g = f \cdot g \cdot f \cdot g$. Multiplying by g^{-1} on the right and by f^{-1} on the left now gives $f \cdot g = g \cdot f$.

- *Ch. 3, Ex. 21* (a) Since $e = f \cdot g_1$, we obtain by multiplying with g_2 on both sides of the equality sign that $g_1 \cdot e = g_1 \cdot (f \cdot g_2)$. However, by group axiom two, we have $g_1 \cdot e = g_1$. Therefore, we obtain that $g_1 = g_1 \cdot (f \cdot g_2)$. Now we use the group axioms to simplify $g_1 \cdot (f \cdot g_2)$:

$$g_1 = g_1 \cdot (f \cdot g_2) = (g_1 \cdot f) \cdot g_2 = e \cdot g_2 = g_2.$$

  The first equality we have already shown, the second follows from the associativity (group axiom one). The third equality follows, since g_1 was an inverse to f, while the fourth equality follows from group axiom two.

(b) First of all we have

$$\begin{aligned}(g^{-1}\cdot f^{-1})\cdot(f\cdot g) &= ((g^{-1}\cdot f^{-1})\cdot f)\cdot g \text{ using associativity}\\ &= (g^{-1}\cdot(f^{-1}\cdot f))\cdot g \text{ using associativity again}\\ &= (g^{-1}\cdot e)\cdot g \text{ because } f^{-1} \text{ is the inverse of } f\\ &= g^{-1}\cdot g \text{ because } e \text{ is the identity element}\\ &= e \text{ because } g^{-1} \text{ is the inverse of } g\end{aligned}$$

and

$$\begin{aligned}(f\cdot g)\cdot(g^{-1}\cdot f^{-1}) &= f\cdot(g\cdot(g^{-1}\cdot f^{-1})) \text{ using associativity}\\ &= f\cdot((g\cdot g^{-1})\cdot f^{-1})\cdot g \text{ using associativity again}\\ &= f\cdot(e\cdot f^{-1}) \text{ because } g^{-1} \text{ is the inverse of } g\\ &= f\cdot f^{-1} \text{ because } e \text{ is the identity element}\\ &= e \text{ because } f^{-1} \text{ is the inverse of } f.\end{aligned}$$

This shows that $g^{-1}\cdot f^{-1}$ is an inverse of $f\cdot g$. By the previous part of the exercise, we can conclude that it is the inverse to $f\cdot g$.

- *Ch. 3, Ex. 23* It is clear that g^{-1} is another generator of G as we can rewrite each g^i as $(g^{-1})^{-i}$. We want to show that G cannot have other generators. Suppose that G has another generator f such that $f\neq g$ and $f\neq g^{-1}$. Then $f=g^m$ and $g=f^n$ for some integers m and n. These imply that

$$f=g^m=(f^n)^m=f^{nm}.$$

Since G is infinite cyclic, f cannot have finite order (otherwise G would have the same order as f). Hence nm must equal 1. Since m and n are integers, necessarily $(m,n)=(-1,-1)$ or $(m,n)=(1,1)$. In particular either $f=g$ or $f=g^{-1}$, a contradiction.

- *Ch. 3, Ex. 24* Denote with ℓ the least common multiple of g_1 and g_2. Check first by direct computation that $(g_1,g_2)^\ell=(e_1,e_2)$, so that the order of (g_1,g_2) divides ℓ. Suppose by contradiction that this order is strictly less then ℓ and call it n. Then $(g_1,g_2)^n=(g_1^n,g_2^n)=(e_1,e_2)$ implying that both the order of g_1 and the order of g_2 divide n. Since this gives that ℓ divides n, we get the desired contradiction.

- *Ch. 3, Ex. 25* Let $f\in G$. We want to compute the size of the equivalence class of f, that is, the number of elements $g\in G$ such that $\langle g\rangle=\langle f\rangle$. Note that every such a g is such that $g\in\langle g\rangle=\langle f\rangle$, and hence an element of $\langle f\rangle$. If f has finite order then $\langle f\rangle$ contains a finite number of elements, implying that the equivalence class of f has at most $|\langle f\rangle|<\infty$ elements. If f has infinite order we cannot apply the same strategy. However from exercise 22 we see that if $f\sim g$ then f and g are both generators of the infinite cyclic group $\langle f\rangle$. Hence either $g=f$ or $g=f^{-1}$, and the congruence class contains exactly 2 elements.

- *Ch. 3, Ex. 26* One can simply try to fill the multiplication table using the given information. Doing so one gets

$\cdot$	e	a	b	b^2	ab	ab^2
e	e	a	b	b^2	ab	ab^2
a	a	e	ab	ab^2	b	b^2
b	b	ab^2	b^2	e	a	ab
b^2	b^2	ab	e	b	ab^2	a
ab	ab	b^2	ab^2	a	e	b
ab^2	ab^2	b	a	ab	b^2	e

- *Ch. 3, Ex. 27* We know from the mentioned corollary that for a natural number n it holds $n = \sum_{d|n} \phi(d)$. If $n = p$ then we have only two choices for d, namely $d = p$ and $d = 1$. The equality in the corollary hence gives

$$p = \phi(p) + \phi(1) = \phi(p) + 1.$$

If we choose $n = p^2$ then we have three possibilities for d, namely $d = p^2$, $d = p$ and $d = 1$. Using the same trick we get

$$p^2 = \psi(p^2) + \phi(p) + \phi(1) = \phi(p^2) + p.$$

Choosing $n = pq$ then either $d = pq$ or $d = p$ or $d = q$ or $d = 1$. Hence

$$pq = \phi(pq) + \phi(p) + \phi(q) + \phi(1) = \phi(pq) + p + q - 1.$$

10.4 Chapter 4

- *Ch. 4, Ex. 1* A. FALSE, B. TRUE, C. TRUE.
- *Ch. 4, Ex. 2* A,C.
- *Ch. 4, Ex. 3* A. TRUE, B. FALSE, C. TRUE, D. TRUE.
- *Ch. 4, Ex. 4* A. TRUE, B. FALSE, C. TRUE, D. TRUE, E. FALSE, F. TRUE.
- *Ch. 4, Ex. 5* A,C.
- *Ch. 4, Ex. 6* No. A way is to check that the composition of (12) and (123) is not contained in the set.
- *Ch. 4, Ex. 7* We have $(\mathbb{Z}_8)^* = \{1, 3, 5, 7\}$. All possible subgroups are given by

$$\{1\}, \{1, 3\}, \{1, 5\}, \{1, 7\}, \{1, 3, 5, 7\}.$$

- *Ch. 4, Ex. 9* We know from Examples 4.2.5 that

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

A direct computation therefore gives

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

A more theoretical solution avoiding any computation is the following: In Example 4.2.5 we saw that

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

Since $[G : H] = 3$, the subgroup H has three distinct cosets in G. Since $(1\,2\,3) \in (1\,2\,3)H$, but $(1\,2\,3) \notin H$ and $(1\,2\,3) \notin (1\,2)H$, the coset $(1\,2\,3)H$ consists of the eight elements of S_4 that are not in $H \cup (1\,2)H$.

- *Ch. 4, Ex. 10* We use part (4) of Theorem 4.3.4. It implies that $(1\,2)H = (1\,3\,4)H$ if and only if $(1\,2)^{-1}(1\,3\,4) \in H$. Since $(1\,2)^{-1}(1\,3\,4) = (1\,2)(1\,3\,4) = (1\,3\,4\,2) \notin H$, we see that $(1\,2)H \neq (1\,3\,4)H$.

- *Ch. 4, Ex. 11* Let us write $a := f[1]$, $b := f[2]$, and $c := f[3]$. Since f is a permutation, so is f^{-1}. Therefore a, b, c are three distinct elements from $\{1, 2, \ldots, n\}$. We claim that $f(1\,2\,3)f^{-1} = (a\,b\,c)$. To prove this equality, we study the effect of both $f(1\,2\,3)f^{-1} = (a\,b\,c)$ on all symbols in $\{1, 2, \ldots, n\}$.

  We choose $i \in \{1, 2, \ldots, n\}$. First of all, if $i \notin \{a, b, c\}$, then i is kept fixed by the 3-cycle $(a\,b\,c)$. Now we study what happens with i under the permutation $f \circ (1\,2\,3) \circ f^{-1}$. First i is sent to $f^{-1}[i]$ by the permutation f. Since $i \notin \{a, b, c\}$, we have $f^{-1}[i] \notin \{1, 2, 3\}$. Therefore i is also sent to $f^{-1}[i]$ by the permutation $(1\,2\,3) \circ f^{-1}$. Now we can deduce that $f \circ (1\,2\,3) \circ f^{-1}$ sends i to $f[f^{-1}[i]]$, which is equal to i.

  Now we will show that $f \circ (1\,2\,3) \circ f^{-1}$ maps a to b. Since $a = f[1]$, we see that $f^{-1}[a] = 1$. Therefore we have $((1\,2\,3) \circ f)[a] = ((1\,2\,3))[1] = 2$. This in turn implies that $(f \circ (1\,2\,3) \circ f^{-1})[a] = f[2] = b$. Similarly one shows that $(f \circ (1\,2\,3) \circ f^{-1})[b] = f[3] = c$ and $(f \circ (1\,2\,3) \circ f^{-1})[c] = f[1] = a$.

  Combining all the above, we see that $f^{-1} \circ (123) \circ f$ and (abc) are the same permutations.

  What is left to show is that any 3-cycle in S_n can be written in the form $f \circ (1\,2\,3) \circ f^{-1}$. Given a specific 3-cycle, say $(i\,j\,k)$, what we need to do is to find an $f \in S_n$ such that $f \circ (1\,2\,3) \circ f^{-1} = (i\,j\,k)$. Note that i, j and k are three distinct elements from $\{1, 2, \ldots, n\}$. We have seen above that $f \circ (1\,2\,3) \circ f^{-1}$ is the 3-cycle $(a\,b\,c)$, with $a := f[1]$, $b := f[2]$, and $c := f[3]$. Therefore it is enough to show that there exists a permutation $f \in S_n$ such that $f[1] = i$, $f[2] = j$ and $f[3] = k$. Since i, j and k are distinct this can be done: the sets $\{1, 2, \ldots, n\}\backslash\{i, j, k\}$ and $\{1, 2, \ldots, n\}\backslash\{1, 2, 3\}$ have the same number of elements, there exists a bijection g from the first of the set to the second one (you do not have to show this, but for the die-hards: this follows for example by induction on the number of elements in the set). We can extend g to a permutation f by defining $f[1] = i$, $f[2] = j$ and $f[3] = k$. Let us for example assume $n = 5$ and suppose we want to find f in case $(ijk) = (143)$. In this case g has to be a bijection from $\{2, 5\}$ to $\{4, 5\}$. We choose g to be the bijection defined by $g(2) = 4$ and $g(5) = 5$. The bijection f will then be defined by $f(1) = 1$, $f(4) = 2$, $f(3) = 3$, $f(2) = g(2) = 4$ and $f(5) = g(5) = 5$. In cycle notation we then get $f = (24)$. Indeed we have $(24)(123)(24)^{-1} = (24)(123)(24) = (143)$.

- *Ch. 4, Ex. 16* The possible types of an element in S_6 are:

  1. The identity permutation. It has order 1.
  2. An m-cycle, with $m \in \{2, 3, 4, 5, 6\}$. It has order m.
  3. An element with cycle type $(1, 1, 1, 0, 0, 0)$, that is to say, the composition of a 2- and a 3-cycle that are mutually disjoint. These elements have order 6.

4. An element with cycle type $(0,1,0,1,0,0)$, that is to say, the composition of a 2- and a 4-cycle that are mutually disjoint. These elements have order 4.
5. An element with cycle type $(0,0,2,0,0,0)$, that is to say, the composition of two mutually disjoint 3-cycles. These elements have order 3.
6. An element with cycle type $(2,2,0,0,0,0)$ or $(0,3,0,0,0,0)$, that is to say, the composition of two or three mutually disjoint 2-cycles. These elements have order 2.

Hence the exponent of S_6 is 60, the least common multiple of all orders of elements in S_6.

- *Ch. 4, Ex. 17* If H is a subgroup, then the three requirements from Definition 4.1.1 are satisfied. In particular for any $g \in H$ we have $g^{-1} \in H$. Hence for any $f, g \in H$, we have $f \cdot g^{-1} \in H$.

  Conversely, assume that for any $f, g \in H$, we have $f \cdot g^{-1} \in H$. We need to check that the three conditions from Definition 4.1.1 are satisfied.

  – Choosing $g = f$, we see that $f \cdot f^{-1} \in H$. However, $f \cdot f^{-1} = e$. Hence $e \in H$.
  – Now choose $f = e$. Then we see that for any $g \in H$ we have $e \cdot g^{-1} \in H$. Since $e \cdot g^{-1} = g^{-1}$, we see that for any $g \in H$ also $g^{-1} \in H$.
  – Now choose $f, g \in H$. We have just seen that $g^{-1} \in H$. Hence we may conclude that $f \cdot (g^{-1})^{-1} \in H$. Since $f \cdot (g^{-1})^{-1} = f \cdot g$, we see that for any $f, g \in H$, we have $f \cdot g \in H$.

  Combining the above, we see that H is a subgroup if for any $f, g \in H$, we have $f \cdot g^{-1} \in H$.

- *Ch. 4, Ex. 18* First we determine all possible subgroups H of the groups $(C_n, \cdot)$. It is given that n is a prime number and that $C_n = \{e, r, \ldots, r^{n-1}\}$. We will show that either $H = \{e\}$ or $H = C_n$.

  First of all, a subgroup $H \subseteq C_n$ is either equal to $\{e\}$ or it contains an element of the form r^m with $1 \leq m \leq n-1$. Therefore we are done if we show that in the latter case $H = C_n$. If $r^m \in H$, then also $(r^m)^b \in H$ for any integer b, since H is a subgroup. Since n is a prime number and $1 \leq m \leq n-1$, the greatest common divisor of m and n equals 1. This implies that there exist integers a and b such that

  $$an + bm = 1.$$

  Choosing this particular b, we see that $(r^m)^b = r^{bm} = r^{1-an} = r \cdot (r^n)^{-a} = r \cdot e = r$. Combining the above, we deduce that if $r^m \in H$, then also $r \in H$. Again using that H is a subgroup, this implies that any power of r is in H in this case. The conclusion is that if $H \subseteq C_n$ is a subgroup, then either $H = \{e\}$ or $H = C_n$.

- *Ch. 4, Ex. 19* One can simply check the axioms of the definiton of being a subgroup, by using the fact that they hold for both H and K (and an element in $H \cap K$ is indeed both in H and K). The union $H \cup K$ is not necessarily a subgroup. An example is given by $H = \{e, (12)\}$ and $K = \{e, (13)\}$ in S_3. Then $H \cup K = \{e, (12), (13)\}$ while $(12) \circ (13) = (132) \notin H \cup K$.

- *Ch. 4, Ex. 20* Denote the subset of elements of finite order in G with F. The identity element e of G is contained in F as it has order 1. Let $f, g \in F$ and denote their orders with m and n respectively. Then use the fact that G is abelian to show $(f \cdot g)^{mn} = f^{mn} \cdot g^{mn}$.

Since $f^{mn} = g^{mn} = e$ we get that $f \cdot g$ has finite order dividing mn. This shows that $f \cdot g \in F$. We are left to show that $f^{-1} \in F$. However it is enough to note that (using G abelian again), $e = e^m(f \cdot f^{-1})^m = f^m \cdot (f^{-1})^m = (f^{-1})^m$, so that f^{-1} has finite order dividing m.

- *Ch. 4, Ex. 21* It is enough to check the axioms of being a subgroup, using that the group axioms hold for $(A, \cdot)$ while the operation defined in $A \times A$ is just the same as in A coordinatewise.

- *Ch. 4, Ex. 22* Denote with n the order of g. Then

$$(fgf^{-1})^n = (fgf^{-1})(fgf^{-1}) \dots (fgf^{-1}) = fg(f^{-1}f)g(f^{-1}f) \dots gf^{-1} = fg^m f^{-1} = e.$$

This application of the associative property shows that the order of fgf^{-1} divides n. We need to show that it cannot be smaller. Suppose by contradiction that the order of fgf^{-1} is $m < n$ then

$$e = (fgf^{-1})^m = (fgf^{-1})(fgf^{-1}) \dots (fgf^{-1}) = fg(f^{-1}f)g(f^{-1}f) \dots gf^{-1} = fg^m f^{-1}.$$

Multiplying on the right by f gives $f = fg^m$ and finally multiplying by f^{-1} on the left gives $e = g^m$, which is not possible as $m < n = \text{ord}(g)$.

- *Ch. 4, Ex. 23* Let us assume that $\#G = p$ for some prime p. Proposition 4.4.4 then implies that the order of any group element $g \in G$ is either 1 or p. The only group element having order one is the identity element e. Therefore all other $p - 1$ elements have order p. Such an element $g \in G$ generates the group G, that is to say that $G = \{e, g, g^2, \dots, g^{p-1}\}$. This means that if f_1 and f_2 are two elements from G, then there exist natural numbers i and j such that $f_1 = g^i$ and $f_2 = g^j$. This property is enough to show that a cyclic group is abelian, since

$$f_1 \cdot f_2 = g^i \cdot g^j = g^{i+j} \text{ and } f_2 \cdot f_1 = g^j \cdot g^i = g^{j+i} = g^{i+j}.$$

We can now answer to the second question of the exercise. The answer is negative. The smallest counter example is S_3, which has order 6. It is not abelian, since $(1\,2)(1\,3) = (1\,3\,2)$ and $(1\,3)(1\,2) = (1\,2\,3)$. Therefore it is in particular not cyclic, since any cyclic group is abelian.

- *Ch. 4, Ex. 24* The key is the observation that H itself is both a left (since $H = eH$) and a right coset (since $H = He$). Therefore G can be partitioned (disjoint union) into two left cosets, say H and fH, but also into two right cosets, say H and Hg. Necessarily we have $fH = Hg$, since both cosets consist of all element in G that are not in H.

  First we show this implies that $fH = Hf$. Since $fH = Hg$ and $f \in fH$ (as we can write $f = f \cdot e$), we see that $f \in Hg$. On the other hand $f \in Hf$, since $f = e \cdot f$. Therefore the two right cosets Hg and Hf both contain the element f. Since right cosets are either disjoint or identical, we can conclude $fH = Hg = Hf$.

  Now let $k \in G$ be an arbitrary element. We now wish to show that $kH = Hk$, and this would complete the proof. Note that we can assume that k is not in H as otherwise the claim follows immediately from $kH = H$ and $Hk = H$. The right coset Hk is either equal to H or to Hf (because we have only two right cosets!), while the left coset kH is equal to either H or fH. Since k is not in H, $Hk \neq H$ and $kH \neq H$ implying that $kH = fH$ and $Hk = Hg$. Since we proved already that $fH = Hf = Hg$ we have that $kH = fH = Hg = Hk$, which is what we wanted to prove.

- *Ch. 4, Ex. 25* We want to show that $f(H \cap K) = (H \cap K)f$ for all $f \in G$. Let $f \in G$ be fixed a let $g \in f(H \cap K)$. Then there exists $h \in H \cap K$ such that $g = fh$. Hence $g \in fH$ and $g \in fK$ as h is both in H and K. Since H and K are both normal, $fH = Hf$ and $fK = Kf$, so that there must exist $k \in K$ and $h_1 \in H$ such that $g = fh = h_1 f = kf$.From $h_1 f = kf$, multiplying by f^{-1} on the right, we deduce that $k = h_1$ and hence $k \in H \cap K$. This shows that $g = kf \in (H \cap K)f$ and hence $f(H \cap K) \subseteq (H \cap K)f$. The other inclusion is analogous.

- *Ch. 4, Ex. 26* Recall that the index of $H \cap K$ counts the number of distinct (left) cosets of $H \cap K$ in G. We want to use the fact that the number of distinct cosets of both H and K is finite to deduce that the same holds for $H \cap KL$. Let C be the set of left cosets of $H \cap K$ in G, C_1 the set of left cosets of H in G, and C_2 the set of left cosets of K in G. Hence $|C_i| < \infty$ for $i = 1, 2$ and we want to show $|C| < \infty$.

  Consider the function $\phi : C \to C_1 \times C_2$ defined by

  $$\phi(x(H \cap K)) = (xH, xK).$$

  We first show that this map is well-defined, that is, if $x(H \cap K) = y(H \cap K)$ then $xH = yH$ and $xK = yK$. Recall that $x(H \cap K) = y(H \cap K)$ if and only if $xy^{-1} \in H \cap K$ which means in particular that $xy^{-1} \in H$ and $xy^{-1} \in K$. Since these last two conditions are equivalent to $xH = yH$ and $xK = yK$ respectively, the map is well-defined.

  We now show that ϕ is injective. Suppose that $\phi(x(H \cap K)) = (xH, xK) = (yH, yK) = \phi(y(H \cap K))$. Then $xH = yH$ and $xK = yK$, that is, $xy^{-1} \in H$ and $xy^{-1} \in K$. This implies that $xy^{-1} \in H \cap K$ which is equivalent (as before) to $x(H \cap K) = y(H \cap K)$. Hence ϕ is injective. In particular $|C| \leq |C_1| \cdot |C_2| < \infty$, which is what we wanted to show.

- *Ch. 4, Ex. 27* Recall that H is normal in G if and only if $fH = H$f for all $f \in G$. We start by following the hint. Since the product of left cosets is a left coset, we have that $fH \cdot f^{-1}H = mH$ for some $m \in G$. However $e \in fH \cdot f^{-1}H$ as we can write

  $$e = f \cdot f^{-1} = (fe) \cdot (f^{-1}e) \in fH \cdot f^{-1}H.$$

  This implies that $e = mh$ for some $h \in H$, so that $m = h^{-1} \in H$. Hence $mH = H = fH \cdot f^{-1}H$ for all $f \in G$. Clearly, replacing f with f^{-1} gives $f^{-1}H \cdot fH = H$ (as the above property is true for any $f \in G$). We now show how this yields to $fH = Hf$.

  To do so let first $g \in fH$, so that $g = fh$ for some $h \in H$. Note that $gf^{-1} = (fh) \cdot (f^{-1}e) \in fH \cdot f^{-1}H = H$. Hence there exists $h_1 \in H$ such that $gf^{-1} = h_1$, that is, $g = h_1 f \in Hf$. This proves that $fH \subseteq Hf$. For the other inclusion one can use that $f^{-1}H \cdot fH = H$ instead. Indeed let $g \in Hf$. Then $g = hf$ for some $h \in H$. Furthermore, $f^{-1}g = (f^{-1}h) \cdot (fe) \in f^{-1}H \cdot fH = H$. Hence there exists $h_1 \in H$ such that $f^{-1}g = h_1$. In particular, $g = fh_1 \in fH$ and $Hf \subseteq Hf$.

- *Ch. 4, Ex. 28* (a) From Theorem 4.3.4, we see that $fH = gH$ if and only if $f^{-1}g \in H$, while similarly $Hf^{-1} = Hg^{-1}$ if and only if $Hg^{-1} = Hf^{-1}$ if and only if $f^{-1}g = f^{-1}(g^{-1})^{-1} \in H$. This means that the map σ defined by $\sigma(fH) = Hf^{-1}$ indeed is well defined, since the image Hf^{-1} does not depend on the choice of the representative f of fH.

  (b) Very similarly as in part (a), one can show that the map $\tau : H\backslash G \to G/H$ given by $\tau(Hf) := f^{-1}H$ is well defined. Direct verification shows that τ is the inverse function of σ, since $\sigma(\tau(Hf)) = \sigma(f^{-1}H) = H(f^{-1})^{-1} = Hf$ and $\tau(\sigma(fH)) = \tau(Hf^{-1}) = (f^{-1})^{-1}H = fH$. Hence the cardinalities of G/H and $H\backslash G$ are the same.

10.5 Chapter 5

- *Ch. 5, Ex. 1* A. TRUE, B. FALSE, C. TRUE, D. FALSE.
- *Ch. 5, Ex. 2* B,C.
- *Ch. 5, Ex. 3* A,C.
- *Ch. 5, Ex. 4* A. TRUE, B. FALSE, C. TRUE, D. TRUE, E. TRUE.
- *Ch. 5, Ex. 5* A,B.
- *Ch. 5, Ex. 6* A. TRUE, B. FALSE, C. TRUE.
- *Ch. 5, Ex. 7* We have $|A| = 2^4 = 16$. As in example 5.3.2, we will assume that the colours red and black are used. There are six orbits:
  1. The orbit O_1 of the square whose vertices are all coloured black. We have $|O_1| = 1$, while the corresponding stabilizer is the entire group $C_4 = \{e, r, r^2, r^3\}$.
  2. The orbit O_2 consisting of the squares three of whose vertices are coloured black, and one coloured red. We have $|O_2| = 4$ and for any $a \in O_2$, we have $G_a = \{e\}$.
  3. The orbit O_3 consisting of the squares two of whose vertices are coloured black, and two are coloured red. Moreover, the red-coloured vertices lie on a diagonal of the square, while the black-coloured vertices lie on the other diagonal. We have $|O_3| = 2$ and for any $a \in O_3$, we have $G_a = \{e, r^2\}$.
  4. The orbit O_4 consisting of the squares two of whose vertices are coloured black, and two are coloured red. Moreover, the red-coloured vertices lie on an edge of the square, while the black-coloured vertices lie on the opposite edge. We have $|O_4| = 2$ and for any $a \in O_4$, we have $G_a = \{e\}$.
  5. The orbit O_5 consisting of the squares three of whose vertices are coloured red, and one coloured black. We have $|O_5| = 4$ and for any $a \in O_5$, we have $G_a = \{e\}$.
  6. The orbit O_6 of the square whose vertices are all coloured red. We have $|O_6| = 1$, while the corresponding stabilizer is the entire group $C_4 = \{e, r, r^2, r^3\}$.
- *Ch. 5, Ex. 8* The group of rotation symmetries contains the six elements e, r, r^2, r^3, r^4, r^5. Now we compute for each element how many colourings are fixed.
  - The identity element e keeps all 3^6 colourings fixed.
  - The elements r and r^5 each keep 3 colourings fixed (the colourings where all beads have the same colour.)
  - The elements r^2 and r^4 each keep 3^2 colourings fixed (the colourings where the beads in two triangles independent of each other have the same colour.)
  - The element r^3 keeps 3^3 colourings fixed (the colourings where opposite vertices have the same colour.)

  This gives a total of $\frac{1}{6}\left(3^6 + 2 \cdot 3 + 2 \cdot 3^2 + 3^3\right) = 130$ distinct colourings.

  It is also a good exercise to see what happens if we use the group D_6 instead. In this case we are potentially identifying more colourings with each other than before. Hence the total number of distinct colourings is at most 130, but could be less. It turns out that in this case, there are two types of reflection symmetries: three with symmetry axis joining opposite

vertices and three with symmetry axis joining opposite midpoints of vertices. Similar as above, this gives a total of $\frac{1}{12}\left(3^6 + 2\cdot 3 + 2\cdot 3^2 + 3^3 + 3\cdot 3^4 + 3\cdot 3^3\right) = 92$ distinct colourings. Apparently there are colourings of the regular 6-gon that can be identified if we also allow reflection symmetries, but not if we only allow rotation symmetries.

- *Ch. 5, Ex. 9* The tetrahedron has in total 12 rotational symmetries (and in fact can be identified with the 12 elements from A_4 by looking at their action on the 4 vertices of the tetrahedron):

  1. The identity symmetry e. It fixes all 4^4 possible colourings.
  2. Rotations with rotation axis through the midpoints of opposite edges and rotation angle π. There are 3 such rotations. Each of them fixes 4^2 colourings (sides sharing an edge through which the rotation axe passes, need to get the same colour).
  3. Rotations with rotation axis through a vertex and the midpoint of the opposite side and rotation angle $2\pi/3$ or $4\pi/3$. There are 8 such rotations. Each of them fixes 4^2 colourings (given that the rotation axis passes through a vertex, the three sides on which this vertex lies need to get the same colour; the side opposite the vertex may get another independently chosen colour).

  Using Theorem 5.3.1 we can compute the total number of distinct colourings as follows:

$$\frac{1}{12}\left(4^4 + 3\cdot 4^2 + 8\cdot 4^2\right) = 36.$$

- *Ch. 5, Ex. 10* (a) The group D_5 is the group of symmetries of the regular pentagon, see Chapter 2, Exercise 7. It consists of the identity, four rotations and five reflections. The identity has m^5 fixed points, the four rotations each have m fixed points, while each of the five reflections has m^3 fixed points.

  This can also be seen from the permutations of the edges that the symmetries give rise to. In disjoint cycle notation and including one-cycles, the identity e would have the form $(A)(B)(C)(D)(E)$, a rotation would give rise to a five cycle, while a reflection would give rise to a permutation of the form $(AB)(CD)(E)$. For a given symmetry, the number of disjoint cycles (counting 1-cycles as well) tells one how many degrees of freedom there are in the colourings that are fixed by a symmetry.

  Using Burnside's lemma, this gives a total of

$$\frac{1}{10}\left(m^5 + 5m^3 + 4m\right)$$

  distinct colourings.

  (b) In this case the identify has m^8 fixed points, r, r^3, r^5, r^7 all have m fixed points, r^2, r^6 have m^2 fixed points, while r^4 has m^4 fixed points. The elements s, r^2s, r^4s, r^6s are reflections symmetries in a line connecting two opposite vertices. Hence they each have m^5 fixed colourings. The elements rs, r^3s, r^5s, r^7s are reflections symmetries in a line connecting midpoints of opposite edges. Hence they each have m^4 fixed points.

  Using Burnside's lemma, this gives a total of

$$\frac{1}{16}\left(m^8 + 4m^5 + 5m^4 + 2m^2 + 4m\right)$$

  distinct colourings.

- *Ch. 5, Ex. 11* (a) The two conditions for being a group action can be checked directly, but do not get confused by all the symbols! For example, we have that

$$\begin{aligned}\varphi_{f_1\cdot f_2}[g] =&(f_1\cdot f_2)\cdot g\cdot(f_1\cdot f_2)^{-1} = f_1\cdot f_2\cdot g\cdot f_2^{-1}\cdot f_1^{-1}\\ =&f_1\cdot(f_2\cdot g\cdot f_2^{-1})\cdot f_1^{-1} = \varphi_{f_1}[f_2\cdot g\cdot f_2^{-1}]\\ =&\varphi_{f_1}[\varphi_{f_2}[g]] = (\varphi_{f_1}\circ\varphi_{f_2})[g].\end{aligned}$$

  (b) If $\varphi_f = \mathrm{id}_G$, then for all $g\in G$, $\varphi_f[g] = g$. Writing out what this means, one obtains that for all $g\in G$, $fg = gf$. Conversely, if $g\in G$, $fg = gf$, then rewriting in the opposite direction gives that for all $g\in G$, $\varphi_f[g] = g$

  (c) The orbit if just $\{g\}$, the stabilizer is all of G, since for any $f\in G$ $\phi_f[g] = fgf^{-1} = gff^{-1} = g$.

- *Ch. 5, Ex. 12* The action being transitive implies that the number of orbits of the action on A is 1, because A itself is an orbit (and two orbits do not intersect if they are not equal). The result now follows directly from Burnside's Lemma.

- *Ch. 5, Ex. 13* From the previous exercise we know that the sum of all $|\mathrm{Fix}(g)|$ with $g\in G$ coincides with $|G|$. If we assume by contradiction that every element in G has at least one fixed point then $|\mathrm{Fix}(g)|\geq 1$ for all $g\in G$. We can actually say a bit more, because $e\in G$ fixes all elements in A and by hypothesis $|A|\geq 2$. This means that $|\mathrm{Fix}(g)|\geq 1$ for all $g\in G\setminus\{e\}$ and $|\mathrm{Fix}(e)| = |A|\geq 2$. Hence

$$\sum_{g\in G}|\mathrm{Fix}(g)| = \sum_{g\in G\setminus\{e\}}|\mathrm{Fix}(g)| + |\mathrm{Fix}(e)+ \geq \sum_{g\in G\setminus\{e\}} 1+|A|\geq(|G|-1)+2 = |G|+1.$$

  This is a contradiction to the sum of all $|\mathrm{Fix}(g)|$ with $g\in G$ being equal to $|G|$.

- *Ch. 5, Ex. 14* The situation can be described using a group action $\varphi : S_3\times S_3\to S_A$. The group action is defined as follows: given $f,g\in S_3$, the permutation $\varphi_{f,g}\in S_A$ send a matrix M to the matrix obtained from M by permuting its rows using f and its columns using g. The number of orbits can now be counted using Burnside's lemma. There are several cases:

Case 1 $f = g = \mathrm{id}$. In this case $|\mathrm{Fix}(f,g)| = 2^9$.

Case 2 $f = \mathrm{id}$ and g is a 2-cycle. In this case, the fixed points are the matrices with two identical columns, namely the ones that g interchanges. Hence $|\mathrm{Fix}(f,g)| = 2^6$. The case that $g = \mathrm{id}$ and f is a 2-cycle is similar.

Case 3 $f = \mathrm{id}$ and g is a 3-cycle. In this case, the fixed points are the matrices with three identical columns. Hence $|\mathrm{Fix}(f,g)| = 2^3$. The case that $g = \mathrm{id}$ and f is a 3-cycle is similar.

Case 4 f and g are both 2-cycles. The nine entries of the matrix are pairwise interchanged, while one entry is fixed. This gives that $|\mathrm{Fix}(f,g)| = 2^5$.

Case 5 f is a 2-cycle and g a 3-cycle. In this case, the order of (f,g) is six. The entries in the rows that f interchanges form a 6-cycle when permuted by (f,g), while the remaining three entries form a 3-cycle. Hence $|\mathrm{Fix}(f,g)| = 2^2$. The case f is a 3-cycle and g a 2-cycle is similar.

Case 6 f and g are both 3-cycles. In this case the nine entries in the matrix are permuted by (f,g) in three 3-cycles. Hence $|\mathrm{Fix}(f,g)| = 2^3$ in this case.

Hence we obtain that the number of distinct nonequivalent matrices equals:

$$\frac{1}{36}\left(2^9 + 6\cdot 2^6 + 4\cdot 2^3 + 9\cdot 2^5 + 12\cdot 2^2 + 4\cdot 2^3\right) = 36.$$

- *Ch. 5, Ex. 15*

  The elements of the group of symmetries of the decoration can be identified with elements of S_6. If we numerate the light bulbs from 1 to 6 counterclockwise, the rod flips correspond to the permutation $g_1 = (1\,2)$, $g_2 = (3\,4)$, and $g_3 = (5\,6)$. A counterclockwise rotation of the wheel over $2\pi/3$ then corresponds to the permutation $f = (1\,3\,5)(2\,4\,6)$. There is a total of 24 possible ways to change the decoration combining rotations of the wheel and rod flips in all possible ways, so any symmetry can be written in the form $f^i g_1^j g_2^k g_3^\ell$, with $0 \le i \le 2$, $0 \le j \le 1$, $0 \le k \le 1$, and $0 \le \ell \le 1$. To count the number of distinct decorations, we apply Burnside's lemma. To count the number of fixed points of a given symmetry, we distinguish several cases:

Case 1 $i = 0$, $(j,k,\ell) = (0,0,0)$. In this case $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^6$.

Case 2 $i = 0$ and (j,k,ℓ) contains two 0's and one 1. In this case, the light bulbs at the end of the flipped rod need to have the same colour, but otherwise there are no restrictions, Hence $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^5$ in this case.

Case 3 $i = 0$ and (j,k,ℓ) contains two 1's and one 0. In this case, for each flipped rod the light bulbs attached to it need to have the same colour. Hence $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^4$ in this case.

Case 4 $i = 0$ and $(j,k,\ell) = (1,1,1)$. In this case, for each flipped rod the light bulbs attached to it need to have the same colour. Hence $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^3$ in this case.

Case 5 $i \neq 0$ and $(j,k,\ell) = (0,0,0)$. In this case the six light bulbs are permuted as two 3-cycles, so that $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^2$.

Case 6 $i \neq 0$ and (j,k,ℓ) contains two 0's and one 1. In this case the six light bulbs are permuted as a 6-cycle, so that $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3$.

Case 7 $i \neq 0$ and (j,k,ℓ) contains two 1's and one 0. In this case, the six light bulbs are permuted as two 3-cycles, so that $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3^2$.

Case 8 $i \neq 0$ and $(j,k,\ell) = (1,1,1)$. In this case, the six light bulbs are permuted as a 6-cycle, so that $|\mathrm{Fix}(f^i g_1^j g_2^k g_3^\ell)| = 3$.

Combining all the above we obtain the answer:

$$\frac{1}{24}\left(3^6 + 3\cdot 3^5 + 3\cdot 3^4 + 3^3 + 2\cdot 3^2 + 6\cdot 3 + 6\cdot 3^2 + 2\cdot 3\right) = 76.$$

- *Ch. 5, Ex. 16* Denote with $O_1, \ldots, O_k$ the distinct orbits of the action φ of G on A. From the Orbit-Stabilizer Theorem $|O_i|$ divides $|G|$ for all $i = 1, \ldots, k$. Since $|G| = p^n$ where p is a prime, this means that $|O_i| = p^{n_i}$ with $n_i \ge 0$ for all $i = 1, \ldots, k$. Recall that the orbits $O_1, \ldots, O_k$ form a partition of A, that is, they are mutually disjoint and

$$\bigcup_{i=1}^{k} O_i = A.$$

  This implies that

$$|A| = \sum_{i=1}^{k} |O_i| = \sum_{i=1}^{k} p^{m_i}.$$

Suppose by contradiction that G has no fixed point on A. This is equivalent to say that $O_i > 1$ for all $i = 1, \ldots, k$ as having a fixed point is equivalent to have at least one orbit of size 1. Since $O_i = p^{m_i}$ this means that $m_i \geq 1$ for all $i = 1, \ldots, k$ and hence p divides O_i for all $i = 1, \ldots, k$. However this implies that p divides also $\sum_{i=1}^{k} |O_i| = |A|$ while by hypothesis $\gcd(p, |A|) = 1$. This is a contradiction.

- *Ch. 5, Ex. 17* (a) Since H is a normal subgroup, it holds that $fH = Hf$ for all $f \in G$. If $g \in H$, then for any $f \in G$, it holds that $fgf^{-1} \in fHf^{-1} = Hff^{-1} = H$. Hence $C_g \subseteq H$.

  (b) We know from Proposition 5.2.6 that G is a union of all conjugation classes. Combining this with part (a), we see that H is the union of some of these conjugation classes.

- *Ch. 5, Ex. 18*

  (a) If $g = c_1 \circ \cdots c_\ell$ is the disjoint cycle decomposition of g, then $f \circ g \circ f^{-1} = (f \circ c_1 \circ f^{-1}) \circ \cdots \circ (f \circ c_\ell \circ f^{-1})$. From the hint, it follows directly that if c_i is an m_i-cycle, then so is $f \circ c_i \circ f^{-1}$. Moreover, since f is a bijection any two cycles $f \circ c_i \circ f^{-1}$ and $f \circ c_j \circ f^{-1}$ with $i \neq j$, are mutually disjoint. Hence the cycle type does not change under conjugation.

  What remains to be shown is that any two permutations with the same cycle type are conjugated. If g_1 and g_2 have the same cycle type, say having disjoint cycles decompositions $g_1 = c_1 \circ \cdots c_\ell$ and $g_2 = d_1 \circ \cdots d_\ell$, with $c_i = (a_{i1} \ \ldots \ a_{im_i}), d_i = (b_{i1} \ \ldots \ b_{im_i})$ cycles of the same length, then choose a permutation $f \in S_n$ such that $f[a_{ij}] = b_{ij}$ for $1 \leq i \leq \ell$ and $1 \leq j \leq m_i$. Then $fg_1f^{-1} = g_2$.

  (b) From the previous part, we see that the conjugation classes in S_4 are:

$$C_1 = \{\mathrm{id}\},$$

$$C_2 = \{(1\,2), (1\,3), (1\,4), (2\,3), (2\,4), (3\,4)\},$$

$$C_3 = \{(1\,2\,3), (1\,2\,4), (1\,3\,2), (1\,3\,4), (1\,4\,2), (1\,4\,3), (2\,3\,4), (2\,4\,3)\},$$

$$C_4 = \{(1\,2\,3\,4), (1\,2\,4\,3), (1\,3\,2\,4), (1\,3\,4\,2), (1\,4\,2\,3), (1\,4\,3\,2)\},$$

  and

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

  Of course, $(S_4, \circ)$ contains the trivial normal subgroups $\{\mathrm{id}\}$ and S_4. Suppose that H is a nontrivial normal subgroup of $(S_4, \circ)$. Following the hint, we use that a normal subgroup H is a union of conjugation classes. First of all, any normal subgroup contains the identity element. If H contains the conjugation class of a 2-cycle, then $H = S_4$, since by Exercise 21 from Chapter 2 the 2-cycles generate the whole symmetric group. Therefore we may assume that H does not contain any 2-cycles. If H contains the conjugation class of a 4-cycle, then, since $(1\,2\,3\,4)^2 = (1\,3)(2\,4)$, it also contains the conjugation class of $(1\,3)(2\,4)$. This means that H would already contain $1 + 6 + 3 = 10$ elements, where we also counted the identity element. Since 10 does not divide $|S_4| = 24$, the normal subgroup H needs to contain at least one more conjugation class, but the only one left is that of a 3-cycle. Adding this, we obtain a further 8 elements, but then H would contain exactly 18 elements, in contradiction with Lagrange's theorem. Hence we may now assume that H does not contain any 4-cycle (nor any 2-cycle). If H contains the conjugation class of a 3-cycle, it also contains $(1\,2\,3)(2\,3\,4) = (1\,2)(3\,4)$. In fact, the conjugation classes $C_1, C_3, C_{2,2}$ together form a normal subgroup. This leaves only one possibility: the nontrivial normal subgroup does not contain any 2-,3-, or 4-cycle. In that case $H = C_1 \cup C_{2,2}$ and it turns out that this is a normal subgroup as well.

- *Ch. 5, Ex. 19* (a) Since a transitive group action only has one orbit, namely the entire set A, the orbit-stabilizer theorem yield that for any $a \in A$: $|A| \cdot |G_a| = |O_a| \cdot |G_a| = |G|$. Hence $n = |A|$ divides $|G|$.

  (b) Consider the group action $\psi : G \to A^2$ defined by $\psi_f[(a_1, a_2)] = (\varphi_f[a_1], \varphi_f[a_2])$. If φ is a 2-transitive group action, then the set $A^2 \setminus \{(a, a) \mid a \in A\}$ forms a single orbit under ψ. Using that this orbit contains precisely $n^2 - n = n(n-1)$ elements, the orbit-stabilizer theorem applied to ψ implies similarly as in part (a) that $n(n-1)$ divides $|G|$

  (c) An m-transitive group action $\varphi : G \to A$ satisfies the defining property that for any two m-tuples $(a_1, \ldots, a_m)$ and $(b_1, \ldots, b_m)$ such that all a_i are pairwise distinct and all b_i are pairwise distinct, there exists $f \in G$ such that $\varphi_f[a_1] = b_1, \ldots, \varphi_f[a_m] = b_m$. Then similarly as in part (b) one considers the group action $\psi : G \to A^m$ defined by $\psi_f[(a_1, \ldots, a_m)] = (\varphi_f[a_1], \ldots, \varphi_f[a_m])$. The set of m-tuples with pairwise distinct entries, form a single orbit under ψ containing precisely $n(n-1)\cdots(n-m+1)$ elements. Applying the orbit-stabilizer theorem then implies that $n(n-1)\cdots(n-m+1)$ divides $|G|$.

- *Ch. 5, Ex. 20* (a) Showing that $Z(G)$ is a subgroup is left to the reader. We have already seen in Exercise 11 that $g \in Z(G)$ precisely if $\{g\}$ is an orbit under the conjugation action. This shows that $Z(g)$ is the union of conjugation classes and hence that $fZ(G) = Z(G)f$ for all $f \in G$.

  (b) First of all, observe that G is a disjoint union of conjugation classes. Further, the orbit-stabilizer theorem and the fact that $|G| = p^n$ for some prime number p, imply that the cardinality of any orbit is a power of p as well. In particular, if an orbit contains more than one element, its cardinality is a multiple of p. Since the elements $f \in Z(G)$ are precisely those elements in G such that $\{f\}$ is an orbit, we see that $|G| \equiv |Z(G)| \pmod{p}$. Since moreover, $Z(G)$ contains the identity element, we then see that $|Z(G)| \geq p > 1$.

  (c) Combining the previous two parts, we conclude that the center of any group $(G, \cdot)$ is a normal subgroup of cardinality at least p. If $Z(G) = G$, any element of G commutes with any other element of G. This would mean that the group is abelian. Hence $Z(G)$ is not one of the trivial subgroups (that is, $\{e\}$ or G) of $(G, \cdot)$.

- *Ch. 5, Ex. 21* Let us first consider the reflections. If n is odd, all n of them are of the same type where the line of reflections connects a midpoint of an edge with the opposite vertex. These all have $m^{(n+1)/2}$ fixed points. If n is even, there are two types of reflection symmetries, of each type $n/2$. The first type has line of reflection connecting two opposite vertices, the second has line of reflection connecting the midpoints of two opposite edges. Hence a reflection symmetry of the first type has $m^{n/2+1}$ fixed points, and one of the second type has $m^{n/2}$ fixed points.

  Now consider the rotation symmetries. Using Corollary 3.2.6, we see that there exist precisely $\phi(d)$ rotation symmetries elements of order d. Such a rotation symmetry has $m^{n/d}$ fixed points. Putting all this together, we obtain that if n is odd, there are precisely

$$\frac{1}{2n}\left(nm^{(n+1)/2} + \sum_{d|n} \phi(d) m^{n/d}\right)$$

  distinct colourings, while if n is even, there are precisely

$$\frac{1}{2n}\left((n/2)m^{n/2} + (n/2)m^{n/2+1} + \sum_{d|n} \phi(d) m^{n/d}\right)$$

distinct colourings.

10.6 Chapter 6

- *Ch. 6, Ex. 1* A. FALSE, B. TRUE, C. TRUE, D. TRUE.
- *Ch. 6, Ex. 2* A, B, D.
- *Ch. 6, Ex. 3* A. TRUE, B. FALSE, C. TRUE.
- *Ch. 6, Ex. 4* C.
- *Ch. 6, Ex. 5* We have $\exp(0) = 1$ and $\exp(r+s) = \exp(r)\cdot\exp(s)$ for any real numbers r and s. Hence exp is a group homomorphism. Its kernel is given by $\{r \in \mathbb{R} \mid \exp(r) = 1\} = \{0\}$.
- *Ch. 6, Ex. 6* The map ψ is a group homomorphism, since $\psi(0) = e^0 = 1$ and $\psi(x+y) = e^{i(x+y)} = e^{ix} \cdot e^{iy}$ for any two complex numbers. Since $e^{ir} = \cos(r) + i\sin(r)$, the kernel of ψ consists of those real numbers r satisfying $\cos(r) + i\sin(r) = 1$. This implies that $\ker\psi = 2\pi\mathbb{Z}$.
- *Ch. 6, Ex. 7* In Ex. 5, we considered the group homomorphism $\exp : \mathbb{R} \to \mathbb{R}\backslash\{0\}$ defined by $\exp(r) = e^r$. The kernel of this group homomorphism is $\{0\}$, while its image is $\mathbb{R}_{>0}$, the set of positive real numbers. We conclude using the isomorphism theorem that the groups $(\mathbb{R}/\{0\}, +)$ and $(\mathbb{R}_{>0}, \cdot)$ are isomorphic with isomorphism $\overline{\exp} : \mathbb{R}/\{0\} \to \mathbb{R}_{>0}$ defined by $\overline{\exp}(r + \{0\}) = e^r$. In Ex. 6, we considered the group homomorphism $\psi : \mathbb{R} \to \mathbb{C}\backslash\{0\}$ defined by $\psi(r) = e^{ir}$. Its image is $C := \{z \in \mathbb{C} \mid |z| = 1\}$, while $\ker\psi = 2\pi\mathbb{Z}$. The isomorphism theorem implies that the groups $(\mathbb{R}/2\pi\mathbb{Z}, +)$ and $(C, \cdot)$ are isomorphic with isomorphism $\overline{\psi} : \mathbb{R}/2\pi\mathbb{Z} \to C$ defined by $\overline{\psi}(r + 2\pi\mathbb{Z}) = e^{ir}$.
- *Ch. 6, Ex. 8* We have for example that:

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

  This cannot be equal to a coset of H, since each coset of H contains precisely two elements.
- *Ch. 6, Ex. 9*

  (a) Checking that H is a subgroup can be done by direct computation. Further $sr^2 = r^{-2}s = r^2s$, since $r^4 = e$. This immediately implies that r^2 commutes with any element in D_4. Hence $fH = \{f, fr^2\} = \{f, r^2f\} = Hf$ for any $f \in D_4$.

  (b) We know that $[D_4 : H] = |D_4|/|H| = 4$. Hence there are four distinct cosets. Since the four cosets $eH = H = \{e, r^2\}$, $rH = \{r, r^3\}$, $sH = \{s, r^2s\}$, and $rsH = \{rs, r^3s\}$ are distinct, we can conclude that the elements e, r, s, rs for a complete set of representatives of all possible cosets of H in D_4.

  (c) Working out all possibilities, we obtain the following multiplication table for $(D_4/H, \circ)$:

$\circ$	H	rH	sH	rsH
H	H	rH	sH	rsH
rH	rH	H	rsH	sH
sH	sH	rsH	H	rH
rsH	rsH	sH	rH	H

(d) Computing the multiplication tables of the groups $(D_4/H, \circ)$ and $(\mathbb{Z}_2 \times \mathbb{Z}_2, +_2)$, we obtain that:

$\circ$	H	rH	sH	rsH
H	H	rH	sH	rsH
rH	rH	H	rsH	sH
sH	sH	rsH	H	rH
rsH	rsH	sH	rH	H

$+_2$	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$(0,0)$	$(0,0)$	$(1,0)$	$(0,1)$	$(1,1)$
$(1,0)$	$(1,0)$	$(0,0)$	$(1,1)$	$(0,1)$
$(0,1)$	$(0,1)$	$(1,1)$	$(0,0)$	$(1,0)$
$(1,1)$	$(1,1)$	$(0,1)$	$(1,0)$	$(0,0)$

Comparing these tables, we see that the map $\chi : D_4/H \to \mathbb{Z}_2 \times \mathbb{Z}_2$ defined by $\psi(H) = (0,0)$, $\psi(rH) = (1,0)$, $\psi(sH) = (0,1)$, and $\psi(rsH) = (1,1)$ is an isomorphism of groups.

- *Ch. 6, Ex. 10*

  Consider the map $\psi : D_4 \to \mathbb{Z}_2 \times \mathbb{Z}_2$ defined by $\psi(r^i s^j) = (i \bmod 2, j \bmod 2)$. Since $r^i r^k s^\ell = r^{i+k} s^\ell$ and $r^i s r^k s^\ell = r^{i-k} s^{1+\ell}$, this defines a group homomorphism. Further $\ker \psi = \{e, r^2\} = H$ and $\mathrm{im}\psi = \mathbb{Z}_2 \times \mathbb{Z}_2$. The isomorphism theorem now implies that the groups $(D_4/H, \circ)$ and $(\mathbb{Z}_2 \times \mathbb{Z}_2, +_2)$ are isomorphic.

- *Ch. 6, Ex. 11* (a) No. Indeed for example

$$\det \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \det \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} = 0,$$

  while

$$\det \left(\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right) = \det \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = 1.$$

  (b) Yes. The map is not injective but it is surjective. Indeed if $r \in \mathbb{R} \setminus \{0\}$ then r coincides with the determinant of the matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & r \end{pmatrix}.$$

  The kernel is given by $\{A \in GL(2, \mathbb{R}) \mid \det A = 1\}$. This set is usually denoted by $SL(2, \mathbb{R})$.

  (c) No. Indeed $\psi(1+2) = 9$ while $\psi(1) + \psi(2) = 1 + 4 = 5$.

  (d) Yes. The map is injective as $r^5 = 1$ implies $r = 1$ in $\mathbb{Q}$ but not surjective. Indeed $2 \in \mathbb{Q} \setminus \{0\}$ is not a 5-th power in $\mathbb{Q}$.

- *Ch. 6, Ex. 12* (a) We need to show that if $[x]_{24} = [y]_{24}$ for $x, y \in \mathbb{Z}$ then $[x]_8 = [y]_8$. If $[x]_{24} = [y]_{24}$ then $x \equiv y \pmod{24}$, that is, $x - y = 24k$ for some $k \in \mathbb{Z}$. Since $x.y = 24k = 8 \cdot (3k)$ we deduce that $x \equiv y \pmod 8$ and hence $[x]_8 = [y]_8$.

  (b) Let $[x]_8 \in \mathbb{Z}_8$. Without loss of generality we can assume $x = 0, \ldots, 7$. Then $\psi([x]_{24}) = [x]_8$.

  (c) $\ker \psi = \{[0]_{24}, [8]_{24}, [16]_{24}\}$.

  (d) The isomorphism theorem implies that $(\mathbb{Z}_{24}/\{[0]_{24}, [8]_{24}, [16]_{24}\}, +_{24})$ is isomorphic to $(\mathbb{Z}_8, +_8)$

- *Ch. 6, Ex. 13* We have

$$(\psi_2 \circ \psi_1)(e_1) = \psi_2(\psi_1(e_1)) = \psi_2(e_2) = e_3$$

and

$$\begin{aligned}(\psi_2 \circ \psi_1)(f \cdot_1 g) &= \psi_2(\psi_1(f \cdot_1 g)) \\ &= \psi_2(\psi_1(f) \cdot_2 \psi_1(g)) \\ &= \psi_2(\psi_1(f)) \cdot_3 \psi_2(\psi_1(g)) \\ &= (\psi_2 \circ \psi_1)(f) \cdot_3 (\psi_2 \circ \psi_1)(g).\end{aligned}$$

Hence $\psi_2 \circ \psi_1$ is a group homomorphism from G_1 to G_3. Now assume that ψ_1 and ψ_2 are in fact group isomorphisms. Then they have inverses $\psi_1^{-1} : G_2 \to G_1$ and $\psi_2^{-1} : G_3 \to G_2$. We have already showed that $\psi_1 \circ \psi_2$ is a group homomorphism. A direct verification shows that $\psi_1^{-1} \circ \psi_2^{-1}$ is the inverse of $\psi_2 \circ \psi_1$, showing that $\psi_2 \circ \psi_1$ is a bijection by Lemma 2.1.4. Hence $\psi_2 \circ \psi_1$ is a group isomorphism, if ψ_1 and ψ_2 are group isomorphisms.

- *Ch. 6, Ex. 14* Assume first that $\ker(\psi) = \{e_1\}$ and that $\psi(f) = \psi(g)$. Since $\psi(g)^{-1} = \psi(g^{-1})$, we obtain that

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

  Hence $fg^{-1} \in \ker(\psi)$. Since we assumed that the kernel of ψ only contains the identity element of G_1, we conclude that $fg^{-1} = e_1$. Hence $f = g$, which implies that ψ is injective.

  Conversely, now assume that ψ is injective. If $f \in \ker(\psi)$, then $\psi(f) = e_2 = \psi(e_1)$. The injectivity of ψ then implies that $f = e_1$. Therefore $\ker(\psi) = \{e_1\}$.

- *Ch. 6, Ex. 15* Let $f, g \in \psi(G_1)$. By Chapter 2, we can conclude that $\psi(G_1)$ is a subgroup if we can show that $f \cdot_2 g^{-1} \in \psi(G_1)$. Since $f, g \in \psi(G_1)$, there exist $k, \ell \in G_1$ such that $\psi(k) = f$ and $\psi(\ell) = g$. Then we have

  $$\psi(k \cdot_1 \ell^{-1}) = \psi(k) \cdot_2 \psi(\ell^{-1}) = \psi(k) \cdot_2 \psi(\ell)^{-1} = f \cdot_2 g^{-1}.$$

  Hence $f \cdot_2 g^{-1} \in \psi(G_1)$, which is what we wanted to show.

- *Ch. 6, Ex. 16* Cosets of $\{e\}$ are of the form $\{g\}$ for $g \in G$. The isomorphism is given by $\psi(\{g\}) = g$. If $\psi : G_1 \to G_2$ is an injective group homomorphism, then $\ker(\psi) = \{e_1\}$. The isomorphism theorem then implies that $(G_1/\{e_1\}, \cdot_1)$ is isomorphic to $(\mathrm{im}(\psi), \cdot_2)$. By the previous $(G_1/\{e_1\}, \cdot_1)$ and $(G_1, \cdot_1)$ are isomorphic. Combining the two isomorphisms, we see that the groups $(G_1, \cdot_1)$ and $(\mathrm{im}(\psi), \cdot_2)$ are isomorphic.

- *Ch. 6, Ex. 17* First of all, by Lagrange's theorem (Theorem 4.4.1), the order of a subgroup can be 1, 2, 4, or 8. In case the order is 1, reps. 8, the subgroup is $\{1\}$, resp. Q_8, which is a normal subgroup. If H is a subgroup of order 4, Exercise 24 implies that H is a normal subgroup. This leaves subgroups of $(Q_8, \cdot)$ of order 2. Since -1 is the only element of Q_8 of order 2, we have $H = \{1, -1\}$ in this case. Since both 1 and -1 commute with any other element of Q_8, this is also a normal subgroup.

- *Ch. 6, Ex. 18* A group $(G, \cdot)$ is isomorphic to itself, since $\mathrm{id}_G : G \to G$ is an isomorphism. Hence being isomorphic is a reflexive relation. If $\psi : G_1 \to G_2$ is a group isomorphism, then so is $\psi^{-1} : G_2 \to G_1$. Hence being isomorphic is a symmetric relation. Transitivity follows from Exercise 13.

- *Ch. 6, Ex. 19* First of all, Exercise 13 implies that composition is an operator on the set $\mathrm{Aut}(G)$. The fact that it is an associative operator follows from Lemma 2.1.2. The identity isomorphism $\mathrm{id}_G : G \to G$ is the identity element, while the inverse of an automorphism ψ exists, since any isomorphism has an inverse.

- *Ch. 6, Ex. 20* (a) The given operator is associative, since

$$\begin{aligned}((n_1,h_1)\cdot_\psi(n_2,h_2))\cdot_\psi(n_3,h_3) &= (n_1\cdot_1\psi(h_1)(n_2),h_1\cdot_2 h_2)\cdot_\psi(n_3,h_3)\\ &= (n_1\cdot_1\psi(h_1)(n_2)\cdot_1\psi(h_1\cdot_2 h_2)(n_3),(h_1\cdot_2 h_2)\cdot_2 h_3)\\ &= (n_1\cdot_1\psi(h_1)(n_2)\cdot_1\psi(h_1)(\psi_{h_2}(n_3)),h_1\cdot_2(h_2\cdot_2 h_3))\\ &= (n_1\cdot_1\psi(h_1)(n_2\cdot_1\psi_{h_2}(n_3)),h_1\cdot_2(h_2\cdot_2 h_3))\\ &= (n_1\cdot_1\psi(h_1)(n_2\cdot_1\psi_{h_2}(n_3)),h_1\cdot_2(h_2\cdot_2 h_3))\\ &= (n_1,h_1)\cdot_\psi(n_2\cdot_1\psi_{h_2}(n_3),h_2\cdot_2 h_3)\\ &= (n_1,h_1)\cdot_\psi((n_2,h_2)\cdot_\psi(n_3,h_3)).\end{aligned}$$

  Further (e_1,e_2) is the identity element, while $(n_1,h_1)^{-1}=(\psi(h_1^{-1})(n_1^{-1}),h_1^{-1})$.

  (b) If $\psi(h)=\mathrm{id}_N$ for any $h\in H$, then $(n_1,h_1)\cdot_\psi(n_2,h_2)=(n_1\cdot_1 n_2,h_1\cdot_2 h_2)$. Hence in this case the semidirect product is the usual direct product of groups.

  (c) Note that the map $\iota: C_n\to C_n$ such that $r^i\mapsto r^{-i}$ is an automorphism of C_n. Now let $\psi:\{1,-1\}\to\mathrm{Aut}(C_n)$ be defined as $1\mapsto\mathrm{id}_{C_n}$ and $-1\mapsto\iota$. Then ψ is a group homomorphism. We claim that the semidirect product $(C_n\rtimes_\psi\{1,-1\},\cdot_\psi)$ is isomorphic to the dihedral group $(D_n,\circ)$ by identifying the element $(r,1)$ with r and the element $(e,-1)$ with s. First of all, note that $C_n\rtimes_\psi\{1,-1\}$ and D_n both contain $2n$ elements and that the proposed identification gives rise to a bijection. To see that it gives an isomorphism of groups, the main fact that we need to verify, is that $(r,1)$ and $(e,-1)$ satisfy the same relations as r and s, namely $(r,1)^n=e$, $(e,-1)^2=e$ and $(e,-1)(r,1)=(r^{-1},1)(e,-1)$. Using induction on i, one obtains that $(r,1)^i=(r^i,1)$, whence $(r,1)$ has the same order as r. In particular, $r^n=e$. Further $(e,-1)^2=(e\iota(-1)(e),(-1)^2)=(e,1)$ and $(e,-1)\cdot_\psi(r,1)=(e\circ\iota(-1)(r),-1)=(r^{-1},-1)=(r^{-1}\circ\iota(1)(e),-1)=(r^{-1},1)\cdot_\psi(e,-1)$.

- *Ch. 6, Ex. 21* (a) Let $f,g\in HN$. By Lemma 4.1.2, we can conclude that HN is a subgroup if we can show that $f\cdot g^{-1}\in HN$. By definition of HN we can write $f=h_1n_1$ and $g=h_2n_2$ for some $h_1,h_2\in H$ and $n_1,n_2\in N$. Since $g^{-1}=n_2^{-1}h_2^{-1}$ we get $f\cdot g^{-1}=h_1n_1n_2^{-1}h_2^{-1}=h_1(n_1n_2^{-1}h_2^{-1})$. Since N is a normal subgroup in G (left and right cosets coincide!) there must exists $n_3\in N$ such that $n_1n_2^{-1}h_2^{-1}=h_2^{-1}n_3$. So $f\cdot g^{-1}=h_1(n_1n_2^{-1}h_2^{-1})=h_1h_2^{-1}n_3\in HN$.

  (b) Let $h\in H\cap N$. Recall that N is normal in G, which means that for all $n\in N$ and all $g\in G$, $gng^{-1}\in N$. This applies in particular when $n=h$ and $g\in H$, giving the definition of $H\cap N$ to be normal in H.

  (c) Following the hint we consider the map $\psi: H\to HN/N$ with $\psi(h)=hN$. It is easy to see that this map is a group homomorphism. The kernel of ψ is given by $\{h\in H\mid\psi(h)=N\}=\{h\in H\mid h\in N\}=H\cap N$. Also the map is surjective. Indeed if $dN\in HN/N$ is arbitrary then there must exists $h\in H$ and $n\in N$ such that $d=hn$. Hence $dN=hnN=hN$. This shows that $\psi(h)=dN$ and ψ is surjective. The desired result follows now directly from the isomorphism theorem.

- *Ch. 6, Ex. 22* (a) Since for all $f\in G$ it holds that $fN=Nf$, it certainly holds for all $f\in K$ that $fN=Nf$.

  (b) Consider the map $\psi: G/N\to G/K$ given by $fN\mapsto fK$. First of all, this is a well-defined map: if $fN=gN$, then $f^{-1}g\in N$. Since $N\subseteq K$, we then obtain that $f^{-1}g\in K$, which in turn implies that $fK=gK$. A direct verification shows that ψ is a homomorphism of groups. Further: $\ker\psi=\{fN\mid fK=K\}=\{fN\mid f\in K\}=K/N$. Theorem 6.1.9 then implies that K/N is a normal subgroup of $(G/N,\cdot)$.

(c) Consider the same group homomorphism $\psi : G/N \to G/K$ as in the previous part. Since $\psi(fN) = fK$, we see that $\mathrm{im}\psi = G/K$. Since we already know that $\ker\psi = K/N$, the isomorphism theorem for groups implies that $((G/N)/(K/N), \cdot)$ and $(G/K, \cdot)$ are isomorphic groups.

10.7 Chapter 7

- *Ch. 7, Ex. 1* A. TRUE, B. FALSE, C. TRUE, D. TRUE, E. FALSE.
- *Ch. 7, Ex. 2* C.
- *Ch. 7, Ex. 3* D.
- *Ch. 7, Ex. 4* A. TRUE, B. FALSE.
- *Ch. 7, Ex. 5* A.
- *Ch. 7, Ex. 6* B.
- *Ch. 7, Ex. 7* D.
- *Ch. 7, Ex. 8* A. FALSE, B. TRUE, C. TRUE, D. TRUE.
- *Ch. 7, Ex. 9* A. TRUE, B. FALSE, C. TRUE, D. FALSE, E. TRUE.
- *Ch. 7, Ex. 10*
  - No, because $(\mathbb{N}, +)$ is not a group [see the exercises in Chapter 3].
  - $(D, +, \cdot)$ is a commutative ring. The zero-element is the zero-matrix and the identity element for multiplication is the identity matrix. It is not a domain, nor a field. Indeed the diagonal matrix
    $$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$$
    is a zero-divisor and it does not admit a multiplicative inverse.
  - $(\{0_R\}, +_R, \cdot_R)$ is a ring, though a somewhat artificial one. Note that the neutral element for multiplication is 0_R. The ring is commutative and is also a domain, since there are no zero-divisors (recall that a zero-divisor is by definition not equal to 0_R). The ring is not a

    field. A commutative ring $(R, +_R, \cdot_R)$ is a field if $R^* = R\backslash\{0_R\}$. However, 0_R is a unit in the zero-ring, since $0_R \cdot_R 0_R = 0_R$ and the identity element for the multiplication is 0_R in the zero-ring as well.
- *Ch. 7, Ex. 11*
  - Since $0_R +_R r = r$ for all $r \in R$, this equation is also true for $r = 0_R$. Therefore $0_R +_R 0_R = 0_R$ as suggested in the hint. Now let $x \in R$ be an arbitrarily chosen ring element. Multiplying both sides of the equation $0_R +_R 0_R = 0_R$ from the right with x, one obtains that
    $$(0_R +_R 0_R) \cdot_R x = 0_R \cdot_R x.$$

By the distributive law this implies that

$$0_R \cdot_R x +_R 0_R \cdot_R x = 0_R \cdot_R x.$$

Now we add the additive inverse of the ring element $0_R \cdot_R x$ on both sides of this equation, obtaining

$$-0_R \cdot_R x +_R (0_R \cdot_R x +_R 0_R \cdot_R x) = -0_R \cdot_R x + 0_R \cdot_R x.$$

The right-hand side simplifies to 0_R, while using the associative law for the addition, the left-hand side becomes:

$$-0_R \cdot_R x +_R (0_R \cdot_R x +_R 0_R \cdot_R x) = (-0_R \cdot_R x +_R 0_R \cdot_R x) +_R 0_R \cdot_R x$$

$$= 0_R +_R 0_R \cdot_R x = 0_R \cdot_R x.$$

Hence $0_R \cdot_R x = 0_R$. In a similar way, one can show that $x \cdot_R 0_R = 0_R$.

– It is not part of the rings axioms that the elements 0_R and 1_R are distinct. The goal of this exercise is to show that the only ring in which they are the same is the zero-ring. If $1_R = 0_R$, then for any $x \in R$, it will hold that $1_R \cdot_R x = 0_R \cdot_R x$. But then $x = 1_R \cdot_R x = 0_R \cdot_R x = 0_R$, where we used the previous part of the exercise in the last equality. This shows that every $x \in R$ would coincide with 0_R and hence $R = \{0_R\}$.

- *Ch. 7, Ex. 12* The operations $+_R$ and $\cdot_R$ make sense on $Z(R)$ as $Z(R)$ is a subset of R. We first observe that both 0_R and 1_R are contained in $Z(R)$. Indeed for all $r \in R$, $1_R \cdot_R r = r \cdot_R 1_R = r$ by the definition of 1_R, while $r \cdot_R 0_R = 0_R \cdot_R r = 0_R$ from the previous exercise. Note also that if $(Z(R), +_R, \cdot_R)$ is a ring then it is commutative.

  We need to show that $+_R$ and $\cdot_R$ are well-defined operations on $Z(R)$, that is, if $a, b \in Z(R)$ then both $a +_R b$ and $a \cdot_R b$ are in $Z(R)$. Let $r \in R$ be arbitrary. Using the distributive laws:

  $$(a +_R b) \cdot_R r = a \cdot_R r +_R b \cdot_R r.$$

  Since $a, b \in Z(R)$ we know that $a \cdot_R r = r \cdot_R a$ and $b \cdot_R r = r \cdot_R b$ and hence (using also the associative law and distributive law)

  $$a \cdot_R r +_R b \cdot_R r = (a \cdot_R r) +_R (b \cdot_R r) = (r \cdot_R a) +_R (r \cdot_R b) = r \cdot_R (a +_R b).$$

  This shows that for all $r \in R$, $(a +_R b) \cdot_R r = r \cdot_R (a +_R b)$ and hence $a +_R b \in Z(R)$.

  Analogously, let $r \in R$ be arbitrary. From $a \cdot_R b = b \cdot_R a$ and the associative law we have

  $$(a \cdot_R b) \cdot_R r = (b \cdot_R a) \cdot_R r = b \cdot_R (a \cdot_R r).$$

  Since $a \cdot_R r = r \cdot_R a$ and $b \cdot_R r = r \cdot_R b$ we get again from the associative law,

  $$b \cdot_R (a \cdot_R r) = b \cdot_R (r \cdot_R a) = (b \cdot_R r) \cdot_R a = (r \cdot_R b) \cdot_R a = r \cdot_R (a \cdot_R b).$$

  This shows $(a \cdot_R b) \cdot_R r = e \cdot_R (a \cdot_R b)$ and so $a \cdot_R b \in Z(R)$.

  The associative property as well as the distributive laws are trivially satisfied since they hold in $(R, +_R, \cdot_R)$ (in $Z(R) \subseteq R$ we are just considering less elements!). To show that $(Z(R), +_R)$ is an abelian group (we know that the identity element will be 0_R) one can use that $(R, +_R)$ is an abelian group when checking the axioms. Note that $Z(R) = R$ if and only if $(R, +_R, \cdot_R)$ is a commutative ring.

- *Ch. 7, Ex. 13* Note that 1 is a root of $X^3 + X^2 + X + 1 \in \mathbb{Z}_2[X]$ as $1 + 1 + 1 + 1 = 4 \equiv 0 \pmod 2$. From Proposition 152 we know that $X^3 + X^2 + X + 1 = p(X) \cdot (X - 1) = p(X) \cdot (X + 1)$ for some polynomial $p(X) \in \mathbb{Z}_2[X]$. Using the division algorithm one gets $X^3 + X^2 + X + 1 = (X^2 + 1) \cdot (X + 1)$. Since 1 is also a root of $X^2 + 1$, we must be able to write $X^2 + 1 = P_1(X) \cdot (X + 1)$ fo some $P_1(X) \in \mathbb{Z}_2[X]$. Using again the division algorithm one gets $X^2 + 1 = (X + 1)^2$ and hence $X^3 + X^2 + X + 1 = (X + 1)^3$.

- *Ch. 7, Ex. 14* Using the Extended Euclidian Algorithm one gets that $\lambda = -15$ and $\gamma = 4$. Hence
$$-15 \cdot 13 = 1 - 4 \cdot 49 \equiv 1 \pmod{49}.$$
This shows that $-15 \equiv 34 \pmod{49}$ is the multiplicative inverse of 13 in $\mathbb{Z}_{49}$ as the product of -15 and 13 is the identity element for $\cdot_{49}$. Hence 13 is a unit in $\mathbb{Z}_{49}$. An example of zero-divisor in $\mathbb{Z}_{49}$ is 7 because $7 \not\equiv 0 \pmod{49}$ but $7 \cdot 7 = 49 \equiv 0 \pmod{49}$.

- *Ch. 7, Ex. 15*
  - Let $u \in R$ be arbitrary. We need to show that $(-1_R) \cdot_R u +_R u = u +_R (-1_R) \cdot_R u = 0_R$. First we see that,
$$(-1_R) \cdot_R u +_R u = (-1_R) \cdot_R u +_R (1_R \cdot_R u),$$
and from the distributive law
$$(-1_R) \cdot_R u +_R (1_R \cdot_R u) = (-1_R +_R 1_R) \cdot_R u = 0_R \cdot_R u.$$
From Exercise 11 $0_R \cdot_R u = u \cdot_R 0_R = 0_R$ and hence we get $(-1_R) \cdot_R u +_R u = 0_R$. The proof of $u +_R (-1_R) \cdot_R u = 0_R$ is analogous.
  - By definition of $u \in R$ being a unit we know that there exists $v \in R \setminus \{0_R\}$ such that $u \cdot_R v = v \cdot_R u = 1_R$. We want to show that also $-u$ is a unit by proving that $(-u) \cdot_R (-v) = u \cdot_R v = 1_R$ and similarly $(-v) \cdot_R (-u) = v \cdot_R v = 1_R$. To this aim we show here that $(-u) \cdot_R (-v) = u \cdot_R v$ as the proof of $(-v) \cdot_R (-u) = v \cdot_R u$ is completely analogous.
  We first show that $-(u \cdot_R v) = (-u) \cdot_R v$. This is just a consequence of the distributive/associative laws and Exercise 11 as
$$(-u) \cdot_R v +_R (u \cdot_R v) = (-u +_R u) \cdot_R v = 0_R \cdot_R v = 0_R.$$
and
$$(u \cdot_R v) +_R (-u) \cdot_R v = (u +_R (-u)) \cdot_R v = 0_R \cdot_R v = 0_R.$$
To prove that $(-u) \cdot_R (-v) = u \cdot_R v$ we see that since $-(u \cdot_R v) = (-u) \cdot_R v$,
$$\begin{aligned}(-u) \cdot_R (-v) +_R -(u \cdot_R v) &= (-u) \cdot_R (-v) +_R (-u) \cdot_R v \\ &= -u \cdot_R (-v +_R v) = -u \cdot_R 0_R = 0_R.\end{aligned}$$
  - If $u \in R \setminus \{0_R\}$ is a zero-divisor then there exists $v \in R \setminus \{0_R\}$ such that $u \cdot_R v = v \cdot_R u = 0_R$ (recall that the ring is assumed to be commutative in this exercise). Hence using the commutative ring axioms,
$$(u \cdot_R x) \cdot_R v = u \cdot_R (x \cdot_R v) = u \cdot_R (v \cdot_R x) = (u \cdot_R v) \cdot_R x = 0_R \cdot_R x = 0_R.$$
Since $v \neq 0_R$ this shows that either $u \cdot_R x = 0_R$ or $u \cdot_R x$ is a zero-divisor.

– First of all, using the distributive laws and the first point of this exercise, one shows that $x^2 - 1_R = (x +_R 1_R) \cdot (x +_R -1_R)$. Hence $x^2 = 1_R$ implies that $(x +_R 1_R) \cdot (x +_R -1_R) = 0_R$. Since R is a domain, and hence it has no zero-divisors, we get that either $x +_R 1_R = 0_R$ or $x +_R -1_R = 0_R$. This proves that $x = \pm 1_R$.

- *Ch. 7, Ex. 16* Since $r^2 = r$ for all $r \in R$, this is in particular true for $1_R + 1_R$. Hence from the distributive/associative laws

$$1_R +_R 1_R = (1_R +_R 1_R)^2 = (1_R +_R 1_R) \cdot_R (1_R +_R 1_R) = 1_R +_R 1_R +_R 1_R +_R 1_R$$
$$= (1_R +_R 1_R) +_R (1_R +_R 1_R).$$

Adding $-(1_R +_R 1_R)$ to both sides one gets $0_R = 1_R +_R 1_R$ which shows that $1_R = -1_R$ as suggested in the hint. Let now $x, y \in R$. We need to show that $x \cdot_R y = y \cdot_R x$. To this aim, following the hint we have that

$$x +_R y = (x +_R y)^2 = x^2 +_R y^2 + x \cdot_R y +_R y \cdot_R x.$$

From $x^2 = x$ and $y^2 = y$ we get that

$$x +_R y = x +_R y + x \cdot_R y +_R y \cdot_R x.$$

Adding $-(x +_R y)$ to both sides we obtain $0_R = x \cdot_R y +_R y \cdot_R x$, and hence $x \cdot_R y = -(y \cdot_R x)$. From the first item of Exercise 11 we know that $-(y \cdot_R x) = (-1_R) \cdot_R (y \cdot_R x)$, but since in this case $-1_R = 1_R$ we obtain $x \cdot_R y = (-1_R) \cdot_R (y \cdot_R x) = y \cdot_R x$. This proves the commutativity property.

- *Ch. 7, Ex. 17* If $x \in R^*$, and $x \cdot_R y = 0_R$, then $y = x^{-1} \cdot_R x \cdot_R y = x^{-1} \cdot_R 0_R = 0_R$. Similarly $y \cdot_R x = 0_R$ implies $y = 0_R$ as well. Hence x cannot be a zero-divisor.

- *Ch. 7, Ex. 18* There are several possibilities and one of them is the following. We know that if R is a domain, a polynomial in $R[X]$ of degree two can have at most two roots in R. Therefore to solve the exercise, we need to use that $\mathbb{Z}_6$ is not a domain. Indeed, it has zero-divisors, since for example $2 \cdot_6 3 = 0$. Now consider the polynomial

$$p(X) := (X + 2)(X + 3) = X^2 + 5X.$$

Since $2 \cdot_6 3 = 0$, 0 is a root of $p(X)$, but also $-2 = 4$ and $-3 = 3$ are roots, since $X + 2$ and $X + 3$ are factors of $p(X)$. This shows that $p(X)$ has at least three roots in $\mathbb{Z}_6$.

Remark: In fact, since also $p(X) = X(X + 5)$, also $-5 = 1$ is a root of $p(X)$. It turns out that $p(X)$ has exactly 4 roots in $\mathbb{Z}_6$: The polynomial $3X^2 + 3X$ has even all six elements of $\mathbb{Z}_6$ as roots!

- *Ch. 7, Ex. 19* We can write $n = n_1 d$ and $a = a_1 d$ for some integers a_1 and a_2. Suppose by contradiction that $c \in \mathbb{Z}_n$ is a root of $aX - b \in \mathbb{Z}_n[X]$. Then $ac - b \equiv 0 \pmod{n}$ that is, there exists $m \in \mathbb{Z}$ such that $ac = b + mn$. Hence $a_1 dc = b + mn_1 d$, and $(a_1 - mn_1)d = b$. Since $b \neq 0$ we get that d divides b, which is not possible as d divides n and $\gcd(n, b) = 1$.

- *Ch. 7, Ex. 20*

– Recall that $m \in \mathbb{Z}_\ell$ is a zero-divisor if and only if $m \neq 0 \pmod{p^\ell}$ and $\gcd(m, p^\ell) > 1$. Hence if we define $m = p$ we see that m is a zero-divisor in $\mathbb{Z}_{p^\ell}$. Indeed as $\ell > 1$, $p \not\equiv 0 \pmod{p^\ell}$, $p^{\ell-1} \not\equiv 0 \pmod{p^\ell}$ but $p \cdot p^{\ell-1} = p^\ell \equiv 0 \pmod{p^\ell}$.

– Recall that $m \in \mathbb{Z}_{p^\ell} = \{0, 1, \ldots, p^{\ell-1}\}$ is a unit if and only if $\gcd(m, p^\ell) = 1$ and the non-units are exactly all the $m \in \{0, 1, \ldots, p^{\ell-1}\}$ such that $\gcd(m, p^\ell) \neq 1$, that is, all the $m \in \{0, 1, \ldots, p^{\ell-1}\}$ which are divisible by p. To count such m's we write $m = m_1 p$ where $m_1 = 1, \ldots, p^{\ell-1}$ and observe that in this way we obtain exactly $p^{\ell-1}$ distinct values.

- *Ch. 7, Ex. 21* The element $1 + \sqrt{2}$ is a unit, since $(1 + \sqrt{2}) \cdot (-1 + \sqrt{2}) = 1$. Since for any ring $(R, +, \cdot)$, $(R^*, \cdot)$ is a group, any power of $1 + \sqrt{2}$ is a unit as well. Alternatively, just note that for any $n \in \mathbb{N}$:

$$(1 + \sqrt{2})^n \cdot (-1 + \sqrt{2})^n = \Big((1 + \sqrt{2}) \cdot (-1 + \sqrt{2})\Big)^n = 1.$$

We claim that the unit $1+\sqrt{2}$ has infinite (multiplicative) order. Suppose that $(1+\sqrt{2})^m = 1$ for some $m \in \mathbb{Z}$, then $|1 + \sqrt{2}|^m = 1$. However, since $|1 + \sqrt{2}| > 1$, this would imply that $m = 0$. This proves the claim. Hence all units of type $(1 + \sqrt{2})^n$ with $n \in \mathbb{N}$ are distinct and $\mathbb{Z}[\sqrt{2}]$ contains infinitely many units.

- *Ch. 7, Ex. 22*

– This follows by taking lengths of the complex numbers, since $|e + fi| = |(a + bi)(c + di)| = |a + bi| \cdot |c + di|$. Alternatively, one can argue as follows: if $(a + bi)(c + di) = (e + fi)$, then also $(a - bi)(c - di) = (e - fi)$. Hence

$$(a^2 + b^2)(c^2 + d^2) = (a + bi)(c + di)(a - bi)(c - di) = (e + fi)(e - fi) = e^2 + f^2.$$

If $(a + bi)(c + di) = 0$ then $(a^2 + b^2)(c^2 + d^2) = 0$.

– Using that $\mathbb{Z}$ has no zero-divisors, we see that either $a^2 + b^2 = 0$ or $c^2 + d^2 = 0$. In the first case $a = b = 0$ and in the second $c = d = 0$.

– If $a+bi$ is a unit in $\mathbb{Z}[i]$, then there exist integers c and d such that $(a+bi)(c+di) = 1$. Using the first part of the exercise, we see that $(a^2+b^2)(c^2+d^2) = 1$. Since a, b, c, d are integers and a square of an integer cannot be negative, this implies that $a^2 + b^2 = 1$. Still using that a and b are integers, this gives rise to exactly 4 possibilities for (a, b), namely $(1, 1)$, $(1, -1)$, $(-1, 1)$ and $(-1, -1)$.

- *Ch. 7, Ex. 23*

– The desired equalities simply follow by a direct computation of all the matrix products involved using the explicit description of $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{k}$.

– We first need to show that the usual matrix addition and multiplication are well-defined operations in $\mathbb{H}$, that is, if $a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}, a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k} \in \mathbb{H}$ then also $(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) + (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \in \mathbb{H}$ and $(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) \cdot (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) \in \mathbb{H}$. This can be done for example by proving directly (using the previous part of the exercise) that the following equalities are satisfied

$$(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) + (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) =$$
$$(a_1 + a_2)\mathbf{1} + (b_1 + b_2)\mathbf{i} + (c_1 + c_2)\mathbf{j} + (d_1 + d_2)\mathbf{k}$$

and

$$(a_1\mathbf{1} + b_1\mathbf{i} + c_1\mathbf{j} + d_1\mathbf{k}) \cdot (a_2\mathbf{1} + b_2\mathbf{i} + c_2\mathbf{j} + d_2\mathbf{k}) =$$

$$\alpha\mathbf{1}+\beta\mathbf{i}+\gamma\mathbf{j}+\delta\mathbf{k}$$

where

$$\begin{cases}\alpha = a_1a_2 - b_1b_2 - c_1c_2 - d_1d_2,\\ \beta = a_1b_2 + a_1b_1 + c_1d_2 - d_1c_2,\\ \gamma = a_1c_2 - b_1d_2 + c_1a_2 - d_1b_2,\\ \delta = a_1d_2 + b_1c_2 - c_1b_2 - d_1a_2.\end{cases}$$

The identity element for addition is the zero-matrix $0\mathbf{1}+0\mathbf{i}+0\mathbf{j}+0\mathbf{k}$ and if the think the addition in this ring is the usual addition of real numbers coordinatewise (where the coordinates are the coefficients of $\mathbf{1}$, $\mathbf{i}$, $\mathbf{j}$ and $\mathbf{j}$) we see that $(\mathbb{H},+)$ is an abelian group because $(\mathbb{R},+)$ is an abelian group.

The identity element for the matrix multiplication in $\mathbb{H}$ is given by the identity matrix $\mathbf{1}$. This can be seen by observing that if $(a_2,b_2,c_2,d_2)=(1,0,0,0)$ in the formula above then $(\alpha,\beta,\gamma,\delta)=(a_1,b_1,c_1,d_1)$ so that $(a_1\mathbf{1}+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})\cdot\mathbf{1}=(a_1\mathbf{1}+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})$. The proof of $\mathbf{1}\cdot(a_1\mathbf{1}+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})\cdot\mathbf{1}=(a_1\mathbf{1}+b_1\mathbf{i}+c_1\mathbf{j}+d_1\mathbf{k})$ is analogous. The associativite and distributive laws follows from the assumption that they are satisfied by the usual addition of multiplication of matrices (and we are just restricting our attention to a set $\mathbb{H}$ of special matrices, so the properties are satisfied for free).

– To prove that every non-zero element in $\mathbb{H}$ (i.e. all but the zero-matrix) admits a multiplicative inverse, one first show that the hint is true by a direct computation. Suppose we want to find the multiplicative inverse of the non-zero element $a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}\in\mathbb{H}$ with $a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}\neq 0$. From the hint we know that $(a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k})(a\mathbf{1}-b\mathbf{i}-c\mathbf{j}-d\mathbf{k})=a^2+b^2+c^2+d^2\in\mathbb{R}$. Hence one can simply check that

$$(a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k})\left(\frac{a\mathbf{1}-b\mathbf{i}-c\mathbf{j}-d\mathbf{k}}{a^2+b^2+c^2+d^2}\right)=\mathbf{1},$$

where dividing the matrix $a\mathbf{1}-b\mathbf{i}-c\mathbf{j}-d\mathbf{k}$ by $a^2+b^2+c^2+d^2\in\mathbb{R}$ means that every entry in $a\mathbf{1}-b\mathbf{i}-c\mathbf{j}-d\mathbf{k}$ is divided by $a^2+b^2+c^2+d^2\in\mathbb{R}$. Doing so, we can see that the matrix

$$\frac{a\mathbf{1}-b\mathbf{i}-c\mathbf{j}-d\mathbf{k}}{a^2+b^2+c^2+d^2}$$

is the multiplicative inverse of $a\mathbf{1}+b\mathbf{i}+c\mathbf{j}+d\mathbf{k}$.

10.8 Chapter 8

- *Ch. 8, Ex. 1* A. FALSE, B.TRUE, C. FALSE, D. TRUE, E. FALSE, F. TRUE, G. FALSE, H. TRUE.
- *Ch. 8, Ex. 2* C.
- *Ch. 8, Ex. 3* A. TRUE, B. TRUE, C. TRUE, D. FALSE, E. TRUE.
- *Ch. 8, Ex. 4* A.
- *Ch. 8, Ex. 5* B.

- *Ch. 8, Ex. 6* It is not, since it is not closed under addition: $1 +_8 2$ is for example not in $\{0,1,2,4\}$. First of all note that any ideal I of a ring $(R, +_R, \cdot_R)$ contains 0_R. Also, any ring $(R, +_R, \cdot_R)$ has the ideals $\{0_R\}$ and R. If an ideal contains 1_R, it has to be the entire ring, since if $1_R \in I$, then $r = r \cdot_R 1_R \in I$ for all $r \in R$. More generally, if an ideal contains a unit u, it has to be the entire ring, since $r = r \cdot_R u^{-1} \cdot_R u \in I$ for all $r \in R$.

  Since $(\mathbb{Z}_8)^* = \{1,3,5,7\}$, this means that any ideal not equal to $\mathbb{Z}_8$ has to be contained in $\{0,2,4,6\}$. A direct computation shows that $\{0,2,4,6\}$ in fact is an ideal of the ring $(\mathbb{Z}_8, +_8, \cdot_8)$. If an ideal contains 2 or 6, it is equal to $\{0,2,4,6\}$. Indeed, if $2 \in I$, then $2 \cdot_8 2 = 4 \in I$ and $3 \cdot_8 2 = 6 \in I$. If $6 \in I$, then $2 \cdot_8 6 = 4 \in I$ and $3 \cdot_8 6 = 2 \in I$. This means that the only ideal contained in, but not equal to $\{0,2,4,6\}$, has to be either $\{0\}$ or $\{0,4\}$. A direct computation shows that $\{0,4\}$ in fact is an ideal. This gives a complete list of ideals in the ring $(\mathbb{Z}_8, +_8, \cdot_8)$, namely $\{0\}$, $\{0,4\}$, $\{0,2,4,6\}$, and $\mathbb{Z}_8$ itself.

- *Ch. 8, Ex. 7* R is a ring in which the additive identity element is the zero-matrix while the multiplicative identity element is the identity matrix. Also $\mathbb{Z} \times \mathbb{Z}$ is a ring with coordinatewise addition and multiplication: for $(a_1, c_1), (a_2, c_2) \in \mathbb{Z} \times \mathbb{Z}$ we define $(a_1, c_1) + (a_2, c_2) = (a_1 + a_2, c_1 + c_2)$ and $(a_1, c_1) \cdot (a_2, c_2) = (a_1 a_2, c_1 c_2)$. The zero-element in the ring is $(0,0)$ while the multiplicative identity element is $(1,1)$. We prove that ϕ is a ring homomorphism. First of all we need to show that ϕ is a group homomorphism between $(R, +)$ and $(\mathbb{Z} \times \mathbb{Z}, +)$. This is true because

$$\phi \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} = (0,0),$$

  and for two elements in R with entries $a_1, b_1, c_1 \in \mathbb{Z}$ and $a_2, b_2, c_2 \in \mathbb{Z}$ respectively one has

$$\phi\left(\begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix} + \begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix} \right) = \phi \begin{pmatrix} a_1 + a_2 & b_1 + b_2 \\ 0 & c_1 + c_2 \end{pmatrix}$$

$$= (a_1 + a_2, c_1 + c_2) = (a_1, c_1) + (a_2, c_2) = \phi \begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix} + \phi \begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix}.$$

  Now we check that the identity matrix is mapped to the identity element of $\mathbb{Z} \times \mathbb{Z}$, which is the pair $(1,1)$.This is true because the diagonal elements of the identity matrix are indeed both equal to 1. Finally two elements in R with entries $a_1, b_1, c_1 \in \mathbb{Z}$ and $a_2, b_2, c_2 \in \mathbb{Z}$ respectively one has

$$\phi\left(\begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix} \cdot \begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix} \right) = \phi \begin{pmatrix} a_1 a_2 & a_1 b_2 + b_1 c_2 \\ 0 & c_1 c_2 \end{pmatrix}$$

$$= (a_1 a_2, c_1 c_2) = (a_1, c_1) \cdot (a_2, c_2) = \phi \begin{pmatrix} a_1 & b_1 \\ 0 & c_1 \end{pmatrix} \cdot \phi \begin{pmatrix} a_2 & b_2 \\ 0 & c_2 \end{pmatrix}.$$

  This shows that ϕ is a ring homomorphism. The surjectivity follows from the fact that an arbitrary pair $(a, c) \in \mathbb{Z} \times \mathbb{Z}$ is (for example) the image of the matrix

$$\begin{pmatrix} a & 0 \\ 0 & c \end{pmatrix}.$$

  Finally

$$\ker \phi := \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mid b \in \mathbb{Z} \right\},$$

which is a copy of $\mathbb{Z}$ inside R.

- *Ch. 8, Ex. 8*
  - One has
  $$7X^{13} - 11X^9 + 5X^5 - 2X^3 + 3 =$$
  $$(7X^9 + 101X^5 + 1621X) \cdot (X^4 - 16) - 2X^3 + 25936X + 3,$$
  so the quotient is $7X^9 + 101X^5 + 1621X$ and the reminder is $-2X^3 + 25936X + 3$.
  - As suggested also in Exercise 5, if we want to compute the standard form of a coset $g(X)+\langle f(X)\rangle$, once we have computed the division with reminder $g(X) = q(X)f(X)+r(X)$, the standard form is just given by the coset $r(X) + \langle f(X)\rangle$. Hence from the previous part of the exercise, the standard form of $7X^{13} - 11X^9 + 5X^5 - 2X^3 + 3 + \langle X^4 - 16\rangle$ is $-2X^3 + 25936X + 3 + \langle X^4 - 16\rangle$.
  - We see that $X^4 - 16 = (X^2 - 4)(X^2 + 4) = (X - 2)(X + 2)(X^2 + 4)$. Hence following the hint $p_1(X) = (X - 2)(X^2 + 4)$ while $p_2(X) = (X + 2)(X^2 + 4)$. Hence, the cosets $p_i(X) + \langle X^4 - 16\rangle$ with $i = 1, 2$ are not the zero-coset, but
  $$(X + 2 + \langle X^4 - 16\rangle) \cdot p_1(X) + \langle X^4 - 16\rangle) = (X + 2)p_1(X) + \langle X^4 - 16\rangle$$
  $$= X^4 - 16 + \langle X^4 - 16\rangle = 0_R,$$
  and
  $$(X - 2 + \langle X^4 - 16\rangle) \cdot (p_2(X) + \langle X^4 - 16\rangle) = (X - 2)p_2(X) + \langle X^4 - 16\rangle$$
  $$= X^4 - 16 + \langle X^4 - 16\rangle = 0_R.$$
  This shows that $X + 2 + \langle X^4 - 16\rangle$ and $X - 2 + \langle X^4 - 16\rangle$ are zero-divisors in R.
  This exercise is meant to give an intuition of a more general property of quotient rings of polynomial rings. If $(R, +, \cdot)$ is a commutative ring and $g(X) \in R[X]$ is a polynomial dividing another polynomial $f(X) \in R[X]$ (as here $X \pm 2$ divides $X^4 - 16$) then $g(X)+\langle f(X)\rangle$ is a zero-divisor in $R[X]/\langle f(X)\rangle$. This is true because we can write $f(X) = p(X)g(X)$ and hence $(g(X) + \langle f(X)\rangle)(p(X) + \langle f(X)\rangle) = f(X) + \langle f(X)\rangle = 0_{R[X]/\langle f(X)\rangle}$. This also implies that if $f(X)$ is a polynomial that can be written as $f(X) = f_1(X)f_2(X)$ where $f_i(X)$ with $i = 1, 2$ is not constant, then $R[X]/\langle f(X)\rangle$ cannot be a field. We will call polynomials $f(X)$ of this type *reducible polynomials*.

- *Ch. 8, Ex. 9* We use the strategy written in Exercise 5. Using the division with reminder one gets
$$3X^3 + 5X^2 + 4X + 2 = (X + 4)(3X^2 + 1) + 3X + 5,$$
where $r(X) = 3X + 5$ is the reminder of the division of $g(X) = 3X^3 + 5X^2 + 4X + 2$ and the generator of the ideal $f(X) = 3X^2 + 1$. We conclude that the standard form of the coset $g(X) + \langle f(X)\rangle$ is $3X + 5 + \langle 3X^2 + 1\rangle$.

- *Ch. 8, Ex. 10* First we show that $I \cap J$ is an ideal if both I and J are ideals. We start by showing that $(I \cap J, +_R)$ is a group. Since $(I, +_R)$ and $(J, +_R)$ are groups inside $(R, +_R)$, both I and J contain the neutral element 0_R for the addition. This means that $0_R \in I \cap J$. Further, if $x, y \in I \cap J$, then $x +_R y \in I$ (since $(I, +_R)$ is a group) and $x +_R y \in J$ (since $(J, +_R)$ is a group). Therefore $x +_R y \in I \cap J$. Similarly, if $x \in I \cap J$, then $-x \in I$ and $-x \in J$, implying that $-x \in I \cap J$. We may now conclude that $I \cap J, +_R)$ is a group. Next we show that if $r \in R$ and $x \in I \cap J$, then $r \cdot_R x \in I \cap J$. However, since I is an ideal and $x \in I \cap J \subseteq I$, we have $r \cdot_R x \in I$. Similarly, $r \cdot_R x \in J$. Therefore $r \cdot_R x \in I \cap J$ if $r \in R$ and $x \in I \cap J$. This concludes the proof that $I \cap J$ is an ideal of R.

  It is not true in general that $I \cup J$ is an ideal of R given that I and J are ideals of R. Consider for example $R = \mathbb{Z}$, $I = 2\mathbb{Z} = \langle 2 \rangle$ (all multiples of 2) and $J = 3\mathbb{Z} = \langle 3 \rangle$ (all multiples of 3). Then $2\mathbb{Z} \cup 3\mathbb{Z}$ is not an ideal of $\mathbb{Z}$. We have for example that $2 \in 2\mathbb{Z} \cup 3\mathbb{Z}$ and $3 \in 2\mathbb{Z} \cup 3\mathbb{Z}$, but their sum 5 is not in $2\mathbb{Z} \cup 3\mathbb{Z}$. Apparently $(2\mathbb{Z} \cup 3\mathbb{Z}, +)$ is not a group and therefore not an ideal of the ring $(\mathbb{Z}, +, \cdot)$ either.

- *Ch. 8, Ex. 11* The fact that I is contained in $[I : R]$ follows from the property $r \cdot_R x \in I$ for all $r \in R$ and $x \in I$ in the definition of I being an ideal of R. We check that $([I : R], +_R)$ is a group. The zero-element 0_R is clearly in $[I : R]$ since for all $r \in R$, $r \cdot_R 0_R = 0_R$ and 0_R is contained in every ideal of R (so in particular I). If $x, y \in [I : R]$ then also $x +_R y \in [I : R]$. Indeed for all $r \in R$, $r \cdot_R (x +_R y) = r \cdot_R x +_R r \cdot_R y \in I$ since $r \cdot_R x, r \cdot_R y \in I$ and I is an ideal. Similarly if $x \in [I : R]$ then also $-x \in [I : R]$. Indeed since for all $r \in R$, $(-r) \cdot_R x \in I$ also $r \cdot_R (-x) \in I$ from $(-r) \cdot_R x = r \cdot_R (-x)$. To complete the proof one needs to show that if $x \in [I : R]$ then $r \cdot_R x \in [I : R]$ for all $r \in R$. However this is true because it is equivalent to require (by definition) that for all $s \in R$, $s \cdot_R r \cdot_R x = (s \cdot_R r) \cdot_R x \in I$ and this property is true because $r \cdot_R s \in R$ and $x \in [I : R]$.

- *Ch. 8, Ex. 12* First, we check that $(I + J, +_R)$ is a group. The zero-element 0_R is clearly in $I + J$ since it is contained in both I and J. If $x, y \in I + J$ then also $x +_R y \in I + J$. Indeed we can write $x = u_1 + v_1$ and $y = u_2 + v_2$ where $u_i \in I$ and $v_i \in J$ for $i = 1, 2$. Hence $x +_R y = (u_1 +_R u_2) +_R (v_1 +_R v_2) \in I + J$ since $u_1 +_R u_2 \in I$ and $v_1 +_R v_2 \in J$ from I and J being ideals. Similarly if $x \in I + J$ then also $-x \in I + J$. Indeed writing again $x = u_1 + v_1$ with $u_1 \in I$ and $v_1 \in J$ we see that $-x = -(u_1 +_R v_1) = (-u_1) +_R (-v_1) \in I + J$. To complete the proof one needs to show that if $x = u_1 + v_1 \in I + J$ then $r \cdot_R x \in I + J$ for all $r \in R$. However this is true because from the distributive property $r \cdot_R x = r \cdot_R (u_1 +_R v_1) = (r \cdot_R u_1) +_R (r \cdot_R v_1) \in I + J$ as $r \cdot_R u_1 \in I$ and $r \cdot_R v_1 \in J$ from I and J being ideals.

- *Ch. 8, Ex. 13*

  - Since $\gcd(12, 20) = 4$ we have $\langle 12, 20 \rangle = \langle 4 \rangle$ from Lemma 166.
  - Using the Euclidean Algorithm we see that $\gcd(X^2 - 1, X^3 - 1) = X - 1$ so from Lemma 170, $\langle X^2 - 1, X^3 - 1 \rangle = \langle X - 1 \rangle$.
  - Again, from Lemma 170 $\langle X^2 + 1, X^5 + X + 1 \rangle = \mathbb{C}$ as $\gcd(X^2 + 1, X^5 + X + 1) = 1$ and an ideal containing 1 is the entire ring.

- *Ch. 8, Ex. 14* We first show that $\langle f(X), g(X) \rangle \subseteq \langle \gcd(f(X), g(X)) \rangle$. An element $p(X)$ of $\langle f(X), g(X) \rangle$ is of the form $p(X) = r(X) \cdot f(X) + s(X) \cdot g(X)$ for some polynomials $r(X), s(X) \in \mathbb{F}[X]$. Since $\gcd(f(X), g(X))$ divides both $f(X)$ and $g(X)$, we see that such $p(X)$ is a multiple of $\gcd(f(X), g(X))$. This implies that $p(X) \in \langle \gcd(f(X), g(X)) \rangle$. Now we show that $\langle \gcd(f(X), g(X)) \rangle \subseteq \langle f(x), g(X) \rangle$. If $p(X) \in \langle \gcd(f(X), g(X)) \rangle$, then there exists a polynomial $r(X) \in \mathbb{F}[X]$ such that $p(X) = r(X) \cdot \gcd(f(X), g(X))$. Also, we know

that using the extended Euclidean algorithm on $f(X)$ and $g(X)$, one can find polynomials $a(X), b(X) \in \mathbb{F}[X]$ such that $a(X) \cdot f(X) + b(X) \cdot g(X) = \gcd(f(X), g(X))$. Inserting this in the expression for $p(X)$, we see that:

$$p(X) = r(X) \cdot \gcd(f(X), g(X)) = r(X)(a(X) \cdot f(X) + b(X) \cdot g(X))$$

$$= r(X) \cdot a(X) \cdot f(X) + r(X) \cdot b(X) \cdot g(X).$$

This shows that $p(X) \in \langle f(x), g(X) \rangle$.

- *Ch. 8, Ex. 15* Let $I \subseteq \mathbb{F}$ be an ideal. If $I \neq \{0_\mathbb{F}\}$, then there exists $c \in I$ different from zero. Since $(\mathbb{F}, +_\mathbb{F}, \cdot_\mathbb{F})$ is a field, the element c has a multiplicative inverse c^{-1}. However, since $c \in I$ and I is an ideal, we may conclude that $1_\mathbb{F} = c^{-1} \cdot_\mathbb{F} c \in I$. In general if an ideal I contains the identity element $1_\mathbb{F}$, it is necessarily equal to the whole ring. In our specific case, we conclude that $I = \mathbb{F}$. All in all, there are therefore only two possible ideals in $\mathbb{F}$, namely $I = \{0_\mathbb{F}\}$ and $I = \mathbb{F}$.

- *Ch. 8, Ex. 16* It is enough to remember that from Theorem 8.1.6 and Definition 8.1.7, given a ring homomorphism $\phi : R \to S$, $\ker \phi$ is an ideal of the ring R. Since in our case $R = \mathbb{F}$ is a field we get from the previous exercises just to possibilities: either $\ker \phi = \{0_\mathbb{F}\}$ or $\ker \phi = \mathbb{F}$. In the latter case, we see that every element in $\mathbb{F}$ is mapped to the zero-element of S and hence $\mathrm{im}\phi = \{0_S\}$.

- *Ch. 8, Ex. 17*

  - Recall that $\langle r_1, \ldots, r_n \rangle = \{a_1 \cdot_R r_1 +_R \ldots +_R a_n \cdot_R r_n \mid a_1, \ldots, a_n \in R\}$. Since I is an ideal, if $r_i \in I$ then all multiples $a_i \cdot_R r_i \in I$ for all $a_i \in R$ and for all $i = 1, \ldots, n$. Since $(I, +_R)$ is a group, the sum $a_1 \cdot_R r_1 +_R \ldots +_R a_n \cdot_R r_n \in I$, completing the proof.
  - Since by definition $u \cdot_R a \in \langle a \rangle$ we see that $\langle u \cdot_R a \rangle \subseteq \langle a \rangle$ from the previous part of the exercise (with $n = 1$). We want to show that $a \in \langle u \cdot_R a \rangle$ so that again $\langle a \rangle \subseteq \langle u \cdot_R a \rangle$ from the previous part of the exercise. However since $u \in R^*$, u admits a multiplicative inverse $u^{-1} \in R$. Since $u \cdot_R a \in \langle u \cdot_R a \rangle$ then also $a = u^{-1} \cdot_R u \cdot_R a \in \langle u \cdot_R a \rangle$.
  - Since $a \in \langle a, a \cdot_R b \rangle$ we have from the previous part of the exercise that $\langle a \rangle \subseteq \langle a, a \cdot_R b \rangle$. However from the definition of principal ideal, also $a \cdot_R b \in \langle a \rangle$ and hence the previous part of the exercise yields again that $\langle a, a \cdot_R b \rangle \subseteq \langle a \rangle$.
  - One implication coincides with the second part of this exercise. Hence we have only to show that if R is an integral domain then $\langle a \rangle = \langle b \rangle$ implies $a = u \cdot_R b$ for some unit $u \in R^*$. Since $\langle a \rangle = \langle b \rangle$ implies both that $a \in \langle b \rangle$ and $b \in \langle a \rangle$, there exist $u, v \in R$ such that $a = u \cdot_R b$ and $b = v \cdot_R a$. Combining these two conditions we get $a = u \cdot_R v \cdot_R a$ and hence $(u \cdot_R v - 1_R) \cdot_R a = 0_R$. Since R is a domain either $a = 0$ or $u \cdot_R v = 1_R$. If $a = 0$ then also $b = v \cdot_R a = 0_R$. Clearly $0_R = u \cdot_R 0_R$ for all $u \in R^*$ and hence in this case the claim is proven. If $u \cdot_R v = 1_R$ we see that v is the multiplicative inverse of u and hence $u \in R^*$ with $a = u \cdot_R b$.

- *Ch. 8, Ex. 18*

  - Let $x, y \in \mathrm{im}\phi$. Then there exists $a, b \in R$ such that $x = \phi(a)$ and $y = \phi(b)$. From the properties of the ring homomorphism ϕ we have that

$$x \cdot_S y = \phi(a) \cdot_S \phi(b) = \phi(a \cdot_R b).$$

Since R is commutative we know that $a \cdot_R b = b \cdot_R a$ and hence

$$\phi(a \cdot_R b) = \phi(b \cdot_R a) = \phi(b) \cdot_S \phi(a) = y \cdot_S x.$$

Since this shows that for all $a, y \in \mathrm{im}\phi$ one has $x \cdot_S y = y \cdot_S x$ we conclude that $(\mathrm{im}\phi, +_S, \cdot_S)$ is commutative.

– Since $(R, +_R, \cdot_R)$ is a field we know that $R^* = R \setminus \{0_R\}$. We need to show that $(\mathrm{im}\phi)^* = \mathrm{im}\phi \setminus \{0_S\}$, that is, every $x \in \mathrm{im}\phi$ with $x \neq 0_S$ admits a multiplicative inverse. Let $x \in \mathrm{im}\phi$ with $x \neq 0_S$. Then there exists $a \in R$ such that $x = \phi(a)$. We note that $a \neq 0_R$ as otherwise $\phi(a) = \phi(0_R) = 0_S$ by the axioms of a ring homomorphism. Since $(R, +_R, \cdot_R)$ and $a \neq 0_R$, a admits a multiplicative inverse a^{-1} in R, that is $a \cdot_R a^{-1} = a^{-1} \cdot_R a = 1_R$. We want to show that $\phi(a^{-1}) \in \mathrm{im}\phi$ is the multiplicative inverse of $x = \phi(a)$. Using the properties of the ring homomorphism ϕ we have

$$\phi(a) \cdot_S \phi(a^{-1}) = \phi(a \cdot_R a^{-1}) = \phi(1_R) = 1_S.$$

A similar computation shows that $\phi(a^{-1}) \cdot_S \phi(a) = 1_S$ and hence $\phi(a^{-1})$ is the multiplicative inverse of $\phi(a)$.

- *Ch. 8, Ex. 19*

  – Following the strategy of Exercise 5, we compute the division with remainder $r(X)$ of $X^3 + X^2$ and $X^2 + x + 1$, predicting that the standard form will be given by the coset $r(X) + \langle X^3 + X + 1 \rangle$. One has

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

  where the last equality follows from the fact that we are working over $\mathbb{F}_2$. Since $\deg(X^2 - X - 1) < 3 = \deg(X^3 + X + 1)$ we deduce that $r(X) = X^2 + X + 1$ and hence the standard form is $X^2 + X + 1 + \langle X^3 + X + 1 \rangle$.

  – Using the Extended Euclidean Algorithm one gets $\gcd(X^3 + X + 1, X^2 + X + 1) = 1$, $r(X) = X + 1$ and $s(X) = X^2$.

- *Ch. 8, Ex. 20*

  – The additive order of the element 1 have to divide 4, the order of the group $(R, +)$. If the order of 1 would be 4, then $(R, +, \cdot)$ would contain zero-divisors, since then $1 + 1 \neq 0$, while $(1 + 1) \cdot (1 + 1) = 1 + 1 + 1 + 1 = 0$. Hence if $(R, +, \cdot)$ is a field, then $1 + 1 = 0$. This also implies $a + a = 0$ and $b + b = 0$. Concretely, one needs to work modulo 2 when adding elements.

  – If $a + 1 = 0$, then $a = -1 = 1$ giving a contradiction. Here we used that 1 is its own additive inverse, since $1 + 1 = 0$ from the previous part of the exercise. If $a + 1 = 1$, then $a = 0$, a contradiction. Similarly if $a + 1 = a$, then $1 = 0$, a contradiction. The only possibility that is left is $a + 1 = b$.

  – If $(R, +, \cdot)$ is a field, then $R^* = R \setminus \{0\} = \{1, a, a + 1\}$. Here we used that from the previous part of the exercise we know that $b = a + 1$. Since $(R^*, \cdot)$ is a group of order three, Proposition 4.4.4 implies that $a^3 = 1$. In fact the order of a is 3, since its order is not 1. Hence the group $(R^*, \cdot)$ is cyclic and a can be chosen as a generator of this group. This implies that $a^2 = a + 1$.

  – Using $1 + 1 = 0$ and $a^2 = b = a + 1$, leads to the following addition and multiplication tables:

$+$	0	1	a	b
0	0	1	a	b
1	1	0	b	a
a	a	b	0	1
b	b	a	1	0

and

$\cdot$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a	0	a	b	1
b	0	b	1	a

One could take these tables and verify that the ring axioms are satisfied by checking all possibilities. The resulting ring is then a field, since the multiplication table shows that all nonzero elements have a multiplicative inverse.

Another way is to realize that the relations $1+1=0$ and $a^2=a+1$ really mean that one is considering polynomials with coefficients in $\mathbb{F}_2$ modulo the ideal generated by X^2+X+1. In other words, the addition and multiplication tables above are those of the quotient ring $(\mathbb{F}_2[X]/\langle X^2+X+1\rangle, +, \cdot)$. The element a is then identified with the element $X+\langle X^2+X+1\rangle$.

- *Ch. 8, Ex. 21*
  - It is not an ideal. Indeed if we multiply a polynomial whose constant term is a multiple of 5 with $2 \in \mathbb{Z}[X]$ the resulting polynomial will have constant term 10.
  - It is not an ideal. Indeed denoting with I the set of polynomials whose coefficient of X^2 is a multiple of 5 we see that $5X^2+1 \in I$, $X^2+1 \in \mathbb{Z}[X]$ but $(5X^2+1)\cdot(X^2+1) = 5X^4+6X^2+1 \notin I$.
  - In other words, the putative I consists of the zero-polynomial and all the nonzero polynomials in $\mathbb{Z}[X]$ that are divisible by X^3. Clearly the sum of polynomials who are divisible by X^3 is still divisible by X^3 and also the additive inverse of a polynomial divisible by X^3 is divisible by X^3. This shows that $(I,+)$ is a group. If we multiply a polynomial which is divisible by X^3 by an arbitrary polynomial in $\mathbb{Z}[X]$, then the resulting polynomial will be still divisible by X^3. This shows that I is an ideal.
  - This is not an ideal. Indeed if we multiply a polynomial in which just even powers of X appears with $X \in \mathbb{Z}[X]$ then just odd powers of X will appear, yielding an element which is not contained in $\mathbb{Z}[X^2]$.
  - Let I be the putative ideal, and let $p(X) = a_0 + a_1X + \ldots + a_nX^n \in I$ where $a_i \in \mathbb{Z}$ for all $i = 1, \ldots, n$. Then $p'(X) = a_1 + 2a_2X + \ldots + na_nX^{n-1}$. Hence the condition $p'(0) = 0$ is equivalent to require that the coefficient of X, that is a_1, is equal to zero. There are several ways to see that this is not an ideal. One way is to note that $1+X^2 \in I$ but $X \cdot (X^2+1) = X^3+X \notin I$ as $(X^3+X)' = 3X^2+1$ that has not 0 as a root. Another way can be the following. Note that identity element $1 \in \mathbb{Z}[X]$ is in 1 as its derivative is the zero-polynomial. This means that if I is an ideal then $I = \mathbb{Z}[X]$. However $X \notin I$ since its derivative is the constant polynomial 1 that do not vanish in 0.

- *Ch. 8, Ex. 22*
  - The element of the ideal $\langle 3\rangle$ are all Gaussian integers of the form $3a+3bi$, where both a and b are integers. This means that a given coset of $c+di+\langle 3\rangle$ has a representative of the form $\gamma+\delta i$ where $\gamma, \delta \in \{-1, 0, 1\}$. This shows that there are at most nine distinct cosets. On the other hand, if $\gamma_1+\delta_1 i+\langle 3\rangle = \gamma_2+\delta_2 i+\langle 3\rangle$ for some $\gamma_1, \gamma_2, \delta_1, \delta_2 \in \{-1, 0, 1\}$, then $(\gamma_1-\gamma_2)+(\delta_1-\delta_2)i \in \langle 3\rangle$. This would imply that 3 divides both $\gamma_1-\gamma_2$ and $\delta_1-\delta_2$. This is only possible if $\gamma_1=\gamma_2$ and $\delta_1=\delta_2$. Hence there are precisely nine distinct cosets.

– This follows by direct computation or a bit of trial and error. For example, one has $(i + \langle 3\rangle) \cdot (-i + \langle 3\rangle) = 1 + \langle 3\rangle$ and $(1 + i + \langle 3\rangle) \cdot (-1 + i + \langle 3\rangle) = -2 + \langle 3\rangle = 1 + \langle 3\rangle$. The last equality holds, since $3 \in \langle 3\rangle$.

- *Ch. 8, Ex. 23*

  – Since $5 = (1 + 2i)(1 - 2i)$, we have

  $$(1 + 2i + \langle 5\rangle)(1 - 2i + \langle 5\rangle) = 5 + \langle 5\rangle = 0 + \langle 5\rangle.$$

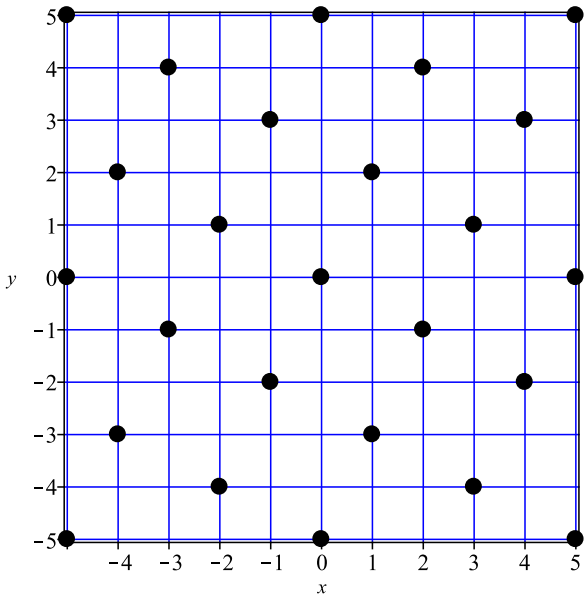
  On the other hand, $1 \pm 2i + \langle 5\rangle \neq 0 + \langle 5\rangle$, since $1 \pm 2i$ is not in $\langle 5\rangle$ (if it were, both real and imaginary part of $1 \pm 2i$ would be multiples of 5). Hence $\mathbb{Z}[i]/\langle 5\rangle$ contains zero-divisors.

  – The elements of the ideal $\langle 1 + 2i\rangle$ are of the form

  $$(a + bi)(1 + 2i) = a(1 + 2i) + b(-2 + i) \quad \text{with } a, b \in \mathbb{Z}.$$

  The elements of the ideal $\langle 1 + 2i\rangle$ form a so-called lattice in the complex plane as illustrated as black dots in Figure 10.1.

Figure 10.1: A plot of some elements in the ideal $\langle 1 + 2i\rangle$.



The elements of a coset of $\langle 1 + 2i\rangle$ is then a translation of this lattice. With this figure in mind, it is not so hard to show that the five cosets $\langle 1 + 2i\rangle$, $\pm 1 + \langle 1 + 2i\rangle$, and $\pm i + \langle 1 + 2i\rangle$ together partition the entire set $\mathbb{Z}[i]$. Hence $\mathbb{Z}[i]/\langle 1 + 2i\rangle$ consists of exactly these five cosets of $\langle 1 + 2i\rangle$.

  – We have

  1. $\psi(0) = 0$ and

  $$\psi((a + bi) + (c + di)) = a + c + 2(b + d) \bmod 5 = \\ (a + 2b \bmod 5) +_5 (c + 2d \bmod 5) = \psi(a + bi) + \psi(c + di),$$

2. $\psi(1) = 1$,
3. for the product of images it holds that

$$\begin{aligned}\psi((a+bi)\cdot(c+di)) &= ac - bd + 2(ad+bc) \bmod 5 \\ &= ac + 4bd + 2(ad+bc) \bmod 5 = (a+2b)(c+2d) \bmod 5 \\ &= \psi(a+bi) \cdot_5 \psi(c+di),\end{aligned}$$

Hence ψ is a ring homomorphism.

– The ring homomorphism ψ is surjective, since $\psi(a) = a$ for any $a \in \mathbb{Z}_5$.

Since $\psi(1+2i) = 5 \bmod 5 = 0$, we see that $1+2i \in \ker(\psi)$. This implies that $\langle 1+2i\rangle \subseteq \ker(\psi)$. On the other hand, if $a+bi \in \ker(\psi)$, then $a+2b \in \ker(\psi)$, since $a+bi = a+2b+bi(1+2i)$. This implies that $a+2b \bmod 5 = 0$ and hence that 5 divides $a+2b$, say $a+2b = 5k$ for some $k \in \mathbb{Z}$. Using that $5 = (1+2i)(1-2i)$, we then obtain that

$$a+bi = 5k + bi(1+2i) = (1+2i)(k(1-2i)+bi)$$

and hence that $a+bi \in \langle 1+2i\rangle$. This shows that $\langle 1+2i\rangle = \ker(\psi)$.

The isomorphism theorem for rings now implies immediately that the rings $(\mathbb{Z}[i]/\langle 1+2i\rangle, +, \cdot)$ and $(\mathbb{Z}_5, +_5, \cdot_5)$ are isomorphic. Hence $(\mathbb{Z}[i]/\langle 1+2i\rangle, +, \cdot)$ is a finite field with 5 elements.

- *Ch. 8, Ex. 24*

– We first prove that $(\phi^{-1}(J), +_R)$ is a group. Note that $0_R \in \phi^{-1}(J)$ as $0_S \in J$ (because J is an ideal of S) and $\phi(0_R) = 0_S$ from the ring homomorphism axioms. Also if $a, b \in \phi^{-1}(J)$ then both $a +_R b$ and $-a$ are in $\phi^{-1}(J)$ as $\phi(a +_R b) = \phi(a) +_R \phi(b) \in I$ and $\phi(-a) = -\phi(a) \in I$ from $(I, +_S)$ being a group. Finally let $r \in R$ and $a \in \phi(J)$ then $r \cdot_R a \in \phi^{-1}(J)$ because $\phi(r \cdot_R a) = \phi(r) \cdot_S \phi(a) \in I$. Indeed I is an ideal of S, $\phi(r) \in S$ (by definition of ϕ) and $\phi(a) \in I$ as $a \in J$.

– We first show that $(\phi(I), +_S)$ is a group. From the axioms of a ring homomorphism we have that $0_S = \phi(0_R) \in \phi(I)$. Let $x, y \in \phi(I)$. Then there exists $a, b \in I$ such that $x = \phi(a)$ and $y = \phi(b)$. We show that both $x +_S y$ and $-x$ are contained in $\phi(I)$. This is true because from ϕ being a ring homomorphism $x +_R y = \phi(a) +_s \phi(b) = \phi(a +_R b) \in \phi(I)$ and $-x = -\phi(a) = \phi(-a) \in \phi(I)$, as $(I, +_R)$ as a group. The surjectivity of the map ϕ is used to show that for all $s \in S$ and $x = \phi(a) \in \phi(I)$ with $a \in I$, the product $a \cdot_S x \in \phi(I)$. From the surjectivity of ϕ we know that there exists $r \in R$ such that $s = \phi(r)$. Hence $s \cdot_S x = \phi(r) \cdot_S \phi(a) = \phi(r \cdot_R a) \in \phi(I)$ because $r \cdot_R a \in I$ as I is an ideal. This proves the first claim of the exercise.

An example where the property fails if ϕ is not surjective is the following. Let $\phi : \mathbb{Z} \to \mathbb{Q}$ be the natural injection $\phi(n) = n$. This is a non-surjective ring homomorphism. Since $(\mathbb{Q}, +, \cdot)$ is a field we know from Exercise 15, that it admits only two ideals, namely $\{0\}$ and $\mathbb{Q}$ itself. On the other hand in $\mathbb{Z}$ we can consider the principal ideal $I := \langle 2\rangle$ consisting of exactly all even integers. Since the image of this ideal with respect to ϕ is neither $\{0\}$ nor the entire $\mathbb{Q}$ we conclude that $\phi(I)$ cannot be an ideal of $\mathbb{Q}$.

- *Ch. 8, Ex. 25*

– The equality of equivalence classes $[(a,b)] = [(a',b')]$ implies that $a \cdot b' = b \cdot a'$ and $[(c,d)] = [(c',d')]$ implies that $c \cdot d' = d \cdot c'$. In order to show that $[(a,b)] \cdot [(c,d)] = [(a',b')] \cdot [(c',d')]$ we need to show that $[(a \cdot c, b \cdot d)] = [(a' \cdot c', b' \cdot d')]$ equivalently, $a \cdot c \cdot b' \cdot d' = b \cdot d \cdot a' \cdot c'$. We start from the left hand side. We see that

$$\begin{aligned} a \cdot c \cdot b' \cdot d' &= a \cdot b' \cdot c \cdot d' \quad \text{by commutativity} \\ &= b \cdot a' \cdot c \cdot d' \\ &= b \cdot a' \cdot d \cdot c' \\ &= b \cdot d \cdot a' \cdot c' \quad \text{by commutativity.} \end{aligned}$$

This is the required equality.

– The equality of equivalence classes $[(a,b)] = [(a',b')]$ implies that $a \cdot b' = b \cdot a'$ and $[(c,d)] = [(c',d')]$ implies that $c \cdot d' = d \cdot c'$.
To show that $[(a,b)] + [(c,d)] = [(a',b')] + [(c',d')]$ we need to show that $[(a \cdot d + b \cdot c, b \cdot d)] = [(a' \cdot d' + b' \cdot c', b' \cdot d')]$. Equivalently $(a \cdot d + b \cdot c) \cdot b' \cdot d' = (a' \cdot d' + b' \cdot c') \cdot b \cdot d$ if and only if $a \cdot d \cdot b' \cdot d' + b \cdot c \cdot b' \cdot d' = a' \cdot d' \cdot b \cdot d + b' \cdot c' \cdot b \cdot d$.
We start from the left hand side:

$$\begin{aligned} a \cdot d \cdot b' \cdot d' + b \cdot c \cdot b' \cdot d' &= a \cdot b' \cdot d \cdot d' + b \cdot b' \cdot c \cdot d' \\ &= a' \cdot b \cdot d \cdot d' + b \cdot b' \cdot d \cdot c' \\ &= a' \cdot d' \cdot b \cdot d + b' \cdot c' \cdot b \cdot d \end{aligned}$$

as required.

– Let $X = [(a,b)]$, $Y = [(c,d)]$ and $Z = [(e,f)]$ in F. Then

$$\begin{aligned} (X+Y) \cdot Z &= ([(a,b)] + [(c,d)]) \cdot [(e,f)] \\ &= [(a \cdot d + b \cdot c, b \cdot d)] \cdot [(e,f)] \\ &= [((a \cdot d + b \cdot c) \cdot e, b \cdot d \cdot f)] \quad \text{by distributivity in D} \\ &= [(a \cdot d \cdot e + b \cdot c \cdot e, b \cdot d \cdot f)] \cdot X \cdot Z + Y \cdot Z \\ &= [(a,b)] \cdot [(e,f)] + [(c,d)] \cdot [(e,f)] \\ &= [(a \cdot e, b \cdot f)] + [(c \cdot e, d \cdot f)] \\ &= [(a \cdot e \cdot d \cdot f + b \cdot f \cdot c \cdot e, b \cdot f \cdot d \cdot f)], \end{aligned}$$

but on the other hand

$$[(a \cdot d \cdot e + b \cdot c \cdot e, b \cdot d \cdot f)] = [(a \cdot e \cdot d \cdot f + b \cdot f \cdot c \cdot e, b \cdot f \cdot d \cdot f)].$$

Equivalently,

$$\begin{aligned} (a \cdot d \cdot e + b \cdot c \cdot e) \cdot b \cdot f \cdot d \cdot f = & \\ &= (b \cdot d \cdot f) \cdot (a \cdot e \cdot d \cdot f + b \cdot f \cdot c \cdot e) \\ &= a \cdot d \cdot e \cdot b \cdot f \cdot d \cdot f + b \cdot c \cdot e \cdot b \cdot f \cdot d \cdot f \\ &= b \cdot d \cdot f \cdot a \cdot e \cdot d \cdot f + b \cdot d \cdot f \cdot b \cdot f \cdot c \cdot e. \end{aligned}$$

Hence $(X+Y) \cdot Z = X \cdot Z + Y \cdot Z$.

– The function $\phi : D \to F$ with $\phi(a) = [(a, 1)]$ is such that

$$\begin{aligned}\phi(a \cdot b) &= [(a \cdot b, 1)] \\ &= [(a \cdot b \cdot x^2, x^2)] \\ &= [(a \cdot x \cdot (b \cdot x), x^2)] \\ &= [(a \cdot x, x)] \cdot [(b \cdot x, x)] \\ &= [(a, 1)] \cdot [(b, 1)] = \phi(a) \cdot \phi(b); \end{aligned} \tag{10.1}$$

while $\phi(a + b) = [(a + b, 1)] = [(a, 1)] + [(b, 1)] = \phi(a) + \phi(b)$.
By definition, $\ker \phi = \{a \in D | \phi(a) = [(0, 1)]\}$ as one can easily check that $[(0, 1)]$ is the zero-element in F. So, $a \in \ker \phi$ if and only if $[(a, 1)] = [(a \cdot x, x)] = [(0, 1)] = [(0y, y)]$ if and only if $a \cdot x \cdot y = 0 \cdot y \cdot x = 0$. Since $x \neq 0$ and $y \neq 0$ and D is an integral domain we have $a = 0$. This shows $\ker \phi = \{0\}$.

10.9 Chapter 9

- *Ch. 9, Ex. 1* A. FALSE, B. TRUE, C. TRUE.
- *Ch. 9, Ex. 2* A.
- *Ch. 9, Ex. 3* A. FALSE, B. TRUE, C. TRUE.
- *Ch. 9, Ex. 4* B.
- *Ch. 9, Ex. 5* A. TRUE, B. FALSE, C. TRUE
- *Ch. 9, Ex. 6* B.
- *Ch. 9, Ex. 7*
  - The kernel consists of all polynomials that have the element a as a root. Using the second part of Proposition 7.4.3, one may conclude that $\ker(\psi) = \langle X - a \rangle$. Since for any $b \in \mathbb{F}$, we have $\psi(b) = b$, the image of ψ equals $\mathbb{F}$, that is: $\mathrm{im}(\psi) = \mathbb{F}$.
  - This follows from the previous part and the isomorphism theorem for rings (Theorem 8.3.5).
- *Ch. 9, Ex. 8*
  - Using the division algorithm on $X^3 + 2X^2 + 4X + 3$ and $X^2 + 3X + 2$, one obtains that $X^3 + 2X^2 + 4X + 3 = (X + 4) \cdot (X^2 + 3X + 2) + 0$. Hence $X^3 + 2X^2 + 4X + 3 + \langle X^2 + 3X + 2 \rangle = 0 + \langle X^2 + 3X + 2 \rangle$, which is the zero element.
  - Since $\gcd(X^2 + X, X^2 + 3X + 2) = X + 1$, Proposition 9.1.2 implies that r is a zero-divisor. We have $X^2 + 3x + 2 = (X + 1)(X + 2)$ in $\mathbb{F}_5[X]$. Therefore

$$\begin{aligned}(X + 2 + \langle X^2 + 3X + 2 \rangle) \cdot (X^2 + X + \langle X^2 + 3X + 2 \rangle) = \\ X \cdot (X + 1) \cdot (X + 2) + \langle X^2 + 3X + 2 \rangle = \\ 0 + \langle X^2 + 3X + 2 \rangle.\end{aligned}$$

  Hence we can choose $s = X + 2 + \langle X^2 + 3X + 2 \rangle$.

– Since $\gcd(X, X^2+3X+2) = 1$, the element u is a unit. Using the extended Euclidean algorithm, one finds that

$$1 = 3 \cdot (X^2 + 3X + 2) + (2X + 1) \cdot X.$$

Hence $u^{-1} = 2X + 1 + \langle X^2 + 3X + 2\rangle$.

- *Ch. 9, Ex. 9* The quotient ring $\mathbb{F}_3[X]/\langle X^3 + X + 1\rangle$ is a field if and only if the polynomial $X^3 + X + 1 \in \mathbb{F}_3[X]$ is irreducible. However, the polynomial is reducible, since it has 1 as a root. Therefore we can write $X^3 + X + 1$ as a multiple of $X - 1$. Not that $X - 1 = X + 2$, since we work in the finite field with three elements.

  More concretely, we have $X^3 + X + 1 = (X - 1)(X^2 + X - 1)$. As said before this implies that the ring $(\mathbb{F}_3[X]/\langle X^3 + X + 1\rangle, +, \cdot)$ is not a field. In fact, we can use the above factorization of $X^3 + X + 1$ to find zero-divisors in the quotient ring. More precisely, let us denote $I := \langle X^3 + X + 1\rangle$. Since $(X - 1) \cdot (X^2 + X - 1) = X^3 + X + 1 \in I$, we see that

$$(X - 1 + I) \cdot (X^2 + X - 1 + I) = (X - 1) \cdot (X^2 + X - 1) + I = 0 + I.$$

  Therefore, the element $X - 1 + I \in \mathbb{F}_3[X]/I$ is a zero-divisor.

- *Ch. 9, Ex. 10*

  – Since 1 is a root, the polynomial is reducible having $X - 1$ as a factor. Using the division algorithm, one finds that $X^2 + 1 = (X + 1)^2$ in $\mathbb{F}_2[X]$.

  – A quadratic polynomial is either irreducible, or has a root. Since $X^2 + 1$ has no roots in $\mathbb{F}_3$ ($0^2 + 1 = 1, 1^2 + 1 = 2, 2^2 + 1 = 2$ when working modulo three), it is irreducible. Hence the answer is the polynomial $X^2 + 1$ itself.

  – The polynomial has 2 and 3 as roots. Hence $X^2+1 = (X-2)\cdot(X-3) = (X+3)\cdot(X+2)$.

- *Ch. 9, Ex. 11*

  – To find an irreducible, degree two or three polynomial one needs to find a polynomial of degree two or three without roots. The only irreducible degree 2 polynomial with coefficients in $\mathbb{F}_2$ turns out to be $X^2 + X + 1$. The possible irreducible polynomials of degree 3 with coefficients in $\mathbb{F}_2$ are $X^3 + X + 1$ and $X^3 + X^2 + 1$.

  To find an irreducible polynomial of degree 4 it is necessary, but not sufficient that the polynomial has no roots in $\mathbb{F}_2$. The polynomials of degree four with coefficients in $\mathbb{F}_2$ that have no roots are X^4+X+1, X^4+X^2+1, X^4+X^3+1, and $X^4+X^3+X^2+X+1$. If one of these polynomials is reducible, it has to be divisible by $X^2 + X + 1$. It turns out that $X^4 + X^2 + 1 = (X^2 + X + 1)^2$ and hence is reducible. Using the division algorithm, one can show that the two polynomials $X^4 + X + 1$ and $X^4 + X^3 + 1$ are not divisible by $X^2 + X + 1$ and hence irreducible.

  – Since the degree of $f(X)$ is required to be two, we only need to find a quadratic polynomial that has no roots in $\mathbb{F}_3$. A possible answer is $f(X) = X^2 + X + 2$. Other possible answers are $X^2 + 1$ and $X^2 + 2X + 2$. Let us choose $X^2 + X + 2$. Using the division algorithm, one obtains that $X^9 - X = X^9 + 2X = (X^7 + 2X^6 + 2X^5 + 2X^3 + X^2 + X) \cdot (X^2 + X + 2)$.

- *Ch. 9, Ex. 12*

  – Using Lemma 9.3.1, one sees that R contains $2^3 = 8$ distinct elements.

- The element $X + \langle X^3 \rangle$ is a zero-divisor, since it is not the zero element and

$$(X + \langle X^3 \rangle)(X^2 + \langle X^3 \rangle) = X^3 + \langle X^3 \rangle = 0 + \langle X^3 \rangle.$$

- $(X + 1 + \langle X^3 \rangle)^{-1} = X^2 + X + 1 + \langle X^3 \rangle$.
- We need to determine all non-zero polynomials $r(X)$ of degree up to 2 such that $\gcd(r(X), X^3)$ has degree zero. This simply means that X should not divide $r(X)$. Hence, we obtain that

$$R^* = \{1 + \langle X^3 \rangle, 1 + X + \langle X^3 \rangle, 1 + X^2 + \langle X^3 \rangle, 1 + X + X^2 + \langle X^3 \rangle\}.$$

  Hence R^* contains 4 elements.

- *Ch. 9, Ex. 13*

  - Let w be a primitive element in $\mathbb{F}_q$. Then every element $a \in \mathbb{F}_q^*$ can be written as $a = w^n$ for some $n = 1, \dots q - 1$. We see that $a = w^n$ with $n = 1, \dots, q - 1$ is a quadratic residue in $\mathbb{F}_q$ if and only if there exists $m = 1, \dots, q-1$ such that $a = w^n = (w^m)^2 = w^{2m}$. This implies that n must be of type $2m$ for some $m = 1, \dots, (q-1)/2$. On the other hand, if $n = 2k$ with $k = 1, \dots, (q-1)/2$ then w^k is a root in $\mathbb{F}_q$ of $X^2 - a$ since $(w^k)^2 = w^{2k}) = w^n = a$, and hence a is a quadratic residue in $\mathbb{F}_q$. This shows that the quadratic residues in $\mathbb{F}_q$ are exactly those elements in $\mathbb{F}_q^*$ of type w^{2m} with $m = 1, \dots, (q-1)/2$. If we show that $|\{w^{2m} \mid m = 1, \dots, (q-1)/2\}| = (q-1)/2$ then the exercise is complete. This is equivalent to show that the elements of type w^{2n} with $n = 1, \dots, (q-1)/2$ are all distinct.

    So suppose by contradiction that $w^{2m_1} = w^{2m_2}$ with $m_1 \neq m_2$ and $m_1, m_2 = 1, \dots, (q-1)/2$. We can assume that $m_1 > m_2$ up to change the labels. Then $w^{2m_2}(w^{2(m_1 - m_2)} - 1) = 0$. Since $w^{2m_2} \neq 0$ and $\mathbb{F}_q$ is a field (in particular a domain) we get that $w^{2(m_1-m_2)} = 1$. This is not possible because w is a primitive element and $0 < m_1 - m_2 \leq (q-1)/2 - 1 < (q-1)/2$ (so in particular $2(m_1 - m_2) < q - 1$ while $q - 1$ is the order of w).

  - From the previous part of the exercise a is a quadratic residue if and only if $a = w^{2m}$ for some $m = 1, \dots, (q-1)/2$. These are exactly the elements a in $\mathbb{F}_q^*$ such that $a^{(q-1)/2} = 1$. Indeed if $a = w^n \in \mathbb{F}_q^*$ is such that $a^{(q-1)/2} = 1$ then $1 = a^{(q-1)/2} = w^{n(q-1)/2}$ and since the order of w is $q - 1$ we get that n must be even.

    From this, we also deduce that $a = w^n \in \mathbb{F}_q^*$ is not a quadratic residue in $\mathbb{F}_q$ if and only if $b := a^{(q-1)/2} \neq 1$. However $b^2 = a^{q-1} = 1$ as the order of a divides $q - 1$. Hence either $b = 1$ or $b = -1$. Since $a^{(q-1)/2} = b \neq 1$ we get $b = -1$. We deduce that a is not a quadratic residue if and only if $a^{(q-1)/2} = -1$.

- *Ch. 9, Ex. 14* To check that $(\mathbb{F}_2[X]/\langle X^4 + X + 1 \rangle, +, \cdot)$ is a field, it is sufficient to check that the polynomial $X^4 + X + 1 \in \mathbb{F}_2[X]$ is irreducible. First of all, it has no roots in $\mathbb{F}_2$, so it cannot be written as the product of a polynomial of degree 1 and one of degree 3. This is not enough to conclude that $X^4 + X + 1$ is irreducible, since it still may be possible to write it as the product of two polynomials of degree two. We may assume that each of these degree two polynomials are irreducible, since otherwise, we could find a factor of $X^4 + X + 1$ of degree one again. A direct computation shows that the only polynomial of degree two in $\mathbb{F}_2[X]$ not having any roots in $\mathbb{F}_2$ is $X^2 + X + 1$. Since $(X^2 + X + 1)^2 = X^4 + X^2 + 1 \neq X^4 + X + 1$, we may now conclude that see that $X^4 + X + 1$

is irreducible. All in all, we have shown that $(\mathbb{F}_2[X]/\langle X^4+X+1\rangle, +, \cdot)$ is a field. It is a field with 16 elements.

Now we wish to find a primitive element. We know that a primitive element exists and therefore that the group $\left(\left(\mathbb{F}_2[X]/\langle X^4+X+1\rangle\right)^*, \cdot\right)$ is a cyclic group. This group has order 15. A primitive element therefore is an element of order 15 in this group. The order of any element will divide the order of the group. Therefore any element will have order 1, 3, 5 or 15. To prove that an element has order 15, we therefore only have to make sure it does not have order 1, 3 or 5. Note that only the 1-element has order 1. We try the element $X + \langle X^4+X+1\rangle$. It does not have order 1, since it is not equal to $1 + \langle X^4+X+1\rangle$. It does not have order 3 either, since

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

Finally it does not have order five either, since

$$(X + \langle X^4+X+1\rangle)^5 = X^5 + \langle X^4+X+1\rangle = X^2 + X + \langle X^4+X+1\rangle.$$

Apparently, the element $X + \langle X^4+X+1\rangle$ has order 15 and is a primitive element of the field.

- *Ch. 9, Ex. 15* Since $\deg(f(X)) = 3$ and $f(X)$ is irreducible we see that the field $\mathbb{F} := \mathbb{F}_3[X]/\langle f(X)\rangle$ has order $3^3 = 27$. Hence $\mathbb{F}^* = \mathbb{F}\setminus\{0_\mathbb{F}\}$ contains exactly $26 = 2\cdot 13$ elements. Note that $X + \langle f(X)\rangle, 2X + \langle f(X)\rangle \in \mathbb{F}^*$, since X and $2X$ are nonzero polynomials of degree strictly less than the degree of $f(X)$. This implies, from Lagrange's Theorem, that $X + \langle f(X)\rangle$ has either order 2, or 13 or is a primitive element.

  Following the hint, let us first show that $X + \langle f(X)\rangle$ cannot have order 2. The same proof works also for $2X + \langle f(X)\rangle = -X + \langle f(X)\rangle$. In fact if this would be the case, then

  $$(X + \langle f(X)\rangle)^2 = X^2 + \langle f(X)\rangle = 1 + \langle f(X)\rangle,$$

  so that $X^2 - 1 + \langle f(X)\rangle = \langle f(X)\rangle = 0_\mathbb{F}$. This is again not possible, since X^2-1 has degree 2 and therefore is not divisible by $f(X)$, which has degree 3.

  This shows that the order of $X + \langle f(X)\rangle$ is either 13 or $X + \langle f(X)\rangle$ is a primitive element in R. Our aim is to show that if $X + \langle f(X)\rangle$ has order 13 then $2X + \langle f(X)\rangle = -X + \langle f(X)\rangle$ has order 26 (that is, it is a primitive element). Suppose so that $X + \langle f(X)\rangle$ has order 13, that is, $(X + \langle f(X)\rangle)^{13} = X^{13} + \langle f(X)\rangle = 1 + \langle f(X)\rangle$. Then $-X + \langle f(X)\rangle$ will be a primitive element, provided that we show that it cannot have order 13 as well. However,

  $$\begin{aligned}(-X + \langle f(X)\rangle)^{13} &= (-1)^{13}X^{13} + \langle f(X)\rangle = (-1 + \langle f(X)\rangle)\cdot(X^{13} + \langle f(X)\rangle)\\ &= -1 + \langle f(X)\rangle.\end{aligned}$$

  Since $-1 + \langle f(X)\rangle \neq 1 + \langle f(X)\rangle$, the exercise is complete.

- *Ch. 9, Ex. 16* Since $32 = 2^5$, "all" we need to do is to find an irreducible polynomial $p(X) \in \mathbb{F}_2[X]$ of degree 5. Once we have found such a $p(X)$, the quotient ring $(\mathbb{F}_2[X]/\langle p(X)\rangle, +, \cdot)$ gives the field we are looking for.

  To find an irreducible polynomial of degree 5 in $\mathbb{F}_2[X]$, it is sufficient to find a polynomial of degree five, that has no factors of degree 1 or 2. Equivalently, it is sufficient to find a polynomial of degree five, that has no *irreducible* factors of degree 1 or 2. The only

irreducible factors of degree 1 are X and $X+1$, while a direct computation shows that the only irreducible polynomial in $\mathbb{F}_2[X]$ of degree two equals X^2+X+1. Hence we need to find a polynomial of degree 5 in $\mathbb{F}_2[X]$ that is not divisible by X, $X+1$, or X^2+X+1. Moreover, we know that a polynomial $p(X) \in \mathbb{F}_2[X]$ is divisible by X if and only $p(0)=0$ and similarly that it is divisible by $X+1$ if and only if $p(1)=0$ (note $-1=1$, since we work modulo two).

A first try could now be $p(X)=X^5+X+1$, since $p(0)\neq 0$ and $p(1)\neq 1$. However, we still need to check whether or not X^2+X+1 divides X^5+X+1. Using the division algorithm, it turns out that $X^5+X+1=(X^2+X+1)(X^3+X^2+1)$, so this choice of $p(X)$ does not work.

A second try could be $p(X)=X^5+X^2+1$. We still have that X and $X+1$ do not divide X^5+X^2+1 and the division algorithm gives that $X^5+X^2+1=(X^2+X+1)(X^3+X^2)+1$. Hence X^2+X+1 does not divide X^5+X^2+1 either. We conclude that X^5+X^2+1 has no irreducible factors of degree one or two. Hence X^5+X^2+1 is an irreducible polynomial of degree five.

Since $32-1=31$ is a prime number, it is easy to find a primitive element. More precisely: any element different from the one-element in $\mathbb{F}_{32}$ has order 31.

- *Ch. 9, Ex. 17*
  - Let m be a natural number. We note that summing $1+\langle f(X)\rangle$ with itself m times one gets the coset $m+\langle f(X)\rangle$. Since $m \in \mathbb{Z}_p[X]$ is a constant polynomial, $m+\langle f(X)\rangle$ cannot be equal to $\langle f(X)\rangle$ unless m itself is not zero modulo p. So the first time one gets $m+\langle f(X)\rangle$ is when $m=p$, proving that p is the characteristic of the field $\mathbb{F}_q$.
  - If $q=2$ then the sum of elements in $\mathbb{F}_2=\mathbb{Z}_2$ is $0+1=1$. So we suppose $q\neq 2$. Let w be a primitive element in $\mathbb{F}_q$, so that $\mathbb{F}_q=\{0\}\cup\{w^n \mid n=1,\dots,q-1\}$. Then using the hint and recalling that $w^{q-1}=1$:

$$\sum_{a\in\mathbb{F}_q} a = 0+\sum_{n=1}^{q-1} w^n = \sum_{n=1}^{q-1} w^n = -1+\sum_{n=0}^{q-1} w^n =$$

$$-1+\frac{w^q-1}{w-1} = -1+\frac{w\cdot w^{q-1}-1}{w-1} = -1+\frac{w-1}{w-1} = 0.$$

    For the part of the exercise dealing with the product, we see first that,

$$\prod_{a\in\mathbb{F}_q^*} a = \prod_{n=1}^{q-1} w^n = w^{\sum_{n=1}^{q-1} n} = w^{q(q-1)/2}.$$

    Following the hint and divide the cases q even and q odd. Assume first that q is even. From the previous part of the exercise we know that $\mathbb{F}_q$ has characterstic 2, so that $1_{\mathbb{F}_q}=-1_{\mathbb{F}_q}$. Then from the equality obtained above:

$$\prod_{a\in\mathbb{F}_q^*} a = w^{q(q-1)/2} = (w^{q-1})^{q/2} = (1_{\mathbb{F}_q})^{q/2} = 1_{\mathbb{F}_q} = -1_{\mathbb{F}_q}.$$

    Now suppose q is odd. Then

$$\prod_{a\in\mathbb{F}_q^*} a = w^{q(q-1)/2} = (w^{(q-1)/2})^q = (-1_{\mathbb{F}_q})^q.$$

    Since q is odd, $(-1_{\mathbb{F}_q})^q = -1_{\mathbb{F}_q}$.

- *Ch. 9, Ex. 18* In any commutative ring Newton's binomial theorem holds:

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

where n is a non-negative integer. This follows for example by induction on n. Using that in $\mathbb{F}_p$, and hence also in $\mathbb{F}_p[X]/\langle f(X)\rangle$, we have "$p = 0$", we obtain that for $a, b \in \mathbb{F}_p[X]/\langle f(X)\rangle$ we have

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

In the last equality we used the fact that p divides $\binom{p}{i}$ for $1 \leq i \leq p-1$.

- *Ch. 9, Ex. 19* We first show that ϕ is a ring homomorphism. To this end note that $\phi(0) = 0^p = 0$ and for all $a, b \in \mathbb{F}_p[X]/\langle f(X)\rangle$ one has $\phi(a+b) = (a+b)^p = a^p + b^p = \phi(a) + \phi(b)$ from Exercise 18. Also $\phi(1) = q^p = 1$ and $\phi(a \cdot b) = a^p \cdot b^p = \phi(a) \cdot \phi(b)$. It remains to show that ϕ is injective so that the surjectivity follows from Lemma 2.1.6. Assume that $\phi(a) = \phi(b)$. This is equivalent to $a^p - b^p = 0$. If $p = 2$ then $1 = -1$ and hence $0 = a^2 - b^2 = a^2 + b^2 = (a+b)^2 = (a-b)^2$ from Exercise 18. This implies that $a = b$ and ϕ is injective. Assume now that $p > 2$, that is, p is odd. Then $-b^p = (-b)^p$ and hence $0 = a^p - b^p = a^p + (-b)^p = (a-b)^p$ from Exercise 18. Again this yields $a = b$ and ϕ is injective.

- *Ch. 9, Ex. 20* Let w be a generator of the cyclic group $\mathbb{F}^*$. Since $1_\mathbb{F}$ is a unit, so is $-1_\mathbb{F}$ (recall the exercises from the previous chapters!). Also, as suggested in the hint, $1_\mathbb{F} \neq -1_\mathbb{F}$ since the characteristic of $\mathbb{F}$ is not 2. Hence we should be able to find a natural number n such that $w^n = -1_\mathbb{F}$ and hence $w^{2n} = 1_\mathbb{F}$. This implies that the order of w divides $2n \in \mathbb{N}$ and since w is primitive, $\mathbb{F}^*$ contains a finite number of elements. Using that $\mathbb{F} = \{0\} \cup \mathbb{F}^*$ we see that $\mathbb{F}$ is a finite field.

Appendix: The Greek alphabet

An overview of the Greek alphabet. Note that some letters can be written in several ways.

Name	upper case	lower case	name	upper case	lower case
alpha	A	α	nu	N	ν
beta	B	β	xi	Ξ	ξ
gamma	Γ	γ	omicron	O	o
delta	Δ	δ	pi	Π	π, ϖ
epsilon	E	ϵ, ε	rho	P	ρ, ϱ
zeta	Z	ζ	sigma	Σ	σ, ς
eta	H	η	tau	T	τ
theta	Θ	θ, ϑ	upsilon	Υ	υ
iota	I	ι	phi	Φ	ϕ, φ
kappa	K	κ	chi	X	χ
lambda	Λ	λ	psi	Ψ	ψ
mu	M	μ	omega	Ω	ω

Appendix: A small dictionary for mathematical terms

English	Danish
associative	associativ
bijective, bijection	bijektiv, bijektion
co-domain	dispositionsmængde
commutative	kommutativ
composite, composition	sammensat, sammensætning
congruent (congruence)	kongruent (kongruens)
coset	sideklasse
cube	kube, terning, hexaeder
cycle	cykel
cyclic group	cyklisk gruppe
degree	grad
dihedral group	diedergruppe
disjoint	disjunkt
dodecahedron	dodekaeder
domain, integral domain	integritetsområde, område
domain (of a function)	definitionsmængde
double root	dobbeltrod
edge	kant
equation	ligning
even, odd permutation	lige, ulige permutation
equivalence class	ækvivalensklasse
equivalence relation	ækvivalensrelation
extension field	udvidelseslegeme
face, facet	sideflade
field	legeme
fraction	brøktal
generator	frembringer
-gon (four-gon, five-gon, etc.)	-kant (firekant, femkant, osv.)
group	gruppe
group action	gruppevirkning
group homomorphism	gruppehomomorfi
group isomorphism	gruppeisomorfi
homomorphism of groups	gruppeisomorfi
homomorphism of rings	ringisomorfi
icosahedron	ikosaeder
ideal	ideal
image	billede
indeterminate	ubekendte

English	Danish
injective	injektiv
integer	heltal
integral domain, domain	integritetsområde, område
inverse element	inverst element
isomorphism of groups	gruppeisomorfi
isomorphism of rings	ringisomorfi
kernel	kern
left coset	venstresideklasse
monic	monisk
natural number	naturligt tal
octahedron	oktaeder
orbit	bane
permutation	permutation
PID, principal ideal domain	hovedidealområde
polynomial	polynom
preimage	urbillede
primitive element	primitivt element
principal ideal	hovedideal
quotient group	kvotientgruppe, faktorgruppe
reflexive	refleksiv
representative	repræsentant
right coset	højresideklasse
ring	ring
ring homomorphism	ringhomomorfi
ring isomorphism	ringisomorfi
root of a polynomial	rod i et polynom
set	mængde
side	sideflade
stabilizer	stabilisator
subfield	underlegeme
subgroup	undergruppe
surjective	surjektiv
symmetric	symmetrisk
tetrahedron	tetraeder
transitive	transitiv
unit	enhed
vertex (vertices)	hjørne(r)
zero-divisor	nuldivisor

Dansk	**Engelsk**
associativ	associative
bane	orbit
bijektiv, bijektion	bijective, bijection
billede	image
brøktal	fraction
kube	cube
cykel	cycle
cyklisk gruppe	cyclic group
definitionsmængde	domain (of a function)
diedergruppe	dihedral group
disjunkt	disjoint
dispositionsmængde	co-domain
dobbeltrod	double root, root of multiplicity two
dodekaeder	dodecahedron
enhed	unit
faktorgruppe	quotient group
frembringer	generator
grad	degree
gruppe	group
gruppehomomorfi	group homomorphism, homomorphism of groups
gruppeisomorfi	group isomorphism, isomorphism of groups
gruppevirkning	group action
heltal	integer
hexaeder	cube
hjørne(r)	vertex (vertices)
hovedideal	principal ideal
hovedidealområde	PID, principal ideal domain
højresideklasse	right coset
ideal	ideal
ikosaeder	icosahedron
injektiv	injective
integritetsområde	integral domain, domain
inverst element	inverse element
kant	edge
-kant (firekant, femkant, osv.)	-gon (four-gon, five-gon, etc.)
kern	kernel
kommutativ	commutative
kongruent, kongruens	congruent, congruence
kvotientgruppe	quotient group
legeme	field
lige, ulige permutation	even, odd permutation
ligning	equation
monisk	monic
mængde	set
naturligt tal	natural number
nuldivisor	zero-divisor
oktaeder	octahedron
område	integral domain, domain
permutation	permutation
polynom	polynomial
primitivt element	primitive element

Dansk	**Engelsk**
refleksiv	reflexive
repræsentant	representative
ring	ring
ringhomomorfi	ring homomorphism, homomorphism of rings
ringisomorfi	ring isomorphism, isomorphism of rings
rod i et polynomium	root of a polynomial
sammensat, sammensætning	composite, composition
sideflade, side	face, facet, side
sideklasse	coset
stabilisator	stabilizer
surjektiv	surjective
symmetrisk	symmetric
terning	cube, dice
tetraeder	tetrahedron
transitiv	transitive
ubekendte	indeterminate
udvidelseslegeme	extension field
undergruppe	subgroup
underlegeme	subfield
urbillede	preimage
venstresideklasse	left coset
ækvivalensklasse	equivalence class
ækvivalensrelation	equivalence relation

Index