Algebra

Tauno Metsänkylä — Marjatta Näätänen

Finnish original, 2010

English translation, 2022: Nerissa Shakespeare

Preface

This is the English translation of the algebra text in Finnish written by Tauno Metsänkylä and Marjatta Näätänen in 2010 called *Algebra*. The translation was done in the summer of 2022 to serve as optional lecture material for the *Abstract Algebra* course at Aalto University.

For editorial comments, suggestions or corrections, please contact

Camilla Hollanti camilla.hollanti@aalto.fi

Nerissa Shakespeare
September 2022

Contents

Chapter 1

Preliminaries

1.1 Propositions and quantifiers

Propositions

The basic building blocks of formal logic are *propositions* or *expressions*. Propositions are denoted by $A, B, C, \ldots$ We focus only on the *truth value*; a proposition must be either *true* or *false*, and cannot both at the same time. If we cannot determine whether a proposition is true or false, then it is not a valid proposition. We introduce a shorthand, denoting $+$ for true and $-$ for false.

Example 1.1. "2 is an even number" is true.

Just like with natural languages, we can form new expressions from given ones.

1) The *negation* operator "not", denoted by $\neg A$, is the opposite of the proposition A. From the use of the word "not", it is natural to require that if a proposition A is true then its negation $\neg A$ is false and vice versa. The truth value presented in a table is

A	$\neg A$
$+$	$-$
$-$	$+$

2) The *conjunction* "and", denoted by $A \wedge B$, is true only if both A and B are true:

A	B	$A \wedge B$
$+$	$+$	$+$
$+$	$-$	$-$
$-$	$+$	$-$
$-$	$-$	$-$

Example 1.2. Let A be "3 is an odd number" and B be "$2 + 2 = 5$". Then $A \wedge B$ is false.

3) *Disjunction* "or" is denoted by $A \vee B$. The natural language has two meanings for or, the exclusive "either-or" and the inclusive "either-or- or both". Disjunction has the latter meaning as shown by its truth table:

A	B	$A \vee B$
$+$	$+$	$+$
$+$	$-$	$+$
$-$	$+$	$+$
$-$	$-$	$-$

4) *Implication* is denoted by $A \Rightarrow B$ and read as "if A, then B", "A implies B", "B follows from A" or "A only if B". Determining the truth values is trickier when A is false. By convention the implication is then true. If A is true, the implication is true precisely when B is true:

A	B	$A \Rightarrow B$
$+$	$+$	$+$
$+$	$-$	$-$
$-$	$+$	$+$
$-$	$-$	$+$

It must be noted here that there need not be any causal relationship; $A \Rightarrow B$ can be true even if the propositions A and B are not causally linked.

Example 1.3. $A =$"Laura is wearing a red shirt" and $B =$"The time is 12 o'clock".

5) *Equivalence* "A is equivalent to B", "A and B are logically equivalent" or "A if and only if B" is denoted by $A \Leftrightarrow B$. Here $A \Leftrightarrow B$ is true if A and B have the same truth value, and false if A and B have different truth values:

A	B	$A \Leftrightarrow B$
$+$	$+$	$+$
$+$	$-$	$-$
$-$	$+$	$-$
$-$	$-$	$+$

The tables presented above can be used to determine truth values of other expressions: if we know the truth values of the expressions A and B, the truth values of new expressions formed from A and B can be determined by their tables.

Example 1.4. Let V denote the expression $(A \Rightarrow B) \Leftrightarrow (\neg B \Rightarrow \neg A)$. We write the truth table for V:

A	B	$\neg A$	$\neg B$	$A \Rightarrow B$	$\neg B \Rightarrow \neg A$	V
$+$	$+$	$-$	$-$	$+$	$+$	$+$
$+$	$-$	$-$	$+$	$-$	$-$	$+$
$-$	$+$	$+$	$-$	$+$	$+$	$+$
$-$	$-$	$+$	$+$	$+$	$+$	$+$

We result in that V is true regardless of the truth values of A and B. Such an expression is called a *tautology*.

Certain tautologies form rules of inference that are used in mathematical proofs. In the following theorem, each statement can be proved by writing the truth table.

Theorem 1.5. *Let p, q, r be propositions. The following propositions are identically true:*

(1) $\neg(p \vee q) \Leftrightarrow (\neg p \wedge \neg q)$ *(De Morgan's law),*

(2) $\neg(p \wedge q) \Leftrightarrow (\neg p \vee \neg q)$ *(De Morgan's law),*

(3) $[p \wedge (q \vee r)] \Leftrightarrow [(p \wedge q) \vee (p \wedge r)]$ *(distributivity),*

(4) $[p \vee (q \wedge r)] \Leftrightarrow [(p \vee q) \wedge (p \vee r)]$ *(distributivity),*

(5) $(p \Rightarrow q) \Leftrightarrow (\neg p \vee q)$ *(implication),*

(6) $(p \Leftrightarrow q) \Leftrightarrow [(p \Rightarrow q) \wedge (q \Rightarrow p)]$ *(equivalence),*

(7) $[(p \Rightarrow q) \wedge (p \Rightarrow \neg q)] \Leftrightarrow \neg p$,

(8) $(p \Rightarrow q) \Leftrightarrow (\neg q \Rightarrow \neg p)$ *(contrapositive)*,

(9) $[p \wedge (p \Rightarrow q)] \Rightarrow q$,

(10) $[(p \Rightarrow q) \wedge \neg q] \Rightarrow \neg p$

(11) $[(p \Rightarrow q) \wedge (q \Rightarrow r)] \Rightarrow (p \Rightarrow r)$,

(12) $[(p \Leftrightarrow q) \wedge (q \Leftrightarrow r)] \Rightarrow (p \Leftrightarrow r)$.

In the expression $L \Rightarrow M$, L is called the *hypothesis* and M is called the *conclusion*. If $L \Rightarrow M$ is true then "L is true" is a *sufficient condition* for M to be true, and "M is true" is a *necessary condition* for L to be true.

Quantifiers

Example 1.6. Every rational number is a real number. The number 3 is a rational number, therefore 3 is a real number.

The hitherto introduced logic cannot be used to infer formally that the conclusion is correct. We introduce

- the *universal quantifier*: "for all x", "given any x", "for any choice of x", denoted by $\forall x$;
- the *existential quantifier*: "there exists x", "for some x", "there is at least one x", denoted by $\exists x$.

The notation $A(x)$ expresses an argument pertaining to x that is true or false depending on the value of x. It is read as "x has property A". This is called a *one-term predicate* or a *one-term formula*.

Example 1.7. Let $A(x)$ be the expression "$x > 2$". If we substitute the value 1 for x, we get a false proposition; for the value 3 we get a true proposition.

The existential quantifier has another form, $\exists! x$: "there exists exactly one x such that..." In the following example x and y are real numbers:

Example 1.8. a) $\forall x\ (x^2 > x)$, "for all real numbers x we have $x^2 > x$ ". This argument is false. The inequality does not hold for, say, $x = \frac{1}{2}$.

b) $\exists x\ (x^2 > x)$, "there exists some x such that $x^2 > x$". This is true.

c) $\exists! x\ (|x| \leq 0)$, "there exists exactly one x such that $|x| \leq 0$". This is also true since $x = 0$ is the only number that satisfies the inequality.

d) $\forall x\ \exists y\ (x^2 - y = y^2 - x)$, "for each x there exists a y such that $x^2 - y = y^2 - x$". This is true since $y = x$ works, for example.

e) $\exists y\ \forall x\ (x^2 - y = y^2 - x)$, "there exists a y such that for all x we have $x^2 - y = y^2 - x$". Substituting $x = 0$ we see that we must have $y = 0$ or $y = -1$. Yet for $x = 1$ neither of these satisfies the condition. Thus the argument is false.

f)
1. $\forall x\ A(x)$: "For every element x, $A(x)$ is true".
2. $\exists x\ A(X)$: "There exists some x for which $A(x)$ is true".
3. $\forall x\ \neg A(x)$: "For no element x is $A(x)$ true" or "for all elements x, $A(x)$ is false".

4. $\exists x\ \neg A(x)$: "There exists some x for which $A(x)$ is false".

Note that 1 and 4 are negations of each other, likewise 2 and 3.

If a negation and a quantifier, or two quantifiers, are in the same expression, their order is important. In the next examples x denotes a human and $P(x)$ the predicate "x is mortal".

1. $\neg(\forall x\ P(x))$ can be read as follows: "it is not true that every human is mortal" or "there exists some immortal human"; this can thus be written as well as $\exists x\ \neg P(x)$.

2. $\forall x\ \neg P(x)$ means "every human is immortal".

The meaning thus changed when the order of the negation and the quantifier was switched!

Example 1.9. $L(x, y)$ is a two-term formula. The notation $\forall(x, y)\ L(x, y)$ is read as "for all elements x and y, $L(x, y)$ holds". The notation $\forall x\ \exists y\ L(x, y)$ is read as "for all elements x, there exists at least one y such that $L(x, y)$ holds". The notation $\exists y\ \forall x\ L(x, y)$ is read as "there exists at least one y such that $L(x, y)$ holds for all elements x".

Remark 1.10. In mathematical language it is common to omit the notation $\forall x$. For example, when we write $(x + 1)^2 = x^2 + 2x + 1$, where x is considered a real number, we actually mean $\forall x\ \left[(x + 1)^2 = x^2 + 2x + 1\right]$.

Example 1.11. The negation of the expression "the function f has property A at every point" is "there exists some point where f does not have the property A". Beware of the error: "in no point does f have the property A".

Theorem 1.12. *The following propositions are true for all predicates $p(x)$ and $q(x)$:*

$$\begin{aligned}
&\forall x\ \ [p(x) \wedge q(x)] \Leftrightarrow [\forall x\ p(x) \wedge \forall x\ q(x)]\,;\\
&\exists x\ \ [p(x) \wedge q(x)] \Rightarrow [\exists x\ p(x) \wedge \exists x\ q(x)]\,;\\
&\exists x\ \ [p(x) \vee q(x)] \Leftrightarrow [\exists x\ p(x) \vee \exists x\ q(x)]\,;\\
&[\forall x\ p(x) \vee \forall x\ q(x)] \Rightarrow \forall x\ \ [p(x) \vee q(x)]\,.
\end{aligned}$$

The latter two propositions can be obtained from the former ones:

$$\begin{aligned}
&\exists x\ \ [p(x) \vee q(x)] \Leftrightarrow \neg\neg\exists x\ \ [p(x) \vee q(x)]\\
\Leftrightarrow&\neg\forall x\ \neg\,[p(x) \vee q(x)] \Leftrightarrow \neg\forall x\ \ [\neg p(x) \wedge \neg q(x)]\\
\Leftrightarrow&\neg\,[\forall x\ \neg p(x) \wedge \forall x\ \neg q(x)] \Leftrightarrow [\neg\forall x\ \neg p(x)] \vee [\neg\forall x\ \neg q(x)]\\
\Leftrightarrow&\,[\exists x\ \neg\neg p(x)] \vee [\exists x\ \neg\neg q(x)] \Leftrightarrow [\exists x\ p(x) \vee \exists x\ q(x)]\,;
\end{aligned}$$

$$\begin{aligned}
&[\forall x\ p(x) \vee \forall x\ q(x)] \Leftrightarrow [\forall x\ \neg\neg p(x)] \vee [\forall x\ \neg\neg q(x)]\\
\Leftrightarrow&\,[\neg\exists x\ \neg p(x)] \vee [\neg\exists x\ \neg q(x)] \Leftrightarrow \neg\,[\exists x\ \neg p(x) \wedge \exists x\ \neg q(x)]\\
\Rightarrow&\neg\exists x\ \ [\neg p(x) \wedge \neg q(x)] \Leftrightarrow \neg\exists x\ \neg\,[p(x) \vee q(x)]\\
\Leftrightarrow&\neg\neg\forall x\ \ [p(x) \vee q(x)] \Leftrightarrow \forall x\ \ [p(x) \vee q(x)]\,.
\end{aligned}$$

1.2 Mathematical proofs

Mathematics deals with abstract *structures*. Let us take the concepts to be defined as a starting point. We *assume* that these concepts have some basic properties, which are enumerated in *axioms*. An axiom is thus a theorem that is true by convention. Definitions and axioms fix the structure to be considered.

Axioms must not contain *contradictions*: They must not contain conflicting claims, and further logical inference must not conclude in a contradiction. The number of axioms is striven to be minimal: they ought to be *independent*, that is, no axiom can be proven true by the others. The idea of the axiomatic approach is that all properties of the structure to be considered are derived from the axioms.

Later as an example, we will consider the axiomatic definition of the natural numbers. Thereafter we develop mathematical theory pertaining it by proving *theorems* starting from the axioms. A theorem comprises an *assumption* p and an *argument* q; when proving the theorem we deduce that if p is true then q is true as well.

A *direct proof* of a theorem or statement corresponds to the tautology

$$[p \wedge (p \Rightarrow q)] \Rightarrow q,$$

where the idea is that if p is true and the implication $p \Rightarrow q$ can be concluded as true, then the argument q is also true.

Because $(p \Rightarrow q) \Leftrightarrow (\neg q \Rightarrow \neg p)$ is a tautology, the expression

$$[p \wedge (\neg q \Rightarrow \neg p)] \Rightarrow q$$

is also a tautology. This contains the principle of *indirect proofs*: If the assumption p is true and we can prove that the negation of the argument $\neg q$ implies the negation of the assumption $\neg p$, then a contradiction arises if $\neg q$ is true. Both p and $\neg p$ cannot be true simultaneously. The only option is that $\neg q$ is not true, so q is true.

The essential content of an indirect proof thus comprises proving the implication $\neg q \Rightarrow \neg p$. Then we start by considering the negation of the argument $\neg q$, the so-called *antithesis*, and exploring what follows. If a contradiction with the assumption or some derived statement arises, the antithesis cannot be true so its negation, the argument, is true.

Example 1.13. A natural number is said to be *perfect* if it can be written as the sum of its factors. The number itself is not included as a factor. Some examples of perfect numbers are $6 = 1 + 2 + 3$ and $28 = 1 + 2 + 4 + 7 + 14$.

Indirectly we can prove the following argument: A perfect number is not a prime. The assumption p is thus "n is a perfect number" and the argument q is "n is not a prime". The antithesis is "n is a prime".

If n is a prime, then it has only two factors, 1 and n. The number itself is not included in the sum, so the sum equals 1. Since the smallest prime is 2, n is not a perfect number.

The antithesis has thus lead to a contradiction with the assumption; the antithesis is false and so the argument is true.

Note in particular that a theorem cannot be proven by deriving the assumption or some equivalent statement from the argument. This can be seen from the fact that the expressions $p \Rightarrow q$ and $q \Rightarrow p$ are not equivalent, that is, $(p \Rightarrow q) \Leftrightarrow (q \Rightarrow p)$ is not a tautology. As an example, the statement "the sum of the angles of an n-gon is $n\frac{\pi}{3}$" cannot be justified simply by stating the fact that it gives the correct result for $n = 3$.

Instead starting from the argument q and resulting in the assumption or some other true statement p serves as a proof if a chain of equivalences

$$q \Leftrightarrow \ldots \Leftrightarrow p$$

can be formed. Yet even then it is essential that this chain contains the implication chain from right to left, the reasoning from assumption to argument.

Deducing that a proposition of the form $\forall x\ p(x)$ is false is often easier than deducing that it is true. To prove that something is false, it is sufficient to find one element x for which $p(x)$ is false; this is called a *counter example*. If instead the proposition needs to be proven true, all possible elements must be considered.

For example, the prior statement "the sum the angles of an n-gons is given by the formula $n\pi/3$" can be proven false by counter example, it does not hold for squares. Proving the corresponding correct statement "the sum of the angles of an n-gon is $(n-2)\pi$" is more difficult.

Finally, proofs of mathematical statements rarely follow the methods of formal logic, instead natural intuitive logic is applied. Logical symbols should be mainly understood as shorthand. Nonetheless, sometimes the logical structure of a proof may be so complex that understanding it may necessitate the use of formal logic.

1.3 Notations of set theory

The basic objects in the language of mathematics are *sets*, such as sets of numbers, sets of functions, or sets of vectors. You are surely familiar with the following notations for sets of numbers:

$\mathbb{N} = \{0, 1, 2 \ldots\}$, the *natural numbers*,

$\mathbb{Z}$, the *integers*,

$\mathbb{Q}$, the *rational numbers*,

$\mathbb{R}$, the *real numbers*,

$\mathbb{C}$, the *complex numbers*.

The following concepts and notations are everyday tools for the mathematician.

$x \in A$: x is an element of the set A, or x belongs to the set A; the opposite is denoted by $x \notin A$;

$B \subset A$ (or $B \subseteq A$): B is a *subset* of A or B is contained in A;

$B \subsetneq A$: B is a *proper subset* of A, that is, $B \subset A$ and $B \neq A$;

$\{x, y, z, \ldots\}$: a set whose elements are $x, y, z, \ldots$;

$\{x \in A \mid P_1(x), \ldots, P_n(x)\}$ or $\{x \mid P_1(x), \ldots, P_n(x)\}$: the set of elements x (in A) that satisfy the conditions $P_1(x), \ldots, P_n(x)$.

Example 1.14. (i) If $A = \{2, 4, 6, 8, 10\}$, then for example $6 \in A$, $7 \notin A$, $\{4, 8\} \subset A$ and $\{4, 8\} \subsetneq A$.

(ii) Observe that $\{1, 2\} = \{2, 1\} = \{1, 1, 2\}$, for example.

(iii) $\{x \in \mathbb{Z} \mid 0 < x < 5,\ x \text{ is even }\} = \{2, 4\}$.

(iv) $\{x \in \mathbb{R} \mid a < x < b\} =$ is the open interval from point a to point b, denoted by (a, b). It is sometimes denoted by $]a, b[$.

If A and B are subset of some larger set M, then we can define more subsets of M as follows:

$$\text{union} \quad A \cup B = \{x \in M \mid x \in A \text{ or } x \in B\},$$

$$\text{intersection} \quad A \cap B = \{x \in M \mid x \in A \text{ and } x \in B\},$$

$$\text{(set) difference} \quad A \setminus B = \{x \in M \mid x \in A \text{ and } x \notin B\},$$

$$\text{complement} \quad A^c = M \setminus A.$$

The intersection and union of more than two sets or even infinitely many sets is defined similarly.

The set with no elements is called the *empty set* and it is denoted by $\emptyset$. Note that $\emptyset \in A$ whatever the set A is.

Example 1.15. (i) $A \setminus B = A \cap B^c$,

(ii) $\mathbb{Z} \setminus \mathbb{Q} = \emptyset$,

(iii) $\{(x, y) \in \mathbb{R}^2 \mid x = y\} \cap \{(x, y) \in \mathbb{R}^2 \mid x = -y\} = \{(0, 0)\}$,

(iv) The union of all closed real intervals $[n, n + 1]$ where $n = 0, \pm 1, \pm 2$ is equal to $\mathbb{R}$.

We shall denote

– the *power set* of a set A by $\mathcal{P}(A) = \{B \mid B \subset A\}$,

– the *cartesian product* of the sets A and B by

$$A \times B = \{(x, y) \mid x \in A,\ y \in B\}.$$

The set $A \times A$ is also denoted by A^2.

1.4 Congruence

The concept of *congruence* introduced below provides a method for handling divisibility in a similar way to equations.

Definition 1.16. Let m be a positive integer. If $a, b \in \mathbb{Z}$ and $a - b$ is divisible by m, we say that a is congruent to b *modulo* m, denoting

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

This is called a *congruence* of the set $\mathbb{Z}$; the number m is called its *modulus*. The contrary: a is *incongruent* to b modulo m, denoted by $a \not\equiv b \pmod{m}$.

Example 1.17. $38 \equiv 2 \pmod{6}$, $\quad 12 \equiv -3 \pmod{5}$, $\quad 100 \not\equiv 1 \pmod{10}$.

Lemma 1.18. *Let m be a positive integer. For all $a, b, c \in \mathbb{Z}$ we have*

$a \equiv a \pmod{m}$,

$a \equiv b \pmod{m} \implies b \equiv a \pmod{m}$,

$a \equiv b \pmod{m}$ *and* $b \equiv c \pmod{m} \implies a \equiv c \pmod{m}$.

According to the definition $a \equiv b \pmod{m}$ if and only if a is some multiple of m away from b, that is

$$a \equiv b \pmod{m} \qquad \Longleftrightarrow \qquad a = b + mq, \quad q \in \mathbb{Z}.$$

We see from this that the congruence modulo m splits $\mathbb{Z}$ into disjoint sets of the following form:

$$[a] = \{a + mk \mid k \in \mathbb{Z}\}.$$

The set $[a]$ is called the *residue class modulo* m of a; it is also denoted by $\overline{a}$, $[a]_m$, a_m or $a+m\mathbb{Z}$. Numbers belonging to the same residue class give the same remainder when divided by m. By going through all possible division remainders, that is, the numbers $0, 1, \ldots, m-1$, we get a set of representatives for the residue classes, the *smallest nonnegative remainders* modulo m of the integers. We can write the set of all residue classes modulo m, denoted by $\mathbb{Z}_m$, as follows:

$$\mathbb{Z}_m = \{\overline{0}, \overline{1}, \ldots, \overline{m-1}\} = \{0_m, 1_m \ldots, (m-1)_m\}.$$

Sometimes the residue mark, the bar or subscript, is omitted.

Example 1.19. $\mathbb{Z}_3 = \{\overline{0}, \overline{1}, \overline{2}\}$, where

$$\begin{aligned}
0_3 = \overline{0} &= 3\mathbb{Z} = \{3k \mid k \in \mathbb{Z}\} = \{\ldots, -9, -6, -3, 0, 3, 6, 9, \ldots\}, \\
1_3 = \overline{1} &= 1 + 3\mathbb{Z} = \{1 + 3k \mid k \in \mathbb{Z}\} = \{\ldots, -8, -5, -2, 1, 4, 7, 10, \ldots\} \\
2_3 = \overline{2} &= 2 + 3\mathbb{Z} = \{2 + 3k \mid k \in \mathbb{Z}\} = \{\ldots, -7, -4, -1, 2, 5, 8, 11, \ldots\}.
\end{aligned}$$

The set $\mathbb{Z}_3$ can also be presented as $\{\overline{-1}, \overline{0}, \overline{1}\}$ or $\{\overline{7}, \overline{33}, \overline{2}\}$, for example.

Example 1.20. In the special case of $m = 1$, the congruence modulo m is trivial:

$$a \equiv b \pmod{1} \quad \forall\ a, b \in \mathbb{Z}.$$

In particular, we have $\mathbb{Z}_1 = \{\overline{0}\}$ where $\overline{0} = \mathbb{Z}$.

Theorem 1.21. (i) *If* $a \equiv b \pmod{m}$ *and* $c \equiv d \pmod{m}$, *then*

$$a + c \equiv b + d \pmod{m} \qquad \textit{and} \qquad ac \equiv bd \pmod{m}.$$

(ii) *If* $ca \equiv cb \pmod{m}$ *and* $\gcd(c, m) = 1$, *then* $a \equiv b \pmod{m}$.

(iii) *If* $a \equiv b \pmod{km}$ *where* k *is a positive integer, then* $a \equiv b \pmod{m}$.

Proof. (i) The number $(a + c) - (b + d) = (a - b) + (c - d)$ is divisible by m because $m \mid a - b$ and $m \mid c - d$; likewise for $ac - bd = (a - b)c + b(c - d)$.

(ii) It follows from the conditions $m \mid c(a - b)$ and $\gcd(c, m) = 1$ that $m \mid (a - b)$.

(iii) Because $a - b$ is a multiple of km, it is also a multiple of m.

□

Part (i) of the theorem says that congruences modulo m can be added and multiplied, likewise subtracted and raised to a power. In particular, if $P(x)$ is an integer polynomial,

$$P(x) = c_0 + c_1 x + \cdots + c_t x^t \quad (c_i \in \mathbb{Z}\ \forall\ i),$$

then it follows from the congruence $a \equiv b \pmod{m}$ that $P(a) \equiv P(b) \pmod{m}$.

Example 1.22. Let us compute the remainder of $18^2 + 2^{100}$ divided by 11. We compute the powers first and then the sum. Firstly,

$$18^2 \equiv 7^2 = 49 \equiv 5 \pmod{11}.$$

We notice that $2^5 = 32 \equiv 10 \equiv -1 \pmod{11}$, which makes calculating 2^{100} simple:

$$2^{100} = \left(2^5\right)^{20} \equiv (-1)^{20} = 1 \pmod{11}.$$

Thus the remainder is $18^2 + 2^{100} \equiv 5 + 1 = 6 \pmod{11}$.

Example 1.23. If the congruence $3 \equiv 15 \pmod{12}$ is divided by 3 on both sides, we get $1 \equiv 5 \pmod{12}$ which does not hold. Observe that $\gcd(3, 12) \neq 1$. By part (ii) of Theorem 1.21, the requirement $\gcd(c, m) = 1$ is thus imperative.

However, observe that $1 \equiv 5 \pmod{4}$. Based on these observations, we can hypothesise that

$$ca \equiv cb \pmod{m} \qquad \Longrightarrow \qquad a \equiv b \pmod{\frac{m}{\gcd(c, m)}}$$

in general. To prove this, suppose that $m = kx_1$ and $c = kx_2$, where $\gcd(c, m) = k$. We rewrite the initial congruence as

$$kx_2a \equiv kx_2b \pmod{kx_1}.$$

By part (iii) of Theorem 1.21 we can eliminate k from the modulus:

$$kx_2a \equiv kx_2b \pmod{x_1}.$$

Since $k = \gcd(c, m)$, we know that $\gcd(x_1, x_2) = 1 = \gcd(x_1, k)$. Therefore we get by part (ii) of Theorem 1.21 that

$$a \equiv b \pmod{x_1},$$

which proves our hypothesis.

The set of residue classes $\mathbb{Z}_m$ forms an important *algebraic system* when we define addition and multiplication appropriately. It will be discussed later in the theory of groups, rings and fields. Nonetheless as preparation, let us define these operations here:

$$\overline{a} + \overline{b} = \overline{a + b}, \qquad \overline{a} \cdot \overline{b} = \overline{ab}. \tag{1.1}$$

The problem is though that residue classes $\overline{a}$ are represented by a representative a and the choice of the representative is not unique. It is to be shown that the thus defined addition and multiplication are not dependent on the choice of the representatives.

Situations like this where the definition seems to depend on the choice of the representatives of the equivalence classes is common in mathematics. When this seeming dependence has been disproved, we tend to say that the concept in question is *well defined*.

Theorem 1.24. *The addition and multiplication of residue classes defined by the Equations* (1.1) *are well defined.*

Proof. Suppose that $\overline{a} = \overline{a'}$ and $\overline{b} = \overline{b'}$. Then we have $a \equiv a'$ and $b \equiv b' \pmod{m}$. By Theorem 1.21 (i), we get

$$a + b \equiv a' + b', \qquad ab \equiv a'b' \qquad \pmod{m}.$$

It follows that $\overline{a + b} = \overline{a' + b'}$ and $\overline{ab} = \overline{a'b'}$, which proves the argument. □

Example 1.25. Among the set of residue classes modulo 7, we have $\overline{4} + \overline{5} = \overline{2}$. On the other hand, we also have $\overline{4} = \overline{60}$ and $\overline{5} = \overline{75}$, so $\overline{4} + \overline{5} = \overline{60} + \overline{75} = \overline{135}$. Check by computing that $\overline{135} = \overline{2}$.

Congruences are used to study *Diophantine equations*. These are equations for whom we search for integer solution.

Example 1.26. Consider the Diophantine equation $x^2 - 2y^2 = 5$. The congruence $x^2 - 2y^2 \equiv 5 \pmod 8$ has no solutions because the square residues modulo 8 are $0, 1, 4$. Consequently, the equation has no integer solutions.

Example 1.27. Alex bought big cakes at a price of 15€ each and small cakes 11€ each. The bill was 137€. How many of cakes of each type did Alex buy?

The problem to be solved is the Diophantine equation $15x + 11y = 137$. We solve it by transitioning to the congruence $15x \equiv 137 \pmod{11}$

We can immediately see that solving a linear Diophantine equation in two indeterminates $ax + my = c$ is in general equivalent to solving the congruence

$$ax \equiv c \pmod m. \tag{1.2}$$

The next result pertaining to congruences is of use in many situations later.

Theorem 1.28. *If* $\gcd(a, m) = 1$*, the congruence* (1.2) *has a unique solution* $x \in \mathbb{Z}$ *in the interval* $0 \leq x \leq m - 1$.

Proof. By assumption there exists $u, v \in \mathbb{Z}$ such that $au + mv = 1$, and therefore

$$a(uc) + m(vc) = c.$$

The congruence has thus a solution $x = uc$. Furthermore, all solutions x of the congruence are congruent to one another modulo m because by Theorem 1.21 (ii)

$$ax_1 \equiv ax_2 \pmod m \qquad \Longrightarrow \qquad x_1 \equiv x_2 \pmod m.$$

Thus exactly one of the solutions belongs to the interval $0 \leq x \leq m - 1$. □

For small values of m it is often quicker to find the solution by trial and error.

1.5 Maps

The mathematical terms *function* and *mapping* or *map* are synonyms. The latter is more common in algebra. In the following we present a summary of basic facts pertaining to maps.

A map f from a *set* A to a *set* B, denoted by $f\colon A \to B$, connects *every* element x in A *uniquely* to an element $y = f(x)$ in B. Here

- A is the *domain* of f,
- B is the *range* of f,
- y is the *image* of x.

We may also say that f *maps* the element x to the element y, denoting

$$f\colon A \to B, \qquad x \mapsto y,$$

or

$$f\colon A \to B, \qquad f(x) = y.$$

Example 1.29. We denote $\mathbb{R}_+^* = \{x \mid x > 0\}$. Some familiar maps from real analysis are

$$f\colon \mathbb{R} \to \mathbb{R}, \qquad f(x) = \sin x$$
$$g\colon \mathbb{R}_+^* \to \mathbb{R}, \qquad g(x) = \ln x.$$

Example 1.30. The following are some examples from linear algebra.

(i) The determinant map

$$d\colon \mathcal{M}_2(\mathbb{R}) \to \mathbb{R}, \qquad d(A) = \det(A).$$

The determinant map can be defined more generally for $\mathcal{M}_n(\mathbb{R}) \to \mathbb{R}$.

(ii) Denote by S_n the set of all permutations of the numbers $1, 2, \ldots, n$. For a permutation $\alpha \in S_n$, its sign $\operatorname{sign}(\alpha)$ is defined by the map

$$s\colon S_n \to \mathbb{Z}, \qquad s(\alpha) = \operatorname{sign}(\alpha).$$

(iii) If V is a vector space, we can define the map

$$n\colon V \to V, \qquad n(X) = -X.$$

(iv) The absolute value map is defined in the vector space $\mathbb{R}$ by

$$N\colon \mathbb{R} \to \mathbb{R}, \qquad N(X) = |X|.$$

This can be generalised as the norm map for an inner product space V,

$$N\colon V \to \mathbb{R}, \qquad N(X) = \|X\|.$$

Some more terminology related to a map $f\colon A \to B$:

- The set $f(A) = \{f(x) \mid x \in A\}$ is the *image set*, or simply the *image*, of the map f. It is also denoted by $\operatorname{Im}(f)$.
- More generally, if $A_0 \subset A$, the set $f(A_0) = \{f(x) \mid x \in A_0\}$ is the *image (set)* of the subset A_0.
- If $B_0 \subset B$, the set $f^{-1}(B_0) = \{x \in A \mid f(x) \in B_0\}$ is the *preimage* of the set B_0.
- A map f is called *surjective* or a *surjection* if $\operatorname{Im}(f) = B$. Then we say that f is a map from A *onto* B.
- A map f is called *injective* or an *injection* if all elements have different images, that is,

$$x_1, x_2 \in A, \quad x_1 \neq x_2 \quad \Longrightarrow \quad f(x_1) \neq f(x_2).$$

This can also be written in the following form, which is often more convenient:

$$x_1, x_2 \in A, \quad f(x_1) = f(x_2) \quad \Longrightarrow \quad x_1 = x_2.$$

- A map f is called *bijective* or a *bijection* if it is *inversely unique*, if it is an injection and a surjection. Then every element y in B has a unique preimage x in the set A.

Two maps $f_1\colon A \to B$ and $f_2\colon A \to B$ are determined to be equal if

$$f_1(x) = f_2(x) \quad \forall\, x \in A.$$

Then we write $f_1 = f_2$. Then in particular the maps have the same domain and range.

The map

$$f\colon A \to A, \quad f(x) = x$$

is called the *identity map* of the set A and denoted by $f = \mathrm{id}_A$.

The *composition* of two maps $f\colon A \to B$ and $g\colon B \to C$ is defined as

$$g \circ f\colon A \to C, \quad (g \circ f)(x) = g(f(x)).$$

It follows immediately that the composite map is associative, i.e., if we have an additional function h, then

$$(h \circ g) \circ f = h \circ (g \circ f).$$

Composing a map with the identity map is simple:

$$\mathrm{id}_B \circ f = f, \quad f \circ \mathrm{id}_A = f. \tag{1.3}$$

The composition $g \circ f$ is also written as gf.

If a map $f\colon A \to B$ is a bijection, then it has an *inverse map*

$$f^{-1}\colon B \to A, \quad f(x) \mapsto x.$$

Then (think why!)

$$f^{-1} \circ f = \mathrm{id}_A, \quad f \circ f^{-1} = \mathrm{id}_B.$$

More generally every injective map $f\colon A \to B$ defines a bijection $A \to \mathrm{Im}(f), \;\; x \mapsto f(x)$. Then as well the map $\mathrm{Im}(f) \to A, \;\; f(x) \mapsto x$ is called (slightly imprecisely) the inverse map of f, and denoted by f^{-1}.

If a map $f\colon A \to B$ has an inverse map, the notation $f^{-1}(B_0)$ with $B_0 \subset B$ can be interpreted in two ways. Both mean the same set nonetheless.

Theorem 1.31. *Let f be a map $A \to B$. If there exists some map $g\colon B \to A$ such that*

$$g \circ f = \mathrm{id}_A, \quad f \circ g = \mathrm{id}_B,$$

then f is a bijection and $f^{-1} = g$.

Proof. If $y \in B$, then by the condition $f \circ g = \mathrm{id}_B$ we have $f(g(y)) = y$. Hence y has a preimage $g(y)$ and therefore f is surjective.

If $x_1, x_2 \in A$ and $f(x_1) = f(x_2)$, then we also have $g(f(x_1)) = g(f(x_2))$. Because $g \circ f = \mathrm{id}_A$, this equation simplifies to $x_1 = x_2$. Hence f is injective.

By the previous, f is a bijection and its inverse map f^{-1} thus exists. It follows from the equation $f \circ g = \mathrm{id}_B$ that

$$f^{-1} \circ (f \circ g) = f^{-1} \circ \mathrm{id}_B.$$

By applying the associativity property of function composition, the condition $f^{-1} \circ f = \mathrm{id}_A$ and Equations (1.3), we obtain the result $g = f^{-1}$. □

If f is a map $A \to B$ and $A_0 \subset A$, the map

$$g\colon A_0 \to B, \quad g(x) = f(x)$$

is called the *restriction* of f to the *set* A_0, denoted by $g = f|_{A_0}$. We also say that f is the *extension* of the map g to the set A.

Example 1.32. We denote $\mathbb{R}_+ = \{x \in \mathbb{R} \mid x \geq 0\}$. The map

$$f\colon \mathbb{R}_+ \to \mathbb{R}_+, \quad f(x) = \sqrt{x}$$

is a bijection; its inverse map is the restriction of the map

$$h\colon \mathbb{R} \to \mathbb{R}_+, \quad h(x) = x^2$$

to the set $\mathbb{R}_+$.

Example 1.33. For a complex number $z = x + iy$ its absolute value $|z| = \sqrt{x^2 + y^2}$ defines a map $\mathbb{C} \to \mathbb{R}_+, \;\; z \mapsto |z|$, which is the extension of the map $\mathbb{R} \to \mathbb{R}_+, \;\; x \mapsto |x|$.

Lemma 1.34. *If f is a map $X \to Y$, $A \subset X$ and $B \subset Y$, then*

(a)) $f^{-1}(f(A)) \supset A$,

(b)) $f(f^{-1}(B)) \subset B$.

In both cases the inclusion can be proper.

Proof. a) Let $x \in A$. Now we have $f(x) \in f(A)$, so $x \in f^{-1}(f(A))$.

b) Let $y \in f((f^{-1}(B))$. Then there exists $x \in f^{-1}(B)$ for which $y = f(x)$. Because $x \in f^{-1}(B)$, we have $f(x) \in B$. Therefore we get $y \in B$.

An example demonstrating that the inclusion can be proper: Take $X = Y = \{1, 2, 3\}$ and $f\colon X \to Y, \; f(x) = 1$. Let $A = \{3\}$ and $B = Y$. Now $f^{-1}(f(A)) = f^{-1}(\{1\}) = X \neq A$ and $f(f^{-1}(B)) = f(x) = \{1\} \neq B$.

□

1.6 Induction and the natural numbers

The set of the natural numbers $\mathbb{N} = \{0, 1, 2, \ldots\}$ is important because it is the basis upon which more complicated sets, such as the integers, the rational numbers, and the residue classes $\mathbb{Z}_n$, are constructed. Furthermore, inspecting algebraic structures needs certain mappings from the natural numbers to the structure in question.

The natural numbers can be formed with a purely set-theoretic construction. Another way is to define $\mathbb{N}$ axiomatically. Here we use the Peano axioms to define $\mathbb{N}$. Intuitively the first number in $\mathbb{N}$ is $0 \in \mathbb{N}$, and we can define a mapping $s\colon \mathbb{N} \to \mathbb{N}$ by $0 \mapsto 1 \mapsto 1 \mapsto 2 \mapsto \ldots \mapsto n \mapsto n + 1 \mapsto \ldots$

Definition 1.35 (Peano axioms)**.** Let $\mathbb{N}$ be a set, s a mapping $\mathbb{N} \to \mathbb{N}$ and $0 \in \mathbb{N}$. The triple $(\mathbb{N}, s, 0)$ is the set of natural numbers if the following conditions hold:

(P1) s is an injection;

(P2) $0 \notin s(\mathbb{N})$;

(P3) if $A \subset \mathbb{N}$ and

(i) $0 \in A$,

(ii) $n \in A \Rightarrow s(n) \in A$,

then $A = \mathbb{N}$ (induction axiom).

We define addition and multiplication based on these axioms; in particular $s(n) = n + 1$. The *induction principle* follows from Axiom (P3): Assume that we attach an argument $E(n)$ to every natural number, and denote $A = \{n \in \mathbb{N} \mid E(n) \text{ is true } \}$. If $E(0)$ is true and if $E(r) \Rightarrow E(s(r))$, then $E(n)$ is true for all $n \in \mathbb{N}$. The phase $E(r) \Rightarrow E(s(r))$ is called the induction step.

Induction proofs can also be applied to arguments of the form $E(n)\ \forall\ n \geq n_0$, that is, the starting point can be some $n_0 \geq 0$. The following theorem states when we can start from the number 1 instead of 0.

Theorem 1.36. *Let $A \subset \mathbb{N}_+ = \mathbb{N} \setminus \{0\}$. If*

(i) $1 \in A$ *and*

(ii) $k \in A \Rightarrow k + 1 \in A$,

then $A = \mathbb{N}_+$.

Proof. Let $B = A \cup \{0\}$, then we have $0 \in B$. Now let $n \in B$. If $n = 0$ then $1 \in A \subset B$, so $0 + 1 \in B$. If $n \neq 0$, then $n \in A$, so $n + 1 \in A$ and consequently $n + 1 \in B$. Now we have $B = \mathbb{N}$, therefore $A = B \setminus \{0\} = \mathbb{N} \setminus \{0\}$. □

The following illustrates the content of the induction principle: First we state that the argument $E(n_0)$ holds. When the induction has been proved, we get the validity of $E(n_0 + 1)$. Likewise, step by step, we prove $E(n_0 + 2)$ and so forth $E(n)$ for all $n \geq n_0$.

Example 1.37. We prove by induction that

$$1^2 + 2^2 + \cdots + n^2 = \frac{n}{6}(n + 1)(2n + 1),\ n \in \mathbb{N} \setminus \{0\}.$$

Now $E(n)$ is the equation $1^2+2^2+\cdots+n^2 = \frac{n}{6}(n+1)(2n+1)$, and the starting point is $n_0 = 1$.

Proof. 1) For the value $n = 1$, the left side is $1^2 = 1$ and the right side is $\frac{1}{6} \cdot 2 \cdot 3$, hence $E(1)$ is true.

2) When proving the induction step, we *assume* $E(n)$ and *prove* $E(n + 1)$ *based on it*:

$$\begin{aligned} 1^2 + 2^2 + \cdots + n^2 + (n + 1)^2 &= \left[1^2 + 2^2 + \cdots + n^2\right] + (n + 1)^2 \\ &\overset{(*)}{=} \frac{n}{6}(n + 1)(2n + 1) + (n + 1)^2 = \frac{n + 1}{6}\left[(n + 1) + 1\right]\left[2(n + 1) + 1\right], \end{aligned}$$

where the step $(*)$ is justified by the induction assumption $E(n)$.

□

Example 1.38. We prove by induction the formula

$$\sum_{k=0}^{n} k = \frac{n(n + 1)}{2}.$$

Proof. 1) For the value $n = 0$, the left side is 0, and the right side is $\frac{0(0+1)}{6} = 0$. Hence $E(0)$ is true.

2) The induction step:

$$\begin{aligned} 1 + 2 + \cdots + n + (n + 1) &= [1 + 2 + \cdots + n] + (n + 1) \\ &= \frac{n(n + 1)}{2} + (n + 1) = \frac{(n + 1)\left[(n + 1) + 1\right]}{2}. \end{aligned}$$

□

The Peano axioms do not uniquely define the set of natural numbers, there are many sets satisfying the Axioms (P1)–(P3). These are all equally good models for the natural numbers because they are "structurally equivalent". This can be expressed precisely by the following theorem whose proof we will omit.

If $(\mathbb{N}, s, 0)$ and $(\mathbb{N}', s', 0')$ are models of the natural numbers, then there exists a unique bijection $f\colon \mathbb{N} \to \mathbb{N}'$, for which $f \circ s = s' \circ f$ and $f(0) = 0'$.

1.7 Equivalence relations and partitions

Recall that the *Cartesian product* of two sets A_1 and A_2 is the set $A_1 \times A_2$ that is formed by all ordered pairs (a_1, a_2) where $a_1 \in A_1$ and $a_2 \in A_2$. The Cartesian product $A \times A$ is also denoted by A^2.

Every subset R of a Cartesian product $A \times A$ defines a *relation on the set* A: If $(a, b) \in R$ we say that the element a is R-related to the element b, denoted by $a\ R\ b$. Mathematically interesting relations are usually such that some "rules" determine when $a\ R\ b$. This rule is often, though slightly imprecisely, called the relation.

Example 1.39. Examples of some different relations.

(i) On the set $\mathbb{R}^2$: the distance from a point (x_1, y_1) to the point (x_2, y_2) is an integer.

(ii) On the set $\mathbb{R}$: $x < y$.

(iii) On the set $\mathcal{M}_2(\mathbb{R})$: $\det(AB) = 0$.

(iv) The integers x and y can be related by the following relations:
$x \leq y$, $\quad x \mid y$, $\quad \gcd(x, y) = 1$, $\quad x = 2 + y$.

Definition 1.40. A relation R on a set A is called an *equivalence relation* on A if it satisfies:

(E1) $\forall\ a \in A\colon \quad a\ R\ a$ (reflexivity),

(E2) if $a, b \in A$ and $a\ R\ b$, then $b\ R\ a$ (symmetry),

(E3) if $a, b, c \in A$ and $a\ R\ b$ and $b\ R\ c$, then $a\ R\ c$ (transitivity).

These conditions can also be expressed as follows: for all $a, b, c \in A$

$$\begin{aligned} &(a, a) \in R, \\ &(a, b) \in R \quad \Rightarrow \quad (b, a) \in R, \\ &(a, b) \in R \text{ and } (b, c) \in R \quad \Rightarrow \quad (a, c) \in R \end{aligned}$$

The equivalence relation is denoted by $\sim$. If $a \sim b$, we say that a is *equivalent* to b. Due to symmetry, we can also say that a and b are equivalent.

Definition 1.41. Let $\sim$ be an equivalence relation on a set A. The elements equivalent to each $a \in A$ form a subset of A that is called the *equivalence class* of a and denoted by $[a]$, that is,

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

The element a is called the *representative* of the equivalence class $[a]$.

Every equivalence class is formed by elements that are equivalent to one another, that is,

$$a \text{ and } b \text{ belong to the same equivalence class} \quad \Longleftrightarrow \quad a \sim b.$$

Indeed if $a \sim b$, then $b \in [a]$ and by (E1) we have $a \in [a]$. Conversely if a and b belong to $[c]$, then $a \sim c$ and $b \in c$, thus by (E2) and (E3) we get $a \sim b$.

Theorem 1.42. *If $\sim$ is an equivalence relation on a set A, then A can be expressed as a union of disjoint equivalence classes:*

$$A = \bigcup_{a \in D} [a] \qquad ([a] \cap [a'] = \emptyset \ \ \forall\ a, a' \in D,\ a \neq a'),$$

where D is a subset of A that contains exactly one representative from each equivalence class, the so-called set of representatives of the equivalence classes.

Proof. We need to show that for $a, b \in A$ either $[a] = [b]$ or $[a] \cap [b] = \emptyset$ holds. Suppose that $[a] \cap [b] \neq \emptyset$. Then some element $c \in A$ belongs to both of the classes $[a]$ and $[b]$, that is, $c \sim a$ and $c \sim b$. Just like above, it follows that

$$a \sim b.$$

Now let x be an arbitrary element of $[a]$, so $x \sim a$. Combined with what we just got, it follows from (E3) that $x \sim b$, so $x \in [b]$. Thus we have proven that $[a] \subset [b]$. By symmetry, we also have $[b] \subset [a]$. Now combining these, we result in the equality $[a] = [b]$. □

If a set A is the union of its nonempty disjoint subsets, we say that these subsets form a *partition* of A. We can rewrite Theorem 1.42 thus: *If an equivalence relation is defined on the set A, then this equivalence relation forms a partition of A.*

The collection or set of all equivalence classes is called the *quotient set* of A. It is denoted by $A/\sim$, that is,

$$A/\sim = \{[a] \mid a \in A\} = \{[a] \mid a \in D\}.$$

Example 1.43. Some examples of the partitions of a set $A \neq \emptyset$:

(a) $\mathcal{B} = \{A\}$,

(b) $\mathcal{B} = \{\{a\} \mid a \in A\}$,

(c) $\mathcal{B} = \{B, A \setminus B\}$ when $\emptyset \neq B \subsetneq A$.

Theorem 1.44. *If $f\colon A \to E$ is a mapping, the set*

$$\mathcal{O}_f = \{f^{-1}\{e\} \mid e \in f(A)\}$$

is a partition of A. Conversely, every partition of A can be formed like this.

Proof. It is trivial that $f^{-1}\{e\} \neq \emptyset$ for every $e \in f(A)$. Every $a \in A$ belongs to exactly one set in the family $\mathcal{O}_f$, namely the set $f^{-1}\{f(a)\}$. Hence $\mathcal{O}_f$ is a partition of A.

Conversely let $\mathcal{O}$ be a partition of A. Now we can define a mapping $g\colon A \to \mathcal{O}$ by setting $g(a) = O$ when $a \in O \in \mathcal{O}$. The mapping g is a surjection, and every $O \in \mathcal{O}$ satisfies the condition $O = g^{-1}\{\mathcal{O}\}$. Thus $\mathcal{O} = \mathcal{O}_g$. □

The mapping g defined in the proof is called the *canonical surjection* $A \to \mathcal{O}$.

Theorem 1.45. *If $\mathcal{O}$ is a partition of a set X, then the relation*

$$E_{\mathcal{O}} = \{(x, y) \in X \times X \mid \exists\ O \in \mathcal{O}\colon \quad x \in O \text{ and } y \in O\}$$

is an equivalence relation.

Proof. Let $f\colon X \to \mathcal{O}$ be the canonical projection, that is, $f(x) = O \Leftrightarrow x \in O$ where $O \in \mathcal{O}$ is arbitrary. The condition $(x, y) \in E_{\mathcal{O}}$ is equivalent to the condition $f(x) = f(y)$. Thus for all elements $x, y, z \in X$ we have:

$$\begin{aligned} &f(x) = f(x); \\ &f(x) = f(y) \;\Rightarrow\; f(y) = f(x); \\ &f(x) = f(y) \text{ and } f(y) = f(z) \quad \Rightarrow \quad f(x) = f(z). \end{aligned}$$

□

1.8 Order relations

Let A be a set.

Definition 1.46. A relation $\leq$ on the set A is called a *partial order* on A if it satisfies

(O1) $\forall a \in A\colon \quad a \leq a \quad$ *(reflexivity)*,

(O2) if $a, b \in A$, $a \leq b$ and $b \leq a$, then $a = b \quad$ *(antisymmetry)*,

(O3) if $a, b, c \in A$, $a \leq b$ and $b \leq c$, then $a \leq c \quad$ *(transitivity)*.

If in addition it satisfies

(O4) $\forall\ a, b \in A\colon \quad a \leq b$ or $b \leq a \quad$ *(strongly connected)*,

then the relation $\leq$ is called a *total order*, a *full order* or a *linear order*.

A set A equipped with such a relation $\leq$ is called a *partially ordered set* or a *totally ordered set* respectively. A totally ordered set may also be called a chain.

The condition $a \leq b$ is read as "a before b", "a precedes b" or "b follows a". Observe that Postulate (O1) says that every element is both before itself and after itself.

Example 1.47. The usual order of magnitude $x \leq y$ is a total order on $\mathbb{R}$.

Example 1.48. The division relation $a \mid b$ is a partial order on the set of positive integers $\mathbb{Z}_+$. However, it is not a total order.

Example 1.49. The inclusion relation $A \subset B$ is a partial order on the set of all subsets $\mathcal{P}(x)$ of a given set X.

Observe that if $(A, \leq)$ is a partially (or totally) ordered set, so is every subset of A under the relation $\leq$.

Chapter 2

Groups

2.1 Notion of groups

Groups are a basic concept of algebra. The concept was already used by Galois but the modern definition is the result of a long process. As an introduction to groups, let us study the properties of two sets of numbers, and encapsulate the definition of a group. Afterwards we shall introduce some examples of groups from different branches of mathematics. The following chapters continue the topic of group theory.

Let us study the set of integers $\mathbb{Z}$ and the subset of the real numbers $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$. We highlight the common properties in the following table.

CH: In the table the last row is not consistent with the other rows (math display has an empty line above). Also the row titles (e.g., associativity) could stand out a bit better, maybe in italic? NS: Better?

$\mathbb{Z}$, **operation** $+$	$\mathbb{R}^*$, **operation** $\cdot$
1. Stability of the operation:	
$b, c \in \mathbb{Z} \Rightarrow b + c \in \mathbb{Z}$.	$b, c \in \mathbb{R}^* \Rightarrow b \cdot c \in \mathbb{R}^*$.
2. Associativity:	
For all $a, b, c \in \mathbb{Z}$ we have $(a + b) + c = a + (b + c)$.	For all $a, b, c \in \mathbb{R}^*$ we have $(a \cdot b) \cdot c = a \cdot (b \cdot c)$.
3. Existence of a neutral element:	
There exists $0 \in \mathbb{Z}$ such that $0 + a = a + 0 = a \quad \forall\, a \in \mathbb{Z}$.	There exists $1 \in \mathbb{R}^*$ such that $1 \cdot a = a \cdot 1 = a \quad \forall\, a \in \mathbb{R}^*$.
4. Existence of an inverse element:	
For every $a \in \mathbb{Z}$ there is an element $-a \in \mathbb{Z}$ for which $a + (-a) = (-a) + a = 0$.	For every $a \in \mathbb{R}^*$ there is an element $a^{-1} \in \mathbb{R}^*$ for which $a \cdot a^{-1} = a^{-1} \cdot a = 1$.

The listed properties are not the only properties in common between the addition of integers and multiplication of nonzero real numbers. These ones are chosen because they appear in many other contexts as well.

Remark 2.1. The set of rational numbers $\mathbb{Q}$ has all the properties 1–4 under addition, as does the set of real numbers $\mathbb{R}$ and the set of complex numbers $\mathbb{C}$. These properties also hold when these sets are replaced with the set of polynomials in x with coefficients in one of these sets, denoted by $\mathbb{Z}[x]$, $\mathbb{Q}[x]$, $\mathbb{R}[x]$, $\mathbb{C}[x]$ respectively.

Property 4 does not hold for the set of integers under multiplication, because no integer x exists such that, say, $3 \cdot x = 1$. On the other hand, properties 1-4 hold for the sets $\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$ and $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$ under multiplication.

Let us study the previous conditions more rigorously. We have a set, $\mathbb{Z}$ or $\mathbb{R}^*$, and an operation, that is, a method to connect a pair of elements in the set to a third element in the set. If the operation on a set S is denoted by $\circ$, then we can write $s \circ s' = s''$ where $s, s', s'' \in S$. In the previous example the operation $\circ$ is $+$ for $\mathbb{Z}$ and $\cdot$ for $\mathbb{R}^*$. The second property in the table states that the operation is associative.

Note that the table does not include commutativity: $a + b = b + a$. This choice is done on purpose so as to not restrict ourselves to the so called commutative groups. Because the result of the operation can depend on the order of the pair of elements, we should consider the pair as an ordered pair. Now we conclude with a definition for the operation.

Definition 2.2. The operation $\circ$ on a set S connects a unique element $s'' \in S$ for each ordered pair s, s'. This can be denoted by

$$s \circ s' = s'' \quad \text{or} \quad (s, s') \overset{\circ}{\mapsto} s''.$$

This definition can be paraphrased as follows: An operation on a set S is a map $S \times S \to S$.

Now let us define a group.

Definition 2.3. Let G be a nonempty set. The pair $(G, \circ)$ is called a group if is satisfies the following conditions:

(G1) $\circ$ is an operation on G, that is, $a \circ b \in G \ \forall \ a, b, \in G$;

(G2) $(a \circ b) \circ c = a \circ (b \circ c)$;

(G3) There exists an element $e \in G$ (a *neutral element*) such that $e \circ a = a \circ e = a \ \forall \ a \in G$;

(G4) For every $a \in G$ there exists an element $a^{-1} \in G$ (an *inverse element*) such that $a \circ a^{-1} = a^{-1} \circ a = e$.

Definition 2.4. We call $(G, \circ)$ a commutative group or an *Abelian group* if the operation is commutative, that is, $a \circ b = b \circ a \ \forall \ a, b \in G$.

Abelian groups are named after the Norwegian mathematician Niels Henrik Abel. If $(G, \circ)$ is a group, then we may loosely say that G is a group (under operation $\circ$).

Example 2.5. $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are Abelian groups under addition. The neutral element is 0 and the inverse element of an element a is $-a$. Denote $\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$, $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$, $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. These sets are Abelian groups under multiplication. The neutral element is 1, the inverse element of a is $\frac{1}{a}$. (Why are $\mathbb{Q}, \mathbb{R}, \mathbb{C}$ not groups under multiplication? What about $\mathbb{Z}$ or $\mathbb{Z} \setminus \{0\}$?)

We do not denote $Z^* = \mathbb{Z} \setminus \{0\}$ for a good reason, as this notation S^* for some set S is commonly reserved for the so-called *unit group*. Unit groups will be discussed in Section 5.1.

Next, let us study the following five properties in common between the groups $(\mathbb{Z}, +)$ and $(\mathbb{R}^*, \cdot)$:

1. The neutral element is unique in the group.
2. Every element of the group has exactly one inverse element $\overline{a}$.
3. If a and b are elements of the group, then there exists unique elements x and y such that $a \circ x = b$ and $y \circ a = b$.
4. Rules of reduction hold:
   - If $a \circ b = a \circ c$, then $b = c$;
   - if $b \circ a = c \circ a$, then $b = c$.

5. The inverse element of $a \circ b$ is $\overline{b} \circ \overline{a}$

In the case of the group $(\mathbb{Z}, +)$, property 1 says that 0 is the only integer z such that $z + a = a + z = a$ for any $a \in \mathbb{Z}$. Property 2 says that every integer z has exactly one inverse $-z$. Property 3 says that the equations $a + x = b$ and $y + a = b$ can always be solved in $\mathbb{Z}$ no matter how a and b are chosen. Property 4 says that the equation $a + b = a + c$ or $b + a = c + a$ implies $b = c$. Property 5 says that the inverse of $a + b$ is $(-b) + (-a)$.

In case of the group $(\mathbb{R}^*, \cdot)$, property 1 states that 1 is the only nonzero real number z such that $z \cdot a = a \cdot z = a$ for any $a \in \mathbb{R}^*$. Property 2 states that every nonzero real number z has exactly one inverse $\frac{1}{z}$. Property 3 says that the equations $a \cdot x = b$ and $y \cdot a = b$ can always be solved in $\mathbb{R}^*$ no matter how a and b are chosen. Property 4 says that the equation $a \cdot b = a \cdot c$ or $b \cdot a = c \cdot a$ implies $b = c$. Property 5 says that the inverse of $a \cdot b$ is $\frac{1}{b} \cdot \frac{1}{a}$.

All of these mentioned five properties can be derived from Definition 2.3. This means that whenever some set under some operation is a group, these properties hold. Next we will show how properties 1 and 2 follow from Definition 2.3.

Theorem 2.6. *Let G be a group. The neutral element is unique. Similarly, for each $a \in G$ the inverse element a^{-1} is unique.*

Proof. Suppose e and e' are neutral elements of G. Then Postulate (G3) of Definition 2.3 gives $e' = e' \circ e = e$. Suppose a^{-1} and a' are both inverses of $a \in G$. By Postulate (G4) we have $a \circ a' = e$. Multiplying this on the left by a^{-1}, we get $a^{-1} \circ (a \circ a') = a^{-1} \circ e = a^{-1}$ from (G3), and by (G2) we also have $a^{-1} \circ (a \circ a') = (a^{-1} \circ a) \circ a' = e \circ a' = a'$ from (G3). Thus we have $a' = a^{-1}$. □

In group theory the operation on G is often written by using the common multiplication notation as a shorthand:

$$a \circ b = a \cdot b = ab.$$

Then the neutral element is also called the *identity* and denoted by $e = 1 = 1_G$.

Additive notation is sometimes used: $a \circ b = a + b$. In particular, additive notation is used when G is Abelian or the group operation is truly addition. In this case the neutral element is also called the *additive identity* and denoted by $e = 0 = 0_G$. The inverse element a^{-1} of element a is called the additive inverse $-a$.

The phrases $(G, \cdot)$ is a *multiplicative group* and $(G, +)$ is an *additive group* are also used. The number of elements in a group G is called the *order* of G, denoted by $\#G$. (The order is also sometimes denoted by $|G|$.)

Example 2.7. The vectors of an arbitrary vector space V form an additive Abelian group with the zero vector $\vec{0}$ as the additive identity and the additive inverse of a vector $\vec{X}$ being its negation $-\vec{X}$.

Examples of such groups are $\mathbb{R}^n$ and $\mathbb{C}^n$ $(n = 1, 2, \ldots)$, $F(\mathbb{R})$, the set of functions $\mathbb{R} \to \mathbb{R}$, and $\mathbb{R}[x]$, the set of polynomials with real coefficients.

Example 2.8. The group of matrices $\mathcal{M}_{m \times n}(\mathbb{R})$ is an additive Abelian group. The additive identity is the zero matrix and the inverse element of a matrix A is $-A$.

Example 2.9. The group of regular 2×2 matrices

$$GL_2(\mathbb{R}) = \{A \in \mathcal{M}_2(\mathbb{R}) \mid \det(A) \neq 0\}$$

is a multiplicative group. The identity is the identity matrix I_2 and the inverse element of matrix A is its inverse matrix A^{-1}. This group is not Abelian.

Similarly we can define $GL_n(\mathbb{R})$ for general $n = 1, 2, \ldots$. It is called the general linear group.

Example 2.10. Residue classes modulo m form an additive Abelian group $(\mathbb{Z}_m, +)$ when addition is defined as $\overline{a} + \overline{b} = \overline{a+b}$. The additive identity is $\overline{0}$ and the inverse of element $\overline{a}$ is $\overline{-a}$.

This is an example of a finite group: $\#\mathbb{Z}_m = m$.

If $\gcd(a, m) = 1$, the residue class $\overline{a}$ is called a *reduced residue class*. Note that this is well defined, since

$$\gcd(a, m) = 1, \quad \overline{a} = \overline{a'} \quad \Longrightarrow \quad \gcd(a', m) = 1.$$

If we write $a = km + r$ and $a' = k'm + r'$, then $\overline{a} = \overline{a'}$ implies $r = r'$. In addition, we know that $\gcd(km + r, m) = \gcd(r, m)$ for any integers, hence we see that

$$1 = \gcd(a, m) = \gcd(r, m) = \gcd(r', m) = \gcd(a', m).$$

The set of all reduced residue classes $\mod m$

$$\mathbb{Z}_m^* = \{\overline{a} \in \mathbb{Z}_m \mid \gcd(a, m) = 1\}$$

is a group under multiplication of residue classes defined $\overline{a} \cdot \overline{b} = \overline{a \cdot b}$. The identity is the residue class $\overline{1}$ and the inverse of element $\overline{a}$ is the element $\overline{x}$ that satisfies the congruence

$$ax \equiv 1 \pmod{m}.$$

The group $(\mathbb{Z}_m^*, \cdot)$ is called the *multiplicative residue group* $\mod m$. We denote its order by

$$\#\mathbb{Z}_m^* = \varphi(m),$$

and call this function of m *Euler's phi function*, or *Euler's totient function*. If p is a prime, then $\mathbb{Z}_p^* = \{\overline{1}, \overline{2}, \ldots, \overline{p-1}\}$, and $\varphi(p) = p - 1$. As an example for a composite number, say, $\mathbb{Z}_9^* = \{\overline{1}, \overline{2}, \overline{4}, \overline{5}, \overline{7}, \overline{8}\}$.

Example 2.11. Let $J_n = \{1, 2, \ldots, n\}$ be a set. A bijective map $\alpha\colon J_n \to J_n$ is called a permutation of the set J_n. When $\alpha(j) = k_j$ for all elements $j \in J_n$, we can write the permutation α in the form

$$\alpha = \begin{pmatrix} 1 & 2 & \ldots & n \\ k_1 & k_2 & \ldots & k_n \end{pmatrix},$$

where $k_1, k_2, \ldots, k_n$ are the numbers $1, 2, \ldots, n$ in some order.

The set of all permutations of the set $\{1, 2, \ldots, n\}$ forms a group under composition of permutations, called the *symmetric group* of n elements

$$S_n = \{\alpha\colon J_n \to J_n \mid \alpha \text{ is a bijection } \}.$$

The identity is the identity map on J_n and the inverse of α is its inverse map α^{-1}. The composition of bijections is a bijection, so (G1) is satisfied, and composition of functions is associative so (G2) is satisfied as well. If $n > 2$, then S_n is not Abelian.

Note that $\#S_n = n!$. The notation used above works well for calculating compositions of permutations. For example, (observe the order)

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

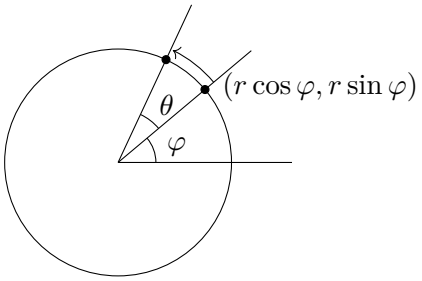
Example 2.12. Consider the linear maps $\varrho\colon \mathbb{R}^2 \to \mathbb{R}^2$ whose matrices are of the form

$$L_\theta = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}.$$

By writing the point $(x_1, x_2) \in \mathbb{R}^2$ in the polar coordinates $(r\cos\varphi, r\sin\varphi)$, we see that

$$\varrho_\theta(x_1, x_2) = (r\cos(\varphi+\theta), r\sin(\varphi+\theta))\,.$$

Hence ϱ_θ is the rotation around the origin by the angle $\theta > 0$.



In the set of maps ϱ_θ, we can define an operation via composition $\varrho_{\theta_1} \circ \varrho_{\theta_2}$. Multiplying their respective matrices, we state that the product is again a rotation with angle $\theta_1 + \theta_2$:

$$\varrho_{\theta_1} \circ \varrho_{\theta_2} = \varrho_{\theta_1+\theta_2}.$$

The matrix of the identity is

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \cos 0 & -\sin 0 \\ \sin 0 & \cos 0 \end{pmatrix},$$

so $\mathrm{id} = \varrho_0$. The formula

$$\varrho_\theta \circ \varrho_{-\theta} = \mathrm{id} = \varrho_{-\theta} \circ \varrho_\theta$$

states that the rotation by angle θ in the negative direction is the inverse element of ϱ_θ since the rotations cancel out.

Example 2.13. Consider the n-sided polygon (n-gon) centred at the origin in the xy-plane. We consider those maps of the plane that map the n-gon to itself, namely reflections and rotations. These mappings form a group under composition, which is called the *symmetry group* of the n-gon, denoted by D_n and has order $2n$. It is also called a *Dihedral group*.

For example, D_4 consists of four rotations, with angles 0, $\pi/2$, π, $3\pi/2$, and four reflections, with the diagonals and the perpendicular bisectors of the edges as reflection axes, see Section 3.6.

If the vertices of the n-gon are labelled as $1, 2, \ldots, n$, then each element of D_n can be represented by a permutation $\alpha \in S_n$.

The concept of symmetry groups can be generalised to shapes in higher dimensions. These groups are important in geometry and physics, where they are used to describe the symmetry of a shape or an object.

Remark 2.14. A pair $(G, \circ)$ that satisfies Postulates (G1) and (G2) of the definition of a group is called a *semigroup*. If a semigroup G has a neutral element in addition, that is, (G3) holds, then G is called a *monoid*.

For example, the sets $\mathbb{Z}$ and $2\mathbb{Z}$ are semigroups under multiplication. Furthermore, the former is a monoid. The theory of semigroups and monoids are not discussed on this course.

Exercises

1. Which of the following subsets of $\mathbb{Z}_{11}$ are groups under multiplication?

   a) $\{\overline{1}, \overline{3}, \overline{4}, \overline{5}, \overline{9}\}$,

   b) $\{\overline{1}, \overline{3}, \overline{4}, \overline{5}, \overline{8}\}$,

   c) $\{\overline{1}, \overline{10}\}$.

2. We define an operation on the real numbers as follows:

   $$x * y = \max\{x, y\}.$$

   Prove that this operation is associative.

3. Prove that the operation of substraction on the integers is not associative.

4. Prove that $(m, n) \mapsto m + n + 1$ is an operation on $\mathbb{N}$. Does it have a neutral element?

5. The operation on the set E has a left neutral element e, which satisfies $e * x = x \ \forall x \in E$, and a right neutral element e', which satisfies $x * e' = x \ \forall x \in E$. Prove that $e = e'$.

6. The operation on the set $E = \{0, 1\}$ is defined by the following table.

	0	1
0	0	1
1	1	0

   Prove that E is a group under this operation.

7. An operation on the set $\mathbb{R} \times \mathbb{R}$ is defined by

   $$(x, y) * (x', y') = (xx', yy').$$

   Prove that this operation is associative but not commutative.

8. Which elements of monoids $(\mathbb{N}, \cdot)$ and $(\mathbb{N}_+, \cdot)$ have inverse elements?

9. Let $(E, *)$ be a monoid. Prove that if an element $x \in E$ has a left inverse element x' (i.e., $x' * x = e$) and a right inverse element x'' (i.e., $x * x'' = e$), then $x' = x''$ (that is, x has an inverse).

10. An operation $*$ on the integers is defined as follows:

    $$x * y = x + y + 1.$$

    Prove that $(\mathbb{Z}, *)$ is a group.

11. An operation on the set $G = \{(x, y) \in \mathbb{R}^2 \mid x \neq 0\}$ is defined by

    $$(x, y) * (x', y') = (xx', yx' + y').$$

    Prove that $(G, *)$ is a group.

12. Define addition $+$ on the set $\mathbb{R}^{\mathbb{R}}$ of maps $f \colon \mathbb{R} \to \mathbb{R}$ as

    $$(f + g)(x) = f(x) + g(x) \ \ \forall x \in \mathbb{R}.$$

    Prove that $\left(\mathbb{R}^{\mathbb{R}}, +\right)$ is a group.

13. Let $(A, \leq)$ be a totally ordered set. Denote for all $x, y \in A$

$$a * b = \begin{cases} a, & \text{when } a \leq b \\ b, & \text{when } b \leq a. \end{cases}$$

Show that $*$ is associative and commutative. When is $(A, *)$ a monoid?

14. Determine the stable subsets and possible submonoids of A in the previous exercise.[1]

15. Determine whether $(\mathbb{Z}, \cdot)$ is a group.

16. Determine whether the set $\{1, -1, i, -i\}$ is a group under multiplication of complex numbers.

17. Is the set of positive rational numbers a group under multiplication of real numbers?

18. Does the matrix $\left(\begin{smallmatrix} 2 & 2 \\ 1 & 1 \end{smallmatrix}\right)$ have an inverse element when the neutral element is the identity matrix? What about when the neutral element is the zero matrix?

2.2 Basic properties

Hereon the group G will be denoted multiplicatively unless otherwise stated.

Since $a(bc) = (ab)c$, the product of three or more elements in a group can be written without parentheses, for example, abc.

The power of an element of G is defined as usual:

$$a^0 = 1, \qquad a^n = a \cdot a \cdots a \quad (n \text{ factors}), \qquad a^{-n} = (a^n)^{-1} \qquad (\forall\ n \geq 1)\,.$$

It is worth to note that $(a^{-1})^n = (a^n)^{-1}$. This follows from

$$a^n(a^{-1})^n = a \cdots a a^{-1} \cdots a^{-1} = 1, \qquad (a^{-1})^n a^n = a^{-1} \cdots a^{-1} a \cdots a = 1.$$

From the definition of exponentiation, we get the nice identities

$$(a^m)^n = a^{mn}, \quad a^m a^n = a^{m+n}, \tag{2.1}$$

where the exponents m, n are nonnegative. Moreover with a hint of work, we see that these identities hold for all exponents. For example, when $m > 0,\ n > 0$

$$(a^m)^{-n} = ((a^m)^n)^{-1} = (a^{mn})^{-1} = a^{-mn},$$

$$a^m a^{-n} = a^m \left(a^{-1}\right)^n = \underbrace{a \cdots a}_{m} \underbrace{a^{-1} \cdots a^{-1}}_{n} = a^{m-n} \quad (m > n)\,.$$

The other cases can be considered similarly.

However, the identity $(ab)^n = a^n b^n$ familiar from basic arithmetic clearly does not hold in general if G is not commutative! In addition, note the identity (compare with matrix calculations)

$$(ab)^{-1} = b^{-1} a^{-1}.$$

Remark 2.15. When using additive notation the power a^n corresponds to the multiple na.

Below is a summary of notations depending on the notation of the group operation.

[1] These concepts have not been defined but one can find their definitions easily online.

$(G,\cdot)$		$(G,+)$	
$a \cdot b$ or ab	product	$a+b$	sum
e or 1	neutral element or identity	0	neutral element or additive identity
a^{-1}	inverse element	$-a$	additive inverse
a^n	power of a	na	multiple of a
ab^{-1}	quotient	$a-b$	difference

Theorem 2.16. *Let G be a group and $a, b \in G$. The equations*

$$ax = b, \qquad ya = b$$

have unique solutions in G, namely $x = a^{-1}b$ and $y = ba^{-1}$.

Proof. Multiplying the equation $ax = b$ on the left by a^{-1}, we get $x = a^{-1}b$. Vice versa, $x = a^{-1}b$ satisfies the equation since $a\,(a^{-1}b)\,b = b$. Similarly, multiplying the equation $ya = b$ on the right by a^{-1}, we get $y = ba^{-1}$. Vice versa, $y = ba^{-1}$ satisfies the equation since $(ba^{-1})\,a = b$. □

We can deduce from Theorem 2.16 or directly that the rules of reduction hold in a group:

$$ac = bc \implies a = b; \qquad ca = cb \implies a = b.$$

A finite group can be described by writing its Cayley table, which is also called the group table. The columns and rows of the table are labelled with the group elements and the element ab is written in the cell of row a and column b. It follows from 2.16 that every column and every row contains each group element exactly once.

Example 2.17. In the case of a group with three elements, 1, a, b, we obtain only a single possible Cayley table when taking the prior condition into account. Postulates (G1), (G3) and (G4) are clearly satisfied. Checking (G2) requires some effort. However, this effort can be skipped by finding a group with this Cayley table. For example, one such group is the additive group $\mathbb{Z}_3 = \{\overline{0}, \overline{1}, \overline{2}\}$.

·	1	a	b
1	1	a	b
a	a	b	1
b	b	1	a

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

Result: There is, essentially, one single group with three elements. This group can be presented in the form $G = \{1, a, a^2\}$, where $a^3 = 1$. Groups like this are called *cyclic* and will be discussed in Section 3.3.

Example 2.18. There are essentially two groups with four elements, $1, a, b, c$, defined by whether $a^2 = 1$ or $a^2 = b$ holds. Their Cayley tables are the following:

·	1	a	b	c
1	1	a	b	c
a	a	b	c	1
b	b	c	1	a
c	c	1	a	b

cyclic group

·	1	a	b	c
1	1	a	b	c
a	a	1	c	b
b	b	c	1	a
c	c	b	a	1

Klein four-group

Definition 2.19. Let G be a group. If $H \subset G$ and H is a group under the operation on G, we call H a *subgroup* of G, and denote $H \leqslant G$.

Note that $1_H = 1_G$. This can be seen by multiplying the equation $1_H 1_H = 1_H$ viewed in G by 1_H^{-1}. Consequently, the inverse of every element a in H is the inverse of a in G.

Example 2.20. The trivial subgroups of a group G are $\{1\}$ and G.

If $H \leqslant G$ and $H \neq G$, we say that H is a *proper* subgroup of G, and denote $H < G$.

Example 2.21. $\mathbb{Z} < \mathbb{Q} < \mathbb{R} < \mathbb{C}$ as additive groups, and $\mathbb{Q}^* < \mathbb{R}^* < \mathbb{C}^*$ as multiplicative groups.

The next theorem provides a short method for testing whether a subset H of group G is a subgroup of G.

Theorem 2.22 (Subgroup criterion)**.** *Let G be a group and $H \subset G$. Then H is a subgroup of G if and only if $H \neq \emptyset$ and*

$$ab^{-1} \in H \quad \forall\ a, b \in H.$$

Proof. Suppose $a \in H$. Now $1 = aa^{-1} \in H$ and further $a^{-1} = 1 \cdot a^{-1} \in H$. From these we see that H satisfies Postulates (G3) and (G4). Because the operation on G is associative, so is the operation on H, and (G2) is satisfied as well.

We shall show that (G1) holds. If $a, b \in H$, then according to what we stated above $b^{-1} \in H$, and by assumption $ab = a\left(b^{-1}\right)^{-1} \in H$. Thus H is a group, thus $H \leqslant G$.

Suppose H is a subgroup of G. By (G3) we know that H is nonempty. By (G4) each element b in H has an inverse element $b^{-1} \in H$. Finally (G1) states that $ab^{-1} \in H$ for any $a, b^{-1} \in H$. □

The subset H is usually given as the set of elements in G with some property P. The use of Theorem 2.22 has two steps, which you can compare with the proof by induction:

1. Show that some elements of G have the property P. The neutral element e of G is often suitable. Thus state $H \neq \emptyset$.

2. Assume that elements a and b have this property P, and show that the element ab^{-1} also has the property P.

Example 2.23. Let G be an Abelian group with neutral element e. We shall show that the subset $H = \{x \in G \mid x^2 = e\}$ is a subgroup of G.

Now, the defining property P of the subset H is the condition $x^2 = e$. First note that $e^2 = e$, so $H \neq \emptyset$. Then we assume that $a, b \in H$. This means that $a^2 = e$ and $b^2 = e$. We need to show that $\left(ab^{-1}\right)^2 = e$. Since G is Abelian, we have

$$\left(ab^{-1}\right)^2 = ab^{-1}ab^{-1} = a^2\left(b^{-1}\right)^2 = a^2\left(b^2\right)^{-1} = ee^{-1} = e.$$

Thus $ab^{-1} \in H$ and by the Subgroup criterion, H is a subgroup of G.

Corollary 2.24. *If $H \leqslant G$ and $K \leqslant G$, then $H \cap K \leqslant G$.*

Proof. Because $1 \in H \cap K$, the intersection is not empty. If $a, b \in H \cap K$, then $a, b \in H$ and $a, b \in K$, so H as a group contains the element ab^{-1} as does K. Thus $ab^{-1} \in H \cap K$. □

This result can be generalised to intersections of arbitrary collections of subgroups. From Theorem 2.22 we can easily form the next, slightly longer test for a subgroup.

Theorem 2.25. *Let G be a group and H some nonempty subset of G. Then H is a subgroup of G if*

$$ab \in H \quad \forall\ a, b \in H \quad \textit{and}$$
$$a^{-1} \in H \quad \forall\ a \in H.$$

Proof. By the Subgroup criterion it is sufficient to show that $a, b \in H$ implies $ab^{-1} \in H$. Let $a, b \in H$. Then $b^{-1} \in H$ by the latter condition and $ab^{-1} \in H$ by the former condition. □

If the subset is a finite subset of the group G, then it can be handled in a slightly simpler way.

Theorem 2.26. *Let G be a group and H a finite subset of G. Then H is a subgroup of G if $H \neq \emptyset$ and*

$$ab \in H \quad \forall\ a, b \in H.$$

Proof. Suppose that $a, b \in H$. We shall show that $ab^{-1} \in H$, and the statement will follow from the Subgroup criterion. By assumption $b^k \in H$ for any $k \geq 1$, and $ab^k \in H$ whenever $k \geq 0$. Because H is finite, we must have $ab^k = ab^j$ for some exponents $k, j \geq 0,\ k \neq j$. We can assume that $k < j$. Now $j - k - 1 \geq 0$ and we see that

$$ab^{-1} = ab^{j-k-1} \in H.$$

□

Example 2.27. We will show that the following subsets are subgroups:

a) $G = GL_n(\mathbb{R})$, $H = \{A \in GL_n(\mathbb{R}) \mid \det(A) = 1\}$:

H clearly contains the identity matrix so $H \neq \emptyset$. Suppose $A, B \in H$. Then we have

$$\det(AB^{-1}) = \det(A)\det(B^{-1}) = \det(A)\det(B)^{-1} = 1 \cdot \frac{1}{1} = 1.$$

Hence $AB^{-1} \in H$, and H is a subgroup of G by the Subgroup criterion.

b) $G = D_n$, $H = \{r \in D_n \mid r \text{ is a rotation}\}$:

The identity, the rotation by 0 degrees, belongs to H so $H \neq \emptyset$. Suppose $\varrho_{\theta_1}, \varrho_{\theta_2} \in H$, that is, $\theta_1 = k\frac{360^\circ}{n}$ and $\theta_2 = l\frac{360^\circ}{n}$ with $k, l \in \mathbb{Z}$. Then

$$\varrho_{\theta_1}\varrho_{\theta_2}^{-1} = \varrho_{\theta_1 - \theta_2} \in H$$

because $\theta_1 - \theta_2 = \frac{360^\circ}{n}(k - l)$ is another integer multiple of $\frac{360^\circ}{n}$, which is a rotation of the n-gon. Hence H is a subgroup of G by the Subgroup criterion.

c) $G = (\mathbb{Z}, +)$, $H = m\mathbb{Z} = \{mk \mid k \in \mathbb{Z}\}$:

Firstly, H is nonempty because $0 = m \cdot 0 \in m\mathbb{Z}$. Suppose that $a = mk$, $b = ml \in H$. Then $a - b = mk - ml = m(k - l) \in m\mathbb{Z} = H$. Thus by the Subgroup criterion H is a subgroup of G.

Theorem 2.28. *If G_1 and G_2 are groups, then their Cartesian product $G_1 \times G_2$ is a group under the following operation:*

$$(a_1, a_2)(b_1, b_2) = (a_1 b_1, a_2 b_2), \quad a_i, b_i \in G_i. \tag{2.2}$$

This group is called the direct product of the groups G_1 and G_2. If we use the additive notation, it is called the direct sum.

Proof. Postulates (G1), (G2), (G3) and (G4) can be checked easily.

(G1): Suppose $a_1, b_1 \in G_1$ and $a_2, b_2 \in G_2$. Then since G_1 and G_2 are groups, $a_1 b_1 \in G_1$ and $a_2 b_2 \in G_2$. Thus

$$(a_1, a_2)(b_1, b_2) = (a_1 b_1, a_2 b_2) \in G_1 \times G_2.$$

(G2): Since the operations of G_1 and G_2 are associative, $(a_1b_1)c_1 = a_1(b_1c_1)$ and $(a_2b_2)c_2 = a_2(b_2c_2)$ for $c_1 \in G_1$ and $c_2 \in G_2$. Thus

$$[(a_1, a_2)(b_1, b_2)](c_1, c_2) = (a_1, a_2)[(b_1, b_2)(c_1, c_2)].$$

For (G3) we note that the identity is $(1, 1)$ and for (G4) we note that $(a_1, a_2)^{-1} = \left(a_1^{-1}, a_2^{-1}\right)$. □

Example 2.29. If $G_1 = G_2 = \mathbb{R}$ under addition, then Theorem 2.28 gives the familiar group $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$, where $(u_1, u_2) + (v_1, v_2) = (u_1 + v_1, u_2 + v_2)$.

Note that in the formula (2.2), the operation a_1b_1 is computed in G_1 and a_2b_2 is computed in G_2. The direct product and sum generalise in a similar way for groups $G_1, G_2, \ldots, G_n$, $n > 2$.

Exercises

1. Is $(\mathbb{Z}_n, +)$ a subgroup of $(\mathbb{Z}, +)$?
2. Solve the pair of equations in $\mathbb{Z}_8$
$$\begin{cases} 1_8x^2 + 4_8y^2 = 0_8 \\ 1_8x + 2_8y = 4_8. \end{cases}$$
3. Consider the set $\mathbb{R}$ under addition. Show that the subset of rational numbers is stable, that is, closed under addition. Is the subset of irrational numbers stable?
4. Show that in the group $(G, \circ)$, the equation
$$a \circ x \circ b \circ c \circ x = a \circ b \circ x$$
has a unique solution and find it.
5. In the three element symmetric group S_3 solve equation
$$\begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \circ S = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}.$$
6. Show that $(x^m)^n = x^{mn}$ for all $m, n \in \mathbb{Z}$ where x is an element of a group G.
7. Suppose that a group $(G, +)$ has property $2(x + y) = 2x + 2y$. Show that G is Abelian.
8. Suppose $(G, +)$ is Abelian and H is the set of elements in G that satisfy $4x = x$. Show that $(H, +)$ is a subgroup of $(G, +)$.
9. Suppose that the group $G = \{0, 1, 2, 3\}$ whose group operation uses addition notation has the Cayley table

	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

Determine all subgroups of $(G, +)$.

10. Let $(G, \circ)$ be a group and

$$H = \{x \in G \mid x \circ a = a \circ x \ \forall a \in G\}.$$

Show that $(H, \circ)$ is a commutative subgroup of $(G, \circ)$.

11. Suppose that H and K are subgroups of a group G. Show that $H \cup K$ is a subgroup of G if and only if $H \subset K$ or $K \subset H$.

12. Let a and b be elements of a group $(G, \cdot)$ satisfying

$$b^6 = e \text{ and } ab = b^4 a.$$

Show that $b^3 = e$ and $ab = ba$.

13. Let $(A, +)$ be an Abelian group and B its subgroup. Show that the set

$$H = \{a \in A \mid na \in B \text{ for some } n \in \mathbb{N} \setminus \{0\}\}$$

is a subgroup of A that contains B.

14. Suppose $(E, \circ)$ is a monoid and that its element a has an inverse in E. Show that the maps

$$f \colon E \to E, \ x \mapsto a \circ x;$$
$$g \colon E \to E \, x \mapsto x \circ a,$$

are bijections.

15. The operation $\circ$ on the set $\mathbb{Z}^* \times \mathbb{Z}$, defined by

$$(i, a) \circ (j, b) = (ij, a + ib), \quad i, j \in \mathbb{Z}^* = \{1, -1\}, \ a, b \in \mathbb{Z},$$

is known to be associative. Show that $(\mathbb{Z}^* \times \mathbb{Z}, \circ)$ is a noncommutative group and compute the inverse of the element $(-1, 5)$.

16. Let $(A, \circ)$ be a finite monoid and let $a, b \in A$ be such that $a \circ b = e_A$. Show that a and b are the inverse elements of each other. (Hint: Show that the map $f \colon A \to A, \ f(x) = b \circ x$, is an injection.)

17. Solve the following equation within the symmetric group S_6:

$$f \circ \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 1 & 6 & 5 & 4 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 3 & 6 & 4 & 1 & 5 \end{pmatrix}.$$

18. Prove that $(\mathbb{Z}, +)$ is an Abelian group by starting from the definition $\mathbb{Z} = \mathbb{N} \times \mathbb{N} / \sim$ and by using analogious properties of $(\mathbb{N}, 0)$.

19. Prove that $\{5, 15, 25, 35\}$ is a group under multiplication modulo 40. What is the neutral element?

2.3 Generating groups; cyclic groups

Let G be a group and S some subset of G. Consider the collection of subgroups of G that contain S. This collection is nonempty because at least G belongs to it. The intersection of subgroups H of G is a subgroup of G, which is proved similarly to Corollary 2.24. We call this subgroup the *subgroup of G generated by S* and denote it by $\langle S\rangle$; that is,

$$\langle S\rangle = \bigcap_{S\subset H\leqslant G} H.$$

The elements of the set S are called the *generators* of the group $\langle S\rangle$.

If there are finitely many generators, that is, $S = \{a_1, \ldots, a_k\}$ for some finite k, we say that the group $\langle S\rangle$ is *finitely generated*, and denote

$$\langle S\rangle = \langle a_1, \ldots, a_k\rangle.$$

By definition $\langle S\rangle$ is the smallest subgroup of G that contains the set S. Any other subset of G that contains S also contains $\langle S\rangle$.

Example 2.30. $\langle\emptyset\rangle = \{1\}$, $\quad\langle 1\rangle = \{1\}$. If $H \leqslant G$, then $\langle H\rangle = H$.

Theorem 2.31. *The subgroup of a group G generated by its subset S consists of all products whose factors are elements of S and the inverses of the elements of S including the empty product 1, that is,*

$$\langle S\rangle = \{a_1a_2\cdots a_m \mid a_i \text{ or } a_i^{-1} \in S \quad \forall\ i,\ m \geq 0\}.$$

Proof. Denote the right-hand side by H. From the Subgroup criterion we see that $H \leqslant G$. Moreover, $S \subset H$. Hence $\langle S\rangle \subset H$ because $\langle S\rangle$ is by definition the smallest subgroup that contains S.

Conversely, every subgroup of G that contains S also contains all the previously stated products, that is, it contains H. Thus H is contained in the intersection of all such subgroups, in other words, $H \subset \langle S\rangle$. These together give $\langle S\rangle = H$ □

Remark 2.32. If G is a finite group, then Theorem 2.31 can be simplified to

$$\langle S\rangle = \{a_1a_2\cdots a_m \mid a_i \in S \quad \forall\ i;\ m \geq 0\}$$

as we can use Theorem 2.26.

Example 2.33. The subset of primes $\mathbb{P}$ of the group $\mathbb{R}^*$ generates the subgroup $\mathbb{Q}_+$, which consists of all positive rational numbers under multiplication.

Example 2.34. If V is a vector space of dimension n, then its basis $\{B_1, \ldots, B_n\}$ generates a subgroup $\{k_1B_1 + \ldots + k_nB_n \mid k_i \in \mathbb{Z} \quad \forall\ i\}$ of the group $(V, +)$. (Why is it not a subspace of V?)

Example 2.35. The dihedral group $D_4 = \langle r, s\rangle$, where r is a rotation by $(\pi/2)$ and s is a suitable reflection.

Example 2.36. The infinite group $\mathbb{Z}$ is finitely generated: $\mathbb{Z} = \{n \cdot 1 \mid n \in \mathbb{Z}\} = \langle 1\rangle = \langle -1\rangle$.

Definition 2.37. A group G is called *cyclic* if G is generated by one element, that is, if there is an element $a \in G$ such that $G = \langle a\rangle$.

Theorem 2.38. *Let $G = \langle a\rangle$ be a cyclic group. If $\#G = n \in \mathbb{N}$, then G is of the form*

$$G = \{1, a, a^2, \ldots a^{n-1}\},$$

and n is the smallest positive integer such that $a^n = 1$. If $\#G = \infty$, then

$$G = \{\ldots, a^{-2}, a^{-1}, 1, a, a^2, \ldots\}$$

and all powers a^m $(m \in \mathbb{Z})$ are distinct, in particular, $a^m \neq 1 \quad \forall\ m \neq 0$.

Proof. By Theorem 2.31

$$G = \{a^m \mid m \in \mathbb{Z}\}. \tag{2.3}$$

First assume that $\#G = n$. Then among the powers a^m there are only m distinct ones, which means that there exist exponents m and k, $m > k$ with $a^m = a^k$. Now $a^{m-k} = 1$ and $m-k > 0$. We choose the smallest positive exponent such that $a^s = 1$ and $s \leq m - k$.

By the division algorithm, every exponent $m \geq 0$ can be written as $m = qs + r$ with $0 \leq r \leq s - 1$. Because

$$a^m = a^{qs+r} = (a^s)^q\, a^r = 1^q a^r = a^r,$$

all elements of G are of the form a^r, $r = 0, \ldots, s - 1$. Moreover, these elements are distinct; otherwise, a similar inference would result in $a^t = 1$ with $0 < t < s$, contradicting the minimality of s. In particular, $\#G = s$, so $s = n$. Moreover, we see that G is exactly as in the claim.

If G is infinite, then all the powers in Equation (2.3) are distinct, since otherwise we would result in the same finite group as above. □

Example 2.39. $\mathbb{Z}_m$ is an additive cyclic group of order m since for all $m \geq 1$:

$$\mathbb{Z}_m = \langle \overline{1} \rangle = \{\overline{0}, \overline{1}, 2 \cdot \overline{1}, \ldots, (m-1) \cdot \overline{1}\}.$$

$\mathbb{Z}$ is an infinite cyclic group with generators 1 and -1.

Example 2.40. The group $\mathbb{Z}_5^*$ is a cyclic group generated by $\overline{2}$ since $\overline{4} = \overline{2}^2$, $\overline{3} = \overline{2}^3$ and $\overline{1} = \overline{2}^4$.

Why is it called a cyclic group? If $G = \langle a \rangle$ has order n, then in the infinite sequence $\ldots, a^{m-1}, a^m, a^{m+1}, \ldots$ any n consecutive elements form a cycle, which repeats when moving along the sequence. This can also be expressed as follows:

$$a^k = a^h \qquad \Longleftrightarrow \qquad k \equiv h \pmod{n}.$$

An infinite cyclic group is a special case in the previous sense, where there is only a single infinitely long cycle.

Definition 2.41. Let G be a group and $a \in G$. The order of the subgroup $\langle a \rangle$ of G is called the *order* of the element a and denoted by $\operatorname{ord}(a)$, that is,

$$\operatorname{ord}(a) = \#\langle a \rangle.$$

It follows immediately from Theorem 2.38 that a is of (finite) order n if and only if n is the smallest positive exponent such that $a^n = 1$. Moreover, all the distinct powers of a are $1, a, a^2, \ldots, a^{n-1}$. Note that $\operatorname{ord}(a) = 1$ if and only if $a = 1$.

Example 2.42. (i) In the group $\mathbb{R}^*$, $\operatorname{ord}(1) = 1$, $\operatorname{ord}(-1) = 2$ and the order of any other element is infinite.

(ii) In the group $\mathbb{Z}_{21}^*$, we have $\operatorname{ord}(\overline{2}) = 6$ and $\operatorname{ord}(\overline{20}) = 2$.

Later we will present how we can simplify the computation of the order of an element.

Exercises

1. Consider the group $(\mathbb{Z}_{10}, +)$. What is $\langle \overline{10} \rangle$?
2. Is $\langle \overline{-1} \rangle = \mathbb{Z}_m$?

3. What is the order of the group $\mathbb{Z}_{12}$? Compute the orders of all its elements.

4. Find all possible generators of $(\mathbb{Z}_8, +)$.

5. Determine the subgroup of $\mathbb{R}$ generated by its subset $\{6, 15, 21\}$.

6. Let $G = \langle a, b\rangle$ with $\text{ord}(a) = \text{ord}(b) = 2$ and $ab = ba$. Show that G is the Klein four-group.

7. Prove that $\langle a\rangle = \langle a^{-1}\rangle$ no matter what group a belongs to.

8. What is the order of the matrix $A = \left(\begin{smallmatrix} 1 & 1 \\ 0 & 1 \end{smallmatrix}\right)$ in $GL_2(\mathbb{R})$?

9. Let x be an element of the group $(G, \cdot)$. Suppose that $x^2 \neq 1$ and $x^6 = 1$. Prove that $x^4 \neq 1$ and $x^5 \neq 1$. What can we say about the order of x?

2.4 Group homomorphisms and isomorphisms

In Section 2.2 we stated that two groups are essentially the same if their Cayley tables differ only by how the elements are labelled. In this chapter, we shall explore this "equality" comparison in more detail, starting with the following definition.

Definition 2.43. Let $(G, \cdot)$ and $(G', *)$ be two groups. Then we say that a map $f : G \to G'$ is a (group) homomorphism if it satisfies the *Homomorphism criterion*

$$f(ab) = f(a) * f(b) \quad \forall\ a, b \in G. \tag{2.4}$$

If multiplication notation is used for both group operations, then the criterion is written as $f(ab) = f(a)f(b) \quad \forall\ a, b \in G$.

Example 2.44. Suppose U and V are vector spaces. Then every linear map $t : U \to V$ is a homomorphism between additive groups U and V because

$$t(X_1 + X_2) = t(X_1) + t(X_2) \quad \forall\ X_1, X_2 \in U.$$

Example 2.45. (i) The map $f : \mathbb{R}^* \to \mathbb{R}^*$ defined by $f(x) = x^2$ is a homomorphism because

$$f(xy) = (xy)^2 = x^2y^2 = f(x)f(y) \quad \forall\ x, y \in \mathbb{R}^*.$$

(ii) Let us denote by $\mathbb{R}_+$ the multiplicative group formed by the positive real numbers. The map $f : \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x) = \ln(x)$ is a homomorphism because

$$f(xy) = \ln(xy) = \ln(x) + \ln(y) = f(x) + f(y) \quad \forall\ x, y \in \mathbb{R}_+.$$

(iii) The map $f : \mathbb{Z} \to \mathbb{Z}_m$ defined by $f(a) = \overline{a}$ is a homomorphism because $\overline{a + b} = \overline{a} + \overline{b}$.

(iv) The map $f : G \to G'$ defined by $f(a) = 1_{G'} \quad \forall\ a \in G$ is called the trivial homomorphism. It is clearly a homomorphism as

$$f(a)f(b) = 1_{G'} = f(ab) \quad \forall\ a, b \in G.$$

A group homomorphism $f : G \to G'$ preserves the identity element and the relationship between inverse elements. That is,

$$f(1_G) = 1_{G'}, \qquad f(a^{-1}) = f(a)^{-1} \quad \forall\, a \in G.$$

The first equation can be obtained directly by multiplying equation

$$f(1_G)f(1_G) = f(1_G \cdot 1_G)$$

by $f(1_G)^{-1}$ on both sides. Note that often the identity elements of *both* groups are denoted by 1. We will mostly use this shorthand notation. The second equation can be obtained from the equations

$$f(a)f(a^{-1}) = f(aa^{-1}) = f(1) = 1, \qquad f(a^{-1})f(a) = 1.$$

Applied to a linear map t, the previous result expresses a familiar fact that t preserves the identity element and the inverse elements, which follows directly from the definition of t in the theory of linear maps.

Definition 2.46. Let $(M, \cdot)$ and $(M', *)$ be two monoids. We say that a map $f : M \to M'$ is a (monoid) homomorphism if it satisfies the Homomorphism criterion (2.4) and $f(1_M) = 1_{M'}$ holds.

The following theorem has an analogy in the theory of linear maps.

Theorem 2.47. *Let $f : G \to G'$ be a group homomorphism.*

(i) *If $H \leqslant G$, then $f(H) \leqslant G'$.*

(ii) *If $H' \leqslant G'$, then $f^{-1}(H') \leqslant G$.*

Proof. (i) The set $f(H)$ is nonempty as $H \neq \emptyset$. Suppose $a', b' \in f(H)$, that is $a' = f(a)$ and $b' = f(b)$ where $a, b \in H$. Then

$$a'b'^{-1} = f(a)f(b)^{-1} = f(ab^{-1}) \in f(H).$$

Now the statement follows from the Subgroup criterion because H is a subgroup so $ab^{-1} \in H$.

(ii) Because H' contains the identity $1_{G'}$ of G' whose preimage is the identity 1_G of G, we have $1_G \in f^{-1}(H')$. In particular, $f^{-1}(H') \neq \emptyset$. The implication $a, b \in f^{-1}(H') \Rightarrow ab^{-1} \in f^{-1}(H')$ is proved once again by using the Homomorphism criterion (try it!).

□

Similarly as with linear maps, note the important special cases $H' = \{1\}$ and $H = G$ of the theorem. The *kernel* of a homomorphism $f : G \to G'$ is

$$\mathrm{Ker}(f) = f^{-1}(\{1\}) = \{a \in G \mid f(a) = 1\},$$

and its *image* is

$$\mathrm{Im}(f) = f(G) = \{f(a) \mid a \in G\}.$$

Example 2.48. Let us determine the kernels and images of the homomorphisms from Example 2.45.

(i) For the map $f : R^* \to R^*$ defined by $f(x) = x^2$, we have

$$\mathrm{Ker}(f) = \{1, -1\} \quad \text{and} \quad \mathrm{Im}(f) = \mathbb{R}^*_+.$$

(ii) For the map $f : \mathbb{R}_+ \to \mathbb{R}$ defined by $f(x) = \ln(x)$, we have

$$\mathrm{Ker}(f) = \{1\} \quad \text{and} \quad \mathrm{Im}(f) = \mathbb{R}.$$

(iii) For the map $f : \mathbb{Z} \to \mathbb{Z}_m$ defined by $f(a) = \overline{a}$, we have

$$\mathrm{Ker}(f) = m\mathbb{Z} \quad \text{and} \quad \mathrm{Im}(f) = \mathbb{Z}_m.$$

(iv) For the trivial map $f : G \to H$ defined by $f(a) = 1_H \quad \forall\, a \in H$, we have

$$\mathrm{Ker}(f) = G \quad \text{and} \quad \mathrm{Im}(f) = \{1\}.$$

The group $\mathrm{Im}(f)$ is called the *homomorphic image* of the group G. This group retains some of the properties of G; however, it loses more properties of G the more elements f maps to the same element of G'. We will handle this rigorously later. For now, we will only consider the case where f is the best possible in this regard.

Definition 2.49. A group homomorphism $f\colon G \to G'$ is called a *(group) isomorphism* if f is bijective. We say that the group G is *isomorphic* to group G' if there exists some isomorphism $f\colon G \to G'$, and denote $G \simeq G'$.

Example 2.50. The map $f\colon \mathbb{R}_+ \to \mathbb{R},\ f(x) = \ln(x)$ from Example 2.45 (ii) is an isomorphism, so $(\mathbb{R}_+, \cdot) \simeq (\mathbb{R}, +)$.

If a homomorphism $f\colon G \to G'$ is injective, then the map $f\colon G \to \mathrm{Im}(f)$ is a bijective homomorphism, that is, an isomorphism. In other words, the homomorphic image $\mathrm{Im}(f)$ of a group G of an injective map f is isomorphic to G. When studying the injectivity of a homomorphism, the following theorem is often useful.

Theorem 2.51. *A group homomorphism $f\colon G \to G'$ is an injection if and only if*

$$\mathrm{Ker}(f) = \{1_G\}.$$

Proof. a) Suppose that f is injective. Because $f(1_G) = 1_{G'}$, we have $\mathrm{Ker}(f) = \{1_G\}$.

b) Suppose now that $\mathrm{Ker}(f) = \{1_G\}$. Let $x, y \in G$ and $f(x) = f(y)$. Now

$$1_{G'} = f(x)f(y)^{-1} = f\left(xy^{-1}\right)$$

and thus $xy^{-1} \in \mathrm{Ker}(f)$. Hence by assumption $x = y$. Thus f is an injection. □

Example 2.52. Let G be a group and $u \in G$. The map

$$f_u\colon G \to G, \quad f_u(a) = uau^{-1},$$

is an isomorphism. Let $a, b \in G$. Firstly, f_u is a homomorphism since

$$f_u(ab) = uabu^{-1} = uau^{-1}ubu^{-1} = f_u(a)f_u(b).$$

Injectivity is easy to check as well. Suppose that $f(a) = f(b)$, that is,

$$uau^{-1} = ubu^{-1}.$$

Multiplying this by u on the right and by u^{-1} on the left, we get $a = b$. Thus f_u is an injection.

Now since f_u is injective and maps $\#G$ elements to $\#G$ elements, f_u must be surjective. As an injection and a surjection, f_u is a bijection, and hence, an isomorphism. An isomorphism from G to itself is called an *automorphism* of G.

Theorem 2.53. *Let $f\colon G \to G'$ and $g\colon G' \to G''$ be group homomorphisms. Then*

(i) *the map $f \circ g$ is a homomorphism;*

(ii) *if f and g are isomorphism, then so is $f \circ g$.*

Proof. (i) $g(f(ab)) = g(f(a)f(b)) = g(f(a))g(f(b))$.

(ii) The bijectivity of map $f \circ g$ can be determined straightforwardly. Thereafter the statement follows from part (i).

□

Theorem 2.54. *Let $f\colon G \to G'$ be a group isomorphism. Then $f^{-1}\colon G' \to G$ is an isomorphism.*

Proof. The map f^{-1} is clearly bijective; we just need to show that it is a homomorphism. Let $a', b' \in G'$, with $a' = f(a),\ b' = f(b)$ and $a, b, \in G$. Now $a'b' = f(a)f(b) = f(ab)$, so

$$f^{-1}\left(a'b'\right) = ab = f^{-1}\left(a'\right) f^{-1}\left(b'\right).$$

□

Remark 2.55. In particular, as a consequence of the theorem we get that the inverse map of a homomorphism, if it exists, is a homomorphism.

The isomorphy of groups is an equivalency relation on any collection of groups. Symmetry can be seen from Theorem 2.54, and transitivity can be seen from Theorem 2.53. To see reflexivity, just note that id_G is an isomorphism $G \to G$.

Isomorphic groups G and G' are structurally the same in the view of group theory: their elements correspond to each other bijectively, and the product of some elements in G corresponds to the product of their images in G'. Particularly, if G and G' are finite, then we can obtain the Cayley table of G' by replacing each element in the Cayley table of G with its image in G'.

Example 2.56. In Examples 2.17 and 2.18 of Section 2.2, we showed that there is exactly one group with three elements and two groups with four elements up to isomorphism. The only group with one element is $\{e\}$, where e is the neutral element. There is only a single group of order two up to isomorphism. This can be seen by writing its Cayley table with the neutral element e and the other element a.

	e	a
e	e	a
a	a	e

Example 2.57. Let us show that the groups $(\mathbb{R}, +)$ and $(\mathbb{R}^*, \cdot)$ are not isomorphic. First, recall that isomorphisms preserve group structure. Suppose $f\colon \mathbb{R}^* \to \mathbb{R}$ is any isomorphism. As a homomorphism f maps the neutral element of $\mathbb{R}^*$ to the neutral element of $\mathbb{R}$, $f(1) = 0$. The order of -1 in $\mathbb{R}^*$ is 2 since $(-1)(-1) = 1$, and

$$f(-1) + f(-1) = f((-1)(-1)) = f(1) = 0.$$

However, in $(\mathbb{R}, +)$ the only solution to the equation $a + a = 0$ is $a = 0$. This yields a contradiction to the bijectivity of f.

One of the basic problems in group theory is the classification of nonisomorphic groups, see Section 3.5

Exercises

1. Determine whether $(\mathbb{Z}, +)$ and $(2\mathbb{Z}, +)$ are isomorphic.
2. Is the map $f\colon (\mathbb{R}_+, \cdot) \to (\mathbb{R} \setminus \{-1\}, \circ)$ a homomorphism when $\cdot$ is ordinary multiplication and $\circ$ is defined by $x \circ y = xy + x + y$?
3. Let E be a set and F its subset. Define an operation $(A, B) \mapsto A \cap B$ on the power set $\mathcal{P}(E)$. Show that the map $f\colon \mathcal{P}(E) \to \mathcal{P}(E), f(A) = A \cup F$, satisfies the Homomorphism criterion and maps neutral element to neutral element ,i.e., that it is a *monoid homomorphism*.
4. Let X, Y be sets and $f\colon X \to Y$ a map. Show that the map

   $$\theta\colon \mathcal{P}(Y) \to \mathcal{P}(X); \qquad \theta(B) = f^{-1}(B);$$

   is a monoid homomorphism $(\mathcal{P}(Y), \cup) \to (\mathcal{P}(x), \cup)$.
5. Let $f\colon G \to H$ be a group homomorphism. Show that $f(x^n) = f(x)^n \ \ \forall x \in G, \ n \in \mathbb{Z}$.
6. Determine the cyclic subgroups of the group $(\mathbb{Z}^* \times \mathbb{Z}^*)$ from Exercise 15. of Section 2.2, and then show that this group is noncyclic.
7. Let G and H be groups. Let A be the subset generated by the set $S \subset G$ and $f\colon G \to H$ a homomorphism. Show that the subgroup $f(A)$ is generated by the set $f(S)$.
8. Let $G = \{a + b\sqrt{2} \mid a, b, \in \mathbb{Q}\}$ and $H = \{\left(\begin{smallmatrix} a & 2b \\ b & a \end{smallmatrix}\right) \mid a, b \in \mathbb{Q}\}$. Are G and H isomorphic under addition?

2.5 Lagrange's theorem

Definition 2.58. Let $H \leqslant G$. Then for each element $a \in G$, the *left coset* of a with respect to the subgroup H is the set

$$aH = \{ah \mid h \in H\}.$$

We define the *right cosets* Ha respectively. When using additive notation we denote the cosets $a + H, \ H + a$.

In the following, we mainly consider left cosets; right cosets behave in the same way. If G is an Abelian group, then $aH = Ha$ for all $a \in G$ and we can exclude the attributes left and right.

The condition

$$b \sim a \qquad \Longleftrightarrow \qquad b \in aH$$

defines an equivalence relation on G (check!). Its equivalence classes are of the form

$$[a] = \{b \in G \mid b \in aH\} = aH.$$

They are exactly the left cosets with respect to H. From this we see that the left cosets with respect to the subgroup H in G form a partition of G

$$G = \bigcup_{a \in D} aH,$$

where a goes through some set D of representative left cosets.

The subgroup H itself is a coset: for instance $H = 1 \cdot H = H \cdot 1$. The condition $b \in aH$ is equivalent to the condition $a^{-1}b \in H$. When dealing with cosets, it is helpful to keep in mind the rule

$$aH = bH \qquad \Longleftrightarrow \qquad a \in bH,$$

which follows from the basic properties of a partition.

Example 2.59. For the group $\mathbb{Z}$, the cosets with respect to the subgroup $\langle m \rangle = m\mathbb{Z}$ are $m\mathbb{Z}$, $1 + m\mathbb{Z}$, $2 + m\mathbb{Z}, \ldots, (m-1) + m\mathbb{Z}$; they are the residue classes mod m ($m \geq 1$).

In the case of additive Abelian groups, cosets are often called residue classes.

Example 2.60. The left cosets with respect to the subgroup $H = \langle s \rangle$ in thes group $D_4 = \langle r, s \rangle = \{1, r, r^2, r^3, s, sr, sr^2, sr^3\}$ are:

$$H = \{1, s\}, \quad rH = \{r, sr^r\}, \quad r^2H = \{r^2, sr^2\}, \quad r^3H = \{r^3, sr\},$$

where note that $rs = sr^3$ Determine the right cosets as well.

The number of left cosets with respect to a subgroup H in a group G is called the *index* of H in G and denoted by $[G : H]$. The index can also be infinite of course. The next theorem is also called the index theorem.

Theorem 2.61 (Lagrange's theorem)**.** *If G is a finite group and $H \leqslant G$, then*

$$[G : H] = \frac{\#G}{\#H}.$$

In particular, the order of any subgroup of a finite group G divides the order of G.

Proof. Every coset aH has as many elements as H because the equation $ah_1 = ah_2$ implies $h_1 = h_2$ for $h_1, h_2 \in H$. We get

$$\#G = \sum_{a \in D} \#(aH) = \sum_{a \in D} \#H = [G : H] \cdot (\#H).$$

□

Remark 2.62. (i) Even if H is infinite, as is G, every coset aH has the same order as H since the map $H \to aH,\ h \mapsto ah$ is a bijection.

(ii) The index $[G : H]$ also states the number of right cosets with respect to H. This follows by showing (do it!) that the map $aH \mapsto Ha^{-1}$ is a bijection from the set of left cosets to the set of right cosets. If G is finite, the result can be obtained more easily by looking at the proof of Theorem 2.61.

Example 2.63. Examples 2.59 and 2.60 give $[\mathbb{Z} : m\mathbb{Z}] = m$ and $[D_4 : \langle s \rangle] = 4$. The latter also follows from Lagrange's theorem.

Example 2.64. Suppose that a group G has finite subgroups H and K whose orders are coprime. What is $H \cap K$? The answer is given by Lagrange's theorem. Because the order of $(H \cap K)$ divides both the order of $\#H$ and $\#G$, it must be 1. Thus $H \cap K = \{1\}$.

Lagrange's theorem is an effective aid when studying which subsets of a given finite group G are subgroups. For instance, it follows from Lagrange's theorem that no proper subset S of G with more than $(\#G)/2$ elements can be a subgroup of G.

Corollary 2.65. *The orders of the elements of a finite group G divide the order of G.*

Proof. This follows directly from Lagrange's theorem because $\mathrm{ord}(a) = \#\langle a \rangle$. □

Example 2.66. Let us compute the orders of the elements of the group $\mathbb{Z}_{13}^*$. 13 is a prime number so the possible orders are 1 and 13. Only the identity can have order 1, so $\text{ord}(1) = 1$. Then all other elements have order 13.

Corollary 2.67. *If G is a finite group, $\#G = g$, and $a \in G$, then $a^g = 1$.*

Proof. Let us denote the order of an element a by h; $a^h = 1$. By Corollary 2.65 we have $h \mid g$, that is, $g = th$ for some $t \in \mathbb{Z}$. Now $a^g = a^{th} = \left(a^h\right)^t = 1$. □

Example 2.68. Applying the previous corollary to the group $\mathbb{Z}_m^*$, we get $\overline{a}^{\varphi(m)} = \overline{1} \ \forall \ \overline{a} \in \mathbb{Z}_m^*$. Writing this as a congruence, we get *Euler's theorem*:

$$a^{\varphi(m)} \equiv 1 \pmod{m} \quad \text{when } \gcd(a, m) = 1.$$

In particular, when $m = p \in \mathbb{P}$, we get *Fermat's little theorem*:

$$a^{p-1} \equiv 1 \pmod{p} \quad \text{when } p \nmid a.$$

For instance, $3^4 \equiv 1 \pmod 5$, $22^{102} \equiv 1 \pmod{103}$ and so forth. Fermat was able to prove this using only methods of number theory (try it yourself!).

Fermat's little theorem can also be written as

$$a^p \equiv a \pmod{a} \ \forall \ a \in \mathbb{Z}.$$

Note that Euler's theorem, which is very important in number theory, was obtained above as just a special case of a general result in group theory.

Corollary 2.69. *If the order of a group G is a prime, then G is cyclic.*

Proof. Choose $a \in G$, $a \neq 1$. Because $\text{ord}(a)$ divides p and it is greater than 1, it must be p. Thus a generates the whole group G: $G = \langle a \rangle$. □

Because cyclic groups of identical orders are isomorphic (think why), it follows from the previous corollary that there is exactly one cyclic group for each given prime up to isomorphy.

Exercises

1. Let $H = 7\mathbb{Z}$. Determine all cosets in $\mathbb{Z}$ with respect to H.
2. Find all subgroups of the group $(\mathbb{Z}_{30}, +)$.
3. Find all generators of the groups $(\mathbb{Z}_6, +)$, $(\mathbb{Z}_8, +)$ and $(\mathbb{Z}_{20}, +)$.
4. Let a be a group element with $\text{ord}(a) = 15$. Compute the orders of the elements a^n, $0 \leq n \leq 15$.
5. Why are all nontrivial subgroups of a group G of order p^2 cyclic when p is a prime? What is the largest possible number of such subgroups of G?
6. Let $H = 3\mathbb{Z}$. Consider the cosets with respect to H in $\mathbb{Z}$. Determine whether the following cosets are equal:
   a) $11 + H$ and $17 + H$,
   b) $-1 + H$ and $5 + H$.
7. Let H and K be subgroups of a group G. Prove that if the group G has elements a, b such that $aH = bK$, then $H = K$.

Chapter 3

Structure of groups

3.1 Factor groups

When studying a given group G, it is often helpful to consider simpler groups, such as groups of smaller order. With this in mind we introduce the concepts *normal subgroup* and *factor group*.

If G is a group and H its subgroup, cosets aH and Ha may not coincide. If they do, then we get an important special case, whose significance Galois noticed around 150 years ago.

Definition 3.1. A subgroup N of a group G is called *normal* if its left and right cosets coincide, that is,

$$aN = Na \quad \forall\ a \in G.$$

In this case, we denote $N \trianglelefteq G$, or if N is a proper subset, $N \triangleleft G$.

If G is Abelian, then all of its subgroups are normal. Note that the definition $aN = Na$ does not imply $na = an$ for all $a \in G,\ n \in N$; instead it implies that

$$\forall\ n \in N \quad \exists n_1 \in N : \qquad an = n_1 a. \tag{3.1}$$

The next criterion is suitable for checking whether a subgroup is normal.

Theorem 3.2 (Normal subgroup criterion)**.** *Let $N \leqslant G$. Then*

$$N \trianglelefteq G \quad \Leftrightarrow \quad ana^{-1} \in N \ \ \forall\ a \in G, n \in N \quad \text{or} \quad aNa^{-1} \subset N \ \ \forall\ a \in G.$$

Proof. ($\Rightarrow$) Suppose $N \trianglelefteq G$. Then by applying condition (3.1) we get $ana^{-1} = n_1 \in N$.

($\Leftarrow$) Let $a \in G$. We need to prove that $aN = Na$.

Let $n \in N$ and denote $ana^{-1} = n_1$. By assumption we have $n_1 \in N$. Hence we get $an = n_1 a \in Na$. Thus $aN \subset Na$.

Now we apply the assumption to the elements a^{-1} and n. Then we get that $a^{-1}na = n_2 \in N$, so $na = an_2 \in aN$. This shows that $Na \subset aN$.

These results combined prove the statement. □

Remark 3.3. If $N \trianglelefteq G$, then it follows from the definition that

$$aNa^{-1} = N \ \ \forall\ a \in G.$$

The normal subgroup criterion thus states that it is enough to prove the inclusion relation $\subset$ instead of equality when checking normality of a subgroup.

Example 3.4. For any group G, the trivial subgroups $\{1\}$ and G are normal. Further, any subgroup H of index 2 is always normal: $aH = H = Ha\ \forall\ a \in H$ and $aH = G \setminus H\ \forall\ a \notin H$, so the left cosets of H are H and $G \setminus H$; likewise the right cosets. Therefore the left and right cosets coincide.

Example 3.5. We show for the subgroups

$$SL_n(\mathbb{R}) = \{A \in GL_n(\mathbb{R}) \mid \det(A) = 1\},$$
$$H = \{A \in GL_n(\mathbb{R}) \mid A \text{ is diagonal}\}$$

of the group $GL_n(\mathbb{R})$ that the former is normal but the latter is not when $n > 1$.

Let $A \in SL_n(\mathbb{R})$ and $P \in GL_n(\mathbb{R})$. Then

$$\det\left(PAP^{-1}\right) = \det(P)\det(A)\det(P)^{-1} = \det(A) = 1.$$

Hence $PAP^{-1} \in SL_n(\mathbb{R})$. Thus by the Normal subgroup criterion $SL_n(\mathbb{R})$ is normal.

We show by counter example that H is not a normal subgroup. Choose $D \in H$ and $P \in GL_n(\mathbb{R})$ as

$$D = \begin{pmatrix} -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}, \quad H = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & -1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

Now,

$$PDP^{-1} = \begin{pmatrix} 0 & -1 & 0 & \cdots & 0 \\ -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 0 \end{pmatrix} \notin H.$$

Hence H is not a normal subgroup of $GL_n(\mathbb{R})$ when $n > 1$.

Remark 3.6. Suppose that $a, x \in G$. The group element $y = axa^{-1}$ is called a *conjugate* of the element x. We also say that we get axa^{-1} from the element x by *conjugating* it by a.

The relation

$$a \sim y \qquad \Longleftrightarrow \qquad \exists a \in G : \quad y = axa^{-1}$$

is an equivalence relation on G. Its equivalence classes

$$[x] = \{axa^{-1} \mid a \in G\}$$

are called the *conjugacy classes* of G.

The Normal subgroup criterion can also be expressed as follows: The subgroup N of a group G is normal if and only if N is closed, that is, stable under conjugation by all elements of G, or if and only if N consists of whole conjugacy classes of G.

If $N \trianglelefteq G$, then the set of cosets in G is denoted by G/N,

$$G/N = \{aN \mid a \in G\} = \{aN \mid a \in D\},$$

where D is some set of representatives.

Theorem 3.7. *Suppose that $N \trianglelefteq G$. The group G/N is a group under the operation defined as follows:*

$$aN \cdot bN = abN.$$

Proof. (G1) follows after we show that the operation is well defined. Suppose $aN = a'N$ and $bN = b'N$. Then $a \in a'N$ and $b \in b'N$, so

$$a = a'n_1, \quad b = b'n_2, \qquad n_1, n_2 \in N.$$

Now $ab = a'n_1b'n_2$. Because the subgroup N is normal, we have $Nb' = b'N$ and hence the element n_1b' can be written as $b'n_2$, where $n_3 \in N$. The result is

$$ab = a'b'n_3n_2 \in a'b'N.$$

This shows that $abN = a'b'N$, that is, $aN \cdot bN = a'N \cdot b'N$.

Associativity (G2) follows from the associativity of the operation on G:

$$(aN \cdot bN) \cdot cN = abN \cdot cN = (ab)\, cN = a\,(bc)\, N = aN \cdot bcN = aN \cdot (bN \cdot cN)\,.$$

The neutral element is N and the inverse of aN is $a^{-1}N$ (check!). □

Definition 3.8. The group $(G/N, \cdot)$ is called the *factor group* (or *quotient group*) of G with respect to N.

Note that $\#\,(G/N) = [G : N]$. If G is Abelian and $H \leqslant G$, then $H \trianglelefteq G$ and the factor group G/H is an Abelian group, since $aH \cdot bH = abH = baH = bH \cdot aH \;\forall\; a, b \in G$.

Example 3.9. Let us form the the Cayley table of the factor group $\mathbb{Z}_{21}^*/H$ with $H = \langle\overline{4}\rangle = \{\overline{1}, \overline{4}, \overline{16}\}$. Group $\mathbb{Z}_{21}^*$ is Abelian so left cosets and right cosets coincide. The left cosets are $1H = 4H = 16H$, $2H = 8H = 11H$, $5H = 17H = 20H$ and $10H = 13H = 19H$. The Cayley table is then

$\cdot$	$1H$	$2H$	$5H$	$10H$
$1H$	$1H$	$2H$	$5H$	$10H$
$2H$	$2H$	$1H$	$10H$	$5H$
$5H$	$5H$	$10H$	$1H$	$2H$
$10H$	$10H$	$5H$	$2H$	$1H$

Example 3.10. The factor group in $\mathbb{R}^*$ with respect to the subgroup $H = \langle -1\rangle = \{+1, -1\}$ is

$$\mathbb{R}^*/H = \{aH \mid a \in \mathbb{R}^*\} = \{aH \mid a \in \mathbb{R}_+\}; \quad aH \cdot bH = abH.$$

Here $aH = \{+a, -a\}$.

Example 3.11. The factor group in the group $\mathbb{Z}$ with respect to the subgroup $\langle m\rangle = m\mathbb{Z}$ $(m \geq 1)$ is

$$\mathbb{Z}/m\mathbb{Z} = \{k + m\mathbb{Z} \mid k \in \mathbb{Z}\} = \{k + m\mathbb{Z} \mid k = 0, 1, \ldots, m-1\},$$

under the operation $(k + m\mathbb{Z}) + (h + m\mathbb{Z}) = (k + h) + m\mathbb{Z}$. This can also be written as

$$\mathbb{Z}/m\mathbb{Z} = \{\overline{0}, \overline{1}, \ldots, \overline{m-1}\}; \quad \overline{k} + \overline{h} = \overline{k+h}.$$

This is the familiar residue group mod m : $\mathbb{Z}/m\mathbb{Z} = \mathbb{Z}_m$. Factor groups can be considered a generalisation of the concept of residue groups.

Definition 3.12. An operation and an equivalence relation $\sim$ defined on a set A are said to be *compatible* if $a \sim a'$ $b \sim b' \quad \Rightarrow \quad ab \sim a'b'$.

Exercises

1. Let $H = \langle 6_{18} \rangle$. Form all the cosets in $\mathbb{Z}_{18}$ with respect to H and write the Cayley table for the group $(\mathbb{Z}_{18}/H, +)$.

2. Let $(G_1, \cdot)$ and $(G_2, \cdot)$ be groups. Show that $G_1 \times \{1_{G_2}\}$ is a subgroup of the group $G = G_1 \times G_2$. Is it a normal subgroup?

3. Suppose H and K are normal subgroups of a group G. Show that $H \cap K$ is also a normal subgroup of G.

4. Determine the subgroups of the group $(\mathbb{Z}_4, +)$, and the factor groups they define.

5. Because $\mathbb{R}$ is an Abelian group, its subgroup $\mathbb{Z}$ is normal and the factor group $\mathbb{R}/\mathbb{Z}$ is thus defined. Explain why the elements of the factor group can be written as $q + \mathbb{Z}$ where $q \in \mathbb{R},\ 0 \leq q < 1$. Write the elements $\left(\frac{1}{2} + \mathbb{Z}\right) + \left(\frac{2}{3} + \mathbb{Z}\right)$ and $-\left(\frac{3}{4} + \mathbb{Z}\right)$ with this form. What is the order of the element $\frac{35}{99} + \mathbb{Z}$?

6. Let G be a group and E an equivalence relation compatible with the operation on G. Show that the equivalence class of the identity element of G is a normal subgroup of G and that the partition of G formed by the cosets of this subgroup corresponds to the equivalence E.

7. Determine the cyclic subgroups that are normal of the group in Exercise 15. of Section 2.2.

8. Prove that $\langle 3 \rangle / \langle 12 \rangle$ is isomorphic to $\mathbb{Z}_4$, where $\langle 3 \rangle \leqslant \mathbb{Z}$ and $\langle 12 \rangle \leqslant \mathbb{Z}$.

9. Suppose H is a normal subgroup of a group G with index k. Prove that $a^k \in H\ \forall a \in G$. (Hint: Consider the factor group.)

10. What is the order of the group $\mathbb{Z}_{60}/\langle \overline{15} \rangle$?

3.2 Homomorphism theorem

Theorem 3.13. *Let $f\colon G \to G'$ be a group homomorphism.*

(i) *If $N \trianglelefteq G$, then $f(N) \trianglelefteq f(G)$.*

(ii) *If $N' \trianglelefteq G'$, then $f^{-1}(N') \trianglelefteq G$.*

(Compare this to Theorem 2.47 where we had $\leqslant$ instead of $\trianglelefteq$. Notice the other difference!)

Proof. (i) By Theorem 2.47 $f(N)$ is a group, moreover $f(N) \leqslant f(G)$. So we only need to show that $byb^{-1} \in f(N)$ for all $b \in f(G)$ and $y \in f(N)$.

We write $b = f(a)$ and $y = f(x)$, where $a \in G$ and $x \in N$. Then

$$byb^{-1} = f(a)f(x)f(a^{-1}) = f(axa^{-1}).$$

Since N is a normal subgroup of G, we have $axa^{-1} \in N$. Hence $f(axa^{-1}) \in f(N)$.

(ii) The proof is similar to above (). □

In particular, if we choose $N' = \{1\}$ we get the following result: The kernel $\mathrm{Ker}(f)$ of a group homomorphism $f\colon G \to G'$ is a normal subgroup of G.

The next theorem states that every homomorphism induces a certain isomorphism, which is why the theorem is also called the Isomorphism theorem. It is one of the basic theorems in group theory.

Theorem 3.14 (Homomorphism theorem)**.** *If $f\colon G \to G'$ is a group homomorphism, then*

$$G/\mathrm{Ker}(f) \simeq \mathrm{Im}(f).$$

More specifically: the homomorphism f induces an isomorphism

$$F\colon G/\mathrm{Ker}(f) \to \mathrm{Im}(f), \qquad F(a \cdot \mathrm{Ker}(f)) = f(a).$$

NS: This theorem has too many names so I stuck with the name used in the Finnish version, although it may be less common in English.

Proof. We denote $K = \mathrm{Ker}(f)$ and consider the map F. First we claim that F is well defined. If $aK = bK$ then $a \in bK$, that is, $a = bk$ for some $k \in K$. Now

$$F(aK) = f(a) = f(bk) = f(b)f(k) = f(b) \cdot 1 = f(b) = f(bK),$$

which proves the claim. Next, we need to show that F is an isomorphism.

Homomorphity: For $aK,\ bK \in G/K$, we have

$$F(aK \cdot bK) = F(abK) = f(ab) = f(a)f(b) = F(aK)F(bK).$$

Injectivity: If $F(aK) = 1$, we have $f(a) = 1$ and $a \in K$. Then $aK = K$ is equal to the identity element of the group G/K. This proves that F is an injection by Theorem 2.51.

Surjectivity follows directly from the definition of F. □

Remark 3.15. If $N \trianglelefteq G$, the map

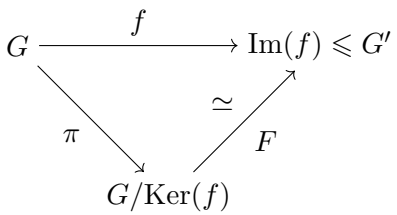
$$\pi\colon G \to G/N, \quad \pi(a) = aN,$$

is called the *(canonical) projection* or *surjection* from the group G to the factor group G/N. It is trivially surjective and also a homomorphism (check).

If we have a given homomorphism f and we choose the group $K = \mathrm{Ker}(f)$ as N, we see that $f(a) = F(aK) = F(\pi(a)) \quad \forall\ a \in G$. Hence

$$f = F \circ \pi.$$

This can be expressed by saying that the diagram below commutes, that the mapping of the elements does not depend on the "path". The notation $\simeq$ denotes that F is an isomorphism.



Remark 3.16. By Theorem 3.14 every homomorphic image $f(G) = \mathrm{Im}(G)$ of a group G is isomorphic to some factor group $G/\mathrm{Ker}(f)$ of G. Vice versa, every factor group G/N is isomorphic, and further, identical, to some homomorphic image of G. From the previous remark we see that $G/N = \mathrm{Im}(\pi)$ where π is the projection $G \to G/N$.

It follows from above that we can find all homomorphic images of G just by finding all of its factor groups. We can thus start from a homomorphism f of G, in which case $\mathrm{Ker}(f)$ is always a normal subgroup of G and the factor group $H/\mathrm{Ker}(f)$ is isomorphic to the image $\mathrm{Im}(f)$; or we can start from the normal subgroup N of G and use the canonical projection $\pi\colon G \to G/N$.

Example 3.17. Let us study the isomorphism induced by the homomorphism $f : \mathbb{Z} \to \mathbb{Z}_m, \quad f(a) = \overline{a}$. The kernel of f is

$$\mathrm{Ker}(f) = \{a \in \mathbb{Z} \mid \overline{a} = 0\} = m\mathbb{Z} = \langle m \rangle$$

and the image is $\{\overline{0}, \overline{1}, \ldots, \overline{m-1}\} = \mathbb{Z}_m$.

Thus by the Homomorphism theorem we get $\mathbb{Z}/\langle m \rangle = \mathbb{Z}_m$.

Example 3.18. The isomorphism induced by the homomorphism $f\colon \mathbb{R}^* \to \mathbb{R}_+, \quad f(x) = |x|$ is

$$F\colon \mathbb{R}^*/\{\pm 1\} \to \mathbb{R}_+, \quad F(x\{\pm 1\}) = x \;\forall\; x > 0.$$

Note that $x\{\pm 1\} = \{\pm x\}$. (Compare with Example 3.10.)

Example 3.19. The homomorphism $f\colon GL_n(\mathbb{R}) \to \mathbb{R}^*, \quad f(A) = \det(A)$ induces the isomorphism

$$GL_n(\mathbb{R})/SL_n(\mathbb{R}) \simeq \mathbb{R}^* \quad \text{(See Example 3.5)}.$$

Since invertible matrices have nonzero determinants, the image of $GL_n(\mathbb{R})$ is contained in $\mathbb{R}^*$. In addition, if $A, B \in GL_n(\mathbb{R})$, then $f(AB) = \det(AB) = \det(A)\det(B) = f(A)f(B)$. By definition, the kernel of f is $SL_n(\mathbb{R})$. Finally, if $a \in \mathbb{R}^*$, then $f(\left(\begin{smallmatrix} a & 0 \\ 0 & 1 \end{smallmatrix}\right)) = a$, so f is surjective.

Example 3.20. The trivial homomorphism $f\colon G \to G, \quad f(a) = 1$ and the identity map (homomorphism) id_G respectively yield isomorphisms

$$G/G \simeq \{1\}, \quad G/\{1\} \simeq G.$$

By applying the Homomorphism theorem to various homomorphisms, we get a number of general isomorphism theorems. The following is one example.

If $H \leqslant G$ and $K \trianglelefteq G$, then the set

$$HK = \{hk \mid h \in H, k \in K\}$$

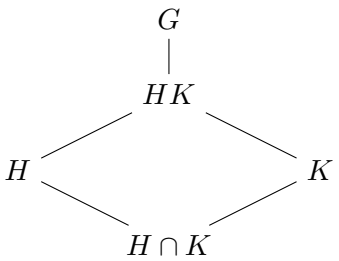
is a subgroup of G (check with the Subgroup criterion; observe that this is generated by the set $H \cup K$). By assumption, K is also a normal subgroup of HK. The map

$$f\colon H \to HK/K, \qquad f(a) = aK,$$

is a homomorphism and $\mathrm{Ker}(f) = H \cap K$, $\mathrm{Im}(f) = HK/K$ (check these yourself). Hence

$$H/\left(H \cap K\right) \simeq HK/K.$$

The diagram below shows why this isomorphism theorem is called the parallelogram rule.



Exercises

1. For the group of integers determine all its

   a) homomorphisms to itself,

   b) automorphisms.

2. Prove that the factor group $(5\mathbb{Z}/20\mathbb{Z}, +)$ and the group $(\mathbb{Z}_4, +)$ are isomorphic.

3. Show that the set $H \coloneqq \{f \in S_4 \mid f(4) = 4\}$ is a subgroup of S_4 that is isomorphic to S_3, and whose cosets $g \circ H$ are

$$X_k \coloneqq \{f \in S_4 \mid f(4) = k\} \ (k = 1, 2, 3, 4).$$

4. Show that the groups $\mathbb{Z}_6$ and S_3 are not isomorphic.

5. Show that the subgroup $T \coloneqq \{z \in \mathbb{C} \mid |z| = 1\}$ of the group $(\mathbb{C}^*, \cdot)$ is isomorphic to the factor group $\mathbb{R}/\mathbb{Z}$ of the group $(\mathbb{R}, +)$. (Hint: $|z|^2 = \overline{z}z$. Use the Homomorphism theorem and Euler's formula $e^{2\pi it} = \cos t + i(\sin t),\ t \in \mathbb{R}$.)

6. Suppose $\overline{G} = \mathbb{R}^{\mathbb{R}} = \{f \mid f : \mathbb{R} \to \mathbb{R}\}$ and $G = \{f \in \overline{G} \mid f \text{ is integrable }\}$. Show that $\overline{G}$ and G are groups under addition and that the map $f \mapsto \int f$ is a homomorphism $G \to \overline{G}$ when we assume that the integration constant is 0. What is the kernel of this map? Is the map a homomorphism if we assume that the integration constant is 1?

7. How many homomorphisms exist between the groups $(\mathbb{Z}_{20}, +)$ and $(\mathbb{Z}_8, +)$? How many of these are surjections?

3.3 Cyclic groups

Recall that a group G is called cyclic if it is generated by one of its elements:

$$G = \langle a \rangle$$

Cyclic groups form the simplest category among groups. In this section we will, among other things, determine all subgroups of a cyclic group.

A cyclic group of order n is denoted by C_n:

$$C_n = \langle c \rangle = \{1, c, c^3, \ldots, c^{n-1}\}; \quad c^n = 1$$

for $n = 1, 2, \ldots$ Similarly, an infinite cyclic group is

$$C_\infty = \langle c \rangle = \{\ldots, c^{-2}, c^{-1}, 1, c, c^2, \ldots\}.$$

When considering these groups, it is often helpful to remember that

$$C_n \simeq \mathbb{Z}/n\mathbb{Z}, \qquad C_\infty \simeq \mathbb{Z},$$

with isomorphisms respectively, say, $c \mapsto \overline{1}$ and $c \mapsto 1$. Here the groups on the right-hand side are additive.

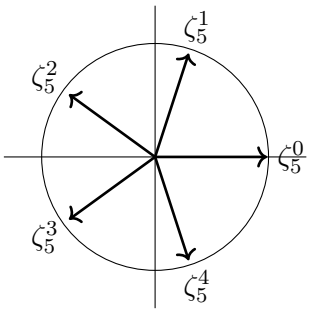
Example 3.21. The complex solutions of the equation $x^n = 1$ are

$$e^{2\pi ik/n} = \cos(2\pi k/n) + i \cdot \sin(2\pi k/n), \qquad k = 0, 1, \ldots, n-1,$$

the so-called *n-th roots of unity*. If we denote the first n-th root of unity by $\zeta_n = e^{2\pi i/n}$, these roots can be written as ζ_n^k, where $k = 0, 1, \ldots, n-1$. They form a subgroup of the group $\mathbb{C}^*$ that is cyclic of order n with generator ζ_n:

$$\langle \zeta_n \rangle \simeq C_n.$$

This is on multiplicative realisation of the group C_n. The figure below shows the 5th ($n = 5$) roots of unity in the complex plane.



Theorem 3.22. *The subgroups of an infinite cyclic group $C_\infty = \langle c \rangle$ are*

$$\langle c^n \rangle, \quad n = 0, 1, 2, \ldots$$

and they are all distinct.

Proof. Anyhow, the group C_∞ has the stated groups $\langle c^n \rangle$ as subgroups.

Suppose that H is a subgroup of C_∞. If $H = \{1\}$, then $H = \langle c^0 \rangle$. If $H \neq \{1\}$, then H contains some element c^n where $n > 0$. We choose the smallest such n. We show that $H = \langle c^n \rangle$.

If $a \in H$, then because $a \in \langle c \rangle$, a is of the form c^m for some $m \in \mathbb{Z}$. By writing $m = kn + r$, where $0 \leq r < n$, we get

$$c^r = c^{m-kn} = c^m \, (c^n)^{-k} \in H.$$

By the minimality of n, we get $r = 0$. Hence $m = kn$ and $a = c^m = (c^n)^k \in \langle c^n \rangle$. Thus we have obtained the result $H \subset \langle c^n \rangle$. The reverse relation $\langle c^n \rangle \subset H$ follows immediately from the fact that $c^n \in H$. Together these prove the statement.

Because c^n is the smallest power of c in the group $\langle c^n \rangle$, we see that

$$\langle c^n \rangle = \langle c^{n'} \rangle, \quad n, n' > 0 \quad \Rightarrow \quad n = n'.$$

Thus, we result in the latter statement. □

Consequently, all subgroups $\neq \{1\}$ of an infinite cyclic group C_∞ are $\simeq C_\infty$.

When applied to the group $\mathbb{Z}$, the previous statement says that $n\mathbb{Z}$ ($n = 0, 1, \ldots$) are all of its subgroups. What about all of the factor groups of $\mathbb{Z}$? Because $\mathbb{Z}$ is Abelian, every subgroup is normal; the factor groups of $\mathbb{Z}$ are the groups $\mathbb{Z}/n\mathbb{Z}$. By Example 3.11 we get the result: All factor groups of the group $\mathbb{Z}$ are the residue groups $\mathbb{Z}_n$ ($n = 1, 2 \ldots$) and $\mathbb{Z}$ itself ($n = 0$).

A similar result holds naturally for any general cyclic group C_∞ as well.

Theorem 3.23. *Let $C_n = \langle c \rangle$ be a cyclic group of order n. For any divisor m of n, there exists exactly one subgroup whose order is m. It is*

$$\langle c^k \rangle = \{1, c^k, c^{2k}, \ldots, c^{(m-1)k}\}, \qquad \text{where } k = n/m.$$

The group C_n has no other subgroups.

Proof. Because $\left(c^k\right)^m = c^n = 1$, the group mentioned in the statement is a subgroup of order m of C_n. Thus it only needs to be shown that this is the only such subgroup.

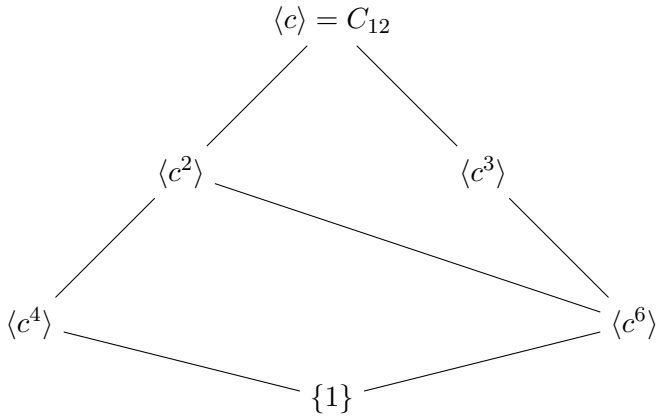
Suppose $H \leqslant C_n$ and $\#H = m$. If $m = 1$, then $H = \{1\} = \langle 1\rangle$ as it should. We assume that $m > 1$. Just like in the proof of Theorem 3.22, we choose the smallest positive exponent k for which $c^k \in H$. A similar reasoning gives $H = \langle c^k\rangle$. Here the order of the element c^k is equal to $\#H = m$, so $km = n$ and thus $k = n/m$.

The latter statement in the theorem follows by Lagrange's theorem. □

As a corollary to Theorems 3.22 and 3.23, we see that all subgroups of a cyclic group are cyclic. It also follows from Theorem 3.23 that if p is a prime number then the group C_p has no other subgroups than the trivial ones.

Example 3.24. Let us determine the subgroups of the group C_{12}. The divisors of 12 are $1, 2, 3, 4, 6, 12$. Applying Theorem 3.23, we get the subgroups $\langle 1\rangle$, $\langle c\rangle = C_{12}$, $\langle c^2\rangle$, $\langle c^3\rangle$, $\langle c^4\rangle$ and $\langle c^6\rangle$.

We can present the subgroups in a subgroup diagram, which is also called a *Hasse diagram*.



The next theorem presents a simple formula for computing the order of an element in a finite group.

Theorem 3.25. *If* $\operatorname{ord}(a) = n$, *then* $\operatorname{ord}(a^m) = \frac{n}{\gcd(n,m)}$.

Proof. Denote $d = \gcd(n, m)$ and $n = n_1 d,\ m = m_1 d$. We need to prove that $\operatorname{ord}(a^m) = n_1$, that is,

$$(a^m)^r = 1 \qquad \Longleftrightarrow \qquad n_1 \mid r.$$

Since ord(a) = n, then $a^{mr} = 1$ if and only if $n \mid mr$. This is equivalent to the condition $n_1 \mid m_1 r$. Because $\gcd(n_1, m_1) = 1$, the statement follows. □

Consequently the generators of a cyclic group $C_n = \langle c\rangle$ are exactly the elements c^m with $\gcd(n, m) = 1$. The number of such generators is $\varphi(n)$.

Example 3.26. Let us compute the orders of the elements in the group $\mathbb{Z}_9^*$. Firstly, $\mathbb{Z}_9^*$ is cyclic since $\langle 2\rangle = \{1, 2, 4, 5, 7, 8\rangle = \mathbb{Z}_9^*$. Thus $\operatorname{ord}(2) = 6$. The order of the identity element is $\operatorname{ord}(1) = 1$. Now, applying Theorem 3.25 we get

$$\operatorname{ord}(4) = \operatorname{ord}(2^2) = \frac{6}{\gcd(6,2)} = 3,$$
$$\operatorname{ord}(5) = \operatorname{ord}(2^5) = \frac{6}{\gcd(6,5)} = 6,$$
$$\operatorname{ord}(7) = \operatorname{ord}(2^4) = \frac{6}{\gcd(6,4)} = 3,$$
$$\operatorname{ord}(8) = \operatorname{ord}(2^3) = \frac{6}{\gcd(6,3)} = 2.$$

Example 3.27. Every group of prime order is cyclic. See Section 2.5 Corollary 2.69.

Exercises

1. List the elements of the subgroup $\langle 3 \rangle$ of the group $(\mathbb{Z}_{18}, +)$.
2. What are the subgroups of the group $(\mathbb{Z}_{30}, +)$? Compute their orders and write down their generators.
3. Find all generators of the cyclic groups $C_6 = \langle a \rangle$, $C_8 = \langle b \rangle$ and $C_{20} = \langle c \rangle$.
4. What elements form the subgroup $H = \langle a, b \rangle$ of $\mathbb{C}^*$ when $a = e^{2\pi i/5}$ and $b = e^{2\pi i/7}$? Determine whether H is cyclic.
5. Draw the Hasse diagram of all subgroups of the cyclic group C_{30}. Choose such a subgroup H of C_{30} that the factor group C_{30}/H has order 3, and construct the Cayley table for this factor group.
6. Suppose that G and H are finite groups whose orders are distinct primes. Determine all homomorphisms $f\colon G \to H$. (Hint: Consider the groups $\mathrm{Ker}(f)$ and $\mathrm{Im}(f)$.)
7. Determine all group homomorphisms $f\colon \mathbb{Z}_9 \to \mathbb{Z}_{15}$. (Hint: The subgroups of $(\mathbb{Z}, +)$ are known to be cyclic. First determine all homomorphisms $g\colon \mathbb{Z} \to \mathbb{Z}_{15}$ and then solve the equations $f \circ j = g$, where j is a projection $\mathbb{Z} \to \mathbb{Z}_9$.)

3.4 Permutation groups

In Section 2.1 we studied an example of the symmetric group of n elements

$$S_n = \{\alpha\colon J_n \to J_n \mid \alpha \text{ is a bijection } \},$$

where $J_n = \{1, 2, \ldots, n\}$ and the bijection, a *permutation*, α was denoted

$$\alpha = \begin{pmatrix} 1 & 2 & \ldots & n \\ k_1 & k_2 & \ldots & k_n \end{pmatrix}.$$

Here we have $\alpha(i) = k_i$ $(i = 1, \ldots, n)$.

Subgroups of the group S_n are called *permutation groups.* Let us switch to a more convenient notation for permutations. We call the permutation of the distinct elements $a_1, \ldots, a_r$, where

$$a_1 \mapsto a_2,\ \ a_2 \mapsto a_3,\ \ \ldots,\ \ a_{r-1} \mapsto a_r,\ \ a_r \mapsto a_1,$$

a *r-length cycle* or simply an *r-cycle* and denote it by

$$(a_1 a_2 \cdots a_r).$$

Every permutation $\alpha \in S_n$ can be written as a product of cycles that have no elements in common. The idea of the proof is to first write a cycle starting from the element 1, then a new cycle starting from one of the remaining elements and so on. For example,

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = (123), \quad \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 4 & 1 & 6 & 2 & 5 & 3 \end{pmatrix} = (142)(36)(5).$$

We can prove that this *cyclic form* is unique up to the order the cycles are written in and what element each cycle starts from, for example,

$$(142)(36)(5) = (36)(214)(5).$$

Furthermore, 1-cycles as the identity map are usually excluded from the product, for example, $(142)(36)(5) = (142)(35)$. Note especially that disjoint cycles commute.

A 2-cycle is called a *transposition*. Note the order of terms in a product of permutations, for example,

$$(23)(12) = (132), \quad (12)(23) = (123).$$

Example 3.28. Let us write the elements of S_4 in cyclic notation, with (1) denoting the identity map.

(1)	(132)	(12)(34)	(1432)
(12)	(123)	(13)(24)	(1243)
(13)	(123)	(14)(23)	(1342)
(14)	(143)		(1423)
(23)	(234)		(1234)
(24)	(134)		(1324)
(34)	(124)		
	(142)		

For example, the set $H_4 = \{(1),\ (12), (34), (12)(34)\}$ is a subgroup of S_4. For any transposition (ij) we have $(ij)^2 = (1)$, and since $(ij)(kl) = (kl)(ij)$, we have $(12)(34)(12)(34) = (1)$. Thus we have closure, each element has an inverse in H_4, and H_4 is nonempty by definition. Thus H_4 is a subgroup of S_4 by Theorem 2.25. The Cayley table of H_4 is

	(1)	(12)	(34)	(12)(34)
(1)	(1)	(12)	(34)	(12)(34)
(12)	(12)	(1)	(12)(34)	(34)
(34)	(34	(12)(34)	(1)	(12)
(12)(34)	(12)(34)	(34)	(12)	(1)

which is the same as the table of the Klein four-group. (See Example 2.18.)

A permutation $\alpha \in S_4$ is said to be of *type* $(r_1, \ldots, r_m)$ if the lengths of its cycles are $r_1, \ldots, r_m$. For example, the type of the previous six element permutation $(142)(36)(5)$ is $(1, 2, 3)$. The numbers $r_1, \ldots, r_m$ (whose order does not matter) form a partition of n, that is,

$$r_1 + \cdots + r_m = n.$$

Theorem 3.29. *If the type of a permutation $\alpha \in S_n$ is $(r_1, \ldots, r_m)$, then*

$$\operatorname{ord}(\alpha) = \operatorname{lcm}(r_1, \ldots, r_m).$$

Proof. For $k = 1, \ldots, r-1$, the permutation $(a_1 a_2 \ldots a_r)^k$ maps the element a_1 to the element a_{k+1}. Thus $(a_1 a_2 \ldots a_r)^k \neq \mathrm{id}_{J_n}$ for these values of k. Because $(a_1 a_2 \ldots a_r)^r = (1)$, we see that the order of the r-cycle is r.

By assumption $\alpha = \alpha_1 \cdots \alpha_m$ where a_j are disjoint r_j-cycles. Since disjoint cycles commute, we have $\alpha^t = \alpha_1^t \cdots \alpha_m^t$. Thus we get

$$\alpha^t = (1) \quad \iff \quad \alpha_j^t = (1)\ (j = 1, \ldots, m) \quad \iff \quad r_j \mid t\ (j = 1, \ldots, m).$$

And the statement follows. □

Remark 3.30. Each conjugacy class of S_n comprises all permutations of the same type (see Remark 3.6). This can be seen by writing

$$\alpha = (a_1 a_2 \dots a_p)(a_{p+1} a \dots a_q) \cdots (\dots a_n),$$
$$\tau = \begin{pmatrix} a_1 & a_2 & \dots & a_p & a_{p+1} & \dots & a_q & \dots & a_n \\ b_1 & b_2 & \dots & b_p & b_{p+1} & \dots & b_1 & \dots & b_n \end{pmatrix}$$

and computing

$$\tau\alpha\tau^{-1} = (b_1 b_2 \dots b_p)(b_{p+1} \dots b_1) \cdots (\dots b_n).$$

Example 3.31. Let us show that H_4 is a normal subgroup of S_4. Recall that a subgroup is normal if and only if it is a union of whole conjugacy classes. Looking at the table in Example 3.28, we see that the conjugacy classes of S_4 are

- the 4-cycles $(abcd)$ with 6 elements
- the 3-cycles $(abc)(d)$ with 8 elements
- the products of two transpositions $(ab)(cd)$ with 3 elements
- the transpositions (ab) with 6 elements
- the identity with 1 element.

The subgroup H_4 has order $4 = 1 + 3$, so it is the union of the identity and the products of two transpositions. Hence, it is normal.

Consider the polynomial in n indeterminates $x_1, \dots, x_n$,

$$\begin{aligned} \Delta = \prod_{i<j} (x_i - x_j) = (x_1 - x_2)(x_1 - x_3) \cdots (x_1 - x_n) \cdot \\ (x_2 - x_3) \cdots (x_2 - x_n) \cdot \\ \cdots\cdots\cdots\cdots \\ (x_{n-1} - x_n). \end{aligned}$$

We assume that $\alpha \in S_n$ and denote by $\alpha(\Delta)$ the polynomial that is obtained by applying the permutation α on the indices of the polynomial Δ.

For example, if $n = 3$ and $\alpha = (13)$, then

$$\alpha(\Delta) = (x_3 - x_2)(x_3 - x_1)(x_2 - x_1) = -\Delta.$$

We see that $\alpha(\Delta) = \pm\Delta$ always holds. This sign is called the *sign* of the permutation α and denoted by $\operatorname{sign}(\alpha)$; then

$$\alpha(\Delta) = \operatorname{sign}(\alpha)\Delta = \begin{cases} +\Delta, & \text{if } \alpha \text{ is even,} \\ -\Delta, & \text{if } \alpha \text{ is odd.} \end{cases}$$

If $\alpha \in S_n$ and $\beta \in S_n$, then

$$\operatorname{sign}(\alpha\beta) = \operatorname{sign}(\alpha)\operatorname{sign}(\beta), \tag{3.2}$$

since

$$\operatorname{sign}(\alpha\beta)\Delta = (\alpha\beta)(\Delta) = \alpha(\beta(\Delta)) = \operatorname{sign}(\alpha)\operatorname{sign}(\beta)\Delta.$$

The map $\operatorname{sign}\colon S_n \to \{\pm 1\}$ is thus a homomorphism.

Example 3.32. a) $\operatorname{sign}(\alpha^{-1}) = \operatorname{sign}(\alpha)$ because

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

b) $\operatorname{sign}(*\alpha*^{-1}) = \operatorname{sign}(*)\operatorname{sign}(\alpha)\operatorname{sign}(*^{-1}) = \operatorname{sign}(*)^2\operatorname{sign}(\alpha) = \operatorname{sign}(\alpha)$.

Theorem 3.33. (i) *Permutations of the same type have the same sign.*

(ii) *Transpositions are odd. More generally: the sign of an r-cycle is $(-1)^{r-1}$.*

Proof. (i) By Remark 3.30, permutations in S_n are of the same type if and only if they are conjugate elements in S_n. The statement now follows from Example 3.32 b).

(ii) The transposition (12) is always odd because $\alpha(\Delta) = -\Delta$. All transpositions are thus odd by part (i). We obtain the more general statement from Equation (3.2) since an r-cycle is always a product of $(r-1)$ transpositions:

$$(a_1 a_2 \dots a_r) = (a_1 a_r)(a_1 a_{r-1}) \cdots (a_1 a_2).$$

□

Because

$$(ab) = (1a)(1b)(1a),$$

the last equation in the previous proof shows that every cycle – and every permutation – can be written as a product of transpositions $(12), (13), \dots, (1n)$. We get the result that the *transpositions* $(12), (13), \dots, (1n)$ *generate* the group S_n. Compare the number of these transpositions to the order of S_n!

The set of all even permutations of n elements

$$A_n = \{\alpha \in S_n \mid \operatorname{sign}(\alpha) = \pm 1\}$$

forms a subgroup of S_n because the product of even permutations is even. It is called the *alternating group* of n elements. We can prove that $A_1 = S_1$ and $[S_n : A_n] = 2 \ \forall\ n \geq 2$. In particular, we have $A_n \trianglelefteq S_n$.

Example 3.34. Let us determine the elements of A_4. In the previous example we wrote down the types of elements in S_4. Now using Theorem 3.33, the types of elements in S_4 are the identity (even), six transpositions (odd), eight 3-cycles (even), six 4-cycles (odd), and three products of two disjoint transpositions (even). Thus

$$A_4 = \{(1), (132), (123), (243), (143), (234), (134), (124), (142), (12)(34), (13)(24), (14)(23)\}.$$

The significance of permutations is illustrated by *Cayley's theorem*: Every finite group is isomorphic to some permutation group.

The idea of the proof is simple: Let $x_1, \dots, x_n$ be the elements of a group G. For $x \in G$, the map t_x defined by $t_x(x_i) = xx_i$ is a permutation of G. We denote $xx_i = x_j$ and state that when the index i goes through the set $\{1, 2, \dots, n\}$, j goes through it as well. Thus t_x defines the permutation $\alpha_x \in S_n$. The map

$$p\colon G \to S_n, \qquad p(x) = \alpha_x,$$

is an injective homomorphism; that is, $G \simeq \operatorname{Im}(p) \leqslant S_n$. (Go through the details.)

Example 3.35. Let us determine the permutation group that C_3 is isomorphic to. First we write the Cayley table for $C_3 = \{1, a, a^2\}$:

	1	a	a^2
1	1	a	a^2
a	a	a^2	1
a^2	a^2	1	a

We then have

$$\lambda_1 = \begin{pmatrix} 1 & a & a^2 \\ 1 & a & a^2 \end{pmatrix}, \quad \sigma_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix} = (1)$$
$$\lambda_a = \begin{pmatrix} 1 & a & a^2 \\ a & a^2 & 1 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix} = (123)$$
$$\lambda_{a^2} = \begin{pmatrix} 1 & a & a^2 \\ a^2 & 1 & a \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix} = (132)$$

where we replaced $1, a, a^2$ with $1, 2, 3$ respectively. Hence $C_3 \simeq \{(1), (123), (132)\} \leqslant S_3$.

Exercises

1. Compute the orders of the following permutations.
   a) (14),
   b) (147),
   c) (14762),
   d) $(147)(62)$.
2. Let $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 2 & 1 & 3 & 5 & 4 & 6 \end{pmatrix}$ and $\beta = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 1 & 2 & 4 & 3 & 5 \end{pmatrix}$. Compute
   a) α^{-1},
   b) $\alpha\beta$,
   c) $\beta\alpha$.
3. Let $\alpha = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ 2 & 1 & 3 & 5 & 4 & 7 & 6 & 8 \end{pmatrix}$. Write α as the product of disjoint cycles.
4. Prove that $[S_n : A_n] = 2$. (Use, e.g., the Homomorphism theorem).

3.5 What next?

In this section we will briefly introduce some ideas which can extend the basic group theory discussed above to solving broader problems.

If $H \lhd G$, the properties of the group G can be returned to the groups H and G/H. This motivates a definition: A group G is called *simple* if its only normal subgroups are $\{1\}$ and G.

Example 3.36. A cyclic group C_p is simple when p is a prime number. By Lagrange's theorem the order of any subgroup of a group divides the order of the group. As p is a prime, the only possible subgroups are $\{1\}$ of order 1 and C_p, both of which are normal.

NS: I wrote the case n=5 but should the whole proof be written out in this example? We can prove for $n \geq 5$ that the alternating group A_n is simple. Let us prove the case $n = 5$. The order of A_5 is 60, and the sizes of its conjugacy classes are 1, 12, 12, 20, and 15. A normal subgroup must contain the conjugacy class of size 1, the identity, and one or more of the other conjugacy classes. Therefore the order of any normal subgroup is some sum of these numbers, including the 1. By Lagrange's theorem the order must also divide the order of A_4. However, no such sum among these numbers divides 60 other than 1 and 60 themselves.

Every finite group G can be constructed of simple groups in the following sense: there exists a sequence of normal subgroups, called the *composition series*,

$$\{1\} \trianglelefteq H_1 \trianglelefteq H_2 \trianglelefteq \cdots \trianglelefteq H_t \trianglelefteq G$$

such that the groups $G/H_t,\ H_t/H_{t-1}, \ldots, H_2/H_1, H_1\ (\simeq H_1/\{1\})$ are simple. These (factor) groups are called the *composition factors* of the group G, and they are unique up to isomorphy.

In order to classify all (finite) groups we need to know all (finite) simple groups. Searching for them has provided researchers an adventure that only reached its end in 1981.

Definition 3.37. A finite group G is called *solvable* if its composition factors are cyclic groups whose orders are primes.

We easily observe that all subgroups of a solvable group are solvable.

Example 3.38. The solvability of the symmetric group S_n:

- $n = 1,\ n = 2:\ \ S_1 \simeq C_1,\ \ S_2 \simeq C_2$; thus S_1 and S_2 are solvable.
- $n = 3:\ \ \{(1)\} \lhd A_3 \lhd S_3$, where $S_3/A_3 \simeq C_2$ and $A_3 \simeq C_3$; thus S_3 is solvable.
- $n = 4:\ \ \{(1)\} \lhd \{(1), (12)(34)\} \lhd H_4 \lhd A_4 \lhd S_4$ by Example 3.31, and hence the composition factors of S_4 are C_2, C_3, C_2, C_2; thus S_4 is solvable.
- $n \geq 5:\ \ \{1\} \lhd A_4 \lhd S_n$, where S_n/A_n is simple ($\simeq C_4$), similarly A_n. These are the composition factors of S_n. Because $\#A_n$ is not a prime number, the group S_n is not solvable for $n \geq 5$.

Every equation

$$p(x) = 0,$$

where $p(x)$ is a polynomial of degree n, can be linked to a certain permutation group $G_p \leqslant S_n$ determined by its roots, which satisfies the following condition: the equation $p(x) = 0$ is solvable if and only if the group G_p is solvable. Here, the solvability of an equation means that its roots can be obtained from the coefficients by using rational operations and taking roots, similar to the solution formula for second degree equations.

For each n we can form such a polynomial $p(x)$ that its linked group G_p is the whole S_n. We see that a *general solution formula* exists for an n-th degree polynomial if and only if the group S_n is solvable. The previous example shows that this occurs if and only if $n < 5$.

This result was attained by Abel and Galois in the early 19th century. It solved a centuries-old problem and signified the beginning of group theory.

In the proof of Cayley's theorem we attached a permutation α_x of the set $\{1, 2, \ldots, n\}$ to each element x of the group $G = \{x_1, x_2, \ldots, x_n\}$. Let V be a n-dimensional (complex) vector space and let fix a basis $\{B_1, \ldots, B_n\}$. When we apply the permutation α_x to the basis vectors,

we get a regular linear map $V \to V$. This in turn corresponds to a regular $n \times n$ matrix A_x. Thus we form the map

$$G \to GL_n(\mathbb{C}), \qquad x \mapsto A_x$$

which is a group homomorphism. This map – or the matrix group obtained as a homomorphic image of G – is called the *regular representation* of group G. Instead of the matrix group $GL_n(\mathbb{C})$, we could have just as well considered a group it is isomorphic to:

$$GL(V) = \{t\colon V \to V \mid t \text{ is a regular linear map }\}.$$

In general, any group homomorphism

$$G \to GL_n(\mathbb{C}) \qquad \text{or} \qquad G \to GL(V)$$

is called a *representation* of group G. Representations allows us to study groups via the theory of matrices or linear maps. This so-called representation theory of groups has proven to be fruitful in both group theory and its applications.

Example 3.39. Let us determine the regular (matrix) representation of the cyclic group C_3. By Example 3.35 we know that C_3 is isomorphic to the subset $\{(1), (123), (132)\}$ of the symmetric group S_3. Using the standard basis for $\mathbb{C}^3$, we get

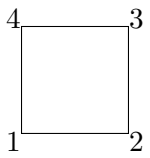
$$\rho((1)) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\rho((123)) = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\rho((132)) = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}.$$

3.6 Symmetry group of the square

In this section we demonstrate the theory in Chapters 2 and 3 by studing the dihedral group D_4 in more detail. We will examine the square

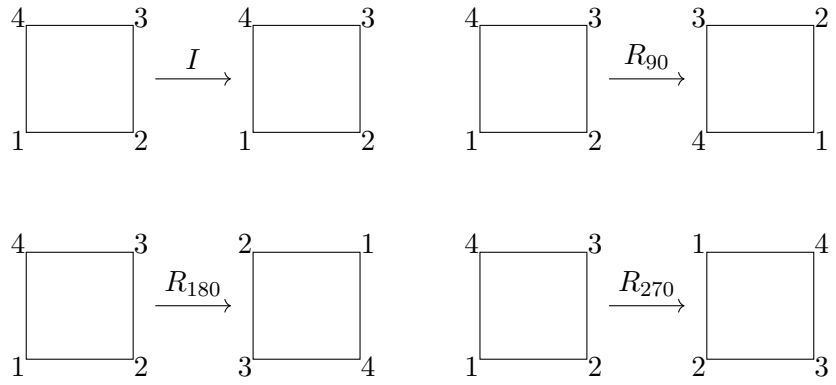


both as a set of points in the plane as well as those rotations and reflections of the plane that map the square to itself.

Rotations

A square, as a set of points in the plane, maps to itself in the following rotations: rotations around the centre of the square by the angle $\alpha = 0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}$. Let us denote these rotations by $I = R_0$ (identity map), R_{90}, R_{180}, R_{270} respectively. By denoting the vertices by $1, 2, 3, 4$ we

obtain a connection to permutations of four elements: for example, $R_{90} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 4 & 1 & 2 & 3 \end{pmatrix} \in S_4$. Nonetheless, we will proceed with geometric observations.



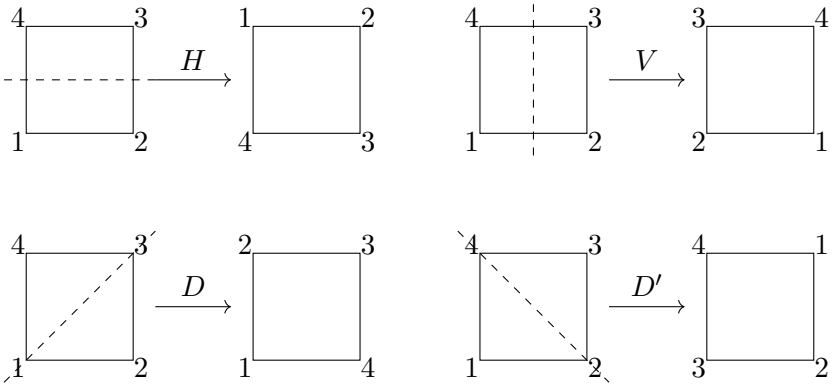
The group action is the composition of maps, for example, $(R_{180}, R_{90}) \mapsto R_{180} \circ R_{90} = R_{270}$. Here, recall that $A \circ B$ denotes a composition where we first apply B then A. This is known to be associative, the neutral element is the identity map I, and the maps $I, R_{90}, R_{180}, R_{270}$ have inverses. For instance, $R_{270}^{-1} = R_{90}$ because $R_{270} \circ R_{90} = R_{90} \circ R_{270} = I$. Hence $(\{I, R_{90}, R_{180}, R_{270}\}, \circ)$ is a group. Check the Cayley table below and note that the group is commutative since the table is symmetric with respect to the diagonal.

	h	g
h		$h \circ g$
g	$g \circ h$	

$\circ$	I	R_{90}	R_{180}	R_{270}
I	I	R_{90}	R_{180}	R_{270}
R_{90}	R_{90}	R_{180}	R_{270}	I
R_{180}	R_{180}	R_{270}	I	R_{90}
R_{270}	R_{270}	I	R_{90}	R_{180}

Reflections

The square has exactly four symmetry axes and reflecting with respect to them maps the square to itself: H is the reflection with respect to the horizontal axis, V the reflection with respect to vertical axis, D the reflection with respect to the diagonal through vertices 1 and 3, and D' the reflection with respect to the diagonal through vertices 2 and 4.



For example, $(\{I, H\}, \circ)$ forms a group whose Cayley table is

$\circ$	I	H
I	I	H
H	H	I

Compare this to the group $(\mathbb{Z}_2, +)$. Is $(\{I, H, D\}, \circ)$ a group?

By taking the compositions of the reflections we get, for instance, $H \circ V = R_{180}$. You should check how the vertices are mapped – what number of vertices is sufficient?. Thus the reflections do not form a group, yet $(\{I, R_{90}, R_{180}, R_{270}, H, V, D, D'\}, \circ)$ is a group, the *symmetry group of the square*, or the *dihedral group* D_4, and we obtain its Cayley table:

$\circ$	I	R_{90}	R_{180}	R_{270}	H	V	D	D'
I	I	R_{90}	R_{180}	R_{270}	H	V	D	D'
R_{90}	R_{90}	R_{180}	R_{270}	I	D	D'	V	H
R_{180}	R_{180}	R_{270}	I	R_{90}	V	H	D'	D
R_{270}	R_{270}	I	R_{90}	R_{180}	D'	D	H	V
H	H	D'	V	D	I	R_{180}	R_{270}	R_{90}
V	V	D	H	D'	R_{180}	I	R_{90}	R_{270}
D	D	H	D'	V	R_{90}	R_{270}	I	R_{180}
D'	D'	V	D	H	R_{270}	R_{90}	R_{180}	I

This group is not commutative since for example, $H \circ R_{270} = D$, $R_{270} \circ H = D'$. The maps $I, R_{90}, R_{180}, R_{270}, H, V, D, D'$ are called the *symmetries* of the square. The group of rotations $G = (\{I, R_{90}, R_{180}, R_{270}\}, \circ)$ is a subgroup of D_4, because earlier G was stated to be a group and $G \subset D_4$.

Subgroups of D_4

a) First we consider the subgroup G formed by the rotations of the square. By comparing the tables below, we state that $(\mathbb{Z}_4, +)$ is isomorphic to the group $(G, \circ)$.

$+$	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

$\circ$	I	R_{90}	R_{180}	R_{270}
I	I	R_{90}	R_{180}	R_{270}
R_{90}	R_{90}	R_{180}	R_{270}	I
R_{180}	R_{180}	R_{270}	I	R_{90}
R_{270}	R_{270}	I	R_{90}	R_{180}

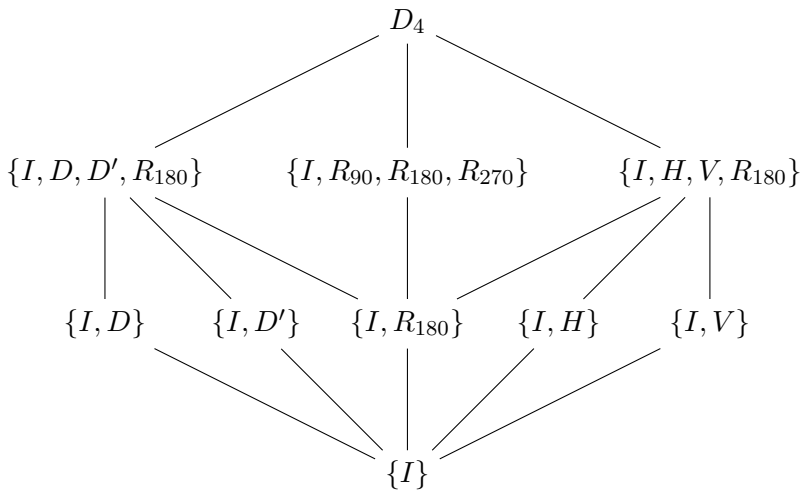
The map $f\colon 0 \mapsto I,\ 1 \mapsto R_{90},\ 2 \mapsto R_{180},\ 3 \mapsto R_{270}$ is a bijection $\mathbb{Z}_4 \to G$ and a group homomorphism, since for all $a, b, \in \mathbb{Z}_4$ we have $f(a + b) = f(a) \circ f(b)$. Note the interpretation of the equation $f(x^{-1}) = f(x)^{-1}$, here $f(-x) = f(x)^{-1}$. For example, $-3 = 1$ since $3 + 1 = 0$, and $f(-3) = f(1) = R_{90} = R_{270}^{-1} = f(3)^{-1}$.

b) G is also an example of a cyclic group, generated by R_{270} for example: using the usual notation $R_{270}^2 = R_{270} \circ R_{270}$, we get $R_{270}^2 = R_{180}$, $R_{270}^3 = R_{90}$, $R_{270}^4 = I$. In addition, R_{90} generates the group G: $(R_{90})^2 = R_{180}$, $(R_{90})^3 = R_{270}$, $(R_{90})^4 = I$. Observe the geometry as well: R_{270} is the rotation by the angle $-\frac{\pi}{4}$ and R_{90} is the rotation by the angle $\frac{\pi}{4}$, that is, R_{270} is a rotation in the negative direction (clockwise) and R_{90} is a rotation in the positive direction (counterclockwise).

c) Let us find all the subgroups of D_4 without relying on geometry. It is easiest to investigate what are the cyclic subgroups of D_4, that is, those subgroups of the type $\{a^n \mid n \in \mathbb{Z}\}$. Anyhow the group has the cyclic subgroup of one element $\{I\}$. Cyclic subgroups with two elements of D_4 can be obtained by choosing a reflection or a rotation by π in addition to the identity map. When the order of the map is 2, the map is its own inverse: $\{I, D\}$, $\{I, D'\}$, $\{I, R_{180}\}$, $\{I, H\}$, $\{I, V\}$. You should observe this from the Cayley table and compare it to Exercise 6. of Section 2.1. For this particular group there are no cyclic subgroups with

three elements but G has a subgroup with four elements. No other subgroups exist. In fact, by taking advantage of the fact $\#D_4 = 8$, we would have directly gotten that D_4 only has subgroups with 1, 2, 4, or 8 elements.

Let us now explore what are the noncyclic subgroups of D_4 that are generated by two elements. Let S be generated by elements a and b. Then we must have $a^n,\ b^n \in S$ and $a^i b^j \in S$ whenever $n, i, j \in \mathbb{Z}$, and in turn these elements form a subgroup. The generators of the subgroup $\{I, D, D', R_{180}\}$ are D, D' or D, R_{180} or D', R_{180}. The generators of the subgroup $\{I, H, V, R_{180}\}$ are R_{180}, H or H, V or R_{180}, V. The generators of the group D_4 are R_{270}, H or D, H or D', H. Below is the Hasse diagram of the group D_4.



Let us consider the statement that D_4 is generated by R_{270} and H:

$$\begin{aligned}
&R_{270}^0 = I, \quad R_{270} = R_{270}, \quad R_{270}^2 = R_{180}, \quad R_{270}^3 = R_{90}, \quad H = H,\\
&R_{270} \circ H = D', \quad R_{270}^2 \circ H = V, \quad R_{270}^3 \circ H = D,\\
&R_{270}^4 = I, \quad H^2 = I.
\end{aligned}$$

All the elements of the group D_n can be written in the unique form

$$R_{270}^j \circ H^i, \qquad j = 0, 1, 2, 3, \quad i = 0, 1.$$

To verify this observe that

$$H \circ R_{270} = R_{270}^3 \circ H, \quad R_{270}^4 = I, \quad H^2 = I,$$

the so-called *relations*. For example,

$$\begin{aligned}
V \circ D' &= (R_{270}^2 \circ H) \circ (R_{270} \circ H) = R_{270}^2 \circ (H \circ R_{270}) \circ H\\
&= R_{270}^2 \circ R_{270}^3 \circ H \circ H = R_{270} \circ H^2 = R_{270} \circ I = R_{270}
\end{aligned}$$

Cosets

Let us determine by direct computation, without utilising general theory, what are the cosets with respect to the subgroup $S = \{I, H\}$. They have the form $x \circ S$, $x \in D_4$:

$$\begin{aligned}
I \circ \{I, H\} &= \{I, H\}, \\
R_{180} \circ \{I, H\} &= \{R_{180} \circ I, R_{180} \circ H\} = \{R_{180}, V\}, \\
R_{270} \circ \{I, H\} &= \{R_{270}, R_{270} \circ H\} = \{R_{270}, D'\}, \\
R_{90} \circ \{I, H\} &= \{R_{90}, R_{90} \circ H\} = \{R_{90}, D\}, \\
H \circ \{I, H\} &= \{H, I\}, \\
V \circ \{I, H\} &= \{V, R_{180}\}, \\
D \circ \{I, H\} &= \{D, R_{90}\}, \\
D' \circ \{I, H\} &= \{D', R_{270}\}.
\end{aligned}$$

Thus the left cosets of S are $\{I, H\}$, $\{R_{180}, V\}$, $\{R_{270}, D'\}$ and $\{R_{90}, D\}$. The group S is not normal because, for instance, $R_{270} \circ \{I, H\} = \{R_{270}, D'\}$ and $\{I, H\} \circ R_{270} = \{R_{270}, D\}$.

Finally we state that we have obtained a partition of four cosets for the group D_4 where each part has two elements. Since $2 \cdot 4 = 8$, as it should be. (By which theorem?)

Centre of D_4

We define the set $N = \{c \in D_4 \mid x \circ y = y \circ x \ \forall\ x \in D_4\}$. From the Cayley table we see that $N = \{I, R_{180}\}$. N is a normal subgroup because $y \circ N = N \circ y,\ \forall\ y \in D_4$. Now, the cosets are

$$\begin{aligned}
I \circ N = R_{180} \circ N &= \{I, R_{180}\}, \\
R_{270} \circ N = R_{90} \circ N &= \{R_{270}, R_{90}\}, \\
H \circ N = V \circ N &= \{H, V\}, \\
D \circ N = D' \circ N &= \{D, D'\}
\end{aligned}$$

and for example, $R_{270} \circ \{I, R_{180}\} = \{R_{270}, R_{90}\} = \{I, R_{180}\} \circ R_{270}$. Since $\{I, R_{180}\}$ is a normal subgroup, we can form the factor group

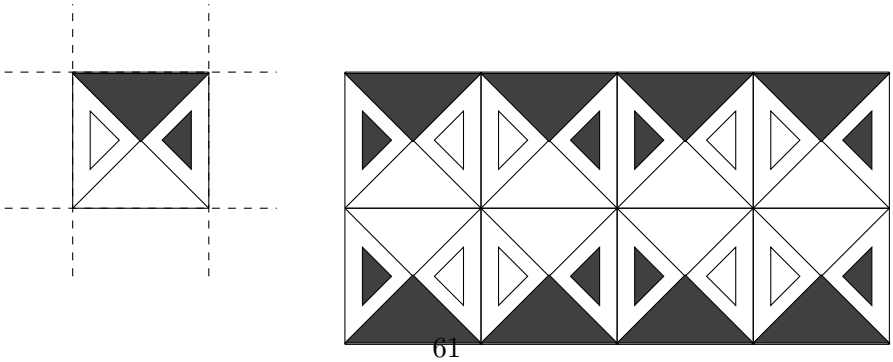
$$D_n/N = \{\{R_{270}, R_{90}\}, \{H, V\}, \{D, D'\}, \{I, R_{180}\}\}.$$

The group action is $(x \circ N) \circ (y \circ N) = (x \circ y) \circ N$.

For example, $\{R_{270}, R_{90}\} \circ \{H, V\} = (R_{270} \circ N) \circ (H \circ N) = (R_{270} \circ H) \circ N = D' \circ N = \{D, D'\}$. The projection $\pi \colon D_4 \to D_4/N$ is a homomorphism, $\pi(x) = x \circ N$ and $\mathrm{Ker}(\pi) = N$.

Ornaments

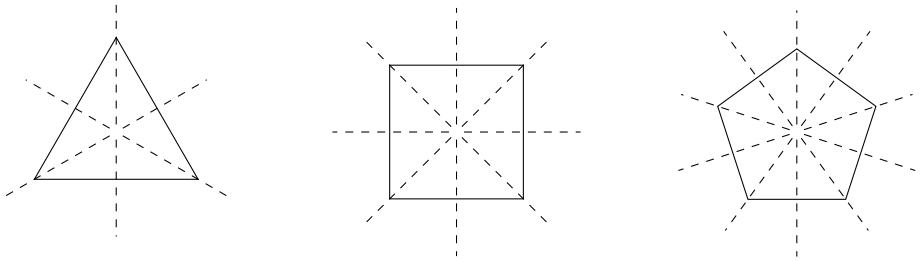
One of the uses of symmetry groups is ornamental decoration.



By reflecting over the edges of the square, we can tile the plane with images of the square as in the picture above. The reflection axes are marked with dashed lines. Observe that by combining the reflections with respect to the vertical and horizontal axes, we get a rotation by the angle π around a vertex of the square, previously $H \circ V = V \circ H = R_{180}$.

D_4 and C_n

D_n is the symmetry group of the regular n-sided polygon when $n \geq 3$. The group D_n has a cyclic subgroup C_n, which is generated by the rotation R by angle $2\pi/n$ around the centre point of the n-gon; that is, $R^n = I$. Furthermore, the reflections with respect to the symmetry axes of the n-gon are contained in the group D_n.



$n = 3$ $n = 4$ $n = 5$

If we denote a reflection with respect to some symmetry axis by H, then the relations $H^2 = I,\ R^n = I$ hold in the group D_n, and the maps H and R generate the group D_n.

Exercises

1. Let G be the symmetry group of the circle. Prove that for each positive integer n there exists an element of G whose order is n. Find some element of G whose order is infinite.

2. What maps belong to the group D_6? For what values of n does $C_n \leqslant D_6$ hold?

Chapter 4

Rings and integral domains

4.1 Rings

Groups are an algebraic system equipped with one operation. Below we describe an object equipped with two operations, a *ring*, whereof one typical example is the set of integers $\mathbb{Z}$ with addition and multiplication. Even in the general definition it is customary to denote and call the operations as addition and multiplication.

Definition 4.1. A set R equipped with two binary operations $+$ and $\cdot$ is called a *ring*, and denoted by $(R, +, \cdot)$, if

(R1) $(R, +)$ is an Abelian group;

(R2) multiplication $\cdot$ is defined on the set R;

(R3) $a(bc) = (ab)c \quad \forall\ a, b, c \in R$ (*associativity* of multiplication);

(R4) there exists an element 1 in R, the *identity*, such that $a \cdot 1 = 1 \cdot a = a\ \forall\ a \in R$;

(R5) $a(b + c) = ab + ac, \quad (a + b)c = ac + bc \quad \forall\ a, b, c \in R$ (*distributivity*)

If multiplication is commutative as well, that is, $ab = ba \quad \forall\ a, b \in R$, we say that R is a *commutative ring*.

The Postulates (R2), (R3) and (R4) together state that $(R, \cdot)$ is a monoid, see Remark 2.14. Sometimes Postulate (R4) is omitted from the definition. Then the ring defined above would be called a *ring with identity*.

The identity 1 is necessarily unique. This follows from Postulate (R4) similarly it does for groups. Postulate″(R1) yields an *additive identity* 0 and the additive inverse $-a$ for the element a. Furthermore, it follows from (R1) that addition is associative and commutative in a ring.

Example 4.2. $\mathbb{Z}, \mathbb{Q}, \mathbb{R}, \mathbb{C}$ are rings under usual addition and multiplication and have identity 1. Yet, for instance, $2\mathbb{Z}$ is not a ring.

The set $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ is also a ring under addition and multiplication, it is called the *Gaussian integers*. These number rings are all commutative.

Example 4.3. The set $\mathcal{M}_n(\mathbb{R})$ of $n \times n$ matrices with real entries forms a ring under addition and multiplication of matrices, and the identity is the identity matrix I_n. Other matrix rings are, for instance, $\mathcal{M}_n(\mathbb{Z})$, $\mathcal{M}_n(\mathbb{Q})$, $\mathcal{M}_n(\mathbb{C})$. You should check that these sets are closed under the operations. These mentioned matrix rings are noncommutative when $n > 1$.

Example 4.4. The set of functions

$$C[a, b] = \{f \colon [a, b] \to \mathbb{R} \mid f \text{ is continuous } \}$$

is a ring under pointwise addition and multiplication of functions:

$$(f+g)(x) = f(x) + g(x), \qquad (fg)(x) = f(x)g(x) \qquad \forall\ x \in [a,b].$$

The identity is the function $f(x) = 1 \quad \forall\ x \in [a,b]$. Remember that $C[a,b]$ is also a vector space. It follows that $(C[a,b],+)$ is an Abelian group; compare with Example 2.7.

The ring $C\,[a,b]$ is commutative. The definition can be generalised by replacing $\mathbb{R}$ with an arbitrary ring R, where commutativity depends on the commutativity of the ring R.

Example 4.5. The set of residue classes mod m, $\mathbb{Z}_m$, forms a ring under addition and multiplication of residue classes, and the identity is $\overline{1}$. $\mathbb{Z}_m$ is a finite commutative ring. It is also called the *quotient ring modulo m*. The term quotient ring will be introduced later in a more general sense.

Example 4.6. The power set $\mathcal{P}(S)$, the collection of all subsets, of a given set S forms a ring when addition is the *symmetric difference* of sets

$$A \triangle B = (A \setminus B) \cup (B \setminus A)$$

and multiplication is the intersection of sets $A \cap B$.

First, let us show that $(\mathcal{P}(S), \triangle)$ is an Abelian group, Postulate (R1). Suppose that $A, B, C \in \mathcal{P}(S)$, that is, A, B and C are subsets of S. As subsets their set-theoretic difference and union are subsets of S. Thus $A \triangle B \subset S$, so $A \triangle B \in \mathcal{P}(S)$ and we have closure.

The symmetric difference is commutative:

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (B \setminus A) \cup (A \setminus B) = B \triangle A.$$

It is also associative, which is easier to see by drawing a Venn diagram:

$$(A \triangle B) \triangle C = A \triangle (B \triangle C).$$

The empty set is the additive identity as

$$A \triangle \emptyset = (A \setminus \emptyset) \cup (\emptyset \setminus A) = A \cup \emptyset = A.$$

Finally, each set is its own inverse since

$$A \triangle A = (A \setminus A) \cup (A \setminus A) = \emptyset \cup \emptyset = \emptyset.$$

Hence $(\mathcal{P}(S), \triangle)$ is a commutative group.

Since $A, B \subset S$, their intersection is another subset of S and thus $A \cap B \in \mathcal{P}(S)$. Thus (R2) holds. Postulate (R3) follows from basic set operation properties:

$$(A \cap B) \cap C = A \cap (B \cap C).$$

The set S also belongs to $\mathcal{P}(S)$ and $S \cap A = A$ for any $A \subset S$, so S is the identity and Postulate (R4) holds. Finally, the distributivity postulate (R5) follows from basic set operation properties

$$\begin{aligned} A \cap (B \triangle C) &= A \cap ((B \setminus C) \cup (C \setminus B)) = (A \cap (B \setminus C)) \cup (A \cap (C \setminus B)) \\ &= ((A \cap B) \setminus (A \cap C)) \cup ((A \cap C) \setminus (A \cap B)) \\ &= (A \cap B) \triangle (A \cap C). \end{aligned}$$

Rings like this $(\mathcal{P}(S), \triangle, \cap)$ are called *Boolean rings*.

Example 4.7. Let $(G, +)$ be some Abelian group. The set

$$\mathrm{End}(G) = \{ f \colon G \to G \mid f \text{ is a group homomorphism } \}$$

forms a ring under pointwise addition and composition of maps:

$$(f+g)(a) = f(a) + g(a), \qquad (f \circ g)(a) = f(g(a)) \quad \forall\, a \in G.$$

The identity is the identity map id_G. Checking the postulates is a good exercise!

A group homomorphism $G \to G$ is called an *endomorphism* of the group G. The notation for the ring $\mathrm{End}(G)$ and the name *endomorphism ring* is due to this. This type of ring is usually not commutative.

The set with one element $R = \{a\}$ forms a ring trivially when we define $a + a = a$ and $aa = a$. Then a is the additive identity R, and hence this ring is also called the *zero ring*.

Henceforth we assume, unless otherwise stated, that the ring R is not the zero ring.

Definition 4.8. We call an element u of a ring R a *unit* if u has an inverse element u^{-1}, that is, if $\exists\ u^{-1} \in D :\ uu^{-1} = u^{-1}u = 1$. The set of all units is denoted by R^*.

Theorem 4.9. *$(R^*, \cdot)$ is a group, the unit group of R.*

Proof. $R^* \neq \emptyset$ because 1 is a unit with inverse element 1. The set R^* is closed under multiplication because the product uv of units u and v has the inverse element $v^{-1}u^{-1}$. The product is associative in the whole R and therefore also in R^*. If $u \in R^*$ then its inverse $u^{-1} \in R^*$ since $\left(u^{-1}\right)^{-1} = u$. □

Example 4.10. The rings $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ have unit groups $\mathbb{Q}^* = \mathbb{Q} \setminus \{0\}$, $\mathbb{R}^* = \mathbb{R} \setminus \{0\}$, $\mathbb{C}^* = \mathbb{C} \setminus \{0\}$. These are the familiar groups introduced in group theory, likewise the unit group of the factor ring $\mathbb{Z}_m$ $(m \geq 2)$ that is $\mathbb{Z}_m^* = \{\overline{a} \in \mathbb{Z}_m \mid \gcd(a, m) = 1\}$.

Some other examples of unit groups:

$$\mathbb{Z}^* = \{-1, +1\}, \qquad \mathcal{M}_n(R)^* = GL_n(\mathbb{R}) = \{A \in \mathcal{M}_n(\mathbb{R}) \mid \det(A) \neq 0\}.$$

Exercises

1. The set $\{0_6, 2_6, 4_6\}$ has an multiplicative identity. What is it?

2. We define two operations on $\mathbb{Z}$ by

   $$x \oplus y = x + y + 1, \qquad x \odot y = x + y - xy.$$

   Prove that $(\mathbb{Z}, \oplus, \odot)$ is a ring.

3. Show that the set $\{m + n\sqrt{2} \mid m, n \in \mathbb{Z}\}$ is a commutative ring under the usual addition and multiplication of numbers.

4. Show that the multiplication

   $$(a, b, c)(x, y, z) = (ax, bx + cy, cz)$$

   extends the group $(\mathbb{Z}^*, +)$ to a noncommutative ring with identity $(1, 0, 1)$.

5. Let $(R_1, +, \cdot)$ and $(R_2, +, \cdot)$ be rings. Prove that $(R_1 \times R_2, +, \cdot)$ is a ring when we define

$$(x_1, x_2) + (y_1, y_2) = (x_1 + y_1, x_2 + y_2),$$
$$(x_1, x_2)(y_1, y_2) = (x_1 y_1, x_2 y_2).$$

6. Determine the units of the ring $\mathbb{Z}[i]$. Is this unit group cyclic?

4.2 Ring arithmetic

Because $(R, +)$ is a group, Abelian moreover, addition in the group obeys the rules familiar from group theory. As is always the case for additive groups, we have the *difference* of elements a and b: $a - b = a + (-b)$.

Multiplication rules can also be inferred from group theory. Observe that the negative powers of an element a are only defined when an inverse element a^{-1} exists.

Now we outline the necessary rules on how multiplication and addition in a ring are connected. The foundation is distributivity (R5). Firstly, we get by induction

$$a(b_1 + \cdots + b_n) = ab_1 + \cdots + ab_n,$$
$$(a_1 + \cdots + a_m)b = a_1 b + \cdots + a_m b,$$

and more generally

$$(a_1 + \cdots + a_m)(b_1 + \cdots + b_n) = a_1 b_1 + a_1 b_2 + \cdots + a_m b_n.$$

Theorem 4.11. *If R is a ring and $a, b, c \in R$, then*

(i) $0 \cdot a = a \cdot 0 = 0$,

(ii) $a(-b) = (-a)b = -(ab)$, $(-a)(-b) = ab$,

(iii) $a(b - c) = ab - ac$, $(a - b)c = ac - bc$.

Proof. (i) By writing $0 \cdot a = (0 + 0)a = 0 \cdot a + 0 \cdot a$ and subtracting $0 \cdot a$ from both sides, we end up with the equation $0 = 0 \cdot a$. The statement $a \cdot 0 = 0$ is proved similarly.

(ii) Because $ab + a(-b) = a(b + (-b)) = a \cdot 0 = 0$, the product $a(-b)$ is the additive inverse of ab. Likewise, we see that $(-a)b$ is also the additive inverse of ab. From the previous equations we get $(-a)(-b) = -((-a)b) = -(-(ab)) = ab$.

(iii) $a(b - c) = ab + a(-c) = ab + (-(ac)) = ab - ac$. The latter statement follows similarly. □

Remark 4.12. Theorem 4.11 can also be proved by considering the maps

$$\rho \colon R \to R, \quad \rho(a) = ab,$$
$$\tau \colon R \to R, \quad \tau(a) = ba,$$

where $b \in R$ is fixed. The maps ρ and τ are homomorphisms – endomorphisms to be precise – of the group $(R, +)$, and the statements (i)–(iii) follow from the basic properties of homomorphisms.

Remark 4.13. It follows from (i) that $1 \neq 0$ always in a ring. If we had $1 = 0$ then $a = a \cdot 1 = a \cdot 0 = 0 \ \forall\ a \in R$, and remember that we excluded $R = \{0\}$.

By (ii) there is no ambiguity in writing $-ab$; the minus sign can be thought of as either part of the product ab or part of element a.

As we stated in group theory, multiples obey the rules of operation:

$$(m+n)a = ma + na, \quad (mn)a = m(na), \quad n(a+b) = na + nb \quad \forall\ m, n \in \mathbb{Z},\ a, b \in R.$$

Observe that the notations $0a$ and $1a$ can be understood in two ways but in either case $0a = 0$ and $1a = a$. For the sake of clarity we may also denote 0_R and 1_R for the additive identity and (multiplicative) identity of the ring R.

Theorem 4.14. *If R is a ring and $a, b \in R$, then*

(iv) $na = (n \cdot 1)a = a(n \cdot 1) \quad \forall\ n \in \mathbb{Z} \quad$ (here $1 = 1_R$),

(v) $n(ab) = (na)b = a(nb) \quad \forall\ n \in \mathbb{Z}$,

(vi) $(ma)(nb) = (mn)(ab) \quad \forall\ m, n \in \mathbb{Z}$.

Proof. (iv) If $n = 0$, then all three products are equal to zero. Suppose $n > 0$. Then

$$(n \cdot 1)a = (1 + \cdots + 1)a = a + \cdots + a = na.$$

Further, $(-n) \cdot 1 = n(-1)$ and $(-n)a = n(-a)$ by the definition of negative multiples; that is,

$$((-n) \cdot 1)a = ((-1) + \cdots + (-1))a = (-a) + \cdots + (-a) = n(-a) = (-n)a.$$

The products $a(n \cdot 1)$ and $a((-n) \cdot 1)$ are handled similarly.

(v) easily reduces to (iv) and part (vi) reduces to (v).

□

This theorem as well can be proven by using the homomorphisms ρ and τ from Remark 4.12.

Example 4.15. $(a+b)^2 = a^2 + ab + ba + b^2$. If R is commutative, then $(a+b)^2 = a^2 + 2ab + b^2$ and in general

$$(a+b)^n = a^n + \binom{n}{1} a^{n-1}b + \cdots + \binom{n}{n-1} ab^{n-1} + b^n.$$

Example 4.16. In the ring $\mathbb{Z}_m$ we have $k\overline{a} = \overline{k} \cdot \overline{a}$ by Theorem 4.14 (iv) for all $k, a \in \mathbb{Z}$. In the ring $\mathbb{Z}_2$ we have $(\overline{a} + \overline{b})^2 = (\overline{a})^2 + (\overline{b})^2$ since $2\overline{ab} = \overline{2} \cdot \overline{ab} = \overline{0} \cdot \overline{ab} = 0$.

Exercises

1. Prove that the formula $x^2 - y^2 = (x+y)(x-y)$ holds in commutative rings.
2. Prove that a ring R is commutative if and only if $(x+y)^2 = x^2 + y^2 + 2xy \ \forall a, y \in R$.
3. Suppose that for a ring $(R, +, \cdot)$ we have $x^2 = x \ \forall x \in R$. Show that $2x = 0 \ \forall x \in R$ and that R is commutative.
4. Which of the following equations hold in an arbitrary ring R?

a) $a^2 - ba = (a - b)a$,

b) $(a + b + 1)(a - b - 1) = a^2 - b^2 - 2b - 1$,

c) $2a \cdot 4b - ab = 7ab$.

4.3 Subrings and ideals

Definition 4.17. A subset S of a ring $(R, +, \cdot)$ is called a *subring* of R if

(SR1) S is a ring under the operations $+$ and $\cdot$, and

(SR2) the identity of S is the identity of R.

If S is a subring of R, then $(S, +)$ is a subgroup of the group $(R, +)$. It follows that the additive identity 0 of R belongs to S and is the additive identity of S.

Example 4.18. $\mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$, here each ring is a (proper) subring of those following it.

Example 4.19. The set of all matrices $\begin{pmatrix} a & 0 \\ 0 & 0 \end{pmatrix}$, where $a \in \mathbb{R}$, forms a ring that is contained in the matrix ring $\mathcal{M}_2(\mathbb{R})$. Yet it is not a subring of $\mathcal{M}_2(\mathbb{R})$ because its identity is $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \neq \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Theorem 4.20 (Subring criterion)**.** *Let R be a ring and $S \subset R$. Then S is a subring of R if and only if the following conditions hold:*

(a) $1_R \in S$,

(b) $a - b \in S \quad \forall\ a, b \in S$,

(c) $ab \in S \quad \forall\ a, b \in S$.

Proof. If S is a subring of R, then it trivially satisfies the conditions. Conversely, by condition (a) we know that $S \neq \emptyset$ and, hence it follows from condition (b) that $(S, +)$ is a subgroup of $(R, +)$. Conditions (a) and (c) ensure that S satisfies the ring postulates (R2) and (R4). The necessary rules of computation in postulates (R3) and (R5) are inherited from the ring R. Thus (SR1) holds, and (SR2) follows directly from condition (a). □

It is often easier to show that a given subset is a ring by using the Subring criterion than by checking the ring postulates.

Example 4.21. The number sets

$$\mathbb{Z}\left[\sqrt{n}\right] = \{a + b\sqrt{n} \mid a, b \in \mathbb{Z}\} \qquad (n = -1, \pm 2, \pm 3, \ldots)$$

are subrings of $\mathbb{C}$, and also subrings of $\mathbb{R}$ when $n > 0$. In the special case $n = -1$ we have the Gaussian integers, see Example 4.2.

When we study the rings $\mathbb{Z}\left[\sqrt{n}\right]$, we usually assume that n is *square free*, that is, n is not divisible by the square of any integer greater than 1. Then $\mathbb{Z}\left[\sqrt{n_1}\right] \neq \mathbb{Z}\left[\sqrt{n_2}\right]$ holds whenever $n_1 \neq n_2$.

Example 4.22. The set of all polynomials with real coefficients

$$\mathbb{R}[x] = \{a_0 + a_1x + \cdots + a_nx^n \mid n \geq 0,\ a_k \in \mathbb{R}\ (k = 0, \ldots, n)\}$$

is a ring under addition and multiplication of polynomials, namely a subring of the function ring

$$C(\mathbb{R}) = \{f \colon \mathbb{R} \to \mathbb{R} \mid f \text{ is continuous }\},$$

see Example 4.4.

Some other examples of polynomial rings are the subrings $\mathbb{Z}[x]$ and $\mathbb{Q}[x]$ of the ring $\mathbb{R}[x]$. Later we will consider the general polynomial ring $R[x]$ where R is any commutative ring.

Example 4.23. Let V be a (real) vector space. The set of linear maps

$$\mathrm{End}_{\mathbb{R}}(V) = \{t \colon V \to V \mid t \in \text{is linear }\}$$

is a subring of $\mathrm{End}(V)$. Here $\mathrm{End}(V)$ denotes the endomorphisms of the group $(V, +)$, see Example 4.7; the addition of maps is defined pointwise and the product is composition.

In group theory we stated that normal subgroups enjoy a special status. In ring theory, normal subgroups correspond to *ideals*.

Definition 4.24. A subset I of a ring R is called an *ideal* of R if

(I1) $(I, +)$ is a subgroup of the group $(R, +)$ and

(I2) $ra \in I$ and $ar \in I\ \ \forall\ r \in R$ and $a \in I$.

If we omit the condition $ar \in I$ from (I2) then we would get a more general concept of a *left ideal*; likewise a *right ideal*.

Every ring has the trivial ideals R itself and the zero ideal $\{0\}$. Is an ideal I of a ring also its subring? If so, then $1 \in I$ and by (I2) we have $r \in I \quad \forall\ r \in R$, that is, $I = R$. Therefore R itself is the only ideal of R that is also a subring.

The previous inference also gives the following: If I is a proper ideal of a ring R, then $I \cap R^* = \emptyset$. If a unit $u \in I$, then the product $uu^{-1} = 1$ also belongs to I, and hence $I = R$.

Example 4.25. In Section 3.3 we showed that the only subgroups of the group $(\mathbb{Z}, +)$ are the groups $m\mathbb{Z},\ m = 0, 1, \ldots$. Because the groups $m\mathbb{Z}$ also satisfy (I2), they are ideals of the ring $\mathbb{Z}$. These are thus *all* the ideals of the ring $\mathbb{Z}$.

Theorem 4.26 (Ideal criterion)**.** *Let R be a ring and $I \subset R$. Then I is an ideal of R if and only if the following hold:*

(a) $I \neq \emptyset$,

(b) $a - b \in I \quad \forall\ a, b \in I$,

(c) $ra \in I$ *and* $ar \in I \quad \forall\ r \in R,\ a \in I$.

Proof. The conditions (a) and (b) are equivalent to (I1) and the condition (c) is the same as (I2). □

Example 4.27. The set of polynomials with real coefficients, whose constant terms are 0, form an ideal of the polynomial ring $\mathbb{R}[x]$. Let us denote this set by

$$I = \{f \in \mathbb{R}\,[x] \mid f(0) = 0\}.$$

It is easy to see that I is nonempty since at least the polynomial $f(x) = x$ belongs to I. Suppose that $f, g \in I$. Then $(f - g)(0) = f(0) - g(0) = 0 - 0 = 0$, so $f - g \in I$ and thus (I1) holds. Finally, let $h \in \mathbb{R}[x]$. Then we have $f(0)h(0) = 0 \cdot h(0) = 0$ and $h(0)f(0) = h(0) \cdot 0 = 0$, and we get $fh \in I$ and $hf \in I$. Since these hold for any $f, g \in I$ and $h \in \mathbb{R}[x]$, (I2) holds. Hence I is an ideal of $\mathbb{R}[x]$ by the Ideal criterion.

Theorem 4.28. *If I and J are ideals of a ring R, then so are their intersection $I \cap J$ and sum*

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

This hold even for more than two ideal, and for intersections even for infinitely many.

Proof. This follows immediately from the Ideal criterion. □

Similarly to how a subset of a group G generates a subgroup of G, a subset S of a ring R generates an ideal of R

$$\langle S \rangle = \bigcap_{S \subset I} I;$$

where I is an ideal of R. Here we used Theorem 4.28. The ideal $\langle S \rangle$ is the smallest ideal of R that contains S.

If S is a finite set, $S = \{a_1, \ldots, a_k\}$, then the ideal $\langle S \rangle$ is said to be *finitely generated*, and denoted by $\langle S \rangle = \langle a_1, \ldots, a_k \rangle$. An ideal that is generated by one element is called a *principal ideal*.

Observe that

$$\langle a_1, \ldots, a_k \rangle = \langle b_1, \ldots, b_h \rangle,$$

if both ideals contain the generators of each other. For example, by the above it follows that $\langle a_1, \ldots, a_k \rangle \subset \langle b_1, \ldots, b_k \rangle$ if $a_i \in \langle b_1, \ldots, b_j \rangle$ $(i = 0, \ldots, k)$.

Example 4.29. The trivial ideals R and $\{0\}$ are principal ideals: $R = \langle 1 \rangle$ and $\{0\} = \langle 0 \rangle$.

Theorem 4.30. *If the ring R is commutative, then*

$$\langle a_1, \ldots, a_k \rangle = \{r_1 a_1 + \cdots + r_k a_k \mid r_i \in R \quad \forall\ i\}.$$

Proof. The proof is similar to the proof of Theorem 2.31. It is essential that the right-hand side is, firstly, an ideal by the Ideal criterion, and secondly, that it is contained in every ideal that the elements $a_1, \ldots, a_k$ are contained in. □

Using a similar notation as we did for cosets with respect to subgroups, when R is commutative we can write

$$\langle a_1, \ldots, a_k \rangle = Ra_1 + \cdots + Ra_k, \qquad \text{in particular } \langle a \rangle = Ra.$$

The notation $Ra_1 + \cdots + Ra_k$ can also be interpreted as the sum of the principal ideals Ra_i.

Example 4.31. The principal ideal generated by the element x of the polynomial ring $\mathbb{R}[x]$ is

$$\langle x \rangle = x\mathbb{R}[x] = \{a_0 x + a_1 x^2 + \cdots + a_n x^{n+1} \mid n \geq 0,\ a_i \in \mathbb{R} \quad \forall\ i\},$$

which is the same ideal as the ideal that appeared in Example 4.27.

Example 4.32. By Example 4.25, the ideals of the ring $\mathbb{Z}$ are the principal ideals

$$\langle m \rangle = m\mathbb{Z} \quad (m = 0, 1, \ldots).$$

Definition 4.33. A ring whose every ideals is a principal ideal is called a *principal ideal ring*, abbreviated by PIR.

Example 4.34. Example 4.32 shows that $\mathbb{Z}$ is a PIR. Thus if $a_1, \ldots, a_k \in \mathbb{Z}$, then there exists a $d \in \mathbb{Z}$, such that

$$\langle a_1, \ldots, a_k \rangle = \langle d \rangle.$$

How can we determine this d? The answer is simple: $d = \gcd(a_1, \ldots, a_k)$. To show this, let us denote $s_i = \frac{a_i}{d}$ for $1 \leq i \leq k$. Each element of $\langle a_1, \ldots, a_k \rangle$ can be written in the form

$$a_1 b_1 + \cdots + a_k b_k = d s_1 b_1 + \cdots + d s_k b_k = d(s_1 b_1 + \cdots + s_k b_k) \in \langle d \rangle$$

where $b_1, \ldots, b_k \in \mathbb{Z}$. Therefore $\langle a_1, \ldots, a_k \rangle \subset \langle d \rangle$.

Conversely, any element $db \in d\mathbb{Z}$ can be written as $db = a_1 b_1 + \cdots + a_k b_k$ because db is a multiple of $\gcd(a_1, \ldots, a_k)$ by definition. Remember that a Diophantine equation $a_1 x_1 + \cdots a_n x_n = c$ has a solution if and only if c is a multiple of $\gcd(a_1, \ldots, a_n)$. Therefore $\langle d \rangle \subset \langle a_1, \ldots, a_k \rangle$.

Thus we get equality $\langle a_1, \ldots, a_k \rangle = \langle d \rangle$.

As an example, $\langle 3, 4 \rangle = \langle 1 \rangle = \mathbb{Z}$ and $\langle 4, 6 \rangle = \langle 2 \rangle = 2\mathbb{Z}$.

Example 4.35. Let $R = \{f \mid f \colon \mathbb{R} \to \mathbb{R}\}$ and $S = \{f \in R \mid f \text{ is differentiable }\}$. Then S is a subring of R but not an ideal.

Exercises

1. Show that $\{0_6, 2_6, 4_6\}$ is a ring, but not a subring of $\mathbb{Z}_6$.
2. Let R be a ring and $C(R) = \{x \in R \mid xy = yx \ \forall y \in R\}$. Prove that $C(R)$ is a subring of R.
3. Prove that $\mathbb{R}^{\mathbb{R}}$ is a ring and that the differentiable functions $\mathbb{R} \to \mathbb{R}$ generate a subring of it. (Like earlier, $\mathbb{R}^{\mathbb{R}} = \{f \mid f \colon \mathbb{R} \to \mathbb{R}\}$ and the operations are like in Example 4.4.)
4. Prove that
$$I = \{f \in \mathbb{R}^{\mathbb{R}} \mid f(1) = 0\}$$
is an ideal of the ring $\mathbb{R}^{\mathbb{R}}$.
5. Let A and B be ideals of a ring R. Prove that $A + B$ is also an ideal of R.
6. Let A and B be ideals of a ring R. Prove that $A \cap B$ is also an ideal of R.
7. Determine the ideals of the ring $\mathbb{Z}_{12}$.
8. Show that the set $\{a + b\sqrt[3]{2} + c\sqrt[3]{4} \mid a, b, c, \in \mathbb{Z}\}$ is a subring of $\mathbb{R}$.
9. Show that M and $\{0_m\}$ are the only ideals of the matrix ring
$$M = \{\begin{pmatrix} a & b \\ c & d \end{pmatrix} \mid a, b, c, d \in \mathbb{R}\}.$$
(Hint: If $I \neq \{0_m\}$ is an ideal of M, then show that $1_m \in I$.)
10. Prove that the Gaussian integers $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$ form a subring of the ring of complex numbers $\mathbb{C}$.
11. Prove that the set $\{\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix} \mid a, b \in \mathbb{Z}\}$ is a subring of the matrix ring $\mathcal{M}_2(\mathbb{Z})$.

4.4 Quotient rings

Earlier, using the normal subgroup N of a group G, we defined the factor group G/N of G. An analogous concept for a ring R is the *quotient ring*, or factor ring, R/I where I is an ideal of R.

Suppose that I is an ideal of R. Then $(I, +)$ is a normal subgroup of $(R, +)$, since normality follows from the commutativity of the group $(R, +)$. Thus we can form a factor group

$$\begin{aligned} R/I =& \{a + I \mid a \in R\} = \{a + I \mid a \in D\}; \\ & (a + I) + (b + I) = (a + b) + I, \end{aligned}$$

where D is some collection of representative residue classes $a + I$ (= cosets, see Example 2.59).

Theorem 4.36. *If I is an ideal of a ring R, then R/I is a ring under the following operations:*

$$(a + I) + (b + I) = (a + b) + I, \quad (a + I)(b + I) = ab + I.$$

Proof. As we noted above $(R/I, +)$ is a group. It is moreover an Abelian group because $(R, +)$ is Abelian by definition.

Second, we prove that the multiplication of residue classes as defined in the statement is well defined. Suppose that $a + I = a' + I$ and $b + I = b' + I$, that is, $a = a' + i_1$ and $b = b' + i_2$ with $i_1, i_2 \in I$. Then

$$ab = (a' + i_1)(b' + i_2) = a'b' + a'i_2 + i_1b' + i_1i_2.$$

Because I is an ideal of R, it contains the products $a'i_2, i_1b'$ and i_1i_2 and thus their sum. Hence $ab = a'b' + i$ for some $i \in I$. This can also be stated as $ab \in a'b' + I$ and implies that $ab + I = a'b' + I$.

The ring postulates (R3)–(R5) revert to the equivalent postulates in R by the definitions of addition and multiplication of residue classes. The identity of the ring R/I is $1 + I$. □

Definition 4.37. The ring R/I is called the *quotient ring* with respect to ideal I. It is also called a *factor ring* or a *residue class ring*.

Remember that for all $a, b \in R$ we have

$$a + I = b + I \qquad \Longleftrightarrow \qquad a \in b + I \qquad \Longleftrightarrow \qquad a - b \in I.$$

Make sure that you understand the following facts of R/I: the additive identity is I $(= 0 + I)$, the (multiplicative) identity is $1 + I$, the additive inverse of $a + I$ is $-a + I$, and if a is a unit, the multiplicative inverse is $a^{-1} + I$. If the ring R is commutative, then so is R/I.

Example 4.38. The quotient ring of the ring $\mathbb{Z}$ with respect to the ideal $m\mathbb{Z}$ is

$$\mathbb{Z}/m\mathbb{Z} = \{a + m\mathbb{Z} \mid a \in \mathbb{Z}\} = \{\overline{0}, \overline{1}, \ldots, \overline{m-1}\},$$

$$\overline{a} + \overline{b} = \overline{a + b}, \qquad \overline{a} \cdot \overline{b} = \overline{ab}.$$

Thus is the familiar ring $\mathbb{Z}_m$, the residue class ring mod m. In the case of $m = 1$ we get the zero ring $\mathbb{Z}/\mathbb{Z} = \{\overline{0}\}$. In general as well $R/R = \{0\}$ of course.

Exercises

1. Form the residue classes of the ideal $I = \langle\sqrt{2}\rangle$ in the ring $\mathbb{Z}\left[\sqrt{2}\right]$.
2. Let $\langle 2_4\rangle$ be the ideal generated by the element 2_4 in the residue class ring $\mathbb{Z}_4$. Form the quotient ring $\mathbb{Z}_4/\langle 2_4\rangle$.
3. Form the quotient ring $\mathbb{Z}\left[\sqrt{10}\right]/I$, where $\mathbb{Z}\left[\sqrt{10}\right] = \{a + b\sqrt{10} \mid a, b \in \mathbb{Z}\}$ and $I = \langle 5, \sqrt{10}\rangle$.

4.5 Ring homomorphisms and isomorphisms

Definition 4.39. Let R and R' be rings. A map $f\colon R \to R'$ is called a *(ring) homomorphism* if it satisfies:

(RH1) $f(a + b) = f(a) + f(b) \quad \forall\ a, b \in R$,

(RH2) $f(ab) = f(a)f(b) \quad \forall\ a, b \in R$,

(RH3) $f(1_R) = 1_{R'}$.

According to the definition, a map $f\colon R \to R'$ is a ring homomorphism if and only if f is a group homomorphism $(R, +) \to (R', +)$ and it has properties (RH2) and (RH3). It follows by the properties of group homomorphisms that a ring homomorphism $f\colon R \to R'$ satisfies

$$f(0_R) = 0_{R'}, \qquad f(-a) = -f(a) \quad \forall\ a \in R.$$

Moreover, it follows from (RH2) and (RH3) that

$$f(a^{-1}) = f(a)^{-1} \quad \forall\ a \in R^*$$

since $f(a)f(a^{-1}) = f(aa^{-1}) = f(1_R) = 1_{R'}$, and likewise $f(a^{-1})f(a) = 1_{R'}$.

Example 4.40. The identity map id_R is a homomorphism $R \to R$. The zero map $f(a) = 0$ for all $a \in R$ is not a homomorphism as it does not satisfy (R4).

Example 4.41. The map $f\colon \mathbb{R}[x] \to \mathbb{R}, \quad f(a_0 + a_1x + \cdots + a_nx^n) = a_0$ is a ring homomorphism. Is the map $g(a_0 + a_1x + \cdots + a_nx^n) = a_0 + a_1 + \cdots + a_n$ also a ring homomorphism?

The next theorem gives analogous results to Theorems 2.47 and 3.13 in group theory.

Theorem 4.42. *Let $f\colon R \to R'$ be a ring homomorphism.*

(i) *If S is a subring of R, then $f(S)$ is a subring of R'.*

(ii) *If S' is a subring of R', then $f^{-1}(S')$ is a subring of R.*

(iii) *If I is an ideal of R, then $f(I)$ is an ideal of the ring $f(R)$.*

(iv) *If I' is an ideal of R', then $f^{-1}(I')$ is an ideal of R.*

Proof. These are proven in a straightforward manner using the Subring criterion, the Ideal criterion and the conditions (RH1)–(RH3). For proving (iii) and (iv) we can use the definition of an ideal and Theorem 2.47 instead of the Ideal criterion. □

In particular, from (iv) we get that the *kernel* of a ring homomorphism

$$\mathrm{Ker}(f) = f^{-1}(\{0\}) = \{a \in R \mid f(a) = 0\}$$

is an ideal of the ring R. Similarly, from (i) we get that the *homomorphic image* of the ring R

$$\mathrm{Im}(f) = f(R) = \{f(a) \mid a \in R\}$$

is a ring, namely a subring of R'.

If we consider the map f as just a group homomorphism $(R, +) \to (R', +)$, its kernel and image are the same as stated above. In particular, note that the kernel is defined with 0 and not 1! Thus from Theorem 2.51 it follows that a ring homomorphims $f\colon R \to R'$ is injective if and only if $\mathrm{Ker}(f) = \{0\}$.

Example 4.43. The kernel and the image of the map f in Example 4.41 are

$$\mathrm{Ker}(f) = \{a_0 + a_1x + \cdots + a_nx^n \in \mathbb{R}[x] \mid a_0 = 0\} = \{a_1x + \cdots + a_nx^n \in \mathbb{R}[x]\},$$

$$\mathrm{Im}(f) = \mathbb{R}.$$

Definition 4.44. A ring homomorphism $f\colon R \to R'$ is called a *(ring) isomorphism* if f is bijective. We say that the rings R and R' are *isomorphic*, and denoting $R \simeq R'$, if some isomorphism $R \to R'$ exists.

Remark 4.45. More terminology related to homomorphisms, some of which have been mentioned earlier:

monomorphism = an injective homomorphism,

epimorphism = a surjective homomorphism,

endomorphism = a homomorphism from an object to itself,

automorphism = an isomorphism from an object to itself.

Example 4.46. The map $f\colon \mathbb{C} \to \mathbb{C}, \quad f(x + iy) = x - iy$ is an automorphism of ring $\mathbb{C}$. First, let us check that f is a ring homomorphism.

$$f(x + iy) + f(u + iv) = x - iy + u - iv = x + u - i(y + v) = f(x + u + i(u + v)),$$

$$\begin{aligned} f((x + iy)(u + iv)) &= f(xu - yv + i(xv + yu)) = xu - yv - i(xv + yu) \\ &= (x - iy)(u - iv) = f(x + iy)f(u + iv), \end{aligned}$$

$$f(1) = 1$$

Thus f is a homomorphism by Definition 4.39.

Since $f(x + iy) = 0$ implies $x = y = 0$, the kernel is $\mathrm{Ker}(f) = \{0\}$, and thus f is injective. As an injective endomorphism, f is necessarily surjective, and hence a ring isomorphism, and further, an automorphism of $\mathbb{C}$.

The analogy between group and ring homomorphisms also applies to Theorems 2.53 and 2.54, which can easily be extended to ring homomorphisms. Consequently, ring isomorphisms are equivalence relations. Isomorphic rings can thus be equated in the view of ring theory.

Theorem 4.47 (Homomorphism theorem for rings)**.** *If $f\colon R \to R'$ is a ring homomorphism then*

$$R/K \simeq \mathrm{Im}(f) \qquad (K = \mathrm{Ker}(f)).$$

More precisely, f induces a ring isomorphism

$$F\colon R/K \to \mathrm{Im}(f), \quad F(a + K) = f(a).$$

Proof. By the Homomorphism theorem for groups, we know that the map F is a group isomorphism $(R/K,+) \to (\mathrm{Im}(f),+)$. Thus it needs only be proven that F satisfies the ring homomorphism conditions (RH2) and (RH3).

(RH2): $F((a+K)(b+K)) = F(ab+K) = f(ab) = f(a)f(b) = F(a+K)F(b+K)$.

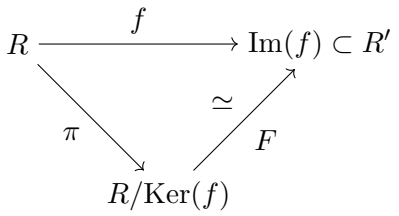
(RH3): $F(1+K) = f(1) = 1$.

□

Just like for groups, the Homomorphism theorem gives the commuting diagram below, where π is the *(canonical) projection*

$$\pi\colon R \to R/I, \quad \pi(a) = a + I,$$

when $I = \mathrm{Ker}(f)$. Check that π truly is a ring homomorphism. Because π is surjective in addition, we see that the homomorphic images of a ring R correspond bijectively to the quotient rings of R.



Example 4.48. Let us determine what isomorphism the homomorphism f in Example 4.41 gives. In Example 4.43 we determined the kernel and image of f. The kernel is

$$K = \mathrm{Ker}(f) = \{a_1 x + \cdots + a_n x^n \in \mathbb{R}[x]\}$$

and the image is $\mathrm{Im}(f) = \mathbb{R}$. Thus the induced ring isomorphism is

$$F\colon \mathbb{R}[x]/K \simeq \mathbb{R}, \qquad F(a+K) = f(a).$$

Example 4.49. Let us show that the map $f\colon \mathbb{Z} \to \mathbb{Z}_m, \quad f(a) = \overline{a}$ is a ring homomorphism for $m \geq 2$. We already know that $(\mathbb{Z},+) \to (\mathbb{Z}_m,+)$ is a group homomorphism so we need to check (RH2) and (RH3).

Firstly, $f(ab) = \overline{ab} = \overline{a}\overline{b} = f(a)f(b)$, and secondly, $f(1) = \overline{1}$, so f is a ring homomorphism. The kernel of f is $m\mathbb{Z}$ and its image is $\{\overline{0}, \overline{1}, \ldots, \overline{m-1}\}$. The isomorphism that f induces is thus

$$F\colon \mathbb{Z}/m\mathbb{Z} \simeq \mathbb{Z}_m \quad F(a+m\mathbb{Z}) = f(a) = \overline{a}.$$

Exercises

1. Show that the condition $x_4 \mapsto (5x)_{10},\ x \in \mathbb{Z}$ gives a well-defined map $\mathbb{Z}_4 \to \mathbb{Z}_{10}$. Is it a ring homomorphism?

2. Suppose that there exists a surjective homomorphism from a commutative ring R to a ring R'. Prove that R' is also commutative.

3. Let $(\mathbb{Z}, \oplus, \odot)$ be a ring under the operations defined as $x \oplus y = x + y - 1$ and $x \odot y = x + y - xy$. Prove that $(\mathbb{Z}, \oplus, \odot)$ and $(\mathbb{Z}, +, \cdot)$ are isomorphic.

4. Let X be a set and $Y \subset X$. Prove that the map between Boolean rings (see Example 4.6) $f\colon \mathcal{P}(X) \to \mathcal{P}(Y),\ f(A) = A \cap Y$ is a ring homomorphism. Determine $\mathrm{Ker}(f)$.

5. Determine all ring homomorphisms $f\colon \mathbb{Z} \to \mathbb{Z}$ and $g\colon \mathbb{Q} \to \mathbb{Q}$.

6. The set $A = \left\{ \begin{pmatrix} a & b \\ 0 & d \end{pmatrix} \mid a, b, d \in \mathbb{R} \right\}$ is a ring. Show that the set $I = \left\{ \begin{pmatrix} 0 & b \\ 0 & 0 \end{pmatrix} \mid b \in \mathbb{R} \right\}$ is an ideal of A and that the quotient ring A/I is isomorphic to $\mathbb{R} \times \mathbb{R}$. (Hint: The Homomorphism theorem for rings.)

7. Let $f\colon R \to R$ be a ring homomorphism. Prove that if I is an ideal of R, then $f(I)$ is an ideal of the ring $f(R)$.

8. Does a map $f\colon 2\mathbb{Z} \to 3\mathbb{Z}$ which satisfies conditions (RH1) and (RH2) exist? Does such a bijection exist?

9. Determine whether the map $f\colon \mathbb{Z}_4 \to \mathbb{Z}_{12},\ x_4 \mapsto (3x)_4$ $(\forall x \in \mathbb{Z})$ is well defined. Does it satisfy the conditions (RH1) and (RH2)?

10. Prove that $\mathbb{Z}\left[\sqrt{2}\right] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$ and $H = \left\{ \begin{pmatrix} a & 2b \\ b & a \end{pmatrix} \mid a, b \in \mathbb{Z} \right\}$ are isomorphic rings.

4.6 Integral domains; characteristics

The number rings, such as $\mathbb{R}$, have the important property that the product of two numbers is equal to zero if and only if at least one of the factors is zero. This is used for solving equations, such as

$$x^4 = 1 \qquad \Longleftrightarrow \qquad (x-1)(x+1)(x-i)(x+i) = 0 \qquad \Longleftrightarrow \qquad x = \pm 1, \pm i.$$

Not all rings have this property: for example, in the ring $\mathbb{Z}_{12}$ we have $\overline{3} \cdot \overline{4} = 0$.

Example 4.50. Let us solve the equation $x^3 + x = 0$ in the ring $\mathbb{Z}_{10}$. We want to find those elements of $\mathbb{Z}_{10}$ that satisfy $x^3 = -x$. Clearly $\overline{0}$ is one.

x	$\overline{0}$	$\overline{1}$	$\overline{2}$	$\overline{3}$	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{7}$	$\overline{8}$	$\overline{9}$
x^3	$\overline{0}$	$\overline{1}$	$\overline{8}$	$\overline{7}$	$\overline{4}$	$\overline{5}$	$\overline{6}$	$\overline{3}$	$\overline{2}$	$\overline{9}$

From this table we see that the solutions are $\overline{0}, \overline{3}, \overline{5}, \overline{7}, \overline{8}$.

Definition 4.51. A nonzero element of a ring R is called a *zero divisor* if there exists a nonzero $b \in R$ such that

$$ab = 0 \qquad \text{or} \qquad ba = 0.$$

Example 4.52. The matrix $\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$ is a zero divisor of the ring $\mathcal{M}_2(\mathbb{R})$ since

$$\begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \begin{pmatrix} 6 & 2 \\ -3 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Example 4.53. Let us show that $\overline{a}$ is a zero divisor of the ring $\mathbb{Z}_m$ if and only if $\overline{a}$ is nonzero and $\gcd(a, m) > 1$. First suppose that $\gcd(\overline{a}, m) > 1$. Since $\mathbb{Z}_m$ is cyclic and generated by $\overline{1}$, by Theorem 3.25 we know that the order of an element $\overline{a}$ is $\operatorname{ord}(\overline{a}) = \operatorname{ord}(a \cdot \overline{1}) = \frac{m}{\gcd(a,m)}$. Thus by assumption the order of $\overline{a}$ is less than m. Therefore there exists a nonzero element $\overline{b} = \operatorname{ord}(\overline{a})$ such that $\overline{a}\overline{b} = \overline{0}$, and hence $\overline{a}$ is a zero divisor.

Suppose now that $\overline{a}$ is a zero divisor. Then there exists a nonzero element $\overline{b}$ such that $\overline{a}\overline{b} = \overline{0}$. We assume that $\overline{b}$ is the smallest such element. Then $\operatorname{ord}(\overline{a}) = \overline{b} < m$ which implies that $\gcd(\overline{a}, m) > 1$ by Theorem 3.25.

Definition 4.54. A ring R is called an *integral domain* if

(D1) R is commutative, and

(D2) R has no zero divisors.

Example 4.55. All the number rings ($\mathbb{Z}, \mathbb{Q}$, etc.) are integral domains.

Example 4.56. It follows from Example 4.53 that the residue class ring $\mathbb{Z}_m$ is an integral domain if and only if m is a prime number.

Laws of cancellation hold in any integral domain D: If $a \in D,\ a \neq 0$ then

$$ab = ac \qquad \Longrightarrow \qquad b = c \qquad \forall\ b, c \in D.$$

The left equation can be written in the form $a(b - c) = 0$; the statement follows from this because a is not a zero divisor. Note that here a need not have an inverse.

Example 4.57. Let us solve the equation $x^3 + 10x = 0$ in the rings $\mathbb{Z}_5$ and $\mathbb{Z}_7$. As 5 and 7 are prime number, these rings are both integral domains by Example 4.56.

In $\mathbb{Z}_5$ the equation reduces to $x^3 = 0$ which thus only has the solution $x = \overline{0}$. In $\mathbb{Z}_7$ the equation reduces to $x^3 + 3x = 0$ or $x(x^2 + 3) = 0$. This has the solutions $x = \overline{0},\ x = \overline{2}$ or $x = \overline{5}$ because $(\overline{2})^2 = (\overline{5})^2 = \overline{4} = \overline{-3}$.

Note that the operations in an integral domain D depend on what multiples of the identity $1 = 1_D$ satisfy $n1 = 1 + \cdots + 1 = 0$. This leads to the next definition.

Definition 4.58. The *characteristic* of an integral domain D is

$$\operatorname{char}(D) = \begin{cases} \text{the smallest positive integer } n \text{ such that } n1_D = 0, \\ 0, \text{ if no such } n \text{ exists.} \end{cases}$$

In other words, $\operatorname{char}(D)$ is the order of the identity element in the group $(D, +)$ unless this order is ∞, when we define $\operatorname{char}(D) = 0$.

Example 4.59. The characteristic of every number ring is 0. The characteristic of the residue class ring $\mathbb{Z}_p$ for any prime p is p.

Remark 4.60. In an integral domain D, *all* nonzero elements a have the same order in the group $(D, +)$; the use of the identity in the definition is just for simplicity. The equation $na = 0$ can be written as $(n1_D)a = 0$ by Theorem 4.14, and because D has no zero divisors, this is equivalent to the equation $n1_D = 0$.

Theorem 4.61. *The characteristic of an integral domain D is either 0 or a prime number.*

Proof. Suppose that $\operatorname{char}(D) = n \neq 0$. We write $n = n_1 n_2$, where n_1 and n_2 are positive integers. Now $n1 = (n_1 1)(n_2 1)$, so by assumption $(n_1 1)(n_2 1) = 0$. However, D has no zero divisors so at least one of the factors must be 0, let us say $n_1 1 = 0$. But now by minimality of n, we have $n_1 = n$. The only factorization of n is thus $n = n \cdot 1$, and therefore n is prime. $\square$

Example 4.62. The binomial coefficients $\binom{p}{k}$ are divisible by p when p is a prime number and $1 \leq k \leq p$ since

$$\binom{p}{k} = \frac{p(p-1)\cdots(p-k+1)}{k\cdots 1}$$

and $\binom{p}{0} = \binom{p}{p} = 1$. From this we deduce that if $\text{char}(D) = p$, then

$$(a+b)^p = a^p + b^p \quad \forall\ a, b \in D.$$

Now let us solve the equation $x^3 + y^3 = 0$ in an integral domain D of characteristic 3. By the above deduction, we have $x^3 + y^3 = (x+y)^3 = 0$ which implies that $x + y = 0$. Thus $x = -y$ since D has no zero divisors.

Exercises

1. Prove that $\mathbb{Z}\left[\sqrt{2}\right] = \{a + b\sqrt{2} \mid a, b \in \mathbb{Z}\}$ is an integral domain.
2. Prove that $\mathcal{M}_2(\mathbb{Z})$ is not an integral domain.
3. Solve the equation $x^2 = x$ in an integral domain.
4. How many solutions does the equation $x^5 - 5x + 6 = 0$ have in $\mathbb{Z}_7$? What about in $\mathbb{Z}_8$?
5. Suppose that D and E are integral domains and $f\colon D \to E$ is a ring homomorphism. Suppose that D has a nonzero characteristic. Prove that D and E have the same characteristic.

Chapter 5

Fields and polynomials

5.1 Fields

In this chapter we will consider algebraic objects called *fields*, which are modelled after the rational numbers, that is, a set of numbers equipped with four "basic" operations. This type of object is more algebraically advanced than groups or rings, and thus has many desired properties. Below we present the basic classification of fields and study how fields can be constructed.

Definition 5.1. A set F together with two operations $+$ and $\cdot$, denoted by $(F, +, \cdot)$, is called a *field* if

(F1) $(F, +, \cdot)$ is a commutative ring ($\neq$ zero ring) and

(F2) Every nonzero element of F has a multiplicative inverse in F, that is, the unit group of F is $F^* = F \setminus \{0\}$.

According to this definition, a 3-tuple $(F, +, \cdot)$ is a field if and only if it fulfils

F1': $(F, +)$ is an Abelian group (the *additive group* of the field),

F2': $(F \setminus \{0\}, \cdot)$ is an Abelian group (the *multiplicative group* of the field),

F3': $a(b + c) = ab + ac$ and $(a + b)c = ac + bc \ \ \forall\, a, b, c \in F$.

If we exclude the commutativity condition for multiplication, we get a so-called *skew field* or a *division ring*, which we shall omit from this course.

Example 5.2. $\mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ are (number) fields. $\mathbb{Z}$ is not a field.

Example 5.3. The set of all rational functions is defined as follows:

$$\mathbb{R}(x) = \left\{ \frac{p(x)}{q(x)} \;\middle|\; p(x), q(x) \in \mathbb{R}\,[x]\,,\ q(x) \neq \text{zero polynomial} \right\}.$$

$\mathbb{R}(x)$ is a field under pointwise addition and multiplication of functions.

Theorem 5.4. (i) *Every field is an integral domain.*

(ii) *Every finite integral domain is a field. (Compare with Example 5.2: $\mathbb{Z}$ is an infinite integral domain.)*

Proof. (i) Compare the condition (F1) of a field with the definition of an integral domain. It needs only be proven that no zero divisors exist in a field F. Suppose that $ab = 0$ where $a, b \in F, a \neq 0$. Then a has an inverse element $a^{-1} \in F$, and this equation gives $b = a^{-1} \cdot 0 = 0$. This proves the statement.

(ii) Let D be a finite integral domain. Now it needs to be proven that each element $a \in D \setminus \{0\}$ has an inverse element in D.

Let us form a subset $D_0 = \{ax \mid x \in D\}$ of D. By applying the cancellation law in the integral domain D, we see that $ax_1 \neq ax_2$ whenever $x_1 \neq x_2$. Because D is finite, it follows that there are as many elements in D_0 as in D, and hence $D_0 = D$. In particular, the identity of D is in D_0, that is,

$$\exists\, x' \in D: \qquad ax' = 1.$$

Observe that D is commutative. This implies that $x' = a^{-1}$.

□

Example 5.5. From the previous theorem and by Example 4.56 it follows that the residue class ring $\mathbb{Z}_m$ is a field if and only if m is a prime number. The residue class field $\mathbb{Z}_p = \{\overline{0}, \overline{1}, \ldots, \overline{p-1}\}$ is an example of a finite field.

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

Addition table of $\mathbb{Z}_5$

$\cdot$	1	2	3	4
1	1	2	3	4
2	2	4	1	3
3	3	1	4	2
4	4	3	2	1

Multiplication table of $\mathbb{Z}_5$

Remark 5.6. A finite field of order p^k, where p is a prime, is denoted by $GF(p^k)$. The notation GF originates from the German term Galois-Feld (Galois field in English) which has fallen out of use. We can prove that for each prime power there exists a unique field $GF(p^k)$ up to isomorphy and no other such finite fields exist.

Later in this course, we will present how fields like $GF(p^k)$ can be constructed.

Because a field is a ring, addition, subtraction and multiplication are defined in the field. *Division* is now defined as usual by setting

$$\frac{a}{b} = ab^{-1}, \qquad \text{in particular} \qquad \frac{1}{b} = b^{-1} \qquad (b \neq 0).$$

Applying the usual rules of computation of commutative rings, we observe that

$$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}, \qquad \frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} \qquad (b \neq 0,\ d \neq 0).$$

Thus we can calculate with quotients in the same way as fractions. Note as well that $\frac{a}{1} = a$.

Example 5.7. Let us compute the sums $\frac{\overline{1}}{\overline{2}} + \frac{\overline{1}}{\overline{4}}$ and $\frac{\overline{1}}{\overline{2}} + \frac{\overline{3}}{\overline{4}}$ in the field $\mathbb{Z}_5$.

$$\frac{\overline{1}}{\overline{2}} + \frac{\overline{1}}{\overline{4}} = \frac{\overline{1} \cdot \overline{4} + \overline{1} \cdot \overline{2}}{\overline{2} \cdot \overline{4}} = \frac{\overline{4} + \overline{2}}{\overline{3}} = \frac{\overline{1}}{\overline{3}} = \overline{1} \cdot (\overline{3})^{-1} = \overline{1} \cdot \overline{2} = \overline{2}$$

$$\frac{\overline{1}}{\overline{2}} + \frac{\overline{3}}{\overline{4}} = \frac{\overline{1} \cdot \overline{4} + \overline{3} \cdot \overline{2}}{\overline{2} \cdot \overline{4}} = \frac{\overline{4} + \overline{6}}{\overline{3}} = \frac{\overline{10}}{\overline{3}} = \overline{0}$$

When there is no chance of ambiguity, the elements of a field F are often denoted by

$$1 + 1 = 2, \quad 1 + 1 + 1 = 3, \ \ldots, \qquad \text{in general} \quad n \cdot 1 = n \quad \forall\, n \in \mathbb{Z}.$$

Since a field is an integral domain, the characteristic $\operatorname{char}(F)$ is defined and it is either 0 or a prime by Theorem 4.61. In particular, note that

If $\operatorname{char}(F) = 0$ *then* F *is infinite. The characteristic of a finite field is thus prime.*

If $\operatorname{char}(F) = p$ *(prime), then*

$$n = 0 \qquad \Longleftrightarrow \qquad p \mid n.$$

Example 5.8. Let us determine the characteristics of the fields in Examples 5.2–5.5. The fields $\mathbb{Q}, \mathbb{R}$ and $\mathbb{C}$ in Example 5.2 have characteristic 0 since no integer n exists such that $n \cdot 1 = 0$. The field of rational functions $\mathbb{R}(x)$ in Example 5.3 is infinite and also has characteristic 0. The fields $\mathbb{Z}_p$ in Example 5.5 are finite and have characteristic p.

Example 5.9. Let us solve the second degree equation $x^2 + ax + b = 0$ in a field F with $\operatorname{char}(F) \neq 2$.

In this case the characteristic is at least three, and thus $2 \cdot a \neq 0$ for any nonzero $a \in F$. Therefore we can use the usual solution formula for second degree equations.

$$\begin{aligned} x^2 + ax + b &= 0 \\ x^2 + ax &= -b \\ x^2 + ax + \left(\frac{a}{2}\right)^2 &= -b + \left(\frac{a}{2}\right)^2 \\ \left(x + \frac{a}{2}\right)^2 &= \frac{a^2 - 4b}{2} \\ x + \frac{a}{2} &= \pm\sqrt{\frac{a^2 - 4b}{4}} \\ x &= \frac{-a \pm \sqrt{a^2 - 4b}}{2} \end{aligned}$$

Where the square root is taken if it exists in F.

Example 5.10. Any number field, a field whose elements are complex numbers, contains the field $\mathbb{Q}$, that is, $\mathbb{Q}$ is the smallest number field. Number fields are fields contained in $\mathbb{C}$. Let F be some field contained by $\mathbb{C}$. Firstly, $1 \in F$ so all elements of the form $1 + \cdots + 1$ belong to F. This implies that $\mathbb{Z}_{\geq 0} \subset F$. Since a field has additive closure, we clearly have $\mathbb{Z} \subset F$. Since $(F \setminus \{0\}, \cdot)$ is a group, all inverses of the natural numbers $\{\frac{1}{n}\}_{n \geq 1}$ belong to F. Subsequently, any rational number is of the form $m \cdot \frac{1}{n}$ with $m, n \in \mathbb{Z},\ n \geq 1$, and hence $\mathbb{Q} \subset F$. Therefore any number field contains $\mathbb{Q}$, and thus $\mathbb{Q}$ is the smallest number field.

Because a field is a ring, we can study its ideals. The next theorem discusses them exhaustively.

Theorem 5.11. *The only ideals of a field* F *are* F *and* $\{0\}$.

Proof. If I is a proper ideal of F, then $I \cap F^* = \emptyset$, see Section 4.3. However, $F^* = F \setminus \{0\}$, and therefore $I = \{0\}$. □

If F and F are fields, then a ring homomorphism $F \to F'$ is called a *field homomorphism* and a ring isomorphism $F \to F'$ is called a *field isomorphism* as well.

Theorem 5.12. *Every field homomorphims* $f \colon F \to F'$ *is an injection.*

Proof. Because the kernel $\operatorname{Ker}(f)$ is an ideal of F, by Theorem 5.11 it is either F or $\{0\}$. In the former, $f(a) = 0 \quad \forall\ a \in F$. However, this is impossible because $f(1) = 1$ by Postulate (RH3). Hence $\operatorname{Ker}(f) = \{0\}$ and thus f is an injection. □

It follows that all homomorphic images of F are isomorphic to F.

Example 5.13. Let us show by Example 4.62 that the map

$$f_p\colon F \to F, \qquad f_p(x) = x^p,$$

is a field homomorphism when $\operatorname{char}(F) = p$.

Firstly, by Example 4.62 we know that for any $a, b \in F$

$$f(a+b) = (a+b)^p = a^p + b^p = f(a) + f(b).$$

Multiplication is commutative in a field, hence

$$f(ab) = (ab)^p = a^p b^p = f(a)f(b).$$

Finally we have $f(1) = 1^p = 1$. Now f is a ring homomorphism by Definition 4.39 and thus a field homomorphism.

Then it follows from Theorem 5.12 that $F \simeq \operatorname{Im}(f_p)$. If F is moreover a finite field, $\operatorname{Im}(f_p)$ contains at least as many elements as F, and thus is F. In this case f_p is an automorphism of the field F. The most simple case of a field that satisfies these conditions is $F = \mathbb{Z}_p$. Then the previous result is trivial: f_p is the identity map of the field, see Euler's theorem in Example 2.68.

Exercises

1. Let $R = \{0_{10}, 2_{10}, 4_{10}, 6_{10}, 8_{10}\} \subset \mathbb{Z}_{10}$. Prove that R is a field.
2. Give an example of a finite field with elements $a, b \neq 0$ that satisfy the equation $a^2 + b^2 = 0$.
3. Solve the following pair of equations in $\mathbb{Z}_7$:
$$\begin{cases} -3x + 2y = 1 \\ \quad x + 3y = -2. \end{cases}.$$
4. Let R be a commutative ring. Show that R is a field precisely when the only ideals of R are the trivial ideals $\{0\}$ and R.
5. Compute the inverse element of 31_{173} in the field $\mathbb{Z}_{173}$ and determine all solutions to the congruence $31x \equiv 5 \pmod{173}$.
6. Prove that the matrices $\begin{pmatrix} a & b \\ -b & a \end{pmatrix}$ form a field which is isomorphic to the field of complex numbers $\mathbb{C}$. Which matrix corresponds to the imaginary unit i?
7. Show that the multiplication operation
$$(a, b)(c, d) = (ac + bd, ad + bc + bd)$$
makes the product group $(\mathbb{Z}_2 \times \mathbb{Z}_2, +)$ a field with four elements.
8. Let K be a field, R a ring and $f\colon K \to R$ is a ring homomorphism. Show that f is an injection.
9. Show that $\mathbb{Z}\left[\sqrt{3}\right]$ is not a field.
10. Show that if f is an element of the function ring $R = \mathbb{R}^{\mathbb{R}}$, $f \neq 0_R$, then f is a zero divisor or a unit.

5.2 Subfields; prime fields

Definition 5.14. A subset K of a field $(F, +, \cdot)$ is called a *subfield* of F if K is a field under the operations $+$ and $\cdot$.

If K is a subfield of a field F, then $(K, +)$ is a subgroup of $(F, +)$, and $(K \setminus \{0\}, \cdot)$ is a subgroup of $(F \setminus \{0\}, \cdot)$. In particular, it follows that the additive identities coincide and likewise the units.

Theorem 5.15 (Subfield criterion)**.** *A subset K of a field F is a subfield of F if and only if it satisfies*

(SF1) *K has at least two elements,*

(SF2) $a - b \in K \ \forall\ a, b, \in K$*,*

(SF3) $\frac{a}{b} \in K \ \forall\ a, b \in K,\ b \neq 0$*.*

Proof. If K is a subfield of F, it trivially satisfies the conditions (SF1)–(SF3). Conversely, if these conditions hold, then it follows from (SF1) and (SF2) that $(K, +)$ is a group, namely a subgroup of $(F, +)$. Similarly, it follows from (SF1) and (SF3) that $(K \setminus \{0\}, \cdot)$ is a group since (SF1) ensures that neither K nor $K \setminus \{0\}$ is empty. In addition K inherits distributivity from the field F, thus K is a field. □

Example 5.16. The set

$$\mathbb{Q}(\sqrt{n}) = \{a + b\sqrt{n} \mid a, b \in \mathbb{Q}\} \quad (n \text{ is a square free integer } \neq 0, 1)$$

is a subfield of $\mathbb{C}$, and a subfield of $\mathbb{R}$ if $n > 0$, that contains the ring $\mathbb{Z}[\sqrt{n}]$.

Example 5.17. If $\mathrm{char}(F) = 2$, then F has $\{0, 1\}$ as a subfield. Let us check the field postulates. The set $\{0, 1\}$ has (exactly) two elements and thus (SF1) holds. Since the characteristic of the field is two, we have $1 + 1 = 0$. This implies that $0 - 1 = -1 = 1$. Hence (SF2) holds. We have $\frac{0}{1} = 0$ and $\frac{1}{1} = 1$ and (SF3) holds. Thus we conclude that $\{0, 1\}$ is a subfield.

Theorem 5.18. *The intersection of subfields of a field F is a subfield of F.*

Proof. Follows directly from the Subfield criterion. □

Lemma 5.19. *Every field homomorphism $f\colon F \to F'$ induces a field isomorphism $F \to \mathrm{Im}(f)$.*

Proof. This follows from Theorem 5.12. □

What kinds of subfields can a field F have? Because every subfield contains the identity of F, it also contains its multiples $n1$. This leads to the following observation.

Lemma 5.20. *Every field F contains an integral domain*

$$D = \{n1 \mid n \in \mathbb{Z}\} \simeq \begin{cases} \mathbb{Z}_p, & \text{if } \mathrm{char}(F) = p, \\ \mathbb{Z}, & \text{if } \mathrm{char}(F) = 0. \end{cases}$$

Proof. We form the map

$$f\colon \mathbb{Z} \to F, \qquad f(n) = n1.$$

This is a ring homomorphism (check), so by the Homomorphism theorem we have $\mathbb{Z}/\mathrm{Ker}(f) \simeq \mathrm{Im}(f) \subset F$. Here $\mathrm{Im}(f) = \{n1 \mid n \in \mathbb{Z}\}$ and

$$\mathrm{Ker}(f) = \{n \in \mathbb{Z} \mid n1 = 0\} = \begin{cases} p\mathbb{Z}, & \text{if } \mathrm{char}(F) = p, \\ \{0\}, & \text{if } \mathrm{char}(F) = 0. \end{cases}$$

The isomorphisms can now be obtained by noting that $\mathbb{Z}/p\mathbb{Z} = \mathbb{Z}_p$ and $\mathbb{Z}/\{0\} \simeq \mathbb{Z}$. Since $\mathbb{Z}_p$ and $\mathbb{Z}$ are moreover integral domains, the statement follows. □

Let us return to the original question on the subfields of a field. Of the following theorems, the first is associated with an important basic concept in field theory, the *quotient field*, which will be introduced later. The second theorem results from this theorem and Lemma 5.20.

Theorem 5.21. *If a field F contains an integral domain D, then F contains the subfield*

$$F_D = \{\frac{a}{b} \mid a, b \in D,\ b \neq 0\}.$$

Furthermore, F_D is contained in every subdomain K of F that satisfies $D \subset K$.

Proof. We use the Subfield criterion. Firstly we see that $D \subset F_D$, so the set F_D contains at least the elements 0 and 1. If $\frac{a}{b} \in F_D$ and $\frac{c}{d} \in F_D$, then their difference $\frac{ad-bc}{bd} \in F_D$ and their quotient $\frac{ad}{bc} \in F_D$, presuming that $\frac{c}{d} \neq 0$. Note that here $c \neq 0$ The latter statement is obvious because every field K contains the quotient $\frac{a}{b}$ of elements a, b of D with $b \neq 0$. □

Example 5.22. If F is a number field, it contains the integral domain $\mathbb{Z}$. All number fields have characteristic 0, and thus it follows from Lemma 5.20 that F contains the integral domain $\mathbb{Z}$.

Alternatively, we know that $1 \in F$ and since $(F, +)$ is a commutative group, so all $1+\cdots+1 \in F$. Hence $\mathbb{Z} \subset F$. Then the field in Theorem 5.21 is $F_{\mathbb{Z}} = \mathbb{Q}$.

Theorem 5.23. *Every field F contains the following field as a subfield:*

$$P \simeq \begin{cases} \mathbb{Z}_p, & \text{if } \operatorname{char}(F) = p, \\ \mathbb{Q}, & \text{if } \operatorname{char}(F) = 0. \end{cases}$$

Proof. We prove that we can choose the field P as the field F_D by Theorem 5.21 where D is an integral domain given by Lemma 5.20.

If $\operatorname{char}(F) = p$, then $D \simeq \mathbb{Z}_p$. Then D is a field itself, so it also contains the quotients of its elements. Thus $F_D = D$ and F thus has a subfield $F_D \simeq \mathbb{Z}_p$.

Suppose that $\operatorname{char}(F) = 0$, in which case

$$D = \{n1 \mid n \in \mathbb{Z}\} \simeq \mathbb{Z}.$$

Now $F_D = \{\frac{n1}{m1} \mid n, m \in \mathbb{Z},\ m \neq 0\}$ and from this we see that $F_D \simeq \mathbb{Q}$. □

Definition 5.24. A field is called *prime field* if it has no proper subfields.

Theorem 5.25. (i) *All prime fields are $\mathbb{Z}_p$, where p is a prime, or $\mathbb{Q}$ up to isomorphism.*

(ii) *Every field F has a unique prime field as its subfield; this is isomorphic to the field $\mathbb{Z}_p$ or $\mathbb{Q}$ depending on whether* $\operatorname{char}(F) = p$ *or 0.*

Proof. (i) The field of rational numbers $\mathbb{Q}$ is a prime field because is the smallest number field, see Example 5.10. The residue class field $\mathbb{Z}_p$ is deduced as a prime field as follows: If K is a subfield of $\mathbb{Z}_p$, then by considering their additive groups, we get by Lagrange's Theorem that $\#F$ divides the prime p. Since $\#F > 1$, it thus must be equal to p. That no other prime fields exist follows from part (ii).

(ii) The existence of the claimed field P was proven in Theorem 5.23. We prove uniqueness. If P_1 and P_2 are prime fields contained in F, then by Theorem 5.18 $P_1 \cap P_2$ is also a field. Because this is a subfield of P_1 and P_2, by the definition of a prime field it is equal to both P_1 and P_2. Thus $P_1 = P_2$.

□

This theorem showed that the characteristic of a field F is crucial when determining its type. From Theorem 5.25 it follows that a field and its subfields contain the same prime field.

Example 5.26. The prime field P contained in the finite field $GF(p^k)$ is $P \simeq \mathbb{Z}_p$. Any subfield K of the finite field $GF(p^K)$ is also a subset of $GF(p^K)$. Therefore, by Lagrange's theorem we must have $\#K \mid p^k$. This implies that $\#P = p^m$ for some $1 \leq m \leq k$. Moreover, since a prime field has no proper subfields by definition, Lagrange's theorem further implies that $\#P = p$. Therefore $P \simeq \mathbb{Z}_p$.

Example 5.27. The prime field contained by the field of functions $\mathbb{R}(x)$, see Example 5.3, is $\mathbb{Q}$. An example of an infinite field of prime characteristic p with prime field $\mathbb{Z}_p$ is presented in Section 5.6.

Exercises

1. Suppose that the characteristic of a field K is 3. Study whether the set $\{a^9 \mid a \in K\}$ is a subfield of K.

5.3 Quotient fields

We often encounter the following problem in algebra: we have a given algebraic object A but it lacks some desired property. Can we construct an object B that has this property and contains A? Because isomorphic objects can be equated in algebra, the containment need not be verbatim; it is enough that B contains some object A' isomorphic to A.

Example 5.28. A classical example of the previous is the extension of number sets: $\mathbb{N} \to \mathbb{Z} \to \mathbb{Q} \to \mathbb{R} \to \mathbb{C}$.

Let us consider the step $\mathbb{Z} \to \mathbb{Q}$ in more depth. Recall that the main points of the contruction of the rational numbers are the following:

- Form the set of all integer pairs $(a, b),\ b \neq 0$.
- Decide that the pairs (a, b) and (c, d) are equivalent if and only if $ad = bc$.
- Introduce the notation $\frac{a}{b}$ to represent the pair (a, b) and all equivalent pairs.
- Denote the set of elements $\frac{a}{b}$ by $\mathbb{Q}$, and name them the *rational numbers*.
- Define addition and multiplication in the set $\mathbb{Q}$.
- Equate the elements $\frac{a}{1}$ with the integers a.

The algebraic treatment of this construction includes proving that the obtained set $\mathbb{Q}$ is a field. The last point in the construction implies that the subset $\{\frac{a}{1} \mid a \in \mathbb{Z}\}$ of $\mathbb{Q}$ is an integral domain and isomorphic to $\mathbb{Z}$.

It is important in algebra that *every* integral domain D can be extended in a similar construction as above to a certain field $Q(D)$, the so-called *quotient field* or *field of fractions* of D.

Let D be an integral domain. We form the set

$$X = \{(a, b) \mid a, b \in D,\ b \neq 0\}$$

and define a relation on this set as follows:

$$(a,b) \sim (c,d) \qquad \Longleftrightarrow \qquad ad = bc.$$

This is an equivalence relation. You should check the conditions (E1)–(E3), and note that we need the law of cancellation. Thus we get a partition of X into equivalency classes $[(a,b)]$, which are called *formal quotients* and denoted by $\frac{a}{b}$:

$$\frac{a}{b} = \{(x,y) \mid (x,y) \sim (a,b)\} = \{(x,y) \mid x,y \in D, y \neq 0,\ xb = ya\}.$$

In particular, note

$$\frac{a}{b} = \frac{c}{d} \qquad \Longleftrightarrow \qquad ad = bc.$$

We denote the set of all formal quotients by $Q(D)$:

$$Q(D) = \left\{ \frac{a}{b} \mid a,b \in D, b \neq 0 \right\}.$$

We define addition and multiplication in the set $Q(D)$:

$$\frac{a}{b} + \frac{c}{d} = \frac{ad+bc}{bd}, \qquad \frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}.$$

Firstly note that $bd \neq 0$. Furthermore, we have to make sure that these operations are well defined. We prove this for addition as multiplication is similar: Let $\frac{a}{b} = \frac{a'}{b'}$ and $\frac{c}{d} = \frac{c'}{d'}$, where $ab' = ba'$ and $cd' = dc'$. We argue that

$$\frac{ad+bc}{bd} = \frac{a'd'+b'c'}{b'd'},$$

that is, $(ad+bc)b'd' = bd(a'd'+b'c')$. This is shown easily:

$$\frac{ad+bc}{bd} = (ab')dd' + bb'(cd') = (ba')dd' + bb'(dc') = \frac{a'd'+b'c'}{b'd'}.$$

Theorem 5.29. *The set $Q(D)$ is a field. Its subset*

$$D' = \left\{ \frac{a}{1} \mid a \in D \right\}$$

is a subring of $Q(D)$; it is an integral domain and isomorphic to D.

Proof. $Q(D)$ is a commutative ring, the additive identity is $\frac{0}{1}$, the identity is $\frac{1}{1}$, and the additive inverse of the element $\frac{a}{b}$ is $\frac{-a}{b}$; checking the ring postulates is straightforward, go through distributivity for example.

If $\frac{a}{b} \neq \frac{0}{1}$, then $a \cdot 1 \neq b \cdot 0$ so $a \neq 0$. Hence $Q(D)$ contains the element $\frac{b}{a}$. This is the inverse of $\frac{a}{b}$ since $\frac{a}{b} \cdot \frac{b}{a} = \frac{ab}{ab} = \frac{1}{1}$. Thus the ring $Q(D)$ is a field.

We form the map

$$j \colon D \to Q(D), \qquad j(a) = \frac{a}{1}.$$

This is a ring homomorphism; check the Postulates (RH1)–(RH3). Furthermore, j is an injection because it follows from equation $\frac{a}{1} = \frac{b}{1}$ that $a \cdot 1 = 1 \cdot b$ and $a = b$. Thus we get that

$$D \simeq \mathrm{Im}(j) = \{\frac{a}{1} \mid a \in D\}.$$

This proves the latter statement. □

It is natural to equate D and D' by denoting the element $\frac{a}{1}$ simply by a. We also say that the map j embeds D into the field $Q(D)$. Then the field $Q(D)$ has

$$ab^{-1} = \frac{a}{1} \cdot \left(\frac{b}{1}\right)^{-1} = \frac{a}{1} \cdot \frac{1}{b} = \frac{a \cdot 1}{1 \cdot b} = \frac{a}{b},$$

thus the notation $\frac{a}{b}$ also denotes the *real* quotient of the elements a and b of the field.

Definition 5.30. The field $Q(D)$ constructed above is called the *quotient field*, or *field of fractions*, of an integral domain D.

Example 5.31. The polynomial ring $\mathbb{R}[x]$ is an integral domain. Firstly, we know that $\mathbb{R}$ is an integral domain. Let $f(x)$ and $g(x)$ be two nonzero polynomials in $\mathbb{R}[x]$, and $a_f,\ b_g$ be their respective leading coefficients. Then by definition $a_b \neq 0$ and $b_g \neq 0$. Further, because $\mathbb{R}$ has no zero divisors, we have $a_f b_g \neq 0$. But the product $a_f b_g$ is the leading coefficient of $f(x)g(x)$, and so $f(x)g(x)$ cannot be the zero polynomial. Consequently, $\mathbb{R}[x]$ has no zero divisors and is thus an integral domain. Its quotient field is the field of rational functions $\mathbb{R}(x)$, see Example 5.3.

If we construct the quotient field $Q(D)$ for an integral domain D which is contained in some field F, then $Q(D)$ is isomorphic to the field

$$F_D = \left\{ \frac{a}{b} \mid a, b \in D,\ b \neq 0 \right\} \tag{5.1}$$

in Theorem 5.21. It is natural to equate $Q(D)$ and F_D, and call F_D the quotient field of the integral domain D as well. Consider $\mathbb{Z}$ in the field of rational numbers $\mathbb{R}$: then the quotient field $\mathbb{Q}$ is equal to $F_{\mathbb{Z}}$.

Hence we see that the concept of quotient fields is simple: Because D is contained in field $F = Q(D)$ anyhow, we can always consider the quotient field in the form (5.1). The construction itself usually need not be considered after it has been done once!

From the previous, we also see that the quotient field of an integral domain D is the smallest field that contains D. Namely if $D \subset F$ for a field F, then $D \subset F_D \subset F$. Finally, let us return to the general question presented at the start of this section. Suppose that the object A can be extended in the way explained above into an object B. Then B is algebraically satisfactory only if it is *unique* in the following sense: if A is isomorphic to an object A_0, then their respective extensions B and B_0 are also isomorphic. It is easy to prove (although omitted here) that the quotient field of an integral domain has this property.

Example 5.32. Let $\mathbb{Z}[i] = \{a + bi \mid a, b \in \mathbb{Z}\}$. The quotient field of $\mathbb{Z}[i]$ is isomorphic to field $\mathbb{Q}(i) = \{r + si \mid r, s \in \mathbb{Q}\}$, which is read as $\mathbb{Q}$ adjoin i.

Suppose F is some field such that $\mathbb{Q} \subset F$ and $i \in F$. Then since $F \setminus \{0\}$ is a commutative multiplicative group, $bi \in F$ for any $b \in \mathbb{Q}$. Also, $(F, +)$ is a group, so $a + bi \in F$ for any $a, b \in \mathbb{Q}$. Thus $\mathbb{Q}(i) \subset F$.

The quotient field of $\mathbb{Z}[i]$ is $\{\frac{a}{b} \mid a, b \in \mathbb{Z}[i],\ b \neq 0\}$. For any $a, b, c, d \in \mathbb{Z}$ we have $(a+bi)(c+di) = (ac - bd) + (ad + bc)i \in \mathbb{Q}(i)$. Thus $\mathbb{Z}[i] \subset \mathbb{Q}(i)$. The quotient field of an integral domain is the smallest field that contains it, therefore $F = \mathbb{Q}(i)$.

5.4 Field extensions

Definition 5.33. If F is a subfield of a field L, we say that L is a *field extension* or an *extension field* of F.

In particular, every field L is a field extension of their prime field P. The subfield F generated by a subset S of a field L is defined similarly as prior concepts:

$$F = \bigcap_{S \subset F} K$$

where K is a subfield of L.

In this form, the definition is not that practical since no simple rule exists for determining the elements of F when S is, say, an arbitrary finite set in L.

Remember that a subfield K of a field L always contains the prime field P of L. Thus it is natural to include P in the generator set S, that is, to choose $S = P \cup S_1$ where S_1 is some subset of L. In fact, it has proven useful to accept other subfields of L as well and establish the following definition.

Definition 5.34. Let F be a subfield of a field L and S some subset of L. Then the set $D \cap S$ generates, in the above sense, a subfield of L which is denoted $F(S)$:

$$F \subset F(S) \subset L.$$

The field $F(S)$ is said to be the field extension of F *generated* by the set S, or the field that we obtain by *adjoining* the set S to the field F.

It follows from the definition that $F(S)$ is the smallest subfield that contains F and S. Observe that $F(S)$ is also the smallest extension field of the field F in L that contains the set S. In fact, only those elements of S outside F are relevant.

If S is finite, $S = \{a_1, \ldots, a_n\}$, we denote $F(S) = F(a_1, \ldots, a_n)$ and we say that $F(S)$ is the *finitely generated* field extension of F. An extension generated by a single element is called *simple*. It is easy to prove that every finitely generated extension can be constructed by consecutive simple extensions.

To summarise, if F and L are fields, $F \subset L$ and $a \in L$, then $F(a)$ is the smallest subfield of L that contains F and a. It is a field extension of F that is obtained by adjoining the element a to the field F.

Example 5.35. If we adjoin the number $\sqrt{2}$ to the field of rational numbers $\mathbb{Q}$, we get a subfield of $\mathbb{R}$:

$$\mathbb{Q}\left(\sqrt{2}\right) = \{a + b\sqrt{2} \mid a, b \in \mathbb{Q}\},$$

compare with Example 5.16.

Example 5.36. If we adjoin the imaginary unit i to the field of real numbers, we get the whole field of complex numbers: $\mathbb{R}(i) = \mathbb{C}$.

Above we discussed extending a field F within some larger field L. The theory of field extensions gets even more interesting when no such field L exists. Then we are in the situation described as a general question in the previous section.

For example, the field of complex numbers $\mathbb{C}$ can be constructed by starting from the field of real numbers $\mathbb{R}$ and the root of the polynomial x^2+1 that we denote by i. The thus constructed field is not a field extension of $\mathbb{R}$ as such; it is formed of "new" elements that can be written in the form (a, b) or $a+bi$ where $a, b \in \mathbb{R}$, and in addition to the rules of operation, $i^2 = -1$ holds. Nonetheless, it contains a subfield R_0 isomorphic to $\mathbb{R}$, which comprises all elements of form $(a, 0)$, or $a + 0i$. As in the previous section, it is natural to equate $\mathbb{R}$ and R_0, that is, denote $a = a + 0i$.

With the same principle, we can construct the extension $F(a)$ for any given field F in general. The properties of the extension then depend on the polynomial in F whose root is adjoined to the field F. Polynomials are fundamentally related to the theory of field extensions.

In this course we only present certain basic results in forming field extensions with the mentioned method and apply them to contruct finite fields. Before that we will discuss some essential theory of polynomials.

5.5 Maximal ideals

In the following, we present how we can form fields from a commutative ring via maximal ideals.

Definition 5.37. An ideal M of a ring R is called *maximal* if it is proper and if no other ideal I of R exists such that $M \subsetneq I \subsetneq R$.

The latter condition of the definition is often convenient rewritten as

$$I \text{ is an ideal of } R \text{ and } M \subsetneq I \qquad \Longrightarrow \qquad I = R.$$

Example 5.38. In the ring $\mathbb{Z}_8$, the ideal $\langle\overline{2}\rangle = \{\overline{0}, \overline{2}, \overline{4}, \overline{6}\}$ is maximal because the index of its additive group in the group $(\mathbb{Z}_8, +)$ is 2.

Example 5.39. The maximal ideals of the ring $\mathbb{Z}$ are the ideals $p\mathbb{Z}$ where p is a prime. Recall that $\mathbb{Z}$ is a PIR, so all its ideals are of the form $m\mathbb{Z}$, $m \geq 0$, and

$$m_1\mathbb{Z} \subsetneq m_2\mathbb{Z} \qquad \Longleftrightarrow \qquad m_1 \mid m_2 \quad \text{and} \quad m_2 < m_1.$$

Compare this example with the knowledge that a residue class ring $\mathbb{Z}/m\mathbb{Z}$ is a field if and only if $m = p$ is a prime. The next theorem generalises this result.

Theorem 5.40. *Let I be an ideal of a commutative ring R. Then*

$$R/I \text{ is a field} \qquad \Longleftrightarrow \qquad I \text{ is a maximal ideal of } R.$$

Proof. ($\Longrightarrow$) Assume that R/I is a field. Because a field has at least two elements, we know $I \neq R$. Thus to prove maximality of I, we need to take another ideal J of R that contains I properly and prove that $J = R$.

Because I is a proper subset of J, there exists $a \in J \setminus I$. Then $a + I \neq I$, that is, $a + I$ is not the additive identity of the field R/I. Thus it has an inverse $b + I$:

$$(a + I)(b + I) = 1 + I.$$

It follows that $1 = ab+i$ where $i \in I$. But now $1 \in J$ since $a \in J$ and $i \in J$; recall Definition 4.24. The statement follows.

($\Longleftarrow$) Assume now that I is a maximal ideal. The quotient ring R/I is anyhow a commutative ring. Thus it needs only be proven that any arbitrary element $a + I \neq I$ has an inverse.

Because $a \notin I$, the sum of the ideals I and Ra contains properly I; see Theorem 4.28. It follows from the maximality of I that $I + Ra = R$. In particular, we have $1 \in I + Ra$ so $1 = i + ra$, where $i \in I$ and $r \in R$. Now we get

$$1 + I = ra + I = (r + I)(a + I).$$

The element $r + I$ is thus the inverse of the element $a + I$ in the ring R/I. □

One objective of the final part of the course is to construct fields via the previous theorem. Polynomial rings have proven to be appropriate: maximality of an ideal returns to divisibility of polynomials.

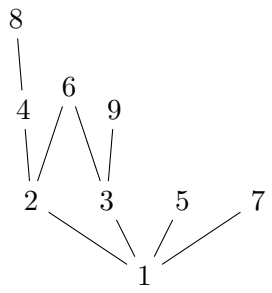
Nonetheless, the remaining part of this section forms its own whole that does not aim at constructing fields. Its purpose is to connect the concept of maximal ideals to more general mathematics.

Recall the definition of an ordered set and its basic properties, see Section 1.8.

Definition 5.41. In a partially ordered set A, an element m is said to be *maximal* if m is not a predecessor of any other element in A, that is, if

$$m \leq a, \quad a \in A \qquad \Longrightarrow \qquad m = a.$$

Example 5.42. The set $\{1, 2, \ldots, 9\}$ ordered with respect to divisibility has maximal elements $5, 6, 7, 8, 9$. Small enough ordered sets such as this are conveniently represented with a Hasse diagram. Drawn below is the Hasse diagram for the partially ordered set $\{1, 2, \ldots, 9\}$.



Example 5.43. This example connects the concept of maximal ideals to the first part of this section: In the set of all proper ideals of a ring R ordered with respect to the inclusion relation $\subset$, the maximal elements are the maximal ideals of R.

The question, when does a ring have (at least) one maximal ideal, leads to a more general question: When does a partially ordered set A have maximal elements?

By considering different examples, we observe that there is no simple answer to this question, consider infinite groups for example. A certain sufficient condition for existence of a maximal element is given by the next so-called *Zorn's lemma*. It can be proven with the famous *Axiom of choice* in group theory. However, we will not present its proof. In fact, the Axiom of choice can conversely be proven from Zorn's lemma, so these two can be considered equivalent axioms in group theory.

To present Zorn's lemma, we need the following concept in a partially ordered set $(A, \leq)$: we say that an element $y \in A$ is an *upper bound* of a subset S of A if $s \leq y \quad \forall\ s \in S$.

Example 5.44. In the set $\mathbb{Z}_+$ ordered with respect to the division relation, an upper bound for the subset $S = \{1, 2, \ldots, 9\}$ is 9!. Find another smaller upper bound.

Lemma 5.45 (Zorn's lemma)**.** *Let A be a partially ordered set. If every totally ordered subset of A has an upper bound in A, then A has at least one maximal element.*

Corollary 5.46. *Let $\mathcal{P}$ be some nonempty collection of subsets of a set X partially ordered with respect to the inclusion relation. Assume that the union $\bigcup_{\alpha \in T} A_\alpha$ of all sets in any totally ordered subcollection $\{A_\alpha \mid \alpha \in T\}$ of $\mathcal{P}$ belongs to $\mathcal{P}$. Then $\mathcal{P}$ has at least one maximal set, that is, a set $M \subset X$ such that*

$$M \subset A, \quad A \in \mathcal{P} \qquad \Longrightarrow \qquad M = A.$$

Here T is an index set that can be uncountable. What does the assumption that $\{A_\alpha \mid \alpha \in T\}$ is totally ordered imply? Consider the simpler case where T is finite or countable, for example, $T = \{1, 2, \ldots\}$. Then this collection is a sequence of increasing sets: $A_1 \subset A_2 \subset \ldots$.

Proof. Because each set A_α is contained in the union $\bigcup_\alpha A_\alpha$, this union is an upper bound of the collection $\{A_\alpha \mid \alpha \in T\}$. The statement follows by applying Zorn's lemma on the collection $\mathcal{P}$. □

Theorem 5.47. *Every ring has at least one maximal ideal. More precisely: every proper ideal I of a ring R is contained in at least one maximal ideal of R.*

Proof. The former statement follows from the latter since R has at least $\{0\}$ as a proper ideal. To prove the latter statement we apply Corollary 5.46 to the collection

$$\mathcal{P} = \{J \mid I \subset J,\ J \text{ is a proper ideal of } R\ \}.$$

$\mathcal{P}$ is nonempty because it contains I. Let $\{J_\alpha \mid \alpha \in T\}$ be some totally ordered collection of sets of $\mathcal{P}$. We need to show that the union $\bigcup_\alpha J_\alpha$ is also contained in $\mathcal{P}$; then $\mathcal{P}$ contains a maximal set M and this is the required maximal ideal.

Firstly, $\bigcup_\alpha J_\alpha$ contains I because each J_α contains I. Secondly, $\bigcup_\alpha J_\alpha$ is an ideal of the ring R as we observe by the Ideal criterion. Let $a, b \in \bigcup_\alpha J_\alpha$. Then for example, $\alpha \in J_\alpha$ and $\beta \in J_\beta$ where $\alpha, \beta \in T$. Because the collection $\mathcal{P}$ is totally ordered, we have $J_\alpha \subset J_\beta$ or $J_\beta \subset J_\alpha$; we can assume that $J_\alpha \subset J_\beta$. Now α and β both belong to the ideal J_β, thus $a - b \in J_\beta$. Hence we have $a - b \in \bigcup_\alpha J_\alpha$. Reason out yourself why both elements ra and ar belong to this union whenever $r \in R$.

Thirdly we need to make sure that the ideal $\bigcup_\alpha J_\alpha$ is proper, that is, $\neq R$. This follows from the fact that the identity element 1 of the ring does not belong to any ideal J_α and thus neither to their union. □

Exercises

1. Find all maximal ideals of the following rings:
   (a) $\mathbb{Z}_8$,
   (b) $\mathbb{Z}_{10}$,
   (c) $\mathbb{Z}_{12}$,
   (d) $\mathbb{Z}_n$.
2. Let $R = \{f \colon \mathbb{R} \to \mathbb{R} \mid f \text{ is continous }\}$. Prove that $A = \{f \in R \mid f(0) = 0\}$ is a maximal ideal of R.
3. Consider the points in the plane $\mathbb{R}^2$ ordered with respect to a product order J:

   $$(x_1, x_2)J(y_1, y_2) \qquad \Longleftrightarrow \qquad (x_1 J_1 y_i) \quad \text{and} \quad (x_2 J_2 y_2) \quad (x_i, y_i \in \mathbb{R}).$$

   Find all maximal elements of the subset

   $$A = \{(x, y) \in \mathbb{R}^2 \mid x, y \in \mathbb{Z},\ x^2 + y^2 \leq 2\}.$$

4. A proper ideal P of a ring R is called a *prime ideal* if P satisfies the condition

$$a, b \in R,\ ab \in P \qquad \Longrightarrow \qquad a \in P \text{ or } b \in P.$$

Prove that a maximal ideal of a commutative ring is a prime ideal. (Hint: Consider the quotient ring R/P.)

5.6 Polynomial rings

In many examples so far we have considered real polynomials $a_0 + a_1x + \cdots + a_nx^n$ and the ring they form. Now we generalise the concept by replacing the real numbers $a_0, a_1, \ldots, a_n$ with elements of a given commutative ring R. Then addition and multiplication of polynomials can be computed exactly as in the case of real numbers by applying the ring operations of R on the coefficients.

The set of all such polynomials is denoted $R[x]$, that is,

$$R[x] = \{a_0 + a_1x + \cdots + a_nx^n \mid n \geq 0,\ a_k \in R \quad (k = 0, \ldots, n)\}.$$

The elements of this set are called *polynomials over the ring* R, more precisely, polynomials *in one indeterminant* x over the ring R.

How do we define equality of two polynomials

$$f(x) = a_0 + a_1x + \cdots + a_nx^n, \tag{5.2}$$

$$g(x) = b_0 + b_1x + \cdots + b_mx^m? \tag{5.3}$$

We considered real polynomials as functions $\mathbb{R} \to \mathbb{R}$ so their equality reverted to the equality of functions: polynomials $f(x), g(x) \in \mathbb{R}[x]$ are equal if their evaluated *values* are equal for all real numbers x.

In the general case, this functional outlook needs to be replaced with an algebraic approach. In this approach a polynomial $a_0 + a_1x + \cdots + a_nx^n$ is simply an expression that is determined by the elements $a_0, \ldots, a_n$ of R. Thus it is natural to say that the polynomials (5.2) and (5.3) are equal if they have the same coefficients, that is,

$$f(x) = g(x) \qquad \Longleftrightarrow \qquad n = m \quad \text{and} \quad a_k = b_k \quad (k = 0, \ldots, n), \quad a_n \neq 0.$$

This equivalence holds even when $f(x), g(x) \in \mathbb{R}[x]$, so no contradiction arises. Consider the real polynomials in the view of linear algebra: $\{1, x, \ldots, x^n\}$ is the basis of a polynomial space P_{n+1} formed by the polynomials of at most nth degree. In general the values of polynomials do not always define a polynomial uniquely, compare with Exercise 4.

Remark 5.48. Defining a polynomial as an expression $a_0 + a_1x + \cdots + a_nx^n$ is not quite precise. For a rigorous definition, consider the polynomial as an infinite *sequence*

$$(a_0, a_1, \ldots, a_n, 0, 0, \ldots),$$

whose elements are 0 starting from some point. So that such sequences are easier to work with, including utilising familiar polynomial computation methods, we introduce the notation $a_0 + a_1x + \cdots + a_nx^n$ where the symbol x^k indicates that a_k is the $(k+1)$-th element in the sequence ,$k \geq 0$, and the symbol x^0 is often not written.

For example,

$$4 + 2x^3 - x^4 = (4, 0, 0, 2, -1, 0, 0, \ldots).$$

Theorem 5.49. *If R is a commutative ring, then the polynomials over R form a commutative ring $(R[x], +, \cdot)$ under addition and multiplication defined as usual. The polynomials $ax^0 = a$, $a \in R$ form a subring of $R[x]$ that can be equated in a natural way with the ring R.*

The ring $R[x]$ is called a *polynomial ring (over R)*. For completeness, the addition and multiplication rules of polynomials are displayed; the polynomials $f(x)$ and $g(x)$ are as in (5.2) and (5.3), $n \leq m$:

$$f(x) + g(x) = (a_0 + b_0) + (a_1 + b_1)x + \cdots + (a_n + b_n)x^n + b_{n+1}x^{n-1} + \cdots + b_m x^m,$$
$$f(x)g(x) = a_0 b_0 + (a_0 b_1 + a_1 b_0)x + (a_0 b_2 + a_1 b_1 + a_2 b_0)x^2 + \cdots + a_n b_m x^{n+m}.$$

The polynomials $ax^a = a$ are called *constant polynomials*.

Proof. We need to verify that addition and multiplication of polynomials obey the postulates of commutative rings. For most of the postulates, this can be observed directly: for example, the additive identity is the constant polynomial 0, *the zero polynomial*, the identity is the constant polynomial 1, and the additive inverse of the polynomial $f(x)$ is $-f(x)$. Checking certain postulates requires more manual work, consider associativity of multiplication.

The latter part of the statement is evident. □

When polynomial operations are thus defined, we observe that the sum and product notation in the original polynomial expression $a_0 + a_1 x + \cdots + a_n x^k$ can be interpreted as these operations. For example, $a_0 + a_1 x$ is the sum of the constant polynomial a_0 and the polynomial $a_1 x$, and $a_k x^k$ is the product of the constant polynomial a_k and the polynomial (monomial) $a^k = 1 \cdot x^k$. Therefore the notation has no ambiguity.

Definition 5.50. If a polynomial $f(x) = a_0 + a_1 x + \cdots + a_n x^n$ has $a_n \neq 0$, then the coefficient a_n is called the *leading coefficient* of $f(x)$ and n is called the *degree* of $f(x)$, denoted by $n = \deg f(x)$. We say that a polynomial $f(x)$ is a *monic polynomial* if its leading term is 1.

We thus define a leading term and a degree ≥ 0 for every nonzero polynomial.

The convention is that the zero polynomial has degree $\deg(0) = -\infty$. One must be extremely careful when digesting the meaning of the symbols $-\infty$ and ∞ in mathematical formulae. Nonetheless, $-\infty$ as a polynomial degree is simply considered a "number" that is smaller than all of the real numbers.

Example 5.51. Let us consider the polynomials $f(x) = 1 + x$ and $g(x) = 1 - x$ over an arbitrary commutative ring R. We observe that

$$f(x) + g(x) = 2, \qquad f(x)g(x) = 1 - x^2.$$

Now we have $\deg f(x) = \deg g(x) = 1$, and further, $\operatorname{def}(f(x) + g(x)) = 0$, or $-\infty$ if $2 = 0$ in R, and $\deg f(x)g(x) = 2$.

Theorem 5.52. *If R is an integral domain, then the polynomial ring $R[x]$ is also an integral domain and*

$$\deg f(x)g(x) = \deg f(x) + \deg g(x) \quad \forall\ f(x), g(x) \in R[x]. \tag{5.4}$$

Proof. Let the leading terms of $f(x)$ and $g(x)$ be a_n and b_m respectively. The term of highest order in the product $f(x)g(x)$ is $a_n b_m x^{n+m}$; here $a_n b_m \neq 0$ because R has no zero divisors. Firstly we see that the degree of the product is $n + m$. But then the product is not the zero polynomial, and therefore $R[x]$ has no zero divisors either. Thus the polynomial ring $R[x]$ is an integral domain.

If $f(x) = 0$ or $g(x) = 0$, then the product is also equal to 0 and both sides of Equation (5.4) are $-\infty$. □

Example 5.53. Because $\mathbb{Z}_p$, p is a prime, is an integral domain, by Theorem 5.52 the polynomial ring $\mathbb{Z}_p[x]$ is also an integral domain and thus has quotient field

$$\mathbb{Z}_p(x) = \left\{ \frac{p(x)}{q(x)} \mid p(x), q(x) \in \mathbb{Z}_p[x],\ q(x) \neq \text{ zero polynomial} \right\}.$$

This is called the *field of rational functions over the field* $\mathbb{Z}_p$, compare with Example 5.3. Note that $\text{char}(\mathbb{Z}_p(x)) = p$ even though $\mathbb{Z}_p(x)$ is an infinite field.

Exercises

1. Compute the polynomials $a + b$, ab and b^5 when $a = 1 + 2x + x^2$ and $b = x^2 + 2$ in the ring $\mathbb{Z}_7[x]$.
2. Determine the degree of the product of the polynomials $1 - 4x + 6x^2$ and $x + 2x^2 + 10x^3$ in the polynomial rings $\mathbb{R}[x]$, $\mathbb{Z}_4[x]$, $\mathbb{Z}_5[x]$.
3. Find the inverse polynomial of $2x + 1$ in $\mathbb{Z}_4[x]$.
4. Determine what values the polynomials $x^4 + x$ and $x^2 + x$ of $\mathbb{Z}_3[x]$ can get.
5. Which polynomials have inverses in $\mathbb{Z}[x]$? (First study the constant polynomials and then use Theorem 5.52).
6. Show that the polynomial $F = 1 - 2x$ has an inverse in the ring $\mathbb{Z}_{16}[x]$. (Hint: Solve $G \in \mathbb{Z}_{16}[x]$ from the equation $FG = 1$.)

5.7 Divisibility of polynomials

Hereon we assume that the coefficients of a polynomial belong to a *field* K.

Recalling Theorem 5.52 a natural question arises: When K is a field, is the polynomial ring $K[x]$ a field as well? The answer is negative. A polynomial $f(x) \in K[x]$ has an inverse in $K[x]$ only when $f(x)$ is a nonzero constant polynomial. This is observed by inferring thusly from Theorem 5.52:

$$f(x)g(x) = 1 \quad \Longrightarrow \quad \deg f(x) + \deg g(x) = 0 \quad \Longrightarrow \quad \deg f(x) = \deg g(x) = 0.$$

Thus a polynomial ring $K[x]$ has a similar algebraic structure to the ring of integers $\mathbb{Z}$: it is an integral domain but not a field. It follows that we get a similar theory for division of polynomials, over a field K, as for the integers. In the following, we present the main points of this theory, many of which are already familiar from the case $K = \mathbb{R}$.

Definition 5.54. Let $a(x), b(x) \in K[x]$. If there exists a polynomial $c(x) \in K[x]$ such that $a(x) = b(x)c(x)$, we say that the polynomial $a(x)$ *is divisible* by the polynomial $b(x)$, denoted by $b(x) \mid a(x)$. We may also say that $b(x)$ *divides* $a(x)$, $b(x)$ is a *divisor* or *factor* of $a(x)$, $a(x)$ is a *multiple* of $b(x)$.

Example 5.55. (1) $(x - 1) \mid (x^2 - 1)$ in $K[x]$ for an arbitrary field K;

(2) $(x + 1) \mid (x^2 + 1)$ in $\mathbb{Z}_2[x]$ since in this ring we have $x^2 + 1 = (x + 1)^2$;

(3) $(x+1) \nmid (x^2+1)$ in $\mathbb{R}[x]$.

Remark 5.56. In the following we often consider examples where K is some residue class field $\mathbb{Z}_p$ (p prime). As in the previous example, the coefficients of polynomials in $\mathbb{Z}_p[x]$ are written without the residue class mark. For example, $2+5x-x^3$ denotes polynomial $\overline{2}+\overline{5}x-x^3$. Note that

$$\begin{aligned} 2+5x-x^3 &= x-x^3 = x+x^3 \qquad \text{in } \mathbb{Z}_2[x], \\ 2+5x-x^3 &= 2-x^2 = 2+4x^3 \qquad \text{in } \mathbb{Z}_5[x]. \end{aligned}$$

The divisibility relation has many basic features in common with the division of integers, for example, we always have $a(x) \mid a(x)$ and

$$\begin{aligned} a(x) \mid b(x),\ b(x) \mid c(x) \qquad &\Longrightarrow \qquad a(x) \mid c(x), \\ a(x) \mid b(x),\ a(x) \mid c(x) \qquad &\Longrightarrow \qquad a(x) \mid (b(x)+c(x)). \end{aligned}$$

Some properties differ slightly, for example,

$$a(x) \mid b(c),\ b(x) \mid a(x) \qquad \Longrightarrow \qquad a(x) = k \cdot b(x), \quad \text{where } k \in K \setminus \{0\}.$$

If you wish to understand what causes the difference to $\mathbb{Z}$, consider the unit groups $\mathbb{Z}^* = \{\pm 1\}$ and $K[x]^* = K \setminus \{0\}$.

When examining the divisibility of a given polynomial $a(x)$ by another polynomial $b(x)$, we can use a similar division with remainder as for integers.

Theorem 5.57 (Polynomial division with remainder)**.** *If $a(x), b(x) \in K[x]$ and $b(x) \neq 0$, there exists unique polynomials $q(x)$, the dividend, and $r(x)$, the remainder, such that*

$$a(x) = q(x)b(x) + r(x), \qquad \deg r(x) < \deg b(x). \tag{5.5}$$

Proof. We choose from the set $S = \{a(x) - k(x)b(x) \mid k(x) \in K[x]\}$ a polynomial

$$r(x) = a(x) - q(x)b(x)$$

whose degree is the smallest possible. If $r(x) = 0$ then its degree is $-\infty$ and Equation (5.5) holds. Otherwise, we need to show that $n = \deg r(x)$ is smaller than $m = \deg b(x)$. Let the leading coefficients of $r(x)$ and $b(x)$ be r_n and b_m respectively. Now, if n is larger than m, we could form a polynomial

$$s(x) = a(x) - \left(q(x) + \frac{r_n}{b_n} \cdot x^{n-m}\right) b(x) = r(x) - \frac{r_n}{b_m} \cdot x^{n-m} b(x),$$

that belongs to the set S and whose n-th coefficient is

$$r_n - \frac{r_n}{b_m} \cdot b_m = 0.$$

From this we see that $s(x)$ is of smaller degree than $r(x)$ which contradicts our choice of $r(x)$. Hence $n < m$.

Uniqueness: If we also have $a(x) = q'(x)b(x) + r'(x)$ where $\deg r'(x) < \deg b(x)$, then

$$r(x) - r'(x) = (q'(x) - q(x))b(x), \qquad \deg(r(x) - r'(x)) < \deg b(x).$$

On the other hand, Theorem 5.52 states that

$$\deg(r(x) - r'(x)) = \deg(q'(x) - q(x)) + \deg b(x),$$

so we get $\deg(q'(x) - q(x)) < 0$. Then, necessarily, we have $q'(x) - q(x) = 0$. It follows that $r(x) - r'(x) = 0 \cdot b(x) = 0$. Thus $q'(x) = q(x)$ and $r'(x) = r(x)$. □

Example 5.58. Applied to the polynomials $a(x) = 2x^3 + x^2 - x - 1$ and $b(x) = x^2 - 2$ in the ring $\mathbb{Q}[x]$, the polynomial division gives

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

Thus we get dividend $q(x) = 2x + 1$ and remainder $r(x) = 3x + 1$.

What if we considered these polynomials $a(x)$ and $b(x)$ over some field $\mathbb{Z}_p$? In a field $\mathbb{Z}_p$, we need to consider reductions modulo p. However, these modular reductions can be applied at any point of the division algorithm, and therefore the formula holds.

In $\mathbb{Z}_2[x]$ we get

$$2x^3 + x^2 - x - 1 = (2x + 1)(x^2 - 2) + (3x + 1) = 1 \cdot (x^2) + (x + 1).$$

In $\mathbb{Z}_3[x]$ we get

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

Example 5.59. Applying the division algorithm to the polynomials $3x^2 + 1$ and $2x + 3$ in $\mathbb{Z}_5[x]$ gives

$$3x^2 + 1 = (4x + 4)(2x + 3) + (4).$$

If $f(x) = a_0 + a_1x + \cdots + a_nx^n \in K[x]$ and $c \in K$, we denote

$$f(c) = a_0 + a_1c + \cdots + a_nc^n.$$

Observe that $f(c)$ is an element of the field K; it can be read as "f evaluated at c". Thus we get the *polynomial map* familiar from analysis

$$K \to K, \qquad c \mapsto f(c).$$

If particularly $f(x) = 0$, we say that c is a zero or a root of the equation $f(x) = 0$.

If $f(a) = f(x) + g(x)$ and $b(x) = f(x) \cdot g(x)$, where $f(x), g(x) \in K[x]$, then we see that

$$a(c) = f(c) + g(c), \quad b(c) = f(c) \cdot g(c).$$

This result is the "substitution principle": Every equation satisfied by polynomials in $K[x]$ is satisfied in the field K when we substitute x for any element $c \in K$.

Theorem 5.60. *Let $f(x) \in K[x]$ and $c \in K$. Then*

$$f(c) = 0 \qquad \Longleftrightarrow \qquad (x - c) \mid f(x).$$

Proof. ($\Longrightarrow$) By the division algorithm, $f(x) = q(x)(x - c) + r(x)$ where $\deg r(x) < 1$. Hence $r(x)$ is a constant polynomial, $r(x) = r \in K$. Now when we substitute $x = c$, we get $r = f(c) = 0$. Thus $f(x)$ is divisible by the polynomial $x - c$.

($\Longleftarrow$) By assumption, $f(x) = (x - c)g(x)$ where $g(x) \in K[x]$. Substituting again $x = c$, we get $f(c) = 0$. □

As a consequence of this theorem, an n-th degree polynomial $f(x)$ over a field K has at most n distinct zeros in K. Suppose that $c_1, \ldots, c_k$ are such zeros. Then by the theorem, $f(x) = (x - c_1)g(x)$ where $g(x) \in K[x]$. Substituting $x = c_2$, we get

$$(c_2 - c_1)g(c_2) = 0.$$

Because $c_2 - c_1 \neq 0$ and fields have no zero divisors, it follows that $g(c_2) = 0$. Applying the theorem again, we get $g(x) = (x - c_2)h(x)$ where $h(x) \in K[x]$, and continuing in the same way we result in

$$(x - c_1)(x - c_2) \cdots (x - c_k) \mid f(x).$$

Thus the polynomial $f(x)$ has degree $\geq k$.

Example 5.61. Let p be a prime. We show that in the polynomial ring $\mathbb{Z}_p[x]$ we have

$$x^{p-1} - 1 = \prod_{a=1}^{p-1}(x - a).$$

By Fermat's little theorem, $x^{p-1} \equiv 0$ for all $x \in \mathbb{Z}_p \setminus \{0\}$. Thus all $a \in \mathbb{Z}_p \setminus \{0\}$ are zeros of the polynomial $x^{p-1} - 1$. These are all distinct and as a consequence of Theorem 5.60, $x^{p-1} - 1$ can have at most $p - 1$ distinct zeros in $\mathbb{Z}_p$. Thus these are all of its zeros. Therefore

$$x^{p-1} - 1 = \prod_{a=1}^{p-1}(x - a).$$

We conclude from this *Wilson's theorem* from number theory: $(p-1)! \equiv -1 \pmod{p}$.

From the previous equation it also follows that the zeros of the polynomial $f(x) = x^p - p$ in $\mathbb{Z}_p[x]$ are *all* the elements of the field $\mathbb{Z}_p$, that is, $f(x)$ is identically zero although it is not the zero polynomial. This situation would be absurd in $\mathbb{R}[x]$; review the general definition for equality of polynomials in Section 5.6.

Definition 5.62. A polynomial $f(x) \in K[x]$ is called *irreducible* if $f(x)$ is neither a constant polynomial nor a product of two polynomials in $K[x]$ of positive degree. We may also say that $f(x)$ is irreducible over the field K.

According to this definition all first degree polynomials are irreducible. If $\deg f(x) > 1$ and $f(x)$ has a zero in K, then it follows from Theorem 5.60 that $f(x)$ is not irreducible over K.

Example 5.63. The polynomial $x^2 + 1$ is irreducible over $\mathbb{R}$ because otherwise it would have a first degree factor $x - c \in \mathbb{R}[x]$, note that $ax + b = a\left(x + \frac{b}{a}\right)$, and thus by Theorem 5.60, a zero $c \in \mathbb{R}$.

However, $x^2 + 1$ is not irreducible over $\mathbb{C}$: $x^2 + 1 = (x + i)(x - i)$.

By applying the reasoning in this example, we see that *second and third degree polynomials in $K[x]$ are irreducible if and only if they have no zeros in K* – consider the ways the number 3 can be written as a sum of positive integers.

If $\deg f(x) > 3$, then $f(x)$ can naturally be written as a product of two irreducible polynomials whose degrees are greater than one; for example,

$$x^4 + 2x^2 + 1 = (x^2 + 1)^2 \in \mathbb{R}[x].$$

Then $f(x)$ is not irreducible but it also has no zeros.

Example 5.64. Let us determine whether $x^3 + 3x + 2$ is irreducible over the fields $\mathbb{Z}_3$ and $\mathbb{Z}_5$. The polynomial $x^3 + 3x + 2$ has degree 3 and thus it is irreducible if and only if it has no zeros in the field.

In $\mathbb{Z}_3$ the polynomial reduces to $x^3 + 3x + 2 = x + 2$. Manually going through the field elements

$$1 + 2 = 0, \quad 2 + 2 = 1, \quad 0 + 2 = 2,$$

we see that it is divisible by $x - 1$, and therefore it is not irreducible.

In $\mathbb{Z}_5$ the polynomial has no immediate modular reductions. Again, manually going through the field elements

$$\begin{aligned}
1^1 + 3 \cdot 1 + 2 &= 6 = 1, \\
2^3 + 3 \cdot 2 + 2 &= 16 = 1, \\
3^3 + 3 \cdot 3 + 2 &= 27 + 9 + 2 = 3, \\
4^3 + 3 \cdot 4 + 2 &= 64 + 12 + 2 = 3, \\
0^3 + 3 \cdot 0 + 2 &= 2
\end{aligned}$$

we see that the polynomial has no zeros and is thus irreducible over $\mathbb{Z}_5$.

Recall the familiar method for finding the rational roots of a polynomial $f(x) \in \mathbb{Q}[x]$. The *Fundamental theorem of algebra* states that every polynomial $f(x) \in \mathbb{C}[x] \setminus \mathbb{C}$ has a root in $\mathbb{C}$; thus $f(x)$ factors into first degree terms in $\mathbb{C}[x]$. The fundamental theorem of algebra is nicer to prove using the theory of complex functions, although more elementary proofs have been found. It also holds that every polynomial $f(x) \in K[x] \setminus K$ can be presented as a product of irreducible polynomials and this representation is unique up to the order of these polynomials and constant terms. We will not prove this here.

Exercises

1. Divide the polynomial $3x^4 + x^3 + 2x^2 + 1$ by the polynomial $x^2 + 4x + 2$ in the ring $\mathbb{R}[x]$. Does the answer change when the division is computed in the ring $\mathbb{Z}_5[x]$?
2. What are the roots of the polynomial $x^2 + 3x + 2$ in $\mathbb{Z}_6$?
3. Study the factorization of the polynomial $x^2 + 1$ over $\mathbb{Z}_5$.
4. Let F be a field and $a \neq 0,\ a \in F$. Prove:
   a) If $af(x)$ is irreducible over F, then $f(x)$ is irreducible over F.
   b) If $f(ax)$ is irreducible over F, then $f(x)$ is irreducible over F.
   c) If $f(x + a)$ is irreducible over F, then $f(x)$ is irreducible over F.
5. Factorize the polynomial $x^4 - 2$ into its prime factors in the rings $\mathbb{Q}[x]$, $\mathbb{R}[x]$, $\mathbb{C}[x]$. (Hint: Uniqueness of the prime factorization.)
6. In the ring $\mathbb{Z}_7[x]$, determine the greatest common divisor $D \in \mathbb{Z}_7[x]$ of the elements $F = 2x^4 + 1$ and $G = x^5 + 2x^4 + x^3 + 5$.
7. Show that the ideal I generated by the subset $\{2, x\}$ in the ring $\mathbb{Z}[x]$ is not a principal ideal. (Hint: Show that the constant term of each nonzero element of I is even, and use an indirect proof.)
8. Find the polynomials $U, V \in \mathbb{Z}_7[x]$ in Exercise 6 such that $D = UF + VG$.
9. Show that the quotient ring $A = \mathbb{Z}_3[x]/I$, where I is the principal ideal generated by the element $F = x^2 + 1$, is a field with nine elements. (Hint: Show that F is irreducible in the ring $\mathbb{Z}_3[x]$, and find the inverse elements by using the polynomial division.)

5.8 How irreducible polynomials form fields

Recall that the quotient ring R/I of a commutative ring R is a field if the ideal I is maximal, Theorem 5.40. In the following, we choose $R = K[x]$ where K is a given field, and we show that the principal ideals of $K[x]$ generated by irreducible polynomials are maximal. We can thus form new fields.

The principal ideal generated by a polynomial $f(x) \in K[x]$ is by Theorem 4.30 of the form

$$\langle f(x) \rangle = \{k(x)f(x) \mid k(x) \in K[x]\}.$$

Note that $\langle f(x) \rangle = \langle cf(x) \rangle$ for any nonzero element c of the field K.

Lemma 5.65. *Every ideal I of a polynomial ring $K[x]$ is principal, that is, $K[x]$ is PIR.*

Proof. Compare this with the proofs for Theorems 3.22 and 3.23. If $I = \{0\}$, it is a principal ideal, namely $\langle 0\rangle$. Suppose that $I \neq \{0\}$. Let $b(x)$ be a nonzero polynomial in I whose order is the smallest possible. We shall show that $I = \langle b(x)\rangle$.

Since $b(x) \in I$, we have $\langle b(x)\rangle \subset I$. Conversely, if $a(x) \in I$, then the division algorithm gives

$$a(x) = q(x)b(x) + r(x), \qquad \deg r(x) < \deg b(x).$$

Now $r(x) = a(x) - q(x)b(x) \in I$, hence $r(x) = 0$ by our choice of $b(x)$. From this we get that $a(x) = q(x)b(x) \in \langle b(x)\rangle$. Thus it follows that $I \subset \langle b(x)\rangle$. □

Theorem 5.66. *If $p(x)$ is an irreducible polynomial over $K[x]$, then the ideal $I = \langle p(x)\rangle$ is a maximal ideal of $K[x]$.*

Proof. Because $\deg p(x) \geq 1$, I contains no nonzero constant polynomials. It is thus a proper ideal. Let J be an ideal of $K[x]$ that has I its proper subset. We need to prove that $J = K[x]$.

By Lemma 5.65, J is a principal ideal, that is, $J = \langle b(x)\rangle$ for some $b(x) \in K[x]$. Because $I \subsetneq J$, the polynomial $p(x)$ is of the form $k(x)b(x)$ where $k(x) \in K[x]$. However, we assumed $p(x)$ to be irreducible; therefore $k(x)$ or $b(x)$ is a constant polynomial. If $k(x)$ is constant, then we see that $\langle p(x)\rangle = \langle b(x)\rangle$, or $I = J$, which contradicts our choice of J. Therefore $b(x)$ must be constant, $b(x) = b \in K \setminus \{0\}$. Then we get

$$J = \{b\} = \{1\} = K[x] \cdot 1 = K[x].$$

□

Corollary 5.67. *If $p(x)$ is an irreducible polynomial over $K[x]$, the quotient ring $K[x]/\langle p(x)\rangle$ is a field.*

What kind of field is $K[x]/\langle p(x)\rangle$? Let us denote $I = \langle p(x)\rangle$ and $d = \deg p(x)$. Like any quotient ring, $K[x]/I$ can be written as

$$K[x]/I = \{f(x) + I \mid f(x) \in K[x]\} = \{f(x) + I \mid f(x) \in D\}$$

where D is some collection of representatives of the cosets. Observe that $f_1(x)$ and $f_2(x)$ belong to the same residue class if and only if $f_1(x) - f_2(x)$ is divisible by $p(x)$. It follows from the division algorithm that

$$D = \{r(x) \in K[x] \mid \deg r(x) < d\}$$

is suitable as the collection of representatives. Thus

$$\begin{aligned} K[x]/I &= \{r(x) + I \mid r(x) \in K[x],\ \deg r(x) < d\} \\ &= \{a_0 + a_1x + \cdots + a_{d-1}x^{d-1} + I \mid a_i \in K \ \ \forall\, i\}. \end{aligned}$$

As usual, the sum and product of the residue classes $r_1(x) + I$ and $r_2(x) + I$ is formed by computing the sum and product of their representatives. The polynomial obtained as the product is returned to the form $a_0 + a_1x + \cdots + a_{d-1}x^{d-1}$ by subtracting a suitable multiple of $p(x)$, by using, say, the division algorithm.

Theorem 5.68. *The field $K[x]/\langle p(x)\rangle$ contains a subfield K' that is isomorphic to the field K. When K is equated with K', the field $K[x]/\langle p(x)\rangle$ becomes an extension of K. In this extension, the polynomial $p(x)$ has a root.*

Proof. As we did prior, we denote $I = \langle p(x) \rangle$. The map

$$j \colon K \to K[x]/I, \qquad j(a) = a + I,$$

is a field homomorphism (check). Its image $K' = \mathrm{Im}(j)$ is by Lemma 5.20 isomorphic to K:

$$K \simeq K' = \{a + I \mid a \in K\}.$$

Equating the fields K and K' implies that the elements a of K are equated to the cosets $a + I$.

In particular, the coefficients of the polynomial $p(x) \in K[x]$ can now be thought of as elements of the field $K[x]/I$. Because the symbol x got a certain meaning in the notation of the elements of this field, it is better to write the indeterminant of the polynomial as y, so $p(y) \in (K[x]/I)\,[y]$, for example. Which of the field elements is the root? Simply $x + I$ because by the rules of computation for cosets

$$p(x + I) = p(x) + I = 0 + I.$$

□

Example 5.69. Let us choose $K = \mathbb{R}$ and $p(x) = x^2 + 1$. Above we showed that we get the following field, now $I = \{x^2 + 1\}$:

$$\mathbb{R}[x]/I = \{a + bx + I \mid a, b \in \mathbb{R}\},$$

$$\begin{aligned}(a + bx + I) + (a' + b'x + I) &= (a + a') + (b + b')x + I,\\ (a + bx + I) \cdot (a' + b'x + I) &= (aa' - bb') + (ab' + a'b)x + I.\end{aligned}$$

When computing the product we subtracted $bb'(x^2 - 1)$ from the coset representative. Furthermore, we know that this field is an extension of $\mathbb{R}$ that contains a root for the equation $x^2 + 1 = 0$.

This thus formed field is none other than the field of complex numbers $\mathbb{C}$. This can be seen immediately if you denote $a + bx + I = (a, b)$ or $a + bx + I = a + bi$.

In the proof of the theorem we obtained $x + I$ as the root of the polynomial $y^2 + 1$, which is, using "better" notation, $(0, 1)$ or $0 + 1 \cdot i = i$, as it should.

Example 5.70. The polynomial $x^2 + x + 1$ is irreducible over the field $\mathbb{Z}_2$ since $x^2 + x + 1 \equiv x + x + 1 \equiv 1 \pmod 2$ and has no roots in $\mathbb{Z}_2$. Thus we get the field

$$\begin{aligned}\mathbb{Z}_2[x]/I &= \{a + bx + I \mid a, b \in \mathbb{Z}_2\} \qquad (I = \langle x^2 + x + 1 \rangle)\\ &= \{I, 1 + I, x + I, 1 + x + I\}.\end{aligned}$$

This is the field $GF(2^2)$ with four elements. If we denote the field elements by $0 = I$ (additive identity), $1 = 1 + I$ (identity), $\alpha = x + I$ and $\beta = 1 + x + I$, we get the following tables for its additive and multiplicative groups:

$+$	0	1	α	β
0	0	1	α	β
1	1	0	β	α
α	α	β	0	1
β	β	α	1	0

$\cdot$	1	α	β
1	1	α	β
α	α	β	1
β	β	1	α

Here the product $\alpha\beta$ was computed as

$$\alpha\beta = x(1 + x) + I = x + x^2 + I = -1 + I = 1 + I = 1.$$

A similar construction can be created by starting from any polynomial $q(x)$ that is irreducible over some field $\mathbb{Z}_p$. The result is then a finite field $GF(p^d)$ where $d = \deg q(x)$. Compare the previous addition table to the Klein four-group in Section 2.2.

Index