5 Algebra and Discrete Mathematics

5.1 Logic

5.1.1 Propositional Calculus

1. Propositions

A *proposition* is the mental reflection of a fact, expressed as a sentence in a natural or artificial language. Every proposition is considered to be true or false. This is the *principle of two-valuedness* (in contrast to many-valued or fuzzy logic, see 5.9.1, p. 413). "True" and "false" are called the *truth value* of the proposition and they are denoted by T (or 1) and F (or 0), respectively. The truth values can be considered as *propositional constants*.

2. Propositional Connectives

Propositional logic investigates the truth of *compositions of propositions* depending on the truth of the components. Only the *extensions* of the sentences corresponding to propositions are considered. Thus the truth of a composition depends *only* on that of the components and on the operations applied. So in particular, the truth of the result of the propositional operations

$$\text{“NOT } A\text{”} \quad (\neg A), \qquad (5.1) \qquad\qquad \text{“}A \text{ AND } B\text{”} \quad (A \wedge B), \qquad (5.2)$$

$$\text{“}A \text{ OR } B\text{”} \quad (A \vee B), \qquad (5.3) \qquad\qquad \text{“IF } A\text{, THEN } B\text{”} \quad (A \Rightarrow B) \qquad (5.4)$$

and

$$\text{“}A \text{ IF AND ONLY IF } B\text{”} (A \Leftrightarrow B) \qquad (5.5)$$

are determined by the truth of the components. Here "logical OR" always means "inclusive OR", i.e., "AND/OR". In the case of implication, for $A \Rightarrow B$ also the following verbal forms are in use:

A implies B, $\quad$ B is necessary for A, $\quad$ A is sufficient for B.

3. Truth Tables

In propositional calculus, the propositions A and B are considered as variables (*propositional variables*) which can have only the values F and T. Then the *truth tables* in **Table 5.1** contain the *truth functions* defining the propositional operations.

Table 5.1 Truth tables of propositional calculus

Negation

A	$\neg A$
F	T
T	F

Conjunction

A	B	$A \wedge B$
F	F	F
F	T	F
T	F	F
T	T	T

Disjunction

A	B	$A \vee B$
F	F	F
F	T	T
T	F	T
T	T	T

Implication

A	B	$A \Rightarrow B$
F	F	T
F	T	T
T	F	F
T	T	T

Equivalence

A	B	$A \Leftrightarrow B$
F	F	T
F	T	F
T	F	F
T	T	T

4. Formulas in Propositional Calculus

Compound expressions (formulas) of propositional calculus can be composed from the propositional variables in terms of a unary operation (negation) and binary operations (conjunction, disjunction, implication and equivalence). These expressions, i.e., the formulas, are defined in an inductive way:

1. Propositional variables and the constants T, F are formulas. (5.6)

2. If A and B are formulas, then $(\neg A)$, $(A \wedge B)$, $(A \vee B)$, $(A \Rightarrow B)$, $(A \Leftrightarrow B)$ (5.7)

I.N. Bronshtein et al., *Handbook of Mathematics*,
DOI 10.1007/978-3-662-46221-8_5

are also formulas.

To simplify formulas parentheses are omitted after introducing *precedence rules*. In the following sequence every propositional operation binds more strongly than the next one in the sequence:

$$\neg,\ \wedge,\ \vee,\ \Rightarrow,\ \Leftrightarrow .$$

Often the notation $\overline{A}$ instead of "$\neg A$" is used, and the symbol $\wedge$ is omitted. By these simplifications, for instance the formula $((A \vee (\neg B)) \Rightarrow ((A \wedge B) \vee C))$ can be rewritten more briefly in the form:

$$A \vee \overline{B} \Rightarrow AB \vee C.$$

5. Truth Functions

Assigning a truth value to every propositional variable of a formula, the assignment is called an *interpretation* of the propositional variables. Using the definitions (truth tables) of propositional operations a truth value can be assigned to a formula for every possible interpretation of the variables. Thus for instance the formula given above determines a truth function of three variables (a *Boolean function* see 5.7.5, p. 413).

A	B	C	$A \vee \overline{B}$	$AB \vee C$	$A \vee \overline{B} \Rightarrow AB \vee C$
F	F	F	T	F	F
F	F	T	T	T	T
F	T	F	F	F	T
F	T	T	F	T	T
T	F	F	T	F	F
T	F	T	T	T	T
T	T	F	T	T	T
T	T	T	T	T	T

■ In this way, every formula with n propositional variables determines an n place (or n ary) truth function, i.e., a function which assigns a truth value to every n tuple of truth values. There are 2^{2^n} n ary truth functions, in particular these are 16 binary ones.

6. Elementary Laws in Propositional Calculus

Two propositional formulas A and B are called *logically equivalent* or *semantically equivalent*, denoted by $A = B$, if they determine the same truth function. Consequently, the logical equivalence of propositional formulas can be checked in terms of truth tables. So there is , e.g., $A \vee \overline{B} \Rightarrow AB \vee C = B \vee C$, i.e., the formula $A \vee \overline{B} \Rightarrow AB \vee C$ does not in fact depend on A, as follows from its truth table above. In particular, there are the following *elementary laws of propositional calculus*:

1. Associative Laws

$$(A \wedge B) \wedge C = A \wedge (B \wedge C), \qquad (5.8a) \qquad\qquad (A \vee B) \vee C = A \vee (B \vee C). \qquad (5.8b)$$

2. Commutative Laws

$$A \wedge B = B \wedge A, \qquad (5.9a) \qquad\qquad A \vee B = B \vee A. \qquad (5.9b)$$

3. Distributive Laws

$$(A \vee B)C = AC \vee BC, \qquad (5.10a) \qquad\qquad AB \vee C = (A \vee C)(B \vee C). \qquad (5.10b)$$

4. Absorption Laws

$$A(A \vee B) = A, \qquad (5.11a) \qquad\qquad A \vee AB = A. \qquad (5.11b)$$

5. Idempotence Laws

$$AA = A, \qquad (5.12a) \qquad\qquad A \vee A = A. \qquad (5.12b)$$

6. Excluded Middle

$$A\overline{A} = \mathrm{F}, \qquad (5.13a) \qquad\qquad A \vee \overline{A} = \mathrm{T}. \qquad (5.13b)$$

7. De Morgan Rules

$$\overline{AB} = \overline{A} \vee \overline{B}, \qquad (5.14a) \qquad\qquad \overline{A \vee B} = \overline{A}\,\overline{B}. \qquad (5.14b)$$

8. Laws for T and F

$$A\mathrm{T} = A, \quad (5.15a) \qquad A \vee \mathrm{F} = A, \quad (5.15b)$$

$$A\mathrm{F} = \mathrm{F}, \quad (5.15c) \qquad A \vee \mathrm{T} = \mathrm{T}, \quad (5.15d)$$

$$\overline{\mathrm{T}} = \mathrm{F}, \quad (5.15e) \qquad \overline{\mathrm{F}} = \mathrm{T}. \quad (5.15f)$$

9. Double Negation

$$\overline{\overline{A}} = A. \quad (5.16)$$

Using the truth tables for implication and equivalence, gives the identities

$$A \Rightarrow B = \overline{A} \vee B \quad (5.17a) \qquad \text{and } A \Leftrightarrow B = AB \vee \overline{A}\,\overline{B}. \quad (5.17b)$$

Therefore implication and equivalence can be expressed in terms of other propositional operations. Laws (5.17a), (5.17b) are applied to reformulate propositional formulas.

■ The identity $A \vee \overline{B} \Rightarrow AB \vee C = B \vee C$ can be verified in the following way: $A \vee \overline{B} \Rightarrow AB \vee C = \overline{A \vee \overline{B}} \vee AB \vee C = \overline{A}\,\overline{\overline{B}} \vee AB \vee C = \overline{A}B \vee AB \vee C = (\overline{A} \vee A)B \vee C = \mathrm{T}B \vee C = B \vee C.$

10. Further Transformations

$$A(\overline{A} \vee B) = AB, \quad (5.18a) \qquad A \vee \overline{A}B = A \vee B, \quad (5.18b)$$

$$(A \vee C)(B \vee \overline{C})(A \vee B) = (A \vee C)(B \vee \overline{C}), \quad (5.18c) \qquad AC \vee B\overline{C} \vee AB = AC \vee B\overline{C}. \quad (5.18d)$$

11. NAND Function and NOR Function As it is known, every propositional formula determines a truth function. Checking the following converse of this statement: Every truth function can be represented as a truth table of a suitable formula in propositional logic. Because of (5.17a) and (5.17b) implication and equivalence can be eliminated from formulas (see also 5.7, p. 395). This fact and the De Morgan rules (5.14a) and (5.14b) imply that one can express every formula, therefore every truth function, in terms of negation and disjunction only, or in terms of negation and conjunction. There are two further binary truth functions of two variables which are suitable to express all the truth functions. They are called the NAND function or Sheffer function (notation "|") and the NOR function or Peirce function (notation "↓"), with the truth tables given in **Tables 5.2** and **5.3**. Comparison of the truth tables for these operations with the truth tables of conjunction and disjunction makes the terminologies NAND function (NOT AND) and NOR function (NOT OR) clear.

Table 5.2 NAND function

A	B	$A\|B$
F	F	T
F	T	T
T	F	T
T	T	F

Table 5.3 NOR function

A	B	$A \downarrow B$
F	F	T
F	T	F
T	F	F
T	T	F

7. Tautologies, Inferences in Mathematics

A formula in propositional calculus is called a *tautology* if the value of its truth function is identically the value T. Consequently, two formulas A and B are called logically equivalent if the formula $A \Leftrightarrow B$ is a tautology. Laws of propositional calculus often reflect inference methods used in mathematics. As an example, consider the *law of contraposition*, i.e., the tautology

$$A \Rightarrow B \Leftrightarrow \overline{B} \Rightarrow \overline{A}. \quad (5.19a)$$

This law, which also has the form

$$A \Rightarrow B = \overline{B} \Rightarrow \overline{A}, \quad (5.19b)$$

can be interpreted in this way: To show that B is a consequence of A is the same as showing that $\overline{A}$ is a consequence of $\overline{B}$. The *Indirect proof* (see also 1.1.2.2, p. 5) is based on the following principle: To show that B is a consequence of A, one supposes B to be false, and under the assumption that A is true, one derives a contradiction. This principle can be formalized in propositional calculus in several ways:

$$A \Rightarrow B = A\overline{B} \Rightarrow \overline{A} \qquad (5.20a) \qquad \text{or} \quad A \Rightarrow B = A\overline{B} \Rightarrow B \quad \text{or} \qquad (5.20b)$$

$$A \Rightarrow B = A\overline{B} \Rightarrow \text{F}. \qquad (5.20c)$$

5.1.2 Formulas in Predicate Calculus

For developing the logical foundations of mathematics one needs a logic which has a stronger expressive power than propositional calculus. To describe the properties of most of the objects in mathematics and the relations between these objects the predicate calculus is needed.

1. Predicates

The objects to be investigated are included into a set, i.e., into the *domain X of individuals (or universe)*, e.g., this domain could be the set $\mathbb{N}$ of the natural numbers. The properties of the individuals, as, e.g., " n is a prime ", and the relations between individuals, e.g., " m is smaller than n ", are considered as *predicates*. An *n place predicate* over the domain X of individual is an assignment P: $X^n \rightarrow \{\text{F,W}\}$, which assigns a truth value to every n tuple of the individuals. So the predicates introduced above on natural numbers are a one-place (or unary) predicate and a two-place (or binary) predicate.

2. Quantifiers

A characteristic feature of predicate logic is the use of *quantifiers*, i.e., that of a *universal quantifier* or *"for every" quantifier* $\forall$ and *existential quantifier* or *"for some" quantifier* $\exists$. If P is a unary predicate, then the sentence "$P(x)$ is true for every x in X" is denoted by $\forall\, x\, P(x)$ and the sentence" There exists an x in X for which $P(x)$ is true "is denoted by $\exists\, x\, P(x)$. Applying a quantifier to the unary predicate P, gives a sentence. If for instance $\mathbb{N}$ is the domain of individual of the natural numbers and P denotes the (unary) predicate "n is a prime", then $\forall\, n\, P(n)$ is a false sentence and $\exists\, n\, P(n)$ is a true sentence.

3. Formulas in Predicate Calculus

The *formulas in predicate calculus* are defined in an inductive way:

1. If $x_1, \ldots, x_n$ are individual variables (variables running over the domain of individual variables) and P is an n-place predicate symbol, then

$$P(x_1, \ldots, x_n) \text{ is a formula } (\textit{elementary formula}). \qquad (5.21)$$

2. If A and B are formulas, then

$$(\neg A),\ (A \wedge B),\ (A \vee B),\ (A \Rightarrow B), (A \Leftrightarrow B),\ (\forall\, x\, A) \text{ and } (\exists\, x\, A) \qquad (5.22)$$

are also formulas.

Considering a propositional variable to be a null-place predicate, the propositional calculus can be considered as a part of predicate calculus. An occurrence of an individual variable x is *bound* in a formula if x is a variable in $\forall\, x$ or in $\exists\, x$ or the occurrence of x is in the scope of these types of quantifiers; otherwise an occurrence of x is *free* in this formula. A formula of predicate logic which does not contain any free occurrences of individual variables is called a *closed formula*.

4. Interpretation of Predicate Calculus Formulas

An *interpretation* of predicate calculus is a pair of

- a set (domain of individuals) and
- an assignment, which assigns an n-place predicate to every n-ary predicate symbol.

For every prefixed value of free variables the concept of the truth evaluation of a formula is similar to the propositional case. The truth value of a closed formula is T or F. In the case of a formula containing free variables, one can associate the values of individuals for which the truth evaluation of the formula is true; these values constitute a relation (see 5.2.3, **1.**, p. 331) on the universe (domain of individuals).

■ Let P denote the two-place relation $\leq$ on the domain $\mathbb{N}$ of individuals, where $\mathbb{N}$ is the set of the natural numbers then

- $P(x,y)$ characterizes the set of all the pairs (x,y) of natural numbers with $x \le y$ (two-place or binary relation on $\mathbb{N}$); here x, y are free variables;
- $\forall y\, P(x,y)$ characterizes the subset of $\mathbb{N}$ (unary relation) consisting of the element 0 only; here x is a free variable, y is a bound variable;
- $\exists x\, \forall y\, P(x,y)$ corresponds to the sentence " There is a smallest natural number "; the truth value is true; here x and y are bound variables.

5. Logically Valid Formulas

A formula is called *logically valid* (or a *tautology*) if it is true for every interpretation. The negation of formulas is characterized by the identities below:

$$\neg\forall x\, P(x) = \exists x\, \neg P(x) \quad \text{or} \quad \neg\exists x\, P(x) = \forall x\, \neg P(x). \tag{5.23}$$

Using (5.23) the quantifiers $\forall$ and $\exists$ can be expressed in terms of each other:

$$\forall x\, P(x) = \neg\exists x\, \neg P(x) \quad \text{or} \quad \exists x\, P(x) = \neg\forall x\, \neg P(x). \tag{5.24}$$

Further identities of the predicate calculus are:

$$\forall x\, \forall y\, P(x,y) = \forall y\, \forall x\, P(x,y), \tag{5.25}$$

$$\exists x\, \exists y\, P(x,y) = \exists y\, \exists x\, P(x,y), \tag{5.26}$$

$$\forall x\, P(x) \wedge \forall x\, Q(x) = \forall x\, (P(x) \wedge Q(x)), \tag{5.27}$$

$$\exists x\, P(x) \vee \exists x\, Q(x) = \exists x\, (P(x) \vee Q(x)). \tag{5.28}$$

The following implications are also valid:

$$\forall x\, P(x) \vee \forall x\, Q(x) \Rightarrow \forall x\, (P(x) \vee Q(x)), \tag{5.29}$$

$$\exists x\, (P(x) \wedge Q(x)) \Rightarrow \exists x\, P(x) \wedge \exists x\, Q(x), \tag{5.30}$$

$$\forall x\, (P(x) \Rightarrow Q(x)) \Rightarrow (\forall x\, P(x) \Rightarrow \forall x\, Q(x)), \tag{5.31}$$

$$\forall x\, (P(x) \Leftrightarrow Q(x)) \Rightarrow (\forall x\, P(x) \Leftrightarrow \forall x\, Q(x)), \tag{5.32}$$

$$\exists x\, \forall y\, P(x,y) \Rightarrow \forall y\, \exists x\, P(x,y). \tag{5.33}$$

The converses of these implications are not valid, in particular, one has to be careful with the fact that the quantifiers $\forall$ and $\exists$ do not commute (the converse of the last implication is false).

6. Restricted Quantification

Often it is useful to restrict quantification to a subset of a given set. So, there is considered

$$\forall x \in X\; P(x) \quad \text{as a short notation of} \quad \forall x\, (x \in X \Rightarrow P(x)) \quad \text{and} \tag{5.34}$$

$$\exists x \in X\; P(x) \quad \text{as a short notation of} \quad \exists x\, (x \in X \wedge P(x)). \tag{5.35}$$

5.2 Set Theory

5.2.1 Concept of Set, Special Sets

The founder of set theory is Georg Cantor (1845–1918). The importance of the notion introduced by him became well known only later. Set theory has a decisive role in all branches of mathematics, and today it is an essential tool of mathematics and its applications.

1. Membership Relation

1. Sets and their Elements The fundamental notion of set theory is the membership relation. A *set* A is a collection of certain different things a (objects, ideas, etc.) that belong together for certain reasons. These objects are called the *elements* of the set. One writes "$a \in A$" or "$a \notin A$" to denote "a is an element of A" or "a is not an element of A", respectively. Sets can be given by enumerating their elements in braces, e.g., $M = \{a, b, c\}$ or $U = \{1, 3, 5, \ldots\}$, or by a defining property possessed exactly by the elements of the set. For instance the set U of the odd natural numbers is defined and denoted by $U = \{x \mid x \text{ is an odd natural number}\}$. For number domains the following notation is generally used:

$\mathbb{N} = \{0, 1, 2, \ldots\}$	set of the natural numbers,
$\mathbb{Z} = \{0, 1, -1, 2, -2, \ldots\}$	set of the integers,
$\mathbb{Q} = \left\{\frac{p}{q} \,\middle|\, p, q \in \mathbb{Z} \wedge q \neq 0\right\}$	set of the rational numbers,
$\mathbb{R}$	set of the real numbers,
$\mathbb{C}$	set of the complex numbers.

2. Principle of Extensionality for Sets Two sets A and B are identical if and only if they have exactly the same elements, i.e.,

$$A = B \Leftrightarrow \forall\, x\, (x \in A \Leftrightarrow x \in B). \tag{5.36}$$

■ The sets $\{3, 1, 3, 7, 2\}$ and $\{1, 2, 3, 7\}$ are the same.

A set contains every element only “once”, even if it is enumerated several times.

2. Subsets

1. Subset If A and B are sets and

$$\forall\, x\, (x \in A \Rightarrow x \in B) \tag{5.37}$$

holds, then A is called a *subset* of B, and this is denoted by $A \subseteq B$. In other words: A is a subset of B, if all elements of A also belong to B. If for $A \subseteq B$ there are some further elements in B such that they are not in A, then A is called a *proper subset* of B, and it is denoted by $A \subset B$ **(Fig. 5.1)**. Obviously, every set is a subset of itself $A \subseteq A$.

■ Suppose $A = \{2, 4, 6, 8, 10\}$ is a set of even numbers and $B = \{1, 2, 3, \ldots, 10\}$ is a set of natural numbers. Since the set A does not contain odd numbers, A is a proper subset of B.

2. Empty Set or Void Set It is important and useful to introduce the notion of *empty set* or *void set*, $\emptyset$, which has no element. Because of the principle of extensionality, there exists only one empty set.

■ **A:** The set $\{x | x \in \mathbb{R} \wedge x^2 + 2x + 2 = 0\}$ is empty.

■ **B:** $\emptyset \subseteq M$ for every set M, i.e., the empty set is a subset of every set M.

For a set A the empty set and A itself are called the *trivial subsets of* A.

3. Equality of Sets Two sets are equal if and only if both are subsets of each other:

$$A = B \Leftrightarrow A \subseteq B \wedge B \subseteq A. \tag{5.38}$$

This fact is very often used to prove that two sets are identical.

4. Power Set The set of all subsets A of a set M is called the *power set* of M and it is denoted by $\mathbb{P}(M)$, i.e., $\mathbb{P}(M) = \{A \mid A \subseteq M\}$.

■ For the set $M = \{a, b, c\}$ the power set is

$$\mathbb{P}(M) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}.$$

It is true that:

a) If a set M has m elements, its power set $\mathbb{P}(M)$ has 2^m elements.

b) For every set M there are $M, \emptyset \in \mathbb{P}(M)$, i.e., M itself and the empty set are elements of the power set of M.

5. Cardinal number The number of elements of a finite set M is called the *cardinal number* of M and it is denoted by $\operatorname{card} M$ or sometimes by $|M|$.

For the the cardinal number of sets with infinitely many elements see 5.2.5, p. 335.

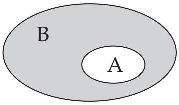
5.2.2 Operations with Sets

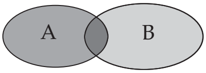
1. Venn diagram

The graphical representations of sets and set operations are the so-called *Venn diagrams*, when representing sets by plane figures. So, **Fig. 5.1**, represents the subset relation $A \subseteq B$.

2. Union, Intersection, Complement

By *set operations* new sets can be formed from the given sets in different ways:





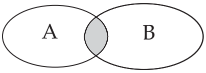


Figure 5.1 Figure 5.2 Figure 5.3

1. Union Let A and B be two sets. The *union set* or the *union* (denoted by $A \cup B$) is defined by

$$A \cup B = \{x \mid x \in A \vee x \in B\}, \tag{5.39}$$

in words "A union B" or "A cup B". If A and B are given by the properties E_1 and E_2 respectively, the union set $A \cup B$ has the elements possessing at least one of these properties, i.e., the elements belonging to at least one of the sets. In **Fig. 5.2** the union set is represented by the shaded region.

■ $\{1,2,3\} \cup \{2,3,5,6\} = \{1,2,3,5,6\}$.

2. Intersection Let A and B be two sets. The *intersection set*, *intersection*, *cut* or *cut set* (denoted by $A \cap B$) is defined by

$$A \cap B = \{x \mid x \in A \wedge x \in B\}, \tag{5.40}$$

in words "A intersected by B" or "A cap B". If A and B are given by the properties E_1 and E_2 respectively, the intersection $A \cap B$ has the elements possessing both properties E_1 and E_2, i.e., the elements belonging to both sets. In **Fig. 5.3** the intersection is represented by the shaded region.

■ With the intersection of the sets of divisors $T(a)$ and $T(b)$ of two numbers a and b one can define the greatest common divisor (see 5.4.1.4, p. 373). For $a = 12$ and $b = 18$ holds $T(a) = \{1,2,3,4,6,12\}$ and $T(b) = \{1,2,3,6,9,18\}$, so $T(12) \cap T(18)$ contains the common divisors, and the greatest common divisor is g.c.d. $(12,18) = 6$.

3. Disjoint Sets Two sets A and B are called *disjoint* if they have no common element; for them

$$A \cap B = \emptyset \tag{5.41}$$

holds, i.e., their intersection is the empty set.

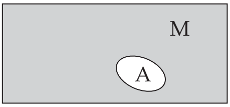
■ The set of odd numbers and the set of even numbers are disjoint; their intersection is the empty set, i.e.,

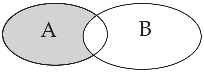
$$\{\text{odd numbers}\} \cap \{\text{even numbers}\} = \emptyset.$$

4. Complement Considering only the subsets of a given set M, then the *complementary set* or the *complement* $C_M(A)$ of A with respect to M contains all the elements of M not belonging to A:

$$C_M(A) = \{x \mid x \in M \wedge x \notin A\}, \tag{5.42}$$

in words "complement of A with respect to M", and M is called the *fundamental set* or sometimes the *universal set*. If the fundamental set M is obvious from the considered problem, then the notation $\overline{A}$ is also used for the complementary set. In **Fig. 5.4** the complement $\overline{A}$ is represented by the shaded region.





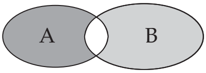


Figure 5.4 Figure 5.5 Figure 5.6

3. Fundamental Laws of Set Algebra

These set operations have analoguous properties to the operations in logic. The *fundamental laws of set algebra* are:

1. Associative Laws

$$(A \cap B) \cap C = A \cap (B \cap C), \quad (5.43) \qquad (A \cup B) \cup C = A \cup (B \cup C). \quad (5.44)$$

2. Commutative Laws

$$A \cap B = B \cap A, \quad (5.45) \qquad A \cup B = B \cup A. \quad (5.46)$$

3. Distributive Laws

$$(A \cup B) \cap C = (A \cap C) \cup (B \cap C), \quad (5.47) \qquad (A \cap B) \cup C = (A \cup C) \cap (B \cup C). \quad (5.48)$$

4. Absorption Laws

$$A \cap (A \cup B) = A, \quad (5.49) \qquad A \cup (A \cap B) = A. \quad (5.50)$$

5. Idempotence Laws

$$A \cap A = A, \quad (5.51) \qquad A \cup A = A. \quad (5.52)$$

6. De Morgan Laws

$$\overline{A \cap B} = \overline{A} \cup \overline{B}, \quad (5.53) \qquad \overline{A \cup B} = \overline{A} \cap \overline{B}. \quad (5.54)$$

7. Some Further Laws

$$A \cap \overline{A} = \emptyset, \quad (5.55) \qquad A \cup \overline{A} = M \ \ (M \text{ fundamental set}), \quad (5.56)$$

$$A \cap M = A, \quad (5.57) \qquad A \cup \emptyset = A, \quad (5.58)$$

$$A \cap \emptyset = \emptyset, \quad (5.59) \qquad A \cup M = M, \quad (5.60)$$

$$\overline{M} = \emptyset, \quad (5.61) \qquad \overline{\emptyset} = M. \quad (5.62)$$

$$\overline{\overline{A}} = A. \quad (5.63)$$

This table can also be obtained from the fundamental laws of propositional calculus (see 5.1.1, p. 323) using the following substitutions: $\wedge$ by $\cap$, $\vee$ by $\cup$, T by M, and F by $\emptyset$. This coincidence is not accidental; it will be discussed in 5.7, p. 395.

4. Further Set Operations

In addition to the operations defined above there are defined some further operations between two sets A and B: the *difference set* or *difference* $A \setminus B$, the *symmetric difference* $A \triangle B$ and the *Cartesian product* $A \times B$.

1. Difference of Two Sets The set of the elements of A, not belonging to B is the *difference set* or *difference* of A and B:

$$A \setminus B = \{x \mid x \in A \wedge x \notin B\}. \quad (5.64a)$$

If A is defined by the property E_1 and B by the property E_2, then $A \setminus B$ contains the elements having the property E_1 but not having property E_2.

In **Fig. 5.5** the difference is represented by the shaded region.

■ $\{1, 2, 3, 4\} \setminus \{3, 4, 5\} = \{1, 2\}$.

2. Symmetric Difference of Two Sets The symmetric difference $A \triangle B$ is the set of all elements belonging to exactly one of the sets A and B:

$$A \triangle B = \{x \mid (x \in A \wedge x \notin B) \vee (x \in B \wedge x \notin A)\}. \quad (5.64b)$$

It follows from the definition that

$$A \triangle B = (A \setminus B) \cup (B \setminus A) = (A \cup B) \setminus (A \cap B), \quad (5.64c)$$

i.e., the symmetric difference contains the elements which have exactly one of the defining properties E_1 (for A) and E_2 (for B).

In **Fig. 5.6** the symmetric difference is represented by the shaded region.

■ $\{1,2,3,4\} \triangle \{3,4,5\} = \{1,2,5\}$.

3. Cartesian Product of Two Sets The *Cartesian product* of two sets $A \times B$ is defined by

$$A \times B = \{(a,b) \mid a \in A \wedge b \in B\}. \tag{5.65a}$$

The elements (a,b) of $A \times B$ are called *ordered pairs* and they are characterized by

$$(a,b) = (c,d) \Leftrightarrow a = c \wedge b = d. \tag{5.65b}$$

The number of the elements of a Cartesian product of two finite sets is equal to

$$\mathrm{card}\,(A \times B) = (\mathrm{card}A)(\mathrm{card}B). \tag{5.65c}$$

■ **A:** For $A = \{1,2,3\}$ and $B = \{2,3\}$ one gets $A \times B = \{(1,2),(1,3),(2,2),(2,3),(3,2),(3,3)\}$ and $B \times A = \{(2,1),(2,2),(2,3),(3,1),(3,2),(3,3)\}$ with $\mathrm{card}A = 3$, $\mathrm{card}B = 2$, $\mathrm{card}(A \times B) = \mathrm{card}(B \times A) = 6$.

■ **B:** Every point of the x, y plane can be defined with the Cartesian product $\mathbb{R} \times \mathbb{R}$ ($\mathbb{R}$ is the set of real numbers). The set of the coordinates x, y is represented by $\mathbb{R} \times \mathbb{R}$, so:

$$\mathbb{R}^2 = \mathbb{R} \times \mathbb{R} = \{(x,y) \mid x \in \mathbb{R}, y \in \mathbb{R}\}.$$

4. Cartesian Product of *n* Sets

From n elements, by fixing an order of sequence (first element, second element, ..., n-th element) an ordered n tuple is defined. If $a_i \in A_i$ $(i = 1, 2, \ldots, n)$ are the elements, the n tuple is denoted by $(a_1, a_2, \ldots, a_n)$, where a_i is called the i-th component.

For $n = 3, 4, 5$ these n tuples are called *triples, quadruples*, and *quintuples*.

The Cartesian product of n terms $A_1 \times A_2 \times \cdots \times A_n$ is the set of all ordered n tuples $(a_1, a_2, \ldots, a_n)$ with $a_i \in A_i$:

$$A_1 \times \ldots \times A_n = \{(a_1, \ldots, a_n) \mid a_i \in A_i \ \ (i = 1, \ldots, n)\}. \tag{5.66a}$$

If every A_i is a finite set, the number of ordered n tuples is

$$\mathrm{card}(A_1 \times A_2 \times \cdots \times A_n) = \mathrm{card}A_1\,\mathrm{card}A_2 \cdots \mathrm{card}A_n. \tag{5.66b}$$

Remark: The n times Cartesian product of a set A with itself is denoted by A^n.

5.2.3 Relations and Mappings

1. *n* ary Relations

Relations define correspondences between the elements of one or different sets. An *n ary relation* or *n-place relation* R between the sets $A_1, \ldots, A_n$ is a subset of the Cartesian product of these sets, i.e., $R \subseteq A_1 \times \ldots \times A_n$. If the sets A_i, $i = 1, \ldots, n$, are all the same set A, then $R \subseteq A^n$ holds and it is called an n ary relation in the set A.

2. Binary Relations

1. Notion of Binary Relations of a Set The two-place *(binary)* relations in a set have special importance.

In the case of a binary relation the notation aRb is also very common instead of $(a,b) \in R$.

■ As an example, the divisibility relation in the set $A = \{1,2,3,4\}$ is considered, i.e., the binary relation

$$T = \{(a,b) \mid a, b \in A \wedge \ a \text{ is a divisor of } b\} \tag{5.67a}$$

$$= \{(1,1),(1,2),(1,3),(1,4),(2,2),(2,4),(3,3),(4,4)\}. \tag{5.67b}$$

2. Arrow Diagram or Mapping Function Finite binary relations R in a set A can be represented by *arrow functions* or *arrow diagrams* or by *relation matrices*. The elements of A are represented as points of the plane and an arrow goes from a to b if aRb holds.

Fig. 5.7 shows the arrow diagram of the relation T in $A = \{1,2,3,4\}$.

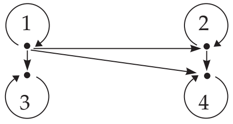


Figure 5.7

	1	2	3	4
1	1	1	1	1
2	0	1	0	1
3	0	0	1	0
4	0	0	0	1

Scheme: Relation matrix

3. Relation Matrix The elements of A are used as row and column entries of a matrix (see 4.1.1, **1.**, p. 269). At the intersection point of the row of $a \in A$ with the column of $b \in B$ there is an entry 1 if aRb holds, otherwise there is an entry 0. The above scheme shows the relation matrix for T in $A = \{1, 2, 3, 4\}$.

3. Relation Product, Inverse Relation

Relations are special sets, so the usual set operations (see 5.2.2, p. 328) can be performed between relations. Besides them, for binary relations, the *relation product* and the *inverse relation* also have special importance.
Let $R \subseteq A \times B$ and $S \subseteq B \times C$ be two binary relations. The product $R \circ S$ of the relations R, S is defined by

$$R \circ S = \{(a, c) \mid \exists\, b\, (b \in B \wedge aRb \wedge bSc)\}. \qquad (5.68)$$

The relation product is associative, but not commutative.

The inverse relation R^{-1} of a relation R is defined by

$$R^{-1} = \{(b, a) \mid (a, b) \in R\}. \qquad (5.69)$$

For binary relations in a set A the following relations are valid:

$$(R \cup S) \circ T = (R \circ T) \cup (S \circ T), \quad (5.70a) \qquad (R \cap S) \circ T \subseteq (R \circ T) \cap (S \circ T), \quad (5.70b)$$

$$(R \cup S)^{-1} = R^{-1} \cup S^{-1}, \quad (5.70c) \qquad (R \cap S)^{-1} = R^{-1} \cap S^{-1}, \quad (5.70d)$$

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}. \quad (5.70e)$$

4. Properties of Binary Relations

A binary relation in a set A can have special important properties:
R is called

$$\begin{aligned}
reflexive,&\ \text{ if } \forall\, a \in A\; aRa, && (5.71a)\\
irreflexive,&\ \text{ if } \forall\, a \in A\; \neg aRa, && (5.71b)\\
symmetric,&\ \text{ if } \forall\, a, b \in A\; (aRb \Rightarrow bRa), && (5.71c)\\
antisymmetric,&\ \text{ if } \forall\, a, b \in A\; (aRb \wedge bRa \Rightarrow a = b), && (5.71d)\\
transitive,&\ \text{ if } \forall\, a, b, c \in A\; (aRb \wedge bRc \Rightarrow aRc), && (5.71e)\\
linear,&\ \text{ if } \forall\, a, b \in A\; (aRb \vee bRa). && (5.71f)
\end{aligned}$$

These relations can also be described by the relation product. For instance: a binary relation is transitive if $R \circ R \subseteq R$ holds. Especially interesting is the *transitive closure* $\operatorname{tra}(R)$ of a relation R. It is the smallest (with respect to the subset relation) transitive relation which contains R. In fact

$$\operatorname{tra}(R) = \bigcup_{n \geq 1} R^n = R^1 \cup R^2 \cup R^3 \cup \cdots, \qquad (5.72)$$

where R^n is the n times relation product of R with itself.

■ Let a binary relation R on the set $\{1, 2, 3, 4, 5\}$ be given by its relation matrix M:

M	1	2	3	4	5
1	1	0	0	1	0
2	0	0	0	1	0
3	0	0	1	0	1
4	0	1	0	0	1
5	0	1	0	0	0

M^2	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	0	1
3	0	1	1	0	1
4	0	1	0	1	0
5	0	0	0	1	0

M^3	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	1	0
3	0	1	1	1	1
4	0	1	0	1	1
5	0	1	0	0	1

Calculating M^2 by matrix multiplication where the values 0 and 1 are treated as truth values and instead of multiplication and addition one performs the logical operations conjunction and disjunction, then, M^2 is the relation matrix belonging to R^2. The relation matrices of R^3, R^4 etc. can be calculated similarly.

$M \vee M^2 \vee M^3$	1	2	3	4	5
1	1	1	0	1	1
2	0	1	0	1	1
3	0	1	1	1	1
4	0	1	0	1	1
5	0	1	0	1	1

The relation matrix of $R \cup R^2 \cup R^3$ (the matrix on the left) can be get by calculating the disjunction elementwise of the matrices M, M^2 and M^3. Since the higher powers of M contains no new 1-s, this matrix already coincides with the relation matrix of $\operatorname{tra}(R)$.

The relation matrix and relation product have important applications in search of path length in graph theory (see 5.8.2.1, p. 404).

In the case of finite binary relations, one can easily recognize the above properties from the arrow diagrams or from the relation matrices. For instance one can recognize the reflexivity from “self-loops” in the arrow diagram, and from the main diagonal elements 1 in the relation matrix. Symmetry is obvious in the arrow diagram if to every arrow there belongs another one in the opposite direction, or if the relation matrix is a symmetric matrix (see 5.2.3, **2.**, p. 331). Easy to see from the arrow diagram or from the relation matrix that the divisibility T is a reflexive but not symmetric relation.

5. Mappings

A *mapping* or *function* f (see 2.1.1.1, p. 48) from a set A to a set B with the notation $f\colon A \to B$ is a rule to assign to every element $a \in A$ exactly one element $b \in B$, which is called $f(a)$.

A mapping f can be considered as a subset of $A \times B$ and so as a binary relation:

$$f = \{(a, f(a)) | a \in A\} \subseteq A \times B\,. \tag{5.73}$$

a) f is called a *injective* or *one to one* mapping, if to every $b \in B$ at most one $a \in A$ with $f(a) = b$ exists.

b) f is called a *surjective mapping* from A *to* B, if to every $b \in B$ at least one $a \in A$ with $f(a) = b$ exists.

c) f is called *bijective*, if f is both injective and surjective.

If A and B are finite sets, between which exists a bijective mapping, then A and B possess the same number of elements (see also 5.2.5, p. 335).

For a bijective mapping $f\colon A \to B$ exists the inverse relation $f^{-1}\colon B \to A$, the so-called *inverse mapping* of f.

The relation product of mappings is used for the one after the other composition of mappings: If f: $A \to B$ and g: $B \to C$ are mappings, then $f \circ g$ is also a mapping from A to C, and is defined by

$$(f \circ g)(a) = g(f(a))\,. \tag{5.74}$$

Remark: Be careful with the order of f and g in this equation (it is treated differently in the literature!).

5.2.4 Equivalence and Order Relations

The most important classes of binary relations with respect to a set A are the equivalence and order relations.

1. Equivalence Relations

A binary relation R with respect to a set A is called an *equivalence relation* if R is reflexive, symmetric, and transitive. For aRb also the notations $a \sim_R b$ or $a \sim b$ are used, if the equivalence relation R is already known, in words a is equivalent to b (with respect to R).

Examples of Equivalence Relations:

■ **A:** $A = \mathbb{Z}$, $m \in \mathbb{N} \setminus \{0\}$. $a \sim_R b$ holds exactly if a and b have the same remainder when divided by m (they are congruent modulo m).

■ **B:** Equality relation in different domains, e.g., in the set $\mathbb{Q}$ of rational numbers: $\frac{p_1}{q_1} = \frac{p_2}{q_2} \Leftrightarrow p_1q_2 = p_2q_1$ (p_1, p_2, q_1, q_2 integer; $q_1, q_2 \neq 0$), where the first equality sign defines an equality in $\mathbb{Q}$, while the second one denotes an equality in $\mathbb{Z}$.

■ **C:** Similarity or congruence of geometric figures.

■ **D:** Logical equivalence of expressions of propositional calculus (see 5.1.1, **6.**, p. 324).

2. Equivalence Classes, Partitions

1. Equivalence Classes An equivalence relation in a set A defines a partition of A into non-empty pairwise disjoint subsets, into *equivalence classes*.

$$[a]_R := \{b \mid b \in A \wedge a \sim_R b\} \tag{5.75}$$

is called an equivalence class of a with respect to R. For equivalence classes the following is valid:

$$[a]_R \neq \emptyset, \quad a \sim_R b \Leftrightarrow [a]_R = [b]_R\,, \quad \text{and} \quad a \not\sim_R b \Leftrightarrow [a]_R \cap [b]_R = \emptyset. \tag{5.76}$$

These equivalence classes form a new set, the *quotient set* A/R:

$$A/R = \{[a]_R \mid a \in A\}. \tag{5.77}$$

A subset $Z \subseteq \mathbb{P}(A)$ of the power set $\mathbb{P}(A)$ is called a *partition* of A if

$$\emptyset \notin Z, \quad X, Y \in Z \wedge X \neq Y \Rightarrow X \cap Y = \emptyset\,, \quad \bigcup_{X \in Z} X = A. \tag{5.78}$$

2. Decomposition Theorem Every equivalence relation R in a set A defines a partition Z of A, namely $Z = A/R$. Conversely, every partition Z of a set A defines an equivalence relation R in A:

$$a \sim_R b \Leftrightarrow \exists X \in Z\,(a \in X \wedge b \in X). \tag{5.79}$$

An equivalence relation in a set A can be considered as a generalization of the equality, where “ insignificant ” properties of the elements of A are neglected, and the elements, which do not differ with respect to a certain property, belong to the same equivalence class.

3. Ordering Relations

A binary relation R in a set A is called a *partial ordering* if R is reflexive, antisymmetric, and transitive. If in addition R is linear, then R is called a *linear ordering* or a *chain*. The set A is called ordered or linearly ordered by R. In a linearly ordered set any two elements are comparable. Instead of aRb also the notation $a \leq_R b$ or $a \leq b$ is used, if the ordering relation R is known from the problem.

Examples of Ordering Relations:

■ **A:** The sets of numbers $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ are completely ordered by the usual $\leq$ relation.

■ **B:** The subset relation is also an ordering, but only a partial ordering.

■ **C:** The *lexicographical order* of the English words is a chain.

Remark: If $Z = \{A, B\}$ is a partition of $\mathbb{Q}$ with the property $a \in A \wedge b \in B \Rightarrow a < b$, then (A, B) is called a *Dedekind cut*. If neither A has a greatest element nor B has a smallest element, so an irrational number is uniquely determined by this cut. Besides the nest of intervals (see 1.1.1.2, p. 2) the notion of Dedekind cuts is another way to introduce irrational numbers.

4. Hasse Diagram

Finite ordered sets can be represented by the *Hasse diagram*: Let an ordering relation $\leq$ be given on a finite set A. The elements of A are represented as points of the plane, where the point $b \in A$ is placed above the point $a \in A$ if $a < b$ holds. If there is no $c \in A$ for which $a < c < b$, one says a and b are *neighbors* or *consecutive members*. Then one connects a and b by a line segment.

A Hasse diagram is a "simplified" arrow diagram, where all the loops, arrowheads, and the arrows following from the transitivity of the relation are eliminated. The arrow diagram of the divisibility relation T of the set $A = \{1, 2, 3, 4\}$ is given in **Fig. 5.7**. T also denotes an ordering relation, which is represented by the Hasse diagram in **Fig. 5.8**.

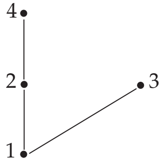


Figure 5.8

5.2.5 Cardinality of Sets

In 5.2.1, p. 327 the number of elements of a finite set was called the cardinality of the set. This notion of cardinality can be extended to infinite sets.

1. Cardinal Numbers

Two sets A and B are called *equinumerous* if there is a bijective mapping between them. To every set A a *cardinal number* $|A|$ or card A is assigned, so that equinumerous sets have the same cardinal number. A set and its power set are never equinumerous, so no " greatest " cardinal number exists.

2. Infinite Sets

Infinite sets can be characterized by the property that they have proper subsets equinumerous to the set itself. The "smallest" infinite cardinal number is the cardinal number of the set $\mathbb{N}$ of the natural numbers. This is denoted by $\aleph_0$ (aleph 0).

A set is called *enumerable* or *countable* if it is equinumerous to $\mathbb{N}$. This means that its elements can be enumerated or written as an infinite sequence $a_1, a_2, \ldots$.

A set is called *non-countable* if it is infinite but it is not equinumerous to $\mathbb{N}$. Consequently every infinite set which is not enumerable is non-countable.

■ **A:** The set $\mathbb{Z}$ of integers and the set $\mathbb{Q}$ of the rational numbers are countable sets.

■ **B:** The set $\mathbb{R}$ of the real numbers and the set $\mathbb{C}$ of the complex numbers are non-countable sets. These sets are equinumerous to $\mathbb{P}(\mathbb{N})$, the power set of the natural numbers, and their cardinality is called the *continuum*.

5.3 Classical Algebraic Structures

5.3.1 Operations

1. *n* ary Operations

The notion of structure has a central role in mathematics and its applications. Next to investigate are algebraic structures, i.e., sets on which operations are defined. An *n ary operation* φ on a set A is a mapping φ: $A^n \to A$, which assigns an element of A to every n tuple of elements of A.

2. Properties of Binary Operations

Especially important is the case $n = 2$, which is called a *binary operation*, e.g., addition and multiplication of numbers or matrices, or union and intersection of sets. A binary operation can be considered as a mapping $* :\ A \times A \to A$, where instead of the notation "$*(a, b)$" in this chapter mostly the *infix form* "$a * b$" will be used. A binary operation $*$ in A is called *associative* if

$$(a * b) * c = a * (b * c), \tag{5.80}$$

and *commutative* if

$$a * b = b * a \tag{5.81}$$

holds for every $a, b, c \in A$.
An element $e \in A$ is called a *neutral element* with respect to a binary operation $*$ in A if

$$a * e = e * a = a \quad \text{holds for every} \quad a \in A. \tag{5.82}$$

3. Exterior Operations

Sometimes exterior operations are to be considered. That are the mappings from $K \times A$ to K, where K is an "exterior" and mostly already structured set (see 5.3.8, p. 365).

5.3.2 Semigroups

The most frequently occurring algebraic structures have their own names. A set H having one associative binary operation $*$, is called a *semigroup*. The notation: is $H = (H, *)$.

Examples of Semigroups:

■ **A:** Number domains with respect to addition or multiplication.

■ **B:** Power sets with respect to union or intersection.

■ **C:** Matrices with respect to addition or multiplication.

■ **D:** The set A^* of all " words " (strings) over an " alphabet " A with respect to concatenation (*free semigroup*).

Remark: Except for multiplication of matrices and concatenation of words, all operations in these examples are also commutative; in this case one talks about a commutative semigroup.

5.3.3 Groups

5.3.3.1 Definition and Basic Properties

1. Definition, Abelian Group

A set G with a binary operation $*$ is called a *group* if

- $*$ is associative,
- $*$ has a neutral element e, and for every element $a \in G$ there exists an *inverse element* a^{-1} such that

$$a * a^{-1} = a^{-1} * a = e. \tag{5.83}$$

A group is a special semigroup.

The neutral element of a group is unique, i.e., there exists only one. Furthermore, every element of the group has exactly one inverse. If the operation $*$ is commutative, then the group is called an *Abelian group*. If the group operation is written as addition, $+$, then the neutral element is denoted by 0 and the inverse of an element a by $-a$.

The number of elements of a finite group is called the *order of the group* (see 5.3.3.2,**3.**, p. 338).

Examples of Groups:

■ **A:** The number of domains (except $\mathbb{N}$) with respect to addition.

■ **B:** $\mathbb{Q} \setminus \{0\}$, $\mathbb{R} \setminus \{0\}$, and $\mathbb{C} \setminus \{0\}$ with respect to multiplication.

■ **C:** $S_M := \{f : M \to M \wedge f \text{bijective}\}$ with respect to composition of mappings. This group is called symmetric. If M is finite having n elements, then S_n is written instead of S_M. S_n has $n!$ elements.

The symmetric group S_n and its subgroups are called *permutation groups*. So, the dieder groups D_n are permutation groups and subgroups of S_n.

■ **D:** The set D_n of all covering transformations of a regular n-gon in the plane is considered. Here a *covering transformation* is the transition between two symmetric positions of the n-gon, i.e., the moving of the n-gon into a superposable position. Denoting by d a rotation by the angle $2\pi/n$ and by σ the reflection with respect to an axis, then D_n has $2n$ elements:

$$D_n = \{e, d, d^2, \ldots, d^{n-1}, \sigma, d\sigma, \ldots, d^{n-1}\sigma\}.$$

With respect to the composition of mappings D_n is a group, the *dihedral group*. Here the equalities $d^n = \sigma^2 = e$ and $\sigma d = d^{n-1}\sigma$ hold.

■ **E:** All the regular matrices (see 4.1.4, p. 272) over the real or complex numbers with respect to multiplication.

Remark: Matrices have a very important role in applications, especially in representation of linear transformations. Linear transformations can be classified by matrix groups.

2. Group Tables or Cayley's Tables

For the representation of finite groups Cayley's tables or group tables are used: The elements of the group are denoted at the row and column headings. The element $a * b$ is the intersection of the row of the element a and the column of the element b.

■ If $M = \{1, 2, 3\}$, then the symmetric group S_M is also denoted by S_3. S_3 consists of all the bijective mappings (permutations) of the set $\{1, 2, 3\}$ and consequently it has $3! = 6$ elements (see 16.1.1, p. 805). Permutations are mostly represented in two rows, where in the first row there are the elements of M and under each of them there is its image. So one gets the six elements of S_3 as follows:

$$\begin{aligned} &\varepsilon = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \quad p_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}, \\ &p_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \quad p_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \quad p_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}. \end{aligned} \tag{5.84}$$

With the successive application of these mappings (binary operations) the following group table is obtained for S_3:

$\circ$	ε	p_1	p_2	p_3	p_4	p_5
ε	ε	p_1	p_2	p_3	p_4	p_5
p_1	p_1	ε	p_5	p_4	p_3	p_2
p_2	p_2	p_4	ε	p_5	p_1	p_3
p_3	p_3	p_5	p_4	ε	p_2	p_1
p_4	p_4	p_2	p_3	p_1	p_5	ε
p_5	p_5	p_3	p_1	p_2	ε	p_4

(5.85)

- From the group table it can be seen that the identity permutation ε is the neutral element of the group.
- In the group table every element appears exactly once in every row and in every column.
- It is easy to recognize the inverse of any group element in the table, i.e., the inverse of p_4 in S_3 is the permutation p_5, because at the intersection of the row of p_4 with the column of p_5 is the neutral element ε.
- If the group operation is commutative (Abelian group), then the table is symmetric with respect to the "main diagonal"; S_3 is not commutative, since, e.g., $p_1 \circ p_2 \neq p_2 \circ p_1$.
- The associative property cannot be easily recognized from the table.

5.3.3.2 Subgroups and Direct Products

1. Subgroups

Let $G = (G, *)$ be a group and $U \subseteq G$. If U is also a group with respect to $*$, then $U = (U, *)$ is called a *subgroup* of G.

A non-empty subset U of a group $(G, *)$ is a subgroup of G if and only if for every $a, b \in U$, the elements $a * b$ and a^{-1} are also in U (*subgroup criterion*).

1. Cyclic Subgroups The group G itself and $E = \{e\}$ are subgroups of G, the so-called *trivial subgroups*. Furthermore, a subgroup corresponds to every element $a \in G$, the so-called *cyclic subgroup* generated by a:

$$< a > = \{\ldots, a^{-2}, a^{-1}, e, a, a^2, \ldots\}. \tag{5.86}$$

If the group operation is addition, then one writes the integer multiple ka as a shorthand notation of the k times addition of a with itself instead of the power a^k, i.e., as a shorthand notation of the k times operation of a by itself,

$$< a > = \{\ldots, (-2)a, -a, 0, a, 2a, \ldots\}. \tag{5.87}$$

Here $< a >$ is the smallest subgroup of G containing a. If $< a > = G$ holds for an element a of G, then G is called cyclic.

There are infinite cyclic groups, e.g., $\mathbb{Z}$ with respect to addition, and finite cyclic groups, e.g., the set $\mathbb{Z}_m$ the residue class modulo m with residue class addition (see 5.4.3, **3.**, p. 377).

■ If the number of elements of a finite G group is a prime, then G is always cyclic.

2. Generalization The notion of cyclic groups can be generalized as follows: If M is a non-empty subset of a group G, then the subgroup of G whose elements can be written in the form of a product of finitely many elements of M and their inverses, is denoted by $< M >$. The subset M is called the *system of generators* of $< M >$. If M contains only one element, then $< M >$ is cyclic.

3. Order of a Group, Left and Right Cosets In group theory the number of elements of a finite group is denoted by ord G. If the cyclic subgroup $< a >$ generated by one element a is finite, then this order is also called the *order of the element* a, i.e., ord $< a >=$ ord a.
If U is a subgroup of a group $(G, *)$ and $a \in G$, then the subsets

$$aU := \{a * u | u \in U\} \quad \text{and} \quad Ua := \{u * a | u \in U\} \tag{5.88}$$

of G are called *left co-sets* and *right co-sets* of U in G. The left or right co-sets form a partition of G, respectively (see 5.2.4, **2.**, p. 334).

All the left or right co-sets of a subgroup U in a group G have the same number of elements, namely ord$U$. From this it follows that the number of left co-sets is equal to the number of right co-sets. This number is called the *index* of U in G. The Lagrange theorem follows from these facts.

4. Lagrange Theorem The order of a subgroup is a divisor of the order of the group.
In general it is difficult to determine all the subgroups of a group. In the case of finite groups the Lagrange theorem as a necessary condition for the existence of a subgroup is useful.

2. Normal Subgroup or Invariant Subgroup

For a subgroup U, in general, aU is different from Ua (however $|aU| = |Ua|$ is valid). If $aU = Ua$ for all $a \in G$ holds, then U is called a *normal subgroup* or *invariant subgroup* of G. These special subgroups are the basis of forming factor groups (see 5.3.3.3, **3.**, p. 339).
In Abelian groups, obviously, every subgroup is a normal subgroup.

Examples of Subgroups and Normal Subgroups:

■ **A:** $\mathbb{R} \setminus \{0\}$, $\mathbb{Q} \setminus \{0\}$ form subgroups of $\mathbb{C} \setminus \{0\}$ with respect to multiplication.

■ **B:** The even integers form a subgroup of $\mathbb{Z}$ with respect to addition.

■ **C:** Subgroups of S_3: According to the Lagrange theorem the group S_3 having six elements can have subgroups only with two or three elements (besides the trivial subgroups). In fact, the group S_3 has the following subgroups: $E = \{\varepsilon\}$, $U_1 = \{\varepsilon, p_1\}$, $U_2 = \{\varepsilon, p_2\}$, $U_3 = \{\varepsilon, p_3\}$, $U_4 = \{\varepsilon, p_4, p_5\}$, S_3.
The non-trivial subgroups U_1, U_2, U_3, and U_4 are cyclic, since the numbers of their elements are primes. But the group S_3 is not cyclic. The group S_3 has only U_4 as a normal subgroup, except the trivial normal subgroups.
Anyway, every subgroup U of a group G with $|U| = |G|/2$ is a normal subgroup of G.
Every symmetric group S_{M} and their subgroups are called *permutation groups.*

■ **D:** Special subgroups of the group $GL(n)$ of all regular matrices of type (n, n) with respect to matrix multiplication:

$SL(n)$	group of all matrices A with determinant 1,
$O(n)$	group of all orthogonal matrices,
$SO(n)$	group of all orthogonal matrices with determinant 1.

The group $SL(n)$ is a normal subgroup of $GL(n)$ (see 5.3.3.3, **3.**, p. 339) and $SO(n)$ is a normal subgroup of $O(n)$.

■ **E:** As subgroups of all complex matrices of type (n, n) (see 4.1.4, p. 272):

$U(n)$	group of all unitary matrices,
$SU(n)$	group of all unitary matrices with determinant 1.

3. Direct Product

1. Definition Suppose A and B are groups, whose group operation (e.g., addition or multiplication) is denoted by $\cdot$. In the Cartesian product (see 5.2.2, **4.**, p. 331) $A \times B$ (5.65a) an operation $*$ can be introduced in the following way:

$$(a_1, b_1) * (a_2, b_2) = (a_1 \cdot a_2, b_1 \cdot b_2). \tag{5.89a}$$

$A \times B$ becomes a group with this operation and it is called the *direct product* of A and B.
(e, e) denotes the unit element of $A \times B$, (a^{-1}, b^{-1}) is the inverse element of (a, b).

For finite groups A, B

$$\operatorname{ord}(A \times B) = \operatorname{ord} A \cdot \operatorname{ord} B \tag{5.89b}$$

holds. The groups $A' := \{(a, e)|a \in A\}$ and $B' := \{(e, b)|b \in B\}$ are normal subsets of $A \times B$ isomorphic to A and B, respectively.

The direct product of Abelian groups is again an Abelian group.

The direct product of two cyclic groups A, B is cyclic if and only if the greatest common divisor of the orders of the groups is equal to 1.

■ **A:** With $Z_2 = \{e, a\}$ and $Z_3 = \{e, b, b^2\}$, the direct product $Z_2 \times Z_3 = \{(e, e), (e, b), (e, b^2), (a, e), (a, b), (a, b^2)\}$, is a group isomorphic to Z_6 (see 5.3.3.3, **2.**, p. 339) generated by (a, b).

■ **B:** On the other hand $Z_2 \times Z_2 = \{(e, e), (e, b), (a, e), (a, b)\}$ is not cyclic. This group has order 4 and it is also called Klein's four group, and it describes the covering operations of a rectangle.

2. Fundamental Theorem of Abelian Groups Because the direct product is a construction which enables to make "larger" groups from "smaller" groups, the question can be reversed: When is it possible to consider a larger group G as a direct product of smaller groups A, B, i.e., when will G be isomorphic to $A \times B$? For Abelian groups, there exists the so-called *fundamental theorem*:

Every finite Abelian group can be represented as a direct product of cyclic groups with orders of prime powers.

5.3.3.3 Mappings Between Groups

1. Homomorphism and Isomorphism

1. Group Homomorphism Between algebraic structures, not arbitrary mappings, but only "structure keeping" mappings are considered:

Let $G_1 = (G_1, *)$ and $G_2 = (G_2, \circ)$ are two groups. A mapping h: $G_1 \rightarrow G_2$ is called a *group homomorphism*, if for all $a, b \in G_1$ holds:

$$h(a * b) = h(a) \circ h(b) \quad (\text{"image of product = product of images"}) \tag{5.90}$$

■ As an example, consider the multiplication law for determinants (see 4.2.2, **7.**, p. 279):

$$\det(AB) = (\det A)(\det B). \tag{5.91}$$

Here on the right-hand side there is the product of non-zero numbers, on the left-hand side there is the product of regular matrices.

If h: $G_1 \rightarrow G_2$ is a group homomorphism, then the set of elements of G_1, whose image is the neutral element of G_2, is called the *kernel* of h, and it is denoted by $\ker h$. The kernel of h is a normal subgroup of G_1.

2. Group Isomorphism If a group homomorphism h is also bijective, then h is called a *group isomorphism*, and the groups G_1 and G_2 are called *isomorphic* to each other (notation: $G_1 \cong G_2$). Then $\ker\, h = E$ is valid.

Isomorphic groups have the same structure, i.e., they differ only by the notation of their elements.

■ The symmetric group S_3 and the dihedral group D_3 are isomorphic groups of order 6 and describe the covering mappings of an equilateral triangle.

2. Cayley's Theorem

The Cayley theorem says that *every* group can be interpreted as a permutation group (see 5.3.3.2, **2.**, p. 338):

Every group is isomorphic to a permutation group.

The permutation group P, whose elements are the permutations π_g $(g \in G)$ mapping a to $G, *g$, is a subgroup of S_G isomorphic to $(G, *)$.

3. Homomorphism Theorem for Groups

The set of co-sets of a normal subgroup N in a group G is also a group with respect to the operation

$$aN \circ bN = abN. \tag{5.92}$$

It is called the *factor group* of G with respect to N, and it is denoted by G/N.

The following theorem gives the correspondence between homomorphic images and factor groups of a group, because of what it is called the homomorphism theorem for groups:
A group homomorphism h: $G_1 \to G_2$ defines a normal subgroup of G_1, namely $\ker h = \{a \in G_1 | h(a) = e\}$. The factor group $G_1/\ker h$ is isomorphic to the homomorphic image $h(G_1) = \{h(a)|a \in G_1\}$. Conversely, every normal subgroup N of G_1 defines a homomorphic mapping nat_N: $G_1 \to G_1/N$ with $nat_N(a) = aN$. This mapping nat_N is called a *natural homomorphism*.

■ Since the determinant construction det: $GL(n) \to \mathbb{R} \setminus \{0\}$ is a group homomorphism with kernel $SL(n)$, $SL(n)$ is a normal subgroup of $GL(n)$ and (according to the homomorphism theorem): $GL(n)/SL(n)$ is isomorphic to the multiplicative group $\mathbb{R}\setminus\{0\}$ of real numbers (for notation see 5.3.3.2, **2.**, p. 338).

5.3.4 Group Representations

5.3.4.1 Definitions

1. Representation

A *representation* $D(G)$ of the *group* G is a map (homomorphism) of G onto the group of non-singular linear transformations D on an n-dimensional (real or complex) vector space V_n:

$$D(G) : a \to D(a), \quad a \in G. \tag{5.93}$$

The vector space V_n is called the *representation space*; n is the dimension of the representation (see also 12.1.3, **2.**, p. 657). Introducing the basis $\{\underline{\mathbf{e}}_i\}$ $(i = 1, 2, \ldots, n)$ in V_n every vector $\underline{\mathbf{x}}$ can be written as a linear combination of the basis vectors:

$$\underline{\mathbf{x}} = \sum_{i=1}^{n} x_i \underline{\mathbf{e}}_i, \quad \underline{\mathbf{x}} \in \mathrm{V}_n. \tag{5.94}$$

The action of the linear transformation $D(a)$, $a \in G$, on $\underline{\mathbf{x}}$ can be defined by the quadratic matrix $\mathbf{D}(a) = (D_{ik}(a))$ $(i, k = 1, 2, \ldots, n)$, which provides the coordinates of the transformed vector $\underline{\mathbf{x}}'$ within the basis $\{\underline{\mathbf{e}}_i\}$:

$$\underline{\mathbf{x}}' = \mathbf{D}(a)\underline{\mathbf{x}} = \sum_{i=1}^{n} x'_i \underline{\mathbf{e}}_i, \qquad x'_i = \sum_{k=1}^{n} D_{ik}(a) x_k. \tag{5.95}$$

This transformation may also be considered as a transformation of the basis $\{\underline{\mathbf{e}}_i\} \to \{\underline{\mathbf{e}}'_i\}$:

$$\underline{\mathbf{e}}'_i = \underline{\mathbf{e}}_i \mathbf{D}(a) = \sum_{k=1}^{n} D_{ki}(a) \underline{\mathbf{e}}_k. \tag{5.96}$$

Thus, every element a of the group is assigned to the *representation matrix* $\mathbf{D} = (D_{ik}(a))$:

$$D(G) : a \to \mathbf{D} = (D_{ik}(a)) \quad (i, k = 1, 2, \ldots, n), a \in G. \tag{5.97}$$

The representation matrix depends on the choice of basis.

■ **A: Abelian Point Group C_n.** A regular polygon (see 3.1.5, p. 138) with n sides has a symmetry such that rotating it around an axis, which is perpendicular to the plane of the figure and goes through its center M (**Fig.5.9**) by an angle $\varphi_k = 2\pi k/n$, $k = 0, 1, \ldots, n-1$ the resulted polygon is identical to the original one (invariance of the system under certain rotations). The rotations $R_k(\varphi_k)$ form the Abelian group of points C_n. C_n is a cyclic group (see 5.3.3.2, p. 337), i.e. every element of the group can be represented as a power of a single element R_1, whose n-th power is the unit element $e = R_0$:

$$C_n = \{e, R_1, R_1^2, \ldots, R_1^{n-1}\}, \quad R_1^n = e\,. \tag{5.98a}$$

Let the center of an equilateral triangle ($n = 3$) be the origin (see **Fig.5.9**), then the angles of rotations and the rotations are in accor-

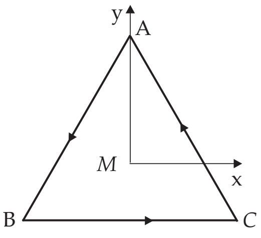


Figure 5.9

dance with (5.98b).

$$\begin{aligned} &k = 0,\ \varphi_0 = 0 \text{ or } 2\pi, \\ &k = 1,\ \varphi_1 = 2\pi/3, \\ &k = 2,\ \varphi_2 = 4\pi/3. \end{aligned} \tag{5.98b}$$

$$\begin{aligned} &R_0 : A \to A, B \to B, C \to C, \\ &R_1 : A \to B, B \to C, C \to A, \\ &R_2 : A \to C, B \to A, C \to B\,. \end{aligned} \tag{5.98c}$$

The rotations (5.98c) satisfy the relations

$$R_2 = R_1^2,\ \ R_1 \cdot R_2 = R_1^3 = R_0 = e\,. \tag{5.98d}$$

They form the cyclic group C_3.
The matrix of rotation (see (3.432), p. 230)

$$\mathbf{R}(\varphi) = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} \tag{5.98e}$$

of a geometric transformation of this triangle (for rotation of this figure in a fixed coordinate system see 3.5.3.3,**3.**, p. 213) gives the representation of group C_3 if φ is substituted by the angles given in (5.98b):

$$\mathbf{D}(e) = \mathbf{R}(0) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix},\ \ \mathbf{D}(R_1) = \mathbf{R}(2\pi/3) = \begin{pmatrix} -1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}, \tag{5.98f}$$

$$\mathbf{D}(R_2) = \mathbf{R}(4\pi/3) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{pmatrix}. \tag{5.98g}$$

The same relations hold for the matrices of this representation given in (5.98f) and (5.98g) as for the group elements R_k (5.98d):

$$\mathbf{D}(R_2) = \mathbf{D}(R_1R_1) = \mathbf{D}(R_1)\mathbf{D}(R_1),\ \ \mathbf{D}(R_1)\mathbf{D}(R_2) = \mathbf{D}(e). \tag{5.98h}$$

■ **B: Dihedral Group D_3**. The equilateral triangle is invariant with respect to rotations by angle π about its bisectors (see **Fig.5.10**). These rotations correspond to reflections S_A, S_B, S_C with respect to a plane being perpendicular to the plane of the triangle and containing one of the rotation axes.

$$\begin{aligned} &S_A: \text{ Rotations } A \to A, B \to C, C \to B; \\ &S_B: \text{ Rotations } A \to C, B \to B, C \to A; \\ &S_C: \text{ Rotations } A \to B, B \to A, C \to C. \end{aligned} \tag{5.99a}$$

For the reflections there is:

$$S_\sigma S_\sigma = e \quad (\sigma = A, B, C). \tag{5.99b}$$

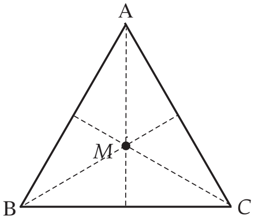


Figure 5.10

The product $S_\sigma S_\tau$ $(\sigma \neq \tau)$ results in one of the rotations R_1, R_2, e.g. using $S_A\, S_B$ for the triangle ΔABC:

$$S_A\, S_B(\Delta ABC) = S_A(\Delta CBA) = \Delta CAB = R_1(\Delta ABC)\,, \tag{5.99c}$$

consequently $S_A\, S_B = R_1$. Here S_A, S_B, S_C correspond to the outcomes on **Fig.5.10**.

The cyclic group C_3 and the reflections S_A, S_B, S_C together form the dihedral group D_3. The reflections do not form a subgroup because of (5.99c). A summary of relations is represented in group-table (5.99d).
Only the signs of the x-coordinates of points B and C are changed at reflection S_A (see **Fig.5.9**). This coordinate transformation is given by the matrix

	e	R_1	R_2	S_A	S_B	S_C
e	e	R_1	R_2	S_A	S_B	S_C
R_1	R_1	R_2	e	S_C	S_A	S_B
R_2	R_2	e	R_1	S_B	S_C	S_A
S_A	S_A	S_B	S_C	e	R_1	R_2
S_B	S_B	S_C	S_A	R_2	e	R_1
S_C	S_C	S_A	S_B	R_1	R_2	e

(5.99d)

$$\mathbf{D}(S_A) = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{5.99e}$$

The matrices representing reflections S_B and S_C can be found in the group-table (5.99d) and from the matrices of representation in (5.98f) and (5.98g)

$$\mathbf{D}(S_B) = \mathbf{D}(R_2)\mathbf{D}(S_A) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix}, \tag{5.99f}$$

$$\mathbf{D}(S_C) = \mathbf{D}(R_1)\mathbf{D}(S_A) = \begin{pmatrix} -1/2 & \sqrt{3}/2 \\ \sqrt{3}/2 & -1/2 \end{pmatrix} \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1/2 & -\sqrt{3}/2 \\ -\sqrt{3}/2 & -1/2 \end{pmatrix}. \tag{5.99g}$$

Matrices (5.98f) and (5.98g) together with matrices (5.99f) and (5.99g) form a representation of the dihedral group D_3.

2. Faithful Representation

A representation is called *faithful* if $G \to D(G)$ is an isomorphism, i.e., the assignment of the element of the group to the representation matrix is a one-to-one mapping.

3. Properties of the Representations

A representation with the representation matrices $\mathbf{D}(a)$ has the following properties ($a, b \in G, \mathbf{I}$ unit matrix):

$$\mathbf{D}(a * b) = \mathbf{D}(a) \cdot \mathbf{D}(b), \quad \mathbf{D}(a^{-1}) = \mathbf{D}^{-1}(a), \quad \mathbf{D}(e) = \mathbf{I}. \tag{5.100}$$

5.3.4.2 Particular Representations

1. Identity Representation

Any group G has a trivial one-dimensional representation (identity representation), for which every element of the group is mapped to the unit matrix $\mathbf{I}$: $a \to \mathbf{I}$ for all $a \in G$.

2. Adjoint Representation

The representation $D^+(G)$ is called *adjoint* to $D(G)$ if the corresponding representation matrices are related by complex conjugation and reflection in the main diagonal:

$$\mathbf{D}^+(G) = \tilde{\mathbf{D}}^*(G). \tag{5.101}$$

3. Unitary Representation

For a *unitary representation* all representation matrices are unitary matrices:

$$\mathbf{D}(G) \cdot \mathbf{D}^+(G) = \mathbf{I}, \tag{5.102}$$

where $\mathbf{E}$ is the unit matrix.

4. Equivalent Representations

Two representations $D(G)$ and $D'(G)$ are called *equivalent* if for each element a of the group the corresponding representation matrices are related by the same similarity transformation with the non-singular matrix $\mathbf{T} = (T_{i,j})$:

$$\mathbf{D}'(a) = \mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T}, \quad D'_{ik}(a) = \sum_{j,l=1}^{n} T^{-1}_{ij} \cdot D_{jl}(a) \cdot T_{lk}, \tag{5.103}$$

where $T^{-1}_{i,j}$ denotes the elements of the inverse matrix $\mathbf{T}^{-1}$ of $\mathbf{T}$. If such a relation does not hold two representations are called *non-equivalent*. The transition from $D(G)$ to $D'(G)$ corresponds to the transformation $T : \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\} \to \{\mathbf{e}'_1, \mathbf{e}'_2, \ldots, \mathbf{e}'_n\}$ of the basis in the representation space V_n:

$$\mathbf{e}' = \mathbf{e}\,T, \quad \mathbf{e}'_i = \sum_{k=1}^{n} T_{ki}\mathbf{e}_k \quad (i = 1, 2, \ldots, n). \tag{5.104}$$

Any representation of a finite group is equivalent to a unitary representation.

5. Character of a Group Element

In the representation $D(G)$ the *character* $\chi(a)$ of the group element a is defined as the trace of the representation matrix $\mathbf{D}(a)$ (sum of the main diagonal elements of the matrix):

$$\chi(a) = \mathrm{Tr}\,(\mathbf{D}) = \sum_{i=1}^{n} D_{ii}(a). \tag{5.105}$$

The character of the unit element e is given by the dimension n of the representation: $\chi(e) = n$. Since the trace of a matrix is invariant under similarity transformations, the group element a has the same character for equivalent representations.

■ Within the shell model of atomic or nuclear physics two out of three particles with space coordinates $\vec{\mathbf{r}}_i$ $(i = 1, 2, 3)$ can be described by the wave function $\varphi_\alpha(\vec{\mathbf{r}})$ while the third particle has the wave function $\varphi_\beta(\vec{\mathbf{r}})$ (configuration $\alpha^2\beta(\vec{\mathbf{r}})$). The wave function ψ of the system is a product of the three one-particle wave functions: $\psi = \varphi_\alpha\varphi_\alpha\varphi_\beta$. In accordance with the possible distributions of the particles $1, 2, 3$ to the wave functions one gets the three functions

$$\psi_1 = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3)\,, \psi_2 = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3)\,, \psi_3 = \varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3)\,, \tag{5.106a}$$

which, when realizing permutations, transform among one another according to 5.3.3.1, **2.**, p. 337. This way one gets for the functions $\psi_1\psi_2\psi_3$ a three dimensional representation of the symmetric group S_3. According to (5.93) the matrix elements of the representation matrices can be found by investigating the action of the group elements (5.84) on the coordinate subscripts in the basis elements e_i. For example:

$$\begin{aligned}
p_1\psi_1 &= p_1\varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3) = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = D_{21}(p_1)\psi_2,\\
p_1\psi_2 &= p_1\varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\beta(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = \varphi_\alpha(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\beta(\vec{\mathbf{r}}_3) = D_{12}(p_1)\psi_1,\\
p_1\psi_3 &= p_1\varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = \varphi_\beta(\vec{\mathbf{r}}_1)\varphi_\alpha(\vec{\mathbf{r}}_2)\varphi_\alpha(\vec{\mathbf{r}}_3) = D_{33}(p_1)\psi_3.
\end{aligned} \tag{5.106b}$$

Altogether one finds:

$$\begin{aligned}
\mathbf{D}(e) &= \begin{pmatrix} 1&0&0\\0&1&0\\0&0&1 \end{pmatrix}, & \mathbf{D}(p_1) &= \begin{pmatrix} 0&1&0\\1&0&0\\0&0&1 \end{pmatrix}, & \mathbf{D}(p_2) &= \begin{pmatrix} 0&0&1\\0&1&0\\1&0&0 \end{pmatrix},\\
\mathbf{D}(p_3) &= \begin{pmatrix} 1&0&0\\0&0&1\\0&1&0 \end{pmatrix}, & \mathbf{D}(p_4) &= \begin{pmatrix} 0&1&0\\0&0&1\\1&0&0 \end{pmatrix}, & \mathbf{D}(p_5) &= \begin{pmatrix} 0&0&1\\1&0&0\\0&1&0 \end{pmatrix}.
\end{aligned} \tag{5.106c}$$

For the characters one has:

$$\chi(e) = 3, \;\; \chi(p_1) = \chi(p_2) = \chi(p_3) = 1, \;\; \chi(p_4) = \chi(p_5) = 0. \tag{5.106d}$$

5.3.4.3 Direct Sum of Representations

The representations $D^{(1)}(G)$, $D^{(2)}(G)$ of dimension n_1 and n_2 can be composed to create a new representation $D(G)$ of dimension $n = n_1 + n_2$ by forming the direct sum of the representation matrices:

$$\mathbf{D}(a) = \mathbf{D}^{(1)}(a) \oplus \mathbf{D}^{(2)}(a) = \begin{pmatrix} \mathbf{D}^{(1)}(a) & \mathbf{0}\\ \mathbf{0} & \mathbf{D}^{(2)}(a) \end{pmatrix}. \tag{5.107}$$

The block-diagonal form of the representation matrix implies that the representation space V_n is the direct sum of two invariant subspaces $\mathrm{V}_{n_1}, \mathrm{V}_{n_2}$:

$$\mathrm{V}_n = \mathrm{V}_{n_1} \oplus \mathrm{V}_{n_2}, \quad n = n_1 + n_2. \tag{5.108}$$

A subspace $\mathrm{V}_m \ (m < n)$ of V_n is called an invariant subspace if for any linear transformation $D(a), \ a \in G$, every vector $\underline{\mathbf{x}} \in \mathrm{V}_m$ is mapped onto an element of V_m again:

$$\underline{\mathbf{x}}' = \mathbf{D}(a)\underline{\mathbf{x}} \quad \text{with} \quad \underline{\mathbf{x}}, \underline{\mathbf{x}}' \in \mathrm{V}_m. \tag{5.109}$$

The *character of the representation* (5.107) is the sum of the characters of the single representations:

$$\chi(a) = \chi^{(1)}(a) + \chi^{(2)}(a). \tag{5.110}$$

5.3.4.4 Direct Product of Representations

If $\underline{\mathbf{e}}_i \ (i = 1, 2, \ldots, n_1)$ and $\underline{\mathbf{e}}'_k \ (k = 1, 2, \ldots, n_2)$ are the basis vectors of the representation spaces V_{n_1} and V_{n_2}, respectively, then the tensor product

$$\underline{\mathbf{e}}_{ik} = \{\underline{\mathbf{e}}_i \underline{\mathbf{e}}_k\} \quad (i = 1, 2, \ldots, n_1; \ \ k = 1, 2, \ldots, n_2) \tag{5.111}$$

forms a basis in the product space $\mathrm{V}_{n_1} \otimes \mathrm{V}_{n_2}$ of dimension $n_1 \cdot n_2$. With the representations $D^{(1)}(G)$ and $D^{(2)}(G)$ in V_{n_1} and V_{n_2}, respectively an $n_1 \cdot n_2$-dimensional representation $D(G)$ in the product space can be constructed by forming the direct or (inner) Kronecker product (see 4.1.5,**9.**, p. 276) of the representation matrices:

$$\mathbf{D}(G) = \mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(G), \quad (D(G))_{ik,jl} = D^{(1)}_{ik}(a) \cdot D^{(2)}_{jl}(a)$$
$$\text{with} \quad i, k = 1, 2, \ldots, n_1; \ \ j, l = 1, 2, \ldots, n_2. \tag{5.112}$$

The character of the Kronecker product of two representations is equal to the product of the characters of the factors

$$\chi^{(1\times 2)}(a) = \chi^{(1)}(a) \cdot \chi^{(2)}(a). \tag{5.113}$$

5.3.4.5 Reducible and Irreducible Representations

If the representation space V_n possesses a subspace $\mathrm{V}_m \ \ (m < n)$ invariant under the group operations the representation matrices can be decomposed according to

$$\mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T} = \begin{pmatrix} \mathbf{D}_1(a) & \mathbf{A} \\ \mathbf{0} & \mathbf{D}_2(a) \end{pmatrix} \begin{array}{l} \{ \ m \quad\quad \text{rows} \\ \{ \ n - m \ \text{rows} \end{array} \tag{5.114}$$

by a suitable transformation $\mathbf{T}$ of the basis in V_n. $\mathbf{D}_1(a)$ and $\mathbf{D}_2(a)$ themselves are matrix representations of $a \in G$ of dimension m and $n - m$, respectively.

A representation $D(G)$ is called *irreducible* if there is no proper (non-trivial) invariant subspace in V_n. The number of non-equivalent irreducible representations of a finite group is finite. If a transformation $\mathbf{T}$ of a basis can be found which makes V_n to a direct sum of invariant subspaces, i.e.,

$$\mathrm{V}_n = \mathrm{V}_1 \oplus \cdots \oplus \mathrm{V}_{n_j}, \tag{5.115}$$

then for every $a \in G$ the representation matrix $\mathbf{D}(a)$ can be transformed into the block-diagonal form ($\mathbf{A} = \mathbf{0}$ in (5.114)):

$$\mathbf{T}^{-1} \cdot \mathbf{D}(a) \cdot \mathbf{T} = \mathbf{D}^{(1)}(a) \oplus \cdots \oplus \mathbf{D}^{(n_j)}(a) = \begin{pmatrix} \mathbf{D}^{(1)}(a) & & 0 \\ & \ddots & \\ 0 & & \mathbf{D}^{(n_j)}(a) \end{pmatrix}. \tag{5.116}$$

by a similarity transformation with $\mathbf{T}$. Such a representation is called *completely reducible*.

Remark: For the application of group theory in natural sciences a fundamental task consists in the classification of all non-equivalent irreducible representations of a given group.

■ The representation of the symmetric group S_3 given in (5.106c), p. 343, is reducible. For example, in the basis transformation $\{\underline{\mathbf{e}}_1, \underline{\mathbf{e}}_2, \underline{\mathbf{e}}_3\} \longrightarrow \{\underline{\mathbf{e}}'_1 = \underline{\mathbf{e}}_1 + \underline{\mathbf{e}}_2 + \underline{\mathbf{e}}_3, \ \ \underline{\mathbf{e}}'_2 = \underline{\mathbf{e}}_2, \ \ \underline{\mathbf{e}}'_3 = \underline{\mathbf{e}}_3\}$ one obtains for the representation matrix of the permutation p_3 (with $\psi_1 = \underline{\mathbf{e}}_1, \psi_2 = \underline{\mathbf{e}}_2, \psi_3 = \underline{\mathbf{e}}_3$):

$$\mathbf{D}(p_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = \begin{pmatrix} \mathbf{D_1}(p_3) & \mathbf{0} \\ \mathbf{A} & \mathbf{D_2}(p_3) \end{pmatrix} \tag{5.117}$$

with $\mathbf{A} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, $\mathbf{D_1}(p_3) = 1$ as the identity representation of S_3 and $\mathbf{D}_2(p_3) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$.

5.3.4.6 Schur's Lemma 1

If $\mathbf{C}$ is an operator commuting with all transformations of an irreducible representation $\mathbf{D}$ of a group $[\mathbf{C}, \mathbf{D}(a)] = \mathbf{C} \cdot \mathbf{D}(a) - \mathbf{D}(a) \cdot \mathbf{C} = 0,\ \ a \in G$, and the representation space V_n is an invariant subspace of $\mathbf{C}$, then $\mathbf{C}$ is a multiple of the unit operator, i.e., a matrix (c_{ik}) which commutates with all matrices of an irreducible representation is a multiple of the matrix $\mathbf{I}$, $\mathbf{C} = \lambda \cdot \mathbf{I},\ \ \lambda \in \mathbb{C}$.

5.3.4.7 Clebsch-Gordan Series

In general, the Kronecker product of two irreducible representations $\mathbf{D}^{(1)}(G)$, $\mathbf{D}^{(2)}(G)$ is reducible. By a suitable basis transformation in the product space $\mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(G)$ can be decomposed into the direct sum of its irreducible parts $\mathbf{D}^{(\alpha)}$ $(\alpha = 1, 2, \ldots, n)$ (*Clebsch–Gordan theorem*). This expansion is called the *Clebsch–Gordan series*:

$$\mathbf{D}^{(1)}(G) \otimes \mathbf{D}^{(2)}(a) = \sum_{\alpha=1}^{n} \oplus\, m_\alpha \mathbf{D}^{(\alpha)}(G). \tag{5.118}$$

Here, m_α is the multiplicity with which the irreducible representation $\mathbf{D}^{(\alpha)}(G)$ occurs in the Clebsch–Gordan series.

The matrix elements of the basis transformation in the product space causing the reduction of the Kronecker product into its irreducible components are called *Clebsch–Gordan coefficients*.

5.3.4.8 Irreducible Representations of the Symmetric Group S_M

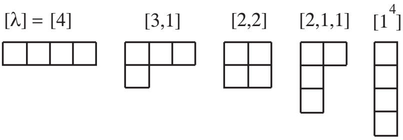
1. Symmetric Group S_M

The non-equivalent irreducible representations of the symmetric group S_M are characterized uniquely by the partitions of M, i.e., by the splitting of M into integers according to

$$[\lambda] = [\lambda_1, \lambda_2, \ldots, \lambda_M], \quad \lambda_1 + \lambda_2 + \cdots + \lambda_M = M, \quad \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_M \ge 0. \tag{5.119}$$

The graphic representation of the partitions is done by arranging boxes in *Young diagrams*.

■ For the group S_4 one obtains five Young diagrams as shown in the figure.



The dimension of the representation $[\lambda]$ is given by

$$n^{[\lambda]} = M! \frac{\prod\limits_{i<j\le k} (\lambda_i - \lambda_j + j - i)}{\Pi_{i=1}^{k} (\lambda_i + k - i)!}. \tag{5.120}$$

The Young diagram $[\tilde{\lambda}]$ conjugated to $[\lambda]$ is constructed by the interchange of rows and columns. In general, the irreducible representation of S_M is reducible if one restricts to one of the subgroups $S_{M-1}, S_{M-2}, \cdots$.

■ In quantum mechanics for a system of identical particles the Pauli principle demands the construction of many-body wave functions that are antisymmetric with respect to the interchange of all coordinates of two arbitrary particles. Often, the wave function is given as the product of a function in space coordinates and a function in spin variables. If for such a case due to particle permutations the spatial part of the wave function transforms according to the irreducible representation $[\lambda]$ of the symmetric group, then it has to be combined with a spin function transforming according to $[\tilde{\lambda}]$ in order to get a total wave function which is antisymmetric if two particles are interchanged.

5.3.5 Applications of Groups

In chemistry and in physics, groups are applied to describe the "symmetry" of the corresponding objects. Such objects are, for instance, molecules, crystals, solid structures or quantum mechanical

systems. The basic idea of these applications is the von Neumann principle:
If a system has a certain group of symmetry operations, then every physical observational quantity of this system must have the same symmetry.

5.3.5.1 Symmetry Operations, Symmetry Elements

A *symmetry operation* s of a space object is a mapping of the space into itself such that the length of line segments remains unchanged and the object goes into a covering position to itself. The set of fixed points of the symmetry operation s is denoted by Fix s, i.e., the set of all points of space which remain unchanged for s. The set Fix s is called the *symmetry element* of s. The Schoenflies symbolism is used to denote the symmetry operation.
Two types of symmetry operations are distinguished: Operations without a fixed point and operations with at least one fixed point.

1. Symmetry Operations without a Fixed Point, for which no point of the space stays unchanged, cannot occur for bounded space objects, but now only such objects are considered. A symmetry operation without a fixed point is for instance a parallel translation.

2. Symmetry Operations with at least One Fixed Point are for instance rotations and reflections. The following operations belong to them.

a) Rotations Around an Axis by an Angle φ: The axis of rotation and also the rotation itself is denoted by C_n for $\varphi = 2\pi/n$. The axis of rotation is then called of n-th order.

b) Reflection with Respect to a Plane: Both the plane of reflection and the reflection itself are denoted by σ. If additionally there is a principal rotation axis, then one draws it perpendicularly and denote the planes of reflections which are perpendicular to this axis by σ_h (h from horizontal) and the planes of reflections passing through the rotational axis are denoted by σ_v (v from vertical) or σ_d (d means dihedral, if certain angles are halved).

c) Improper Orthogonal Mappings: An operation such that after a rotation C_n a reflection σ_h follows, is called an improper orthogonal mapping and it is denoted by S_n. Rotation and reflection commute. The axis of rotation is then called an improper rotational axis of n-th order and it is also denoted by S_n. This axis is called the corresponding symmetry element, although only the symmetry center stays fixed under the application of the operation S_n. For $n = 2$, an improper orthogonal mapping is also called a point reflection or inversion (see 4.3.5.1, p. 287) and it is denoted by i.

5.3.5.2 Symmetry Groups or Point Groups

For every symmetry operation S, there is an inverse operation S^{-1}, which reverses S “back”, i.e.,

$$SS^{-1} = S^{-1}S = \epsilon. \tag{5.121}$$

Here ϵ denotes the identity operation, which leaves the whole space unchanged. The family of symmetry operations of a space object forms a group with respect to the successive application, which is in general a non-commutative *symmetry group* of the objects. The following relations hold:

a) Every rotation is the product of two reflections. The intersection line of the two reflection planes is the rotation axis.

b) For two reflections σ and σ'

$$\sigma\sigma' = \sigma'\sigma \tag{5.122}$$

if and only if the corresponding reflection planes are identical or they are perpendicular to each other. In the first case the product is the identity ϵ, in the second one the rotation C_2.

c) The product of two rotations with intersecting rotational axes is again a rotation whose axis goes through the intersection point of the given rotational axes.

d) For two rotations C_2 and C_2' around the same axis or around axes perpendicular to each other:

$$C_2C_2' = C_2'C_2. \tag{5.123}$$

The product is again a rotation. In the first case the corresponding rotational axis is the given one, in the second one the rotational axis is perpendicular to the given ones.

5.3.5.3 Symmetry Operations with Molecules

It requires a lot of work to recognize every symmetry element of an object. In the literature, for instance in [5.10], [5.13], it is discussed in detail how to find the symmetry groups of molecules if all the symmetry elements are known. The following notation is used for the interpretation of a molecule in space: The symbols above C in **Fig. 5.11** mean that the OH group lies above the plane of the drawing, the symbol to the right-hand side of C means that the group OC_2H_5 is under C.
The determination of the symmetry group can be made by the following method.

1. No Rotational Axis

a) If no symmetry element exists, then $G = \{\epsilon\}$ holds, i.e., the molecule does not have any symmetry operation but the identity ϵ.

■ The molecule hemiacetal **(Fig.5.11)** is not planar and it has four different atom groups.

b) If σ is a reflection or i is an inversion, then $G = \{\epsilon, \sigma\} =: C_s$ or $G = \{\epsilon, i\} = C_i$ hold, and with this it is isomorphic to Z_2.

■ The molecule of tartaric acid **(Fig.5.12)** can be reflected in the center P (inversion).

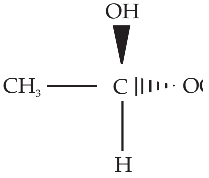


Figure 5.11

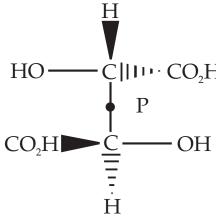


Figure 5.12

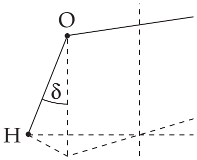


Figure 5.13

2. There is Exactly One Rotational Axis *C*

a) If the rotation can have any angle, i.e., $C = C_\infty$, then the molecule is linear, and the symmetry group is infinite.

■ **A:** For the molecule of sodium chloride (common salt) NaCl there is no horizontal reflection. The corresponding symmetry group of all the rotations around C is denoted by $C_{\infty v}$.

■ **B:** The molecule O_2 has one horizontal reflection. The corresponding symmetry group is generated by the rotations and by this reflection, and it is denoted by $D_{\infty h}$.

b) The rotation axis is of n-th order, $C = C_n$, but it is not an improper rotational axis of order $2n$.
If there is no further symmetry element, then G is generated by a rotation d by an angle π/n around C_n, i.e., $G =< d > \cong Z_n$. In this case G is also denoted by C_n.
If there is a further vertical reflection σ_v, then $G =< d, \sigma_v > \cong D_n$ holds (see 5.3.3.1, p. 336), and G is denoted by C_{nv}.
If there exists an additional horizontal reflection σ_h, then $G =< d, \sigma_v > \cong Z_n \times Z_2$ holds. G is denoted by C_{nh} and it is cyclic for odd n (see 5.3.3.2, p. 337).

■ **A:** For hydrogen peroxide **(Fig.5.13)** these three cases occur in the order given above for $0 < \delta < \pi/2, \delta = 0$ and $\delta = \pi/2$.

■ **B:** The molecule of water H_2O has a rotational axis of second order and a vertical plane of reflection, as symmetry elements. Consequently, the symmetry group of water is isomorphic to the group D_2, which is isomorphic to the Klein four-group V_4 (see 5.3.3.2, **3.**, p. 338).

c) The rotational axis is of order n and at the same time it is also an improper rotational axis of order

$2n$. We have to distinguish two cases.

α) There is no further vertical reflection, so $G \cong Z_{2n}$ holds, and G is denoted also by S_{2n}.

■ An example is the molecule of tetrahydroxy allene with formula $C_3(OH)_4$ **(Fig.5.14)**.

β) If there is a vertical reflection, then G is a group of order $4n$, which is denoted by D_{2n}.

■ $n = 2$ gives $G \cong D_4$, i.e., the dihedral group of order eight. An example is the allene molecule **(Fig.5.15)**.

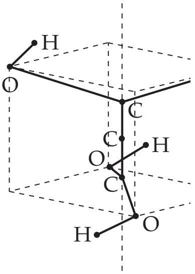


Figure 5.14

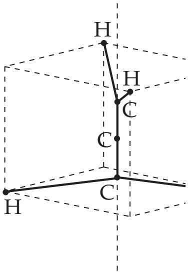


Figure 5.15

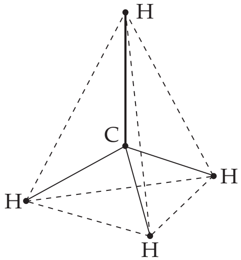


Figure 5.16

3. Several Rotational Axes If there are several rotational axes, then one has to distinguish further cases. In particular, if several rotational axes have an order $n \geq 3$, then the following groups are the corresponding symmetry groups.

a) Tetrahedral group T_d: Isomorphic to S_4, ord$T_d = 24$.

b) Octahedral group O_h: Isomorphic to $S_4 \times Z_2$, ord$O_h = 48$.

c) Icosahedral group I_h: ord$I_h = 120$.

These groups are the symmetry groups of the regular polyhedron discussed in 3.3.3, **Table 3.7**, p. 155, **(Fig.3.63)**.

■ The methane molecule **(Fig.5.16)** has the tetrahedral group T_d as a symmetry group.

5.3.5.4 Symmetry Groups in Crystallography

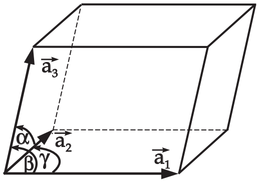


Figure 5.17

1. Lattice Structures

In crystallography the parallelepiped represents, independently of the arrangement of specific atoms or ions, the elementary (unit) cell of the *crystal lattice*. It is determined by three non-coplanar basis vectors $\vec{\mathbf{a}}_i$ starting from one lattice point **(Fig. 5.17)**. The infinite geometric lattice structure is created by performing all *primitive translations* $\vec{\mathbf{t}}_n$:

$$\vec{\mathbf{t}}_{\mathbf{n}} = n_1\vec{\mathbf{a}}_1 + n_2\vec{\mathbf{a}}_2 + n_3\vec{\mathbf{a}}_3, \quad n = (n_1, n_2, n_3) \quad n_i \in \mathrm{Z}. \tag{5.124}$$

Here, the coefficients n_i $(i = 1, 2, \ldots)$ are integers.

All the translations $\vec{\mathbf{t}}_n$ fixing the space points of the lattice $L = \{\vec{\mathbf{t}}_n\}$ in terms of lattice vectors form the translation group T with the group element $T(\vec{\mathbf{t}}_n)$, the inverse element $T^{-1}(\vec{\mathbf{t}}_{\mathbf{n}}) = T(-\vec{\mathbf{t}}_{\mathbf{n}})$, and the composition law $T(\vec{\mathbf{t}}_{\mathbf{n}}) * T(\vec{\mathbf{t}}_{\mathbf{m}}) = T(\vec{\mathbf{t}}_{\mathbf{n}} + \vec{\mathbf{t}}_{\mathbf{m}})$. The application of the group element $T(\vec{\mathbf{t}}_{\mathbf{n}})$ to the position vector $\vec{\mathbf{r}}$ is described by:

$$T(\vec{\mathbf{t}}_{\mathbf{n}})\vec{\mathbf{r}} = \vec{\mathbf{r}} + \vec{\mathbf{t}}_{\mathbf{n}}. \tag{5.125}$$

2. Bravais Lattices

Taking into account the possible combinations of the relative lengths of the basis vectors $\vec{\mathbf{a_i}}$ and the pairwise related angles between them (particularly angles 90° and 120°) one obtains seven different types of *elementary cells* with the corresponding lattices, the *Bravais lattices* (see **Fig. 5.17**, and **Table 5.4**). This classification can be extended by seven *non-primitive elementary cells* and their corresponding lattices by adding additional lattice points at the intersection points of the face or body diagonals, preserving the symmetry of the elementary cell. In this way one may distinguish one-side face-centered lattices, body-centered lattices, and all-face centered lattices.

Tabelle 5.4 Primitive Bravais lattice

Elementary cell	**Relative lengths of basis vectors**	**Angles between basis vectors**
triclinic	$a_1 \neq a_2 \neq a_3$	$\alpha \neq \beta \neq \gamma \neq 90°$
monoclinic	$a_1 \neq a_2 \neq a_3$	$\alpha = \gamma = 90° \neq \beta$
rhombic	$a_1 \neq a_2 \neq a_3$	$\alpha = \beta = \gamma = 90°$
trigonal	$a_1 = a_2 = a_3$	$\alpha = \beta = \gamma < 120° (\neq 90°)$
hexagonal	$a_1 = a_2 \neq a_3$	$\alpha = \beta = 90°, \gamma = 120°$
tetragonal	$a_1 = a_2 \neq a_3$	$\alpha = \beta = \gamma = 90°$
cubic	$a_1 = a_2 = a_3$	$\alpha = \beta = \gamma = 90°$

3. Symmetry Operations in Crystal Lattice Structures

Among the symmetry operations transforming the space lattice to equivalent positions there are point group operations such as certain rotations, improper rotations, and reflections in planes or points. But not all point groups are also crystallographic point groups. The requirement that the application of a group element to a lattice vector $\vec{\mathbf{t_n}}$ leads to a lattice vector $\vec{\mathbf{t'_n}} \in L$ (L is the set of all lattice points) again restricts the allowed point groups P with the group elements $P(R)$ according to:

$$P = \{R : R\vec{\mathbf{t_n}} \in L\}\,, \quad \vec{\mathbf{t_n}} \in L. \tag{5.126}$$

Here, R denotes a proper ($R \in SO(3)$) or improper rotation operator ($R = IR' \in O(3), R' \in SO(3), I$ is the inversion operator with $I\vec{\mathbf{r}} = -\vec{\mathbf{r}}$, $\vec{\mathbf{r}}$ is a position vector). For example, only n-fold rotation axes with $n = 1, 2, 3, 4$ or 6 are compatible with a lattice structure. Altogether, there are 32 crystallographic point groups P.

The symmetry group of a space lattice may also contain operators representing simultaneous applications of rotations and primitive translations. In this way one gets gliding reflections, i.e., reflections in a plane and translations parallel to the plane, and screws, i.e., rotations through $2\pi/n$ and translations by $m\vec{\mathbf{a}}/n$ ($m = 1, 2, \ldots, n-1$, $\vec{\mathbf{a}}$ are basis translations). Such operations are called non-primitive translations $\vec{\mathbf{V}}(R)$, because they correspond to “fractional” translations. For a gliding reflection R is a reflection and for a screw R is a proper rotation.

The elements of the space group G, for which the crystal lattice is invariant are composed of elements P of the crystallographic point group P, primitive translations $T(\vec{\mathbf{t_n}})$ and non-primitive translations $\vec{\mathbf{V}}(R)$:

$$G = \{\{R|\vec{\mathbf{V}}(R) + \vec{\mathbf{t_n}} : R \in P, \quad \vec{\mathbf{t_n}} \in L\}\}. \tag{5.127}$$

The unit element of the space group is $\{e|0\}$ where e is the unit element of R. The element $\{e|\vec{\mathbf{t_n}}\}$ means a primitive translation, $\{R|0\}$ represents a rotation or reflection. Applying the group element

$\{R|\vec{\mathbf{t}}_n\}$ to the position vector $\vec{\mathbf{r}}$ one obtains:

$$\{R|\vec{\mathbf{t_n}}\}\vec{\mathbf{r}} = R\vec{\mathbf{r}} + \vec{\mathbf{t_n}}. \qquad (5.128)$$

4. Crystal Systems (Holohedry)

From the 14 Bravais lattices, $L = \{\vec{\mathbf{t_n}}\}$, the 32 crystallographic point groups $P = \{R\}$ and the allowed non-primitive translations $\vec{\mathbf{V}}(R)$ one can construct 230 space groups $G = \{R|\vec{\mathbf{V}}(R) + \vec{\mathbf{t_n}}\}$. The point groups correspond to 32 crystallographic classes. Among the point groups there are seven groups that are not a subgroup of another point group but contain further point groups as a subgroup. Each of these seven point groups form a *crystal system* (*holohedry*). The symmetry of the seven crystal systems is reflected in the symmetry of the seven Bravais lattices. The relation of the 32 crystallographic classes to the seven crystal systems is given in **Table 5.5** using the notation of Schoenflies.

Remark: The space group G (5.127) is the symmetry group of the "empty" lattice. The real crystal is obtained by arranging certain atoms or ions at the lattice sites. The arrangement of these crystal constituents exhibits its own symmetry. Therefore, the symmetry group G_0 of the real crystal possesses a lower symmetry than G $(G \supset G_0)$, in general.

Table 5.5 Bravais lattice, crystal systems, and crystallographic classes
Notation: C_n – rotation about an n-fold rotation axis, D_n – dihedral group, T_n – tetrahedral group, O_n – octahedral group, S_n – mirror rotations with an n-fold axis.

Lattice type	**Crystal system (holohedry)**	**Crystallographic class**
triclinic	C_i	C_1, C_i
monoclinic	C_{2h}	C_2, C_h, C_{2h}
rhombic	D_{2h}	C_{2v}, D_2, D_{2h}
tetragonal	D_{4h}	$C_4, S_4, C_{4h}, D_4, C_{4v}, D_{2d}, D_{4h}$
hexagonal	D_{6h}	$C_6, C_{3h}, C_{6h}, D_6, C_{6v}, D_{3h}, D_{6h}$
trigonal	D_{3d}	$C_3, S_6, D_3, C_{3v}, D_{3d}$
cubic	O_h	T, T_h, T_d, O, O_h

5.3.5.5 Symmetry Groups in Quantum Mechanics

Linear coordinate transformations that leave the Hamiltonian $\hat{H}$ of a quantum mechanical system (see 9.2.4, **2.**, p. 593) invariant represent a symmetry group G, whose elements g commute with $\hat{H}$:

$$[g, \hat{H}] = g\hat{H} - \hat{H}g = 0, \quad g \in G. \qquad (5.129)$$

The commutation property of g and $\hat{H}$ implies that in the application of the product of the operators g and $\hat{H}$ to a state φ the sequence of the action of the operators is arbitrary:

$$g(\hat{H}\varphi) = \hat{H}(g\varphi). \qquad (5.130)$$

Hence, one has: If $\varphi_{E\alpha}$ $(\alpha = 1, 2, \dots, n)$ are the eigenstates of $\hat{H}$ with energy eigenvalue E of degeneracy n, i.e.,

$$\hat{H}\varphi_{E\alpha} = E\varphi_{E\alpha} \quad (\alpha = 1, 2, \dots, n), \qquad (5.131)$$

then the transformed states $g\varphi_{E\alpha}$ are also eigenstates belonging to the same eigenvalue E:

$$g\hat{H}\varphi_{E\alpha} = \hat{H}g\varphi_{E\alpha} = Eg\varphi_{E\alpha}. \qquad (5.132)$$

The transformed states $g\varphi_{E\alpha}$ can be written as a linear combination of the eigenstates $\varphi_{E\alpha}$:

$$g\varphi_{E\alpha} = \sum_{\beta=1}^{n} D_{\beta\alpha}(g)\varphi_{E\beta}. \tag{5.133}$$

Hence, the eigenstates $\varphi_{E\alpha}$ form the basis of an n-dimensional representation space for the representation $D(G)$ of the symmetry group G of the Hamiltonian $\hat{H}$ with the representation matrices $(D_{\alpha\beta}(g))$. This representation is irreducible if there are no "hidden" symmetries. One can state that the energy eigenstates of a quantum mechanical system can be labeled by the signatures of the irreducible representations of the symmetry group of the Hamiltonian.

Thus, the representation theory of groups allows for qualitative statements on such patterns of the energy spectrum of a quantum mechanical system which are established by the outer or inner symmetries of the system only. Also the splitting of degenerate energy levels under the influence of a perturbation which breaks the symmetry or the selection rules for the matrix elements of transitions between energy eigenstates follows from the investigation of representations according to which the participating states and operators transform under group operations.

The application of group theory in quantum mechanics is presented extensively in the literature (see, e.g., [5.6], [5.7], [5.8], [5.10], [5.11]).

5.3.5.6 Further Applications of Group Theory in Physics

Further examples of the application of particular continuous groups in physics can only be mentioned here (see, e.g., [5.6], [5.10]).

$U(1)$: Gauge transformations in electrodynamics.

$SU(2)$: Spin and isospin multiplets in particle physics.

$SU(3)$: Classification of the baryons and mesons in particle physics. Many-body problem in nuclear physics.

$SO(3)$: Angular momentum algebra in quantum mechanics. Atomic and nuclear many-body problems.

$SO(4)$: Degeneracy of the hydrogen spectrum.

$SU(4)$: Wigner super-multiplets in the nuclear shell model due to the unification of spin and isospin degrees of freedom. Description of flavor multiplets in the quark model including the charm degree of freedom.

$SU(6)$: Multiplets in the quark model due to the combination of flavor and spin degrees of freedom. Nuclear structure models.

$U(n)$: Shell models in atomic and nuclear physics.

$SU(n), SO(n)$: Many-body problems in nuclear physics.

$SU(2) \otimes U(1)$: Standard model of the electro weak interaction.

$SU(5) \supset SU(3) \otimes SU(2) \otimes U(1)$: Unification of fundamental interactions (GUT).

Remark: The groups $SU(n)$ and $SO(n)$ are Lie groups, i.e. continuous groups (see, 5.3.6, p. 351 and e.g., [5.6]).

5.3.6 Lie Groups and Lie Algebras

5.3.6.1 Introduction

Lie groups and *Lie algebras* are named after the Norwegian mathematician Sophus Lie (1842-1899). In this chapter only Lie groups of matrices are considered since they are most important in applications. Main examples of matrix-Lie groups are:

- the group $O(n)$ of orthogonal matrices,

- the subgroup $SO(n)$ of orthogonal matrices of determinants $+1$, i.e. the orthogonal matrices describing rotations in $\mathbb{R}^n$,

- the Euclidean group $SE(n)$, which describes rigid-body motions.

These groups have many applications in computer graphics and in robotics.

The most important relation between a Lie group and the corresponding Lie algebra will be described by the exponential mapping. This relation is explained by the following example.

■ The solution of initial value problems of first order differential equations or of a system of differential equations can be determined with the help of the exponential function.
The initial value problem (5.134a) for $y = y(t)$ has the following solution (5.134b):

$$\frac{dy}{dt} = x\,y \quad (x \text{ const) with } y(0) = y_0\,, \qquad (5.134a) \qquad\qquad y(t) = e^{xt}y_0\,. \qquad (5.134b)$$

Similarly, for the system of first order differential equations with unknown vector $\vec{\mathbf{y}} = \vec{\mathbf{y}}(t)$ and with the constant coefficient matrix $\mathbf{X}$ the initial value problem (5.135a)

$$\frac{d\vec{\mathbf{y}}}{dt} = \left(\frac{dy_1}{dt}, \frac{dy_2}{dt}, \ldots, \frac{dy_n}{dt}\right)^{\mathrm{T}} = \mathbf{X}\vec{\mathbf{y}} \quad (\text{matrix } \mathbf{X} \text{ const) with } \vec{\mathbf{y}}(0) = \vec{\mathbf{y}}_0, \qquad (5.135a)$$

has the solution (5.135b) with the matrix-exponential function $e^{t\mathbf{X}}$:

$$\vec{\mathbf{y}}(t) = e^{\mathbf{X}\,t}\vec{\mathbf{y}}_0\,, \qquad e^{t\mathbf{X}} := \sum_{k=0}^{\infty}\frac{1}{k!}t^k\mathbf{X}^k = I_{n\times n} + \sum_{k=1}^{\infty}\frac{1}{k!}t^k\mathbf{X}^k\,. \qquad (5.135b)$$

The special matrix-exponential function $e^{t\mathbf{X}}$ for a given quadratic $n \times n$ matrix $\mathbf{X}$ has the following properties:

- $e^{0\mathbf{X}} = I_{n\times n}$, where $I_{n\times n}$ denotes the unit matrix.
- $e^{t\mathbf{X}}$ is invertible, because $\det e^{t\mathbf{X}} = e^{t\cdot \mathrm{Spur}\,\mathbf{X}} \neq 0$.
- $e^{t_1\mathbf{X}}e^{t_2\mathbf{X}} = e^{(t_1+t_2)\mathbf{X}} = e^{t_2\mathbf{X}}e^{t_1\mathbf{X}}$ for every $t_1,\, t_2 \in \mathbb{R}$, but in general is $e^{\mathbf{X}_1}e^{\mathbf{X}_2} \neq e^{\mathbf{X}_2}e^{\mathbf{X}_1} \neq e^{\mathbf{X}_1+\mathbf{X}_2}$.
- In particular $e^{-t\mathbf{X}}e^{t\mathbf{X}} = e^{t\mathbf{X}}e^{-t\mathbf{X}} = I_{n\times n}$.
- $\left.\frac{d}{dt}e^{t\mathbf{X}}\right|_{t=0} = \mathbf{X}\,e^{t\mathbf{X}}\big|_{t=0} = \mathbf{X}\,.$

Consequently, the elements $e^{t\mathbf{X}}$ (for a fixed $\mathbf{X}$) form a multiplicative group with respect to matrix multiplication. Since $t \in \mathbb{R}$, the matrices $e^{t\mathbf{X}}$ form a one dimensional group. At the same time it is one of the simplest examples of Lie groups. It will be shown that matrices $\mathbf{X}$ and $t\mathbf{X}$ are elements of the Lie algebra belonging to this Lie group (see 5.3.6.4, p. 356). In this way the exponential function generates the Lie group from the elements of the Lie algebra.

5.3.6.2 Matrix-Lie Groups

For matrix-Lie groups it is not necessary to define Lie groups in general. For general Lie groups there should be introduced the notion of differentiable manifolds, which is not needed here. For matrix-Lie groups the following definitions are important, while in further discussions the main topic will be the *general linear group*.

1. General Linear Group

1. Group A group (see 5.3.3, p. 336) is a set G with a map

$$G \times G \to G\,, \quad (g, h) \mapsto g * h\,, \qquad (5.136a)$$

which is the so called group operation or group multiplication with the following properties:

- Associativity: for every $g, h, k \in G$

$$g * (h * k) = (g * h) * k\,, \qquad (5.136b)$$

- Existence of identity: There is an element $e \in G$, such that for every $g \in G$

$$g * e = e * g = g\,, \qquad (5.136c)$$

- Existence of an inverse: For every $g \in G$ there is an element $h \in G$ such that

$$g * h = h * g = e\,. \tag{5.136d}$$

Remark 1: If $g * h = h * g$ for every$g, h \in G$, then the group is called *commutative*. The matrix groups considered here are not commutative. It follows obviously from the definition, that the product of two elements of the group also belongs to the group, so the group is closed with respect to group multiplication.

Remark 2: Let $M_n(\mathbb{R})$ the vector space of all $n \times n$ matrices with real entries. $M_n(\mathbb{R})$ is obviously not a group with respect to matrix multiplication, since not every $n \times n$ matrix is invertible.

2. Definition of the General Linear Group The set of all real, invertible, $n \times n$ matrices, which obviously form a group with respect to matrix multiplication, is called the *general linear group* and is denoted by $GL(n, \mathbb{R})$.

2. Matrix-Lie Groups

1. Convergence of Matrices A sequence $\{\mathbf{A}_m\}_{m=1}^{\infty}$ of matrices $\mathbf{A}_m = (a_{kl}^{(m)})_{k,l=1}^{n}$ where $\mathbf{A}_m \in M_n(\mathbb{R})$ converges to the $n \times n$ matrix $\mathbf{A}$, if every sequence of entries $\{(a_{kl}^{(m)})\}_{m=1}^{\infty}$ converges to the corresponding matrix entry a_{kl} in the sense of convergence of real numbers.

2. Definition of the Matrix-Lie Groups A matrix-Lie group is a subgroup G of $GL(n, \mathbb{R})$ with the property: Let $\{\mathbf{A}_m\}_{m=1}^{\infty}$ be an arbitrary sequence of matrices from G converging to a matrix $\mathbf{A} \in M_n(\mathbb{R})$ in the sense of convergence in $M_n(\mathbb{R})$. Then either $\mathbf{A} \in G$ or $\mathbf{A}$ is not invertible.

This definition can be also formulated in the following way: A matrix-Lie group is a subgroup which is also a closed subset of $GL(n\,, \mathbb{R})$. (It does not mean, that G must be closed in $M_n(\mathbb{R})$).

3. Dimension of the Matrix-Lie Group The dimension of a matrix-Lie group is defined as the dimension of the corresponding Lie algebra (see 5.3.6.4, p. 356). The matrix-Lie group $GL(n\,, \mathbb{R})$ has dimension n^2.

3. Continuous Groups

Matrix-Lie groups can be introduced also with the help of continuous groups (see [22.22], [5.9], [5.7]).

1. Definition A continuous group is a special infinite group whose elements are given uniquely by a continuous parameter vector $\underline{\varphi} = (\varphi_1, \varphi_2, \ldots, \varphi_n)$:

$$a = a(\underline{\varphi})\,. \tag{5.137}$$

■ Group of rotation matrices in $\mathbb{R}^2$ (see (3.432), p. 230):

$$D = \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix} = a(\varphi) \ \text{ mit } \ 0 \le \varphi \le 2\pi\,. \tag{5.138}$$

The group elements depend only on one real parameter φ.

2. Product The product of two elements $a_1 = a(\underline{\varphi}_1)$, $a_2 = a(\underline{\varphi}_2)$ of a continuous group with elements $a = a(\underline{\varphi})$ is given by

$$a_1 * a_2 = a_3 = a(\underline{\varphi_{\mathbf{3}}}) \ \text{ with} \tag{5.139a}$$

$$\varphi_3 \quad = \underline{f}(\underline{\varphi_{\mathbf{1}}}, \underline{\varphi_{\mathbf{2}}})\,, \tag{5.139b}$$

where the components of $\underline{f}(\underline{\varphi_{\mathbf{1}}}, \underline{\varphi_{\mathbf{2}}})$ are continuously differentiable functions.

■ The product of two rotation matrices $a = a(\varphi_1)$ and $a = a(\varphi_2)$ with $0 \le \varphi_1, \varphi_2 \le 2\pi$ ($a(\varphi)$ as in (5.138), is $a_3 = a(\varphi_1) * a(\varphi_2) = a(\varphi_3)$ with $\varphi_3 = f(\varphi_1, \varphi_2) = \varphi_1 + \varphi_2$. Using the Falk's scheme (see 4.1.4, **5.**, p. 273) and addition theorems one gets:

$$\begin{array}{c|c} & a(\varphi_2) \\ \hline a(\varphi_1) & a(\varphi_3) = a(\varphi_1 + \varphi_2) \end{array}$$ or detailed

$$\begin{array}{cc|cc} & & \cos\varphi_2 & -\sin\varphi_2 \\ & & \sin\varphi_2 & \cos\varphi_2 \\ \hline \cos\varphi_1 & -\sin\varphi_1 & \cos\varphi_1\cos\varphi_2 - \sin\varphi_1\sin\varphi_2 & -\cos\varphi_1\sin\varphi_2 - \sin\varphi_1\cos\varphi_2 \\ \sin\varphi_1 & \cos\varphi_1 & \sin\varphi_1\cos\varphi_2 + \cos\varphi_1\sin\varphi_2 & -\sin\varphi_1\sin\varphi_2 + \cos\varphi_1\cos\varphi_2 \end{array}.$$

3. Dimension The parameter vectors $\underset{\sim}{\varphi}$ are elements of a vector space which is called parameter space. In this parameter space there is a domain which is given as the domain of the continuous group, and it is called the group space. The dimension of this group space is considered as the dimension of the continuous group.

■ **A:** The group of the real quadratic $n \times n$ invertible matrices has the dimension n^2, since every entry can be considered as a parameter.

■ **B:** The group of the rotation matrices (with respect to matrix multiplication) D in (5.138) has dimension 1. The rotation matrices are of type 2×2, but their four entries depend only on one parameter φ $(0 \le \varphi \le 2\pi)$.

4. Lie Groups

1. Definition of the Lie Group A Lie group is a continuous group where all elements of the group are given as continuous functions of the parameters.

2. Special Matrix-Lie Groups and their Dimension

■ **A Group $\boldsymbol{SO(n)}$ of Rotations R:** The group $SO(n)$ of rotations $\mathbf{R}$ acts on the elements $\vec{\mathbf{x}} \in \mathbb{R}^n$ with matrix multiplication as $\vec{\mathbf{x}}' = \mathbf{R}\,\vec{\mathbf{x}} \in \mathbb{R}^n$. $SO(n)$ is an $n(n-1)/2$-dimensional Lie group.

■ **B Special Euclidean Group $\boldsymbol{SE(n)}$:** The special Euclidian group $SE(n)$ consists of elements $g = (\mathbf{R}, \vec{\mathbf{b}})$ with $\mathbf{R} \in SO(n)$ and $\vec{\mathbf{b}} \in \mathbb{R}^n$ and with group multiplication $g_1 \circ g_2 = (\mathbf{R}_1\,\mathbf{R}_2,\ \mathbf{R}_1\vec{\mathbf{b}}_2 + \vec{\mathbf{b}}_1)$. It acts on the elements of Euclidean spaces $\mathbb{R}^n$ as

$$\vec{\mathbf{x}}' = \mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{b}}\,. \tag{5.140}$$

$SE(n)$ is the group of rigid-body motions of n-dimensional Euclidean space, it is an $n(n+1)/2$-dimensional Lie group. Discrete subgroups of $SE(n)$ are e.g. the crystallographic space groups, i.e. the symmetry group of a regular crystal-lattice.

■ **C Scaled Euclidean Group $\boldsymbol{SIM(n)}$:** The scaled Euclidian group $SIM(n)$ consists of all pairs $(e^a\mathbf{R},\ \vec{\mathbf{b}})$ with $a \in \mathbb{R}$, $\mathbf{R} \in SO(n)$, $\vec{\mathbf{b}} \in \mathbb{R}^n$, with group multiplication $g_1 \circ g_2 = (e^{a_1+a_2}\mathbf{R}_1\,\mathbf{R}_2,\ \mathbf{R}_1\vec{\mathbf{b}}_2 + \vec{\mathbf{b}}_1)$. It acts on the elements of $\mathbb{R}^n$ by translation, rotation and dilatation (=stretching or shrinking):

$$\vec{\mathbf{x}}' = e^a\mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{b}}\,. \tag{5.141}$$

The scaled Euclidean group has the dimension $1 + n(n+1)/2$.

■ **D Real Special Linear Group $\boldsymbol{SL(n\,,\mathbb{R})}$:** The real special linear group consists of all (real) $n\times n$ matrices with determinant $+1$. It acts on the elements of $\mathbb{R}^n$ with $\vec{\mathbf{x}}' = \mathbf{L}\vec{\mathbf{x}}$ by rotation, distortion and shearing so that the volume remains the same and parallel lines remain parallel. The dimension is $n^2 - 1$.

■ **E Special Affine Group:** The special affine groups of $\mathbb{R}^n$, which consists of all pairs $(e^a\,\mathbf{L},\ \vec{\mathbf{b}})$ with $L \in SL(n)$ and $\vec{\mathbf{b}} \in \mathbb{R}^n$, acts on the objects in $\mathbb{R}^n$ as rotation, translation, shearing, distortion and dilatation. This Lie group is the most general group of deformations in Euclidean spaces mapping parallel lines into parallel lines; it has dimension $n(n+1)$.

■ **F Group $\boldsymbol{SO(2)}$:** The group $SO(2)$ describes all rotations about the origin in $\mathbb{R}^2$:

$$\mathrm{SO}(2) = \left\{ \begin{pmatrix} \cos\varphi & -\sin\varphi \\ \sin\varphi & \cos\varphi \end{pmatrix},\ \varphi \in \mathbb{R} \right\} \tag{5.142}$$

■ **G Group $\boldsymbol{SL(2)}$:** Every element of $SL(2)$ can be represented as

$$\begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} e^t & 0 \\ 0 & e^{-t} \end{pmatrix} \begin{pmatrix} 1 & \xi \\ 0 & 1 \end{pmatrix} . \tag{5.143}$$

■ **H Group $\boldsymbol{SE(2)}$:** The elements of the group $SE(2)$ can be represented as 3×3 matrices:

$$\begin{pmatrix} \cos\theta & -\sin\theta & x_1 \\ \sin\theta & \cos\theta & x_2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{with } \theta \in \mathbb{R} \text{ and } \vec{\mathbf{x}} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^2 . \tag{5.144}$$

Remark: Beside real matrix-Lie groups complex matrix-Lie groups also can be considered. So, e.g. $SL(n, \mathbb{C})$ is the Lie group of all complex $n \times n$ matrices with determinant $+1$. Similarly there are matrix-Lie groups whose entries are quaternions.

5.3.6.3 Important Applications

1. Rigid Body Movement

1. The group $SE(3)$ is the group of rigid-body motions in the Euclidean space $\mathbb{R}^3$. That is why it is so often applied in control of robots. The 6 independent transformations are defined usually as follows:

1. Translation in x-direction,
2. Translation in y-direction,
3. Translation in z-direction,
4. Rotation about the x-axis,
5. Rotation about the y-axis,
6. Rotation about the z-axis.

These transformations can be represented by 4×4 matrices applied to homogeneous coordinates (see 3.5.4.2, p. 231) in 3 dimensions, i.e. $(x,\, y,\, z)^{\mathrm{T}} \in \mathbb{R}^3$ is represented as a vector $(x,\, y,\, z,\, 1)^{\mathrm{T}}$ with four coordinates (see 3.5.4.2, p. 231).
Matrices corresponding to the transformations 1 until 6 are:

$$\mathbf{M}_1 = \begin{pmatrix} 1&0&0&a \\ 0&1&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&b \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}, \quad \mathbf{M}_3 = \begin{pmatrix} 1&0&0&0 \\ 0&1&0&0 \\ 0&0&1&c \\ 0&0&0&1 \end{pmatrix}, \tag{5.145a}$$

$$\mathbf{M}_4 = \begin{pmatrix} 1&0&0&0 \\ 0&\cos\alpha&-\sin\alpha&0 \\ 0&\sin\alpha&\cos\alpha&0 \\ 0&0&0&1 \end{pmatrix}, \mathbf{M}_5 = \begin{pmatrix} \cos\beta&0&\sin\beta&0 \\ 0&1&0&0 \\ -\sin\beta&0&\cos\beta&0 \\ 0&0&0&1 \end{pmatrix}, \mathbf{M}_6 = \begin{pmatrix} \cos\gamma&-\sin\gamma&0&0 \\ \sin\gamma&\cos\gamma&0&0 \\ 0&0&1&0 \\ 0&0&0&1 \end{pmatrix}. \tag{5.145b}$$

The matrices $\mathbf{M}_4, \mathbf{M}_5, \mathbf{M}_6$ describe the rotations in $\mathbb{R}^3$, consequently $SO(3)$ is a subgroup of $SE(3)$.
The group $SE(3)$ acts on $\vec{\mathbf{x}} = (x,\, y,\, z)^{\mathrm{T}} \in \mathbb{R}^3$ with homogeneous coordinates $(\vec{\mathbf{x}},\, 1)^{\mathrm{T}}$ as follows:

$$\begin{pmatrix} \vec{\mathbf{x}}\,' \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \vec{\mathbf{v}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{x}} \\ 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}\vec{\mathbf{x}} + \vec{\mathbf{v}} \\ 1 \end{pmatrix} \tag{5.146}$$

where $\mathbf{R} \in SO(3)$ is a rotation, and $\vec{\mathbf{v}} = (a,\, b,\, c)^{\mathrm{T}}$ is a translation vector.

2. Affine Transformations of 2-Dimensional Space

The group $GA(2)$ of affine transformations of the 2-dimensional space is a 6-dimensional matrix Lie group with the following 6 dimensions:

1. Translation in x-direction,
2. Translation in y-direction,
3. Rotation about the origin,
4. Stretching or shrinking with respect to the origin,
5. Shearing (stretching with resp. to y, with resp. to x),
6. 45°-shearing with respect to 5.

Also these transformations are described by matrices in homogeneous coordinates $(x, y, 1)^T$ for $(x, y)^T \in \mathbb{R}^2$:

$$\mathbf{M}_1 = \begin{pmatrix} 1&0&a \\ 0&1&0 \\ 0&0&1 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1&0&0 \\ 0&1&b \\ 0&0&1 \end{pmatrix}, \quad \mathbf{M}_3 = \begin{pmatrix} \cos\alpha&-\sin\alpha&0 \\ \sin\alpha&\cos\alpha&0 \\ 0&0&1 \end{pmatrix}, \tag{5.147a}$$

$$\mathbf{M}_4 = \begin{pmatrix} e^{\tau} & 0 & 0 \\ 0 & e^{\tau} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{M}_5 = \begin{pmatrix} e^{\mu} & 0 & 0 \\ 0 & e^{-\mu} & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{M}_6 = \begin{pmatrix} \cosh\nu & \sinh\nu & 0 \\ \sinh\nu & \cosh\nu & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.147b}$$

This group has as essential subgroups the translation group, given by $\mathbf{M}_1$ and $\mathbf{M}_2$, the Euclidean group $SE(2)$, given by $\mathbf{M}_1$, $\mathbf{M}_2$ and $\mathbf{M}_3$, the similarity group, given by $\mathbf{M}_1$, $\mathbf{M}_2$, $\mathbf{M}_3$, $\mathbf{M}_4$.

Application: The group $GA(2)$ can be applied to describe all transformations of a planar object which is recorded under slight angle modifications by a camera moving in the 3 dimensional space.
If also large changes in angles of perspectivity can occur, then group $P(2)$ the group of all transformations of projective spaces can be used. The matrix-Lie group is generated by the matrices $\mathbf{M}_1$ until $\mathbf{M}_6$ and by the two further matrices

$$\mathbf{M}_7 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \beta & 0 & 1 \end{pmatrix}, \qquad \mathbf{M}_8 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \gamma & 1 \end{pmatrix}. \tag{5.147c}$$

These two additional matrices correspond to a change of the horizon or vanishing of an edge of the plane picture.

5.3.6.4 Lie Algebra

1. Real Lie algebra

A *real Lie algebra* $\mathcal{A}$ is a real vector space with an operation

$$[\cdot\,,\,\cdot] \,:\, \mathcal{A} \times \mathcal{A} \to \mathcal{A}, \tag{5.148}$$

which is called the Lie bracket and for which the following properties are valid for all $a, b, c \in \mathcal{A}$:

- $[.\,,\,.]$ is bilinear,
- $[a\,,\,b] = -[b\,,\,a]$, i.e. the operation is skew-symmetric or anticommutative,
- the so called Jacobi identity is valid (as a replacement of the missing associativity)

$$[a\,,\,[b\,,\,c]] + [c\,,\,[a\,,\,b]] + [b\,,\,[c,\,a]] = 0. \tag{5.149}$$

Obviously $[a\,,\,a] = 0$ holds.

2. Lie Bracket

For (real) $n \times n$ matrices $\mathbf{X}$ and $\mathbf{Y}$ a Lie bracket is given by the commutator, i.e.

$$[\mathbf{X},\,\mathbf{Y}] := \mathbf{XY} - \mathbf{YX}. \tag{5.150}$$

3. Special Lie-Algebras

There are associated Lie algebras to matrix-Lie groups.

1. A function $g:\,\mathbb{R} \to GL(n)$ is a *one-parameter subgroup* of $GL(n)$, if

- g is continuous,
- $g(0) = I_{n\times n}$,
- $g(t+s) = g(t)g(s)$ for every $t, s \in \mathbb{R}$.

In particular:

2. If g is a one-parameter subgroup of $GL(n)$, then there exists a uniquely defined matrix $\mathbf{X}$ such that

$$g(t) = e^{t\mathbf{X}} \quad \text{(see 5.3.6.1, p. 351)}. \tag{5.151}$$

3. For every $n \times n$ matrix $\mathbf{A}$ the logarithm $\log \mathbf{A}$ is defined by

$$\log A = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}(A-I)^m\,, \tag{5.152}$$

if this series is convergent. In particular, the series converges if $||\mathbf{A} - \mathbf{I}|| < 1$.

4. Correspondence between Lie Group and Lie Algebra

The correspondence between a matrix-Lie group and the associated Lie algebra is as follows.

1. Let G be a matrix-Lie group. The Lie *algebra of* G, which is denoted by $\mathbf{g}$, is the set of all matrices $\mathbf{X}$ such that $e^{t\mathbf{X}} \in G$ holds for all real numbers t.

In a given matrix-Lie group the elements close to the unit matrix can be represented as $g(t) = e^{t\mathbf{X}}$ with $\mathbf{X} \in \mathbf{g}$, and t close to zero. If the exponential map is surjective, as in the case of $SO(n)$ and $SE(n)$, then the elements of the group can be parameterized with the help of the matrix-exponential function by elements of the corresponding Lie algebra. The matrices $\frac{dg}{dt}g^{-1}$ and $g^{-1}\frac{dg}{dt}$ respectively are called *tangent vectors* or *tangent elements* to $g \in G$. Calculating these elements for $t = 0$, one gets $\mathbf{X}$ itself, i.e. $\mathbf{g}$ is the tangent space $T_{\mathbf{I}}G$ at the identity matrix $\mathbf{I}$.

2. It can be shown that the Lie algebra assigned to a Lie group in this way is a Lie algebra also in the abstract sense.

Let G be a matrix-Lie group with the associated matrix-Lie algebra $\mathbf{g}$ and $\mathbf{X}$ and $\mathbf{Y}$ elements of $\mathbf{g}$. Then:

- $s\mathbf{X} \in \mathbf{g}$ for any real numbers s ,
- $\mathbf{X} + \mathbf{Y} \in \mathbf{g}$,
- $[\mathbf{X},\, \mathbf{Y}] = \mathbf{XY} - \mathbf{YX} \in \mathbf{g}$.

■ **A:** The Lie algebra $\mathbf{so}(2)$ associated to the Lie group $SO(2)$ is calculated from the representation of the elements $g(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ by $SO(2)$ with the help of the tangential elements

$$\left.\frac{dg}{d\theta}\, g^{-1}\right|_{\theta=0} = \left.\begin{pmatrix} -\sin\theta & -\cos\theta \\ \cos\theta & -\sin\theta \end{pmatrix}\begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix}\right|_{\theta=0} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}. \tag{5.153a}$$

Consequently

$$\mathbf{so}(2) = \left\{ s\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad s \in \mathbb{R} \right\}. \tag{5.153b}$$

Conversely, from

$$\mathbf{X} = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} \quad \text{comes} \quad e^{s\mathbf{X}} = \cos s \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \sin s \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} \cos s & -\sin s \\ \sin s & \cos s \end{pmatrix}. \tag{5.153c}$$

■ **B:** The following matrices form a basis for the Lie algebra $\mathbf{so(3)}$:

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.154}$$

Remark: The surjectivity of the exponential mappings $\mathbf{so}(3) \to SO(3)$ and $\mathbf{se}(3) \to SE(3)$ implies the existence of a (many-valued) logarithmic function. Nevertheless this logarithm function can be applied to interpolation.

E.g. if rigid-body motions $\mathbf{B}_1$, $\mathbf{B}_2 \in SE(3)$ are given, then $\log \mathbf{B}_1$, $\log \mathbf{B}_2$ can be calculated which are elements of the Lie algebra $\mathbf{so}(3)$. Then between these logarithms linear interpolation $(1-t)\log \mathbf{B}_1 + t\log \mathbf{B}_2$ can be taken and then the exponential map can be applied in order to get an interpolation between the rigid-body motions $\mathbf{B}_1$ and $\mathbf{B}_2$ by

$$\exp\left((1-t)\log \mathbf{B}_1 + t\log \mathbf{B}_2\right). \tag{5.155}$$

■ **C:** The matrix-Lie algebra $\mathbf{se}(3)$ associated to the matrix-Lie group $SE(3)$ is generated by the matrices:

$$\mathbf{E}_1 = \begin{pmatrix} 0&0&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}, \quad \mathbf{E}_2 = \begin{pmatrix} 0&0&0&0 \\ 0&0&0&1 \\ 0&0&0&0 \\ 0&0&0&0 \end{pmatrix}, \quad \mathbf{E}_3 = \begin{pmatrix} 0&0&0&0 \\ 0&0&0&0 \\ 0&0&0&1 \\ 0&0&0&0 \end{pmatrix}, \tag{5.156a}$$

$$\mathbf{E}_4 = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_5 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{E}_6 = \begin{pmatrix} 0 & -1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{5.156b}$$

5. Inner Product

For a given finite dimensional matrix-Lie group it is always possible to find an orthonormal basis for the associated Lie algebra if a suitable inner product (scalar product) is defined. In this case from any basis of the Lie algebra an orthonormal basis can be obtained by the Gram-Schmidt orthogonalization process (see 4.6.2.2, **4.** p. 316).

In the case of a real matrix-Lie group the Lie algebra consists of real matrices and so an inner product is given by

$$(\mathbf{X},\, \mathbf{Y}) = \frac{1}{2}\mathrm{Spur}\,(\mathbf{X}\mathbf{W}\mathbf{Y}^{\mathrm{T}}) \tag{5.157}$$

with a positive definite real symmetric matrix $\mathbf{W}$.

■ **A:** The group of rigid-body motions $SE(2)$ can be parametrized as

$$g(x_1,\, x_2,\, \theta) = e^{x_1\mathbf{X}_1 + x_2\mathbf{X}_2}\, e^{\theta\,\mathbf{X}_3} = \begin{pmatrix} \cos\theta & -\sin\theta & x_1 \\ \sin\theta & \cos\theta & x_2 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{with} \tag{5.158a}$$

$$\mathbf{X}_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{X}_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.158b}$$

Here $\mathbf{X}_1,\ \mathbf{X}_2,\ \mathbf{X}_3$ form an orthonormal basis of Lie algebra $\mathbf{se}(2)$ with respect to an inner product given by the weight matrix

$$\mathbf{W} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix} \tag{5.158c}$$

■ **B:** A basis of Lie algebra $\mathbf{sl}(2, \mathbb{R})$ is

$$\mathbf{X}_1 = \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{X}_2 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \quad \text{and} \quad \mathbf{X}_3 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}. \tag{5.159}$$

These elements form an orthonormal basis with respect to the weight matrix $\mathbf{W} = \mathbf{I}_{2\times 2} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

5.3.6.5 Applications in Robotics

1. Rigid Body Motion

The special Euclidean group $SE(3)$, which describes the rigid-body motions in $\mathbb{R}^3$, is the semidirect product of group $SO(3)$ (rotation about the origin) and $\mathbb{R}^3$ (translations):

$$SE(3) = SO(3) \times \mathbb{R}^3\,. \tag{5.160}$$

In a direct product the factors have no interaction, but this is a semidirect product since rotations act on translations as it is clear from matrix multiplication:

$$\begin{pmatrix} \mathbf{R}_2 & \vec{\mathbf{t}}_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R}_1 & \vec{\mathbf{t}}_1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R}_2\,\mathbf{R}_1 & \mathbf{R}_2\vec{\mathbf{t}}_1 + \vec{\mathbf{t}}_2 \\ 0 & 1 \end{pmatrix}, \tag{5.161}$$

i.e. the first translation vector is rotated before the second translation vector is added.

2. Theorem of Chasles

This theorem tells that every rigid-body motion which is not a pure translation can be described as a (finite) screw motion. A (finite) screwing motion along an axis through the origin has the form

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} \\ 0 & 1 \end{pmatrix}, \tag{5.162a}$$

where $\vec{\mathbf{x}}$ is a unit vector in the direction of the axis of rotation, θ is the angle of rotation and p is the angular coefficient. Since $\vec{\mathbf{x}}$ is the axis of rotation $\mathbf{R}\vec{\mathbf{x}} = \vec{\mathbf{x}}$, i.e. $\vec{\mathbf{x}}$ is an eigenvector of matrix $\mathbf{R}$ belonging to unit eigenvalue 1 .

When the axis of rotation does not go through the origin, then a point $\vec{\mathbf{u}}$ of the axis of rotation is chosen which is shifted into the origin, then after the screwing it is shifted back:

$$\begin{pmatrix} \mathbf{I} & \vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I} & -\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}. \tag{5.162b}$$

The theorem of Chasles tells that an arbitrary rigid-body motion can be given in the above form, i.e.

$$\begin{pmatrix} \mathbf{R} & \vec{\mathbf{t}} \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix} \tag{5.163}$$

for given $\mathbf{R}$, $\vec{\mathbf{t}}$ and appropriate p and $\vec{\mathbf{u}}$. Assuming that the angle of rotation θ and the axis of rotation $\vec{\mathbf{x}}$ are already known from $\mathbf{R}$

$$\frac{\theta\, p}{2\pi} = \vec{\mathbf{x}} \cdot \vec{\mathbf{t}} \tag{5.164}$$

is valid, so the angular coefficient p can be calculated. Then the solution of a linear system of equations gives $\vec{\mathbf{u}}$:

$$(\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} = \frac{\theta\, p}{2\pi}\vec{\mathbf{x}} - \vec{\mathbf{t}}\,. \tag{5.165}$$

This is a singular system of equations, where $\vec{\mathbf{x}}$ is in its kernel. Therefore the solution $\vec{\mathbf{u}}$ is unique except to a manyfold of $\vec{\mathbf{x}}$. In order to determine $\vec{\mathbf{u}}$ it is reasonable to require that $\vec{\mathbf{u}}$ is perpendicular to $\vec{\mathbf{x}}$. When the rigid body motion is a pure rotation, then it is not possible to determine an appropriate vector $\vec{\mathbf{u}}$.

3. Mechanical Joints

Joints with one degree of freedom can be represented by a one-parameter subgroup of the group $SE(3)$. For the general case of screw joints the corresponding subgroup is

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}, \tag{5.166}$$

where $\vec{\mathbf{x}}$ is the axis of rotation, θ is the angle of rotation, p gives the angular coefficient and $\vec{\mathbf{u}}$ is an arbitrary point on the axis of rotation.

The most often occurring types of joints are the rotational joints which can be described by the following subgroup:

$$A(\theta) = \begin{pmatrix} \mathbf{R} & (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}. \tag{5.167}$$

The subgroup corresponding the shift joints is

$$A(\theta) = \begin{pmatrix} \mathbf{I} & \theta\vec{\mathbf{t}} \\ 0 & 1 \end{pmatrix}, \tag{5.168}$$

where $\vec{\mathbf{t}}$ describes the direction of the shifting.

4. Forward Kinematics

The goal in the case of industrial robots is the moving and control of the end effectors, which is done by joints in a kinematic chain. If all joints are of one parameter and the robot consists e.g. of 6 joints, then every position of the robot can be described by the joint-variables $\vec{\theta}^{\mathrm{T}} = (\theta_1,\, \theta_2,\, \theta_3,\, \theta_4,\, \theta_5,\, \theta_6)$. The output state of the robot is described by the null vector. Then the motions of the robot can be described so that first the farest joint together the end effector are moved and this motion is given by the matrix $A(\theta_6)$. Then the 5-th joint is moved. Since the axis of this joint should not be influenced by the motion of the last joint, this motion is given by the matrix $A(\theta_5)$. In this way all the joints are moved, and the complete motion of the end effector is given by

$$K(\vec{\theta}) = A_1(\theta_1)A_2(\theta_2)A_3(\theta_3)A_4(\theta_4)A_5(\theta_5)A_6(\theta_6). \tag{5.169}$$

5. Vector Product and Lie Algebra

A screw is given by

$$A(\theta) = \begin{pmatrix} \mathbf{R} & \dfrac{\theta\, p}{2\pi}\vec{\mathbf{x}} + (\mathbf{I} - \mathbf{R})\vec{\mathbf{u}} \\ 0 & 1 \end{pmatrix}; \tag{5.170}$$

and it represents rigid body motions parameterized by the angle θ. Obviously, $\theta = 0$ gives the identity transformation. If the derivative is calculated at $\theta = 0$, i.e. the derivative at the identity, then the general element of the Lie algebra is the following:

$$S = \left.\frac{dA}{d\theta}\right|_{\theta=0} = \left.\begin{pmatrix} \dfrac{d\mathbf{R}}{d\theta} & \dfrac{p}{2\pi}\vec{\mathbf{x}} - \dfrac{d\mathbf{R}}{d\theta}\vec{\mathbf{u}} \\ 0 & 0 \end{pmatrix}\right|_{\theta=0} = \begin{pmatrix} \mathbf{\Omega} & \dfrac{p}{2\pi}\vec{\mathbf{x}} - \mathbf{\Omega}\vec{\mathbf{u}} \\ 0 & 0 \end{pmatrix}, \tag{5.171a}$$

where $\mathbf{\Omega} = \dfrac{d\mathbf{R}}{d\theta}(0)$ is a skew symmetric matrix. It can be shown that $\mathbf{R}$ is an orthogonal matrix, so $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{R}\mathbf{R}^{\mathrm{T}} = \mathbf{I}$ holds and therefore

$$\frac{d}{d\theta}(\mathbf{R}\,\mathbf{R}^{\mathrm{T}}) = \frac{d\mathbf{R}}{d\theta}\mathbf{R}^{\mathrm{T}} + \mathbf{R}\frac{d\mathbf{R}^{\mathrm{T}}}{d\theta} = \frac{d\mathbf{I}}{d\theta} = 0\,. \tag{5.171b}$$

Since $\mathbf{R} = \mathbf{I}$ for $\theta = 0$

$$\frac{d\mathbf{R}}{d\theta}(0) + \frac{d\mathbf{R}^{\mathrm{T}}}{d\theta}(0) = 0\,. \tag{5.171c}$$

So every skew symmetric matrix

$$\mathbf{\Omega} = \begin{pmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{pmatrix} \tag{5.171d}$$

can be identified with a vector $\vec{\boldsymbol{\omega}}^{\mathrm{T}} = (\omega_x,\, \omega_y,\, \omega_z)$. In this way the multiplication of any three dimensional vector $\vec{\mathbf{p}}$ by matrix $\mathbf{\Omega}$ corresponds to the vector product with vector $\vec{\boldsymbol{\omega}}$:

$$\mathbf{\Omega}\,\vec{\mathbf{p}} = \vec{\boldsymbol{\omega}} \times \vec{\mathbf{p}}\,. \tag{5.171e}$$

Consequently $\vec{\boldsymbol{\omega}}$ is the angular velocity of the rigid body with an amplitude ω.
Hence a general element of the Lie algebra $\mathbf{se}(3)$ has the form

$$= \begin{pmatrix} \mathbf{\Omega} & \vec{\mathbf{v}} \\ 0 & 0 \end{pmatrix}. \tag{5.171f}$$

These matrices form a 6-dimensional vector space which is often identified with the 6-dimensional vectors of the form

$$\vec{\mathbf{s}} = \begin{pmatrix} \vec{\boldsymbol{\omega}} \\ \vec{\mathbf{v}} \end{pmatrix}. \tag{5.172}$$

5.3.7 Rings and Fields

In this section, there are discussed algebraic structures with two binary operations.

5.3.7.1 Definitions

1. Rings

A set R with two binary operations $+$ and $*$ is called a *ring* (notation: $(R\,,+\,,*)$), if

- $(R,+)$ is an Abelian group,
- $(R,*)$ is a semigroup, and
- the *distributive laws* hold:

$$a * (b + c) = (a * b) + (a * c), \quad (b + c) * a = (b * a) + (c * a). \tag{5.173}$$

If $(R,*)$ is commutative or if $(R,*)$ has a neutral element, then $(R,+,*)$ is called a commutative ring or a ring with identity (ring with unit element), respectively.

A commutative ring with a unit element and without zero divisor is called the *domain of integrity*.

A nonzero element of a ring is called *zero divisor* or *singular element* if there is a nonzero element of the ring such that their product is equal to zero.

In a ring with zero divisor the following implication is generally false: $a * b = 0 \implies (a = 0 \lor b = 0)$.

If R is a ring with a unit element, then the *characteristic of the ring* R is the smallest natural number k such that $k1 = 1 + 1 + \ldots + 1 = 0$ (k times 1 equals to zero), and it is denoted by char $R = k$. If such a k does not exist, then char $R = 0$.

char $R = k$ means that the cyclic subgroup $\langle 1\rangle$ of the additive group $(R,+)$ generated by 1 has order k, so the order of every element is a divisor of k.

If char $R = k$ and for all $r \in R$, then $r + r + \ldots + r$ (k times)is equal to zero. The characteristic of a domain of integrity is zero or a prime.

2. Division Ring, Field

A ring is called *division ring* or *skew field* if $(R \setminus \{0\},*)$ is a group .

If $(R\setminus\{0\},*)$ is commutative, then R is a *field*. So, every field is a domain of integrity and also a division ring. Reversed, every finite domain of integrity and every finite division ring is a field. This statement is a theorem of Wedderburn.

Examples of rings and fields

■ **A:** The number domains $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are commutative rings with identity with respect to addition and multiplication; $\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are also fields. The set of even integers is an example of a ring without identity.

■ **B:** The set M_n of all square matrices of order n with real (or complex) elements is a non-commutative ring with respect to matrix addition and multiplication. It has a unit element which is the identity matrix. M_n has zero divisors, e.g. for $n = 2$, $\begin{pmatrix}1&0\\1&0\end{pmatrix}\begin{pmatrix}0&0\\1&1\end{pmatrix} = \begin{pmatrix}0&0\\0&0\end{pmatrix}$, i.e. both matrices $\begin{pmatrix}1&0\\1&0\end{pmatrix}$ and $\begin{pmatrix}0&0\\1&1\end{pmatrix}$ are zero divisors in M_2.

■ **C:** The set of real polynomials $p(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$ forms a ring with respect to the usual addition and multiplication of polynomials, the polynomial ring $\mathbb{R}[x]$.

More generally, instead of polynomials over $\mathbb{R}$, polynomial rings over arbitrary commutative rings with identity element can be considered.

■ **D:** Examples of finite rings are the *residue class rings* $\mathbb{Z}_n$ modulo n. $\mathbb{Z}_n$ consists of all the classes $[a]_n$ of integers having the same residue on division by n. ($[a]_n$ is the equivalence class defined by the natural number a with respect to the relation $\sim_R$ introduced in 5.2.4, **1.**, p. 334.) The ring operations $\oplus$, $\odot$ on $\mathbb{Z}_n$ are defined by

$$[a]_n \oplus [b]_n = [a + b]_n \quad \text{and} \quad [a]_n \odot [b]_n = [a \cdot b]_n. \tag{5.174}$$

If the natural number n is a prime, then $(\mathbb{Z}_n\,,\oplus\,,\odot)$ is a field. Otherwise $\mathbb{Z}_n$ has zero divisors, e.g. in $\mathbb{Z}_6$ (numbers modulo 6) $[3]_6 \cdot [2]_6 = [0]_6$. Usually $\mathbb{Z}_n$ is considered as $\mathbb{Z}_n = \{0, 1, \ldots , n - 1\}$, i.e. the residue classes are replaced by representatives(see 5.4.3,**3.**, S. 377).

3. Field Extensions

If K and L are fields and $K \subseteq L$, then L is an *extension field* or an *over-field* of K. In this case L can be considered as a vector space over K.

If L is a finite dimensional space over K, then L is called a *finite extension field.* If this dimension is n, then L is called also an *extension of degree* n of K (Notation: $[L : K] = n$).

E.g. $\mathbb{C}$ is a finite extension of $\mathbb{R}$. $\mathbb{C}$ is two-dimensional over $\mathbb{R}$, and $\{1, i\}$ is a basis. $\mathbb{R}$ is an infinite-dimensional space over $\mathbb{Q}$.

For a set $M \subseteq L$, $K(M)$ denotes the smallest field (an over-field of K) which contains the field K and the set M.

Especially important are the simple algebraic extensions $K(\alpha)$, where $\alpha \in L$ is a root of a polynomial from $K[x]$. The polynomial of lowest degree with a leading coefficient 1 having α as a root is called the *minimal polynomial of* α *over* K. If the degree of the minimal polynomials of $\alpha \in L$ is n, then $K(\alpha)$ is an extension of degree n, i.e. the degree of the minimal polynomials is equal to the dimension of L as a vector space over K.

E.g. $\mathbb{C} = \mathbb{R}(i)$ and $i \in \mathbb{C}$ is the root of the polynomial $x^2 + 1 \in \mathbb{R}[x]$, i.e. $\mathbb{C}$ is a simple algebraic extension and $[\mathbb{C} : \mathbb{R}] = 2$.

A field, which does not have any proper subfield, is called a *prime field.*

Every field K contains a smallest subfield, the prime field of K.

Out of isomorphism, $\mathbb{Q}$ (for fields of characteristic 0) and $\mathbb{Z}_p$ (p prime, for fields of characteristic p) are the single prime fields.

5.3.7.2 Subrings, Ideals

1. Subring

Suppose $R = (R, +, *)$ is a ring and $U \subseteq R$. If U with respect to $+$ and $*$ is also a ring, then $U = (U, +, *)$ is called a *subring* of R.

A non-empty subset U of a ring $(R, +, *)$ forms a subring of R if and only if for all $a, b \in U$ also $a + (-b)$ and $a * b$ are in U (subring criterion).

2. Ideal

A subring I is called an *ideal* if for all $r \in R$ and $a \in I$ also $r * a$ and $a * r$ are in I. These special subrings are the basis for the formation of factor rings (see 5.3.7.3, p. 363).

The *trivial subrings* $\{0\}$ and R are always ideals of R. Fields have only trivial ideals.

3. Principal Ideal

If all the elements of an ideal can be generated by one element according to the subring criterion, then it is called a *principal ideal.* All ideals of $\mathbb{Z}$ are principal ideals. They can be written in the form $m\mathbb{Z} = \{mg|g \in \mathbb{Z}\}$ and they are denoted by (m).

5.3.7.3 Homomorphism, Isomorphism, Homomorphism Theorem

1. Ring Homomorphism and Ring Isomorphism

1. Ring Homomorphism: Let $R_1 = (R_1, +, *)$ and $R_2 = (R_2, \circ_+, \circ_*)$ be two rings. A mapping h: $R_1 \to R_2$ is called a *ring homomorphism* if for all $a, b \in R_1$

$$h(a + b) = h(a) \circ_+ h(b) \quad \text{and} \quad h(a * b) = h(a) \circ_* h(b) \tag{5.175}$$

hold.

2. Kernel: The *kernel* of h is the set of elements of R_1 whose image by h is the neutral element 0 of $(R_2, +)$, and it is denoted by $\ker h$:

$$\ker h = \{a \in R_1 | h(a) = 0\}. \tag{5.176}$$

Here $\ker h$ is an ideal of R_1.

3. Ring Isomorphism: If h is also bijective, then h is called a *ring isomorphism*, and the rings R_1 and R_2 are called isomorphic.

4. Factor Ring: If I is an ideal of a ring $(R, +, *)$, then the sets of co-sets $\{a + I|a \in R\}$ of I in the additive group $(R, +)$ of the ring R (see 5.3.3, **1.**, p. 337) form a ring with respect to the operations

$$(a + I) \circ_+ (b + I) = (a + b) + I \quad \text{and} \quad (a + I) \circ_* (b + I) = (a * b) + I. \tag{5.177}$$

This ring is called the *factor ring* of R by I, and it is denoted by R/I.
The factor ring of $\mathbb{Z}$ by a principal ideal (m) is the residue class ring $Z_m = Z_{/(m)}$ (see examples of rings and fields on p. 361).

2. Homomorphism Theorem for Rings

If the notion of a normal subgroup is replaced by the notion of an ideal in the homomorphism theorem for groups, then the *homomorphism theorem for rings* is obtained: A ring homomorphism h: $R_1 \to R_2$ defines an ideal of R_1, namely $\ker h = \{a \in R_1|h(a) = 0\}$. The factor ring $R_1/\ker h$ is isomorphic to the homomorphic image $h(R_1) = \{h(a)|a \in R_1\}$. Conversely, every ideal I of R_1 defines a homomorphic mapping nat_I: $R_1 \to R_2/I$ with $nat_I(a) = a+I$. This mapping nat_I is called a *natural homomorphism.*

5.3.7.4 Finite Fields and Shift Registers

1. Finite Fields

The following statements give an overview of the structure of finite fields.

1. Galois Field GF For every power of primes p^n there exits a unique field with p^n elements (out of an isomorphism), and every finite field has p^n elements. The fields with p^n elements are denoted by GF(p^n) (Galois field).
Note: For $n > 1$ GF(p^n) and $\mathbb{Z}_{p^n}$ are different.

In constructing finite fields with p^n elements (p is prime, $n > 1$), the ring of polynomials over $\mathbb{Z}_p$ (see 5.3.7,**2.**, p. 361, ■ **C**) and irreducible polynomials are needed: $\mathbb{Z}_p[x]$ consists of all polynomials with coefficients from $\mathbb{Z}_p$. The coefficients are calculated modulo p.

2. Algorithm of Division and Euclidean Algorithm In a ring of polynomials $K[x]$ the division algorithm is applicable (dividing polynomials with a remainder), i.e. for $f(x), g(x) \in K[x]$, degf(x) $\leq$ degg(x) there exist $q(x), r(x) \in K[x]$ such that

$$g(x) = q(x) \cdot f(x) + r(x) \text{ and } \deg r(x) < \deg f(x)\,. \tag{5.178}$$

This relation is denoted by $r(x) = g(x)(\text{mod} f(x))$. Repeatedly performed division with remainders is known as the Euclidean algorithm for rings of polynomials and the last nonzero remainder gives the greatest common divisor of $f(x)$ and $g(x)$.

3. Irreducible Polynomials A polynomial $f(x) \in K[x]$ is *irreducible* if it can not be represented as a product of polynomials of lower degrees, i.e. (analogously to the prime numbers in $\mathbb{Z}$) $f(x)$ is a prime in $K[x]$. E.g. for polynomials of second or third degree irreducibility means, that they do not have roots in K.

It can be shown that there are irreducible polynomials of arbitrary degree in $K[x]$. If $f(x) \in K[x]$ is an irreducible polynomial, then

$$K[x]/f(x) := \{p(x) \in K[x] \mid \deg p(x) < \deg f(x)\} \tag{5.179}$$

is a field, where addition and multiplication are performed modulo $f(x)$, i.e. $g(x) * h(x) = g(x) \cdot h(x) \,(\text{mod}\, f(x))$.

If $K = \mathbb{Z}_p$ and $\deg f(x) = n$, then $K[x]/f(x)$ has p^n elements, i.e. GF$(p^n) = \mathbb{Z}_p[x]/f(x)$, where $f(x)$ is an irreducible polynomial of degree n.

4. Calculation Rule in GF(p^n) In GF(p^n) the following useful rule is valid:

$$(a + b)^{p^r} = a^{p^r} + b^{p^r}, r \in \mathbb{N}\,. \tag{5.180}$$

So, in GF(p^n) =$\mathbb{Z}_p[x]/f(x)$ there is an element $\alpha = x$, a root of the polynomial $f(x)$ irreducible in $\mathbb{Z}_p(x)$, and GF$(p^n) = \mathbb{Z}_p[x]/f(x) = \mathbb{Z}_p(\alpha)$. It can be proven that $\mathbb{Z}_p(\alpha)$ is the splitting field of $f(x)$.
The *splitting field* of a polynomial from $\mathbb{Z}_p[x]$ is the smallest extension field of $\mathbb{Z}_p$ which contains all roots of $f(x)$.

5. Algebraic Closure, Fundamental Theorem of Algebra A field K is *algebraically closed* if all roots of the polynomials from $K[x]$ are in K. The *fundamental theorem of algebra* tells that the field $\mathbb{C}$ of complex numbers is algebraically closed. An algebraic extension L of K is called the *algebraic closure* of K if L is algebraically closed. The algebraic closure of a finite field is not finite. So there are infinite fields with characteristic p.

6. Cyclic and Multiplicative Group The multiplicative group $K^* = K \setminus \{0\}$ of a finite field K is cyclic, i.e. there is an element $a \in K$ such that every element of K^* is a power of a: $K^* = \{1, a, a^2, \ldots, a^{q-2}\}$, if K has q elements.

An irreducible polynomial $f(x) \in K[x]$ is called *primitive*, if the powers of x represents all nonzero elements of $L := K[x]/f(x)$, i.e. the multiplicative group L^* of L can be generated by x.

With a primitive polynomial $f(x)$ of degree n it is possible to construct a ,,Table of logarithm" for $\mathrm{GF}(p^n)$ from $\mathrm{GF}(p)[x]$, which makes calculations easier.

■ Construction of field $\mathrm{GF}(2^3)$ and its table of logarithm.

$f(x) = 1 + x + x^3$ is irreducible over $\mathbb{Z}[x]$, since neither 0 nor 1 are roots of it:

$$\mathrm{GF}(2^3) = \mathbb{Z}_2[x]/f(x) = \{a_0 + a_1 x + a_2 x^2 \mid a_0, a_1, a_2 \in \mathbb{Z}_2 \wedge x^3 = 1 + x\}. \tag{5.181}$$

$f(x)$ is primitive, so a table of logarithm can be created for $\mathrm{GF}(2^3)$:

Two expressions are assigned to every polynomial $a_0 + a_1 x + a_2 x^2$ from $\mathbb{Z}_2[x]/f(x)$. The coefficient vector a_0, a_1, a_2 and the so called logarithm which is a natural number i such that $x^i = a_0 + a_1 x + a_2 x^2$ modulo $1 + x + x^3$. The table of logarithm is:

KE	KV	Log.
1	1 0 0	0
x	0 1 0	1
x^2	0 0 1	2
x^3	1 1 0	3
x^4	0 1 1	4
x^5	1 1 1	5
x^6	1 0 1	6

- Addition of the field elements (KE) in GF(8):
- Addition of the coordinate vectors (KV) componentwise mod 2 (in general mod p).
- Multiplication of the field elements (KE) in GF(8):
- Addition of logarithms (Log) mod 7 (in general mod $(p^n - 1)$).

Example: $\dfrac{x^2 + x^4}{x^3 + x^4} = \dfrac{x}{x^6} = x^{-5} = x^2$

Remark: Finite fields are extremely important in *coding theory* as *linear codes*, where vector spaces in form $(\mathrm{GF}(q))^n$ are considered. A subspace of such a vector space is called *linear code* (see 5.4.6.2,**3.**, p. 385). The elements (code words) of a linear code are also n-tuples with elements from a finite field $\mathrm{GF}(q^n)$. In applications in code theory it is important to know the divisors of $X^n - 1$.

The splitting field of $X^n - 1 \in K[X]$ is called the *n-th cyclotomic field* over K.

If the characteristic of K is not a divisor of n and α is a primitive n-th unit root, then:

a) The extension field $K(\alpha)$ is the splitting field of $X^n - 1$ over K.

b) In $K(\alpha)$, the field $X^n - 1$ has exactly n pairwise different roots which form a cyclic group, and among them there are $\varphi(n)$ primitive n-th unit roots, where $\varphi(n)$ denotes the Euler function (5.4.4,**1.**, p. 381). By the k-th powers ($k < n$, g.c.d.(k,n)=1) of a primitive n-th unit root α all unit roots can be got.

2. Applications of Shift Registers

Calculations with polynomials can be performed well by a *linear feedback shift register* (see **Fig.5.18**). With a linear feedback shift register based on the feedback polynomial $f(x) = f_0 + f_1 x + \cdots + f_{r-1} x^{r-1} + x^r$ and from the state polynomial $s(x) = s_0 + s_1 x + \cdots + s_{r-1} x^{r-1}$ one gets the state polynomial $s(x) \cdot x - s_{r-1} \cdot f(x) = s(x) \cdot x \pmod{f(x)}$.

Especially, if $s(x) = 1$, after i steps (i-times applications) the state polynomial is $x^i \pmod{f(x)}$.

■ Demonstration with the example from page 364: The primitive polynomial $f(x) = 1 + x + x^3 \in \mathbb{Z}_2[x]$ is chosen as feedback polynomial. Then the shift register with lengh 3 has the following sequence of states:

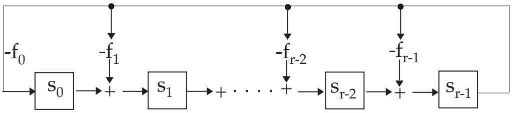


Figure 5.18

$$\begin{array}{rlll}
\text{From the initial state:} & 1\,0\,0 \mathrel{\widehat{=}} 1 & & (\text{mod}\, f(x)) \\
\text{the states follow as:} & 0\,1\,0 \mathrel{\widehat{=}} x & & (\text{mod}\, f(x)) \\
 & 0\,0\,1 \mathrel{\widehat{=}} x^2 & & (\text{mod} f(x)) \\
 & 1\,1\,0 \mathrel{\widehat{=}} x^3 & \equiv 1 + x & (\text{mod}\, f(x)) \\
 & 0\,1\,1 \mathrel{\widehat{=}} x^4 & \equiv x + x^2 & (\text{mod}\, f(x)) \\
 & 1\,1\,1 \mathrel{\widehat{=}} x^5 & \equiv 1 + x + x^2 & (\text{mod}\, f(x)) \\
 & 1\,0\,1 \mathrel{\widehat{=}} x^6 & \equiv 1 + x^2 & (\text{mod}\, f(x)) \\
\hline
 & 1\,0\,0 \mathrel{\widehat{=}} x^7 & \equiv 1 & (\text{mod}\, f(x))
\end{array}$$

The states are considered as coefficient vectors of a state polynomial $s_0 + s_1 x + s_2 x^2$.
In general: A linear feedback shift register with length r gives a sequence of states of maximal length with period $2^r - 1$ if and only if the feedback polynomial is a primitive polynomial of degree r.

5.3.8 Vector Spaces *

5.3.8.1 Definition

A *vector space* over a field F consists of an Abelian group $V = (V, +)$ of “ vectors ” written in additive form, of a field $F = (F, +, *)$ of “ scalars ” and an exterior multiplication $F \times V \to V$, which assigns to every ordered pair (k, v) for $k \in F$ and $v \in V$ a vector $kv \in V$. These operations have the following properties:

(V1) $(u + v) + w = u + (v + w)$ for all $u, v, w \in V$. (5.182)

(V2) There is a vector $0 \in V$ such that $v + 0 = v$ for every $v \in V$. (5.183)

(V3) To every vector v there is a vector $-v$ such that $v + (-v) = 0$. (5.184)

(V4) $v + w = w + v$ for every $v, w \in V$. (5.185)

(V5) $1v = v$ for every $v \in V$, 1 denotes the unit element of F. (5.186)

(V6) $r(sv) = (rs)v$ for every $r, s \in F$ and every $v \in V$. (5.187)

(V7) $(r + s)v = rv + sv$ for every $r, s \in F$ and every $v \in V$. (5.188)

(V8) $r(v + w) = rv + rw$ for every $r \in F$ and every $v, w \in V$. (5.189)

If $F = \mathbb{R}$ holds, then it is called a *real vector space.*

Examples of vector spaces:

■ **A:** Single-column or single-row real matrices of type $(n, 1)$ and $(1, n)$, respectively, with respect to matrix addition and exterior multiplication with real numbers form real vector spaces $\mathbb{R}^n$ (the vector space of column or row vectors; see also 4.1.3, p. 271).

■ **B:** All real matrices of type (m, n) form a real vector space.

■ **C:** All real functions continuous on an interval $[a, b]$ with the operations

$$(f + g)(x) = f(x) + g(x) \quad \text{and} \quad (kf)(x) = k \cdot f(x) \tag{5.190}$$

*In this paragraph, generally, vectors are not printed in bold face.

form a real vector space.
Function spaces have a fundamental role in functional analysis (see Ch. 12, p. 654). For further examples see 12.1.2, p. 655.

5.3.8.2 Linear Dependence

Let V be a vector space over F. The vectors $v_1, v_2, \ldots, v_m \in$ V are called *linearly dependent* if there are $k_1, k_2, \ldots, k_m \in K$ not all of them equal to zero such that $0 = k_1v_1 + k_2v_2 + \cdots + k_mv_m$ holds. Otherwise they are *linearly independent*. Linear dependence of at least two vectors means that one of them is a multiple of the other.

If there is a maximal number n of linearly independent vectors in a vector space V, then the vector space V is called n *dimensional*. This number n is uniquely defined and it is called the *dimension*. Every n linearly independent vectors of V form a *basis*. If such a maximal number does not exist, then the vector space is called *infinite dimensional*. The vector spaces in the above examples are n, $m \cdot n$, and infinite dimensional.
In the vector space $\mathbb{R}^n$, n vectors are independent if and only if the determinant of the matrix, whose columns or rows are these vectors, is not equal to zero.
If $\{v_1, v_2, \ldots, v_n\}$ form a basis of an n-dimensional vector space over F, then every vector $v \in V$ has a *unique* representation $v = k_1v_1 + k_2v_2 + \cdots + k_nv_n$ with $k_1, k_2 \ldots, k_n \in F$.
Every set of linearly independent vectors can be completed into a basis of the vector space.

5.3.8.3 Linear Operators

1. Definition of Linear Operators

Let V and W be two real vector spaces. A mapping $f : V \longrightarrow W$ from V into W is called a *linear mapping* or *linear transformation* or *linear operator* (see also 12.1.5.2, p. 658) from V into W if

$$f(u+v) = fu + fv \quad \text{for all } u, v \in V, \tag{5.191}$$

$$f(\lambda u) = \lambda fu \quad \text{for all } u \in V \text{ and all real } \lambda. \tag{5.192}$$

■ **A:** The mapping $fu := \int\limits_{\alpha}^{\beta} u(t)\,dt$, which transforms the space $\mathcal{C}[\alpha, \beta]$ of continuous real functions into the space of real numbers is linear.

In the special case when $W = \mathbb{R}^1$, as in the previous example, linear transformations are called *linear functionals*.

■ **B:** Let $V = \mathbb{R}^n$ and let W be the space of all real polynomials of degree at most $n-1$. Then the mapping $f(a_0, a_1, \ldots, a_{n-1}) := a_0 + a_1x + a_2x^2 + \cdots + a_{n-1}x^{n-1}$ is linear. In this case each n-element vector corresponds to a polynomial of degree $\le n-1$.

■ **C:** If $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$, then all linear operators f from V into W ($f : \mathbb{R}^n \longrightarrow \mathbb{R}^m$) can be characterized by a real matrix $\mathbf{A} = (a_{ik})$ of type (m, n). The relation $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{y}}$ corresponds to the system of linear equations (4.174a)

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}.$$

2. Sum and Product of Two Linear Operators

Let $f\colon V \longrightarrow W$, $g\colon V \longrightarrow W$ and $h\colon W \longrightarrow U$ be linear operators. Then the

sum $f+g\colon V \longrightarrow W$ is defined as $(f+g)u = fu + gu$ for all $u \in V$ and the (5.193)

product $hf\colon V \longrightarrow U$ is defined as $(hf)u = h(fu)$ for all $u \in V$. (5.194)

Remarks:
1. If f, g and h are linear, then $f+g$ and fh are also linear operators.

2. The product (5.194) of two linear operators represents the consecutive application of these operators f and h.

3. The product of two linear operators is usually non-commutative even if the products exist:

$$hf \neq fh\,. \tag{5.195a}$$

Commutability exists, if

$$hf - fh = 0 \tag{5.195b}$$

holds. In quantum mechanics the left-hand side of this equation $hf - fh$ is called the *commutator*. In the case (5.195a) the operators f and h do not commutate, therefore we have to be very careful about the order.

■ As a particular example of sums and products of linear operators one may think of sums and products of the corresponding real matrices.

5.3.8.4 Subspaces, Dimension Formula

1. Subspace: Let V be a vector space and U a subset of V. If U is also a vector space with respect to the operations of V, then U is called a *subspace* of V.
A non-empty subset U of V is a subspace if and only if for every $u_1, u_2 \in U$ and every $k \in F$ also $u_1 + u_2$ and $k \cdot u_1$ are in U (*subspace criterion*).

2. Kernel, Image: Let V_1, V_2 be vector spaces over F. If f: $V_1 \to V_2$ is a linear mapping, then the linear subspaces *kernel* (notation: ker f) and *image* (notation: im f) are defined in the following way:

$$\ker f = \{v \in V | f(v) = 0\}, \quad \text{im } f = \{f(v) | v \in V\}. \tag{5.196}$$

So, for example, the solution set of a homogeneous linear equation system $\mathbf{A}\underline{\mathbf{x}} = \underline{\mathbf{0}}$ is the kernel of the linear mapping defined by the coefficient matrix $\mathbf{A}$.

3. Dimension: The dimension dim ker f and dim im f are called the *defect* f and *rank* f, respectively. For these dimensions the equality

$$\text{defect } f + \text{rank } f = \dim V, \tag{5.197}$$

is valid and is called the *dimension formula*. In particular, if the defect $f = 0$, i.e., $\ker f = \{0\}$, then the linear mapping f is injective, and conversely. Injective linear mappings are called *regular*.

5.3.8.5 Euclidean Vector Spaces, Euclidean Norm

In order to be able to use notions such as length, angle, orthogonality in abstract vector spaces we introduce *Euclidean vector spaces*.

1. Euclidean Vector Space

Let V be a real vector space. If φ: $V \times V \to \mathbb{R}$ is a mapping with the following properties (instead of $\varphi(v, w)$ one writes $v \cdot w$) for every $u, v, w \in V$ and for every $r \in \mathbb{R}$

(S1) $v \cdot w = w \cdot v,$ (5.198)

(S2) $(u + v) \cdot w = u \cdot w + v \cdot w,$ (5.199)

(S3) $r(v \cdot w) = (rv) \cdot w = v \cdot (rw),$ (5.200)

(S4) $v \cdot v > 0$ if and only if $v \neq 0,$ (5.201)

then φ is called a *scalar product* on V. If there is a scalar product defined on V, then V is called a *Euclidean vector space*.

These properties are used to define a scalar product with similar properties on more general spaces, too (see 12.4.1.1, p. 673).

2. Euclidean Norm

The value

$$\|v\| = \sqrt{v \cdot v} \tag{5.202}$$

denotes the *Euclidean norm* (length) of v. The angle α between v, w from V is defined by the formula

$$\cos\alpha = \frac{v \cdot w}{\|v\| \cdot \|w\|}\,. \tag{5.203}$$

If $v \cdot w = 0$ holds, then v and w are called *orthogonal* to each other.

■ **Orthogonality of Trigonometric Functions:** In the theory of Fourier series (see 7.4.1.1, p. 474), there are functions of the form $\sin kx$ and $\cos kx$. Theses fubctions can be considered as elements of $C[0, 2\pi]$. In the function space $C[a, b]$ the formula

$$f \cdot g = \int_a^b f(x)g(x)\,dx \tag{5.204}$$

defines a scalar product. Since

$$\int_0^{2\pi} \sin kx \cdot \sin lx\,dx = 0 \quad (k \neq l), \tag{5.205}$$

$$\int_0^{2\pi} \cos kx \cdot \cos lx\,dx = 0 \quad (k \neq l), \tag{5.206}$$

$$\int_0^{2\pi} \sin kx \cdot \cos lx\,dx = 0, \tag{5.207}$$

the functions $\sin kx \in \mathbb{C}[0, 2\pi]$ and $\cos lx \in \mathbb{C}[0, 2\pi]$ for every $k, l \in \mathbb{N}$ are pairwise orthogonal to each other. This *orthogonality of trigonometric functions* is used in the calculation of Fourier coefficients in harmonic analysis (see 7.4.1.1, p. 474).

5.3.8.6 Bilinear Mappings, Bilinear Forms

Bilinear mappings can be considered as generalizations of different products between vectors. In that case bilinearity uses the distributivity of the corresponding product with respect to vector addition.

1. Definition

Let U, V, W be vector spaces over the same field K. A mapping $f\colon U \times V \longrightarrow W$ is called *bilinear* if

for every $u \in U$ the mapping $v \mapsto f(u, v)$ is a linear mapping of V into W and
for every $v \in V$ the mapping $u \mapsto f(u, v)$ is a linear mapping of U into W. (5.208)

It means that a mapping $f\colon U \times V \longrightarrow W$ is bilinear, if for every $k \in K$, $u, u' \in U$, and $v, v' \in V$ holds:

$$\begin{aligned} f(u+u', v) &= f(u, v) + f(u', v)\,,\ \ f(ku, v) = kf(u, v) \ \text{ and} \\ f(u, v+v') &= f(u, v) + f(u, v')\,,\ \ f(u, kv) = kf(u, v)\,. \end{aligned} \tag{5.209}$$

If f is replaced by the dot product or vector product or by a multiplication in a field, these relations describe the left sided and right sided distributivity of this multiplication with respect to vector addition.

Especially, if $U = V$, and $W = K$ which is the underlying field, then f is called a *bilinear form.* In this book only the real ($K = \mathbb{R}$) or complex ($K = \mathbb{C}$) cases are considered.

Examles of Bilinearforms

■ **A:** $U = V = \mathbb{R}^n$, $W = \mathbb{R}$, f is the dot product in $\mathbb{R}^n$: $f(u, v) = u^{\mathrm{T}}v = \sum_{i=1}^n u_i v_i$, where u_i and v_i $(i = 1, 2, \ldots, n)$ denote the Cartesian coordinates of u and v.

■ **B:** $U = V = W = \mathbb{R}^3$, f is the cross product in $\mathbb{R}^3$:
$f(u, v) = u \times v = (u_2v_3 - v_2u_3\,,\ v_1u_3 - u_1v_3\,,\ u_1v_2 - v_1u_2)^{\mathrm{T}}$.

2. Special Bilinear Forms

A bilinear form $f\colon V \times V \longrightarrow \mathbb{R}$ is called

- symmetric, if $f(v, v') = f(v', v)$ for every $v, v' \in V$,
- skew-symmetric, if $f(v, v') = -f(v', v)$ for every $v, v' \in V$ and
- positive definite, if $f(v, v) > 0$ for every $v \in V \ \ v \neq 0$.

So an Euclidean dot product in V (see 5.3.8.5, p. 367) can be characterized as a symmetric, positive definite bilinear form. The canonical Euclidean dot product in $\mathbb{R}^n$ is defined as $f(u, v) = u^{\mathrm{T}}v$.

In finite dimensional spaces V a bilinear form can be represented by a matrix: If $f := V \times V \longrightarrow \mathbb{R}$ is a bilinear form, and $B = (b_1, b_2, \ldots, b_n)$ is a basis of V, then the matrix

$$\mathbf{A}_B(f) = (f(b_i, b_j)_{i,j}) \tag{5.210}$$

is the *representation matrix* of f with respect to basis B. The bilinear form then can be written in matrix product form:

$$f(v, v') = v^{\mathrm{T}} \mathbf{A}_B(f) v' \,, \tag{5.211}$$

where v and v' are given with respect to basis B.

The representation matrix is symmetric, if the bilinear form is symmetric. In complex vector spaces (because z^2 can be a negative number) symmetric, positive definite bilinear forms do not have much sense. To define an unitary dot product and also distances and angles with it instead of bilinear form the concept of the so called sesquilinear form is used [5.6], [5.12].

3. Sesquilinear Form

A mapping $f\colon V \times V \longrightarrow \mathbb{C}$ is called *sesquilinear form* if for every $v, v' \in V$ and $k \in \mathbb{C}$:

$$f(u + u', v) = f(u, v) + f(u', v) \,, \quad f(ku, v) = kf(u, v) \ \text{ and}$$
$$f(u, v + v') = f(u, v) + f(u, v') \,, \quad f(u, kv) = k^* f(u, v) \,. \tag{5.212}$$

where k^* denotes the complex conjugate of k. The function is linear in the first argument and ,,semi-linear" in the second argument. Analogously to the real case ,,symmetry" is defined in the following way:

A sesquilinear form $f\colon V \times V \longrightarrow \mathbb{C}$ is called *hermitian* if $f(v, v') = f(v', v)^*$ for every $v, v' \in V$.

In this way a (unitary) dot product is characterized by an hermitian, positive definite sesquilinear form. The canonical unitary dot product in $\mathbb{C}^n$ is defined as $f(u, v) = u^{\mathrm{T}} v^*$.

If V is finite dimensional, then a sesquilinear form can be represented by a matrix (like in the real case):

If $f : V \times V \longrightarrow \mathbb{C}$ is a sesquilinear form, and $B = (b_1, b_2, \ldots, b_n)$ is a basis of V, then the matrix $\mathbf{A}_B(f) = (f(b_i, b_j))_{i,j}$ is the *representation matrix* of f with respect to basis B. The sesquilinear form can be written in matrix product form:

$$f(v, v') = v^{\mathrm{T}} \mathbf{A}_B(f) v' \,, \tag{5.213}$$

where v and v' are given with respect to basis B. A representation matrix is hermitian if and only if the sesquilinear form is hermitian.

5.4 Elementary Number Theory

Elementary number theory investigates divisibility properties of integers.

5.4.1 Divisibility

5.4.1.1 Divisibility and Elementary Divisibility Rules

1. Divisor

An integer $b \in \mathbb{Z}$ is *divisible* by an integer a without remainder iff [*] there is an integer q such that

$$qa = b \tag{5.214}$$

holds. Here a is a divisor of b in $\mathbb{Z}$, and q is the *complementary divisor* with respect to a; b is a *multiple* of a. For "a divides b" we write also $a|b$. For "a does not divide b" we can write $a\not|b$. The divisibility relation (5.214) is a binary relation in $\mathbb{Z}$ (see 5.2.3, **2.**, p. 331). Analogously, divisibility is defined in the set of natural numbers.

2. Elementary Divisibility Rules

(DR1) For every $a \in \mathbb{Z}$ we have $1|a,\ a|a$ and $a|0$. (5.215)

(DR2) If $a|b,$ then $(-a)|b$ and $a|(-b)$. (5.216)

(DR3) $a|b$ and $b|a$ implies $a = b$ or $a = -b$. (5.217)

(DR4) $a|1$ implies $a = 1$ or $a = -1$. (5.218)

(DR5) $a|b$ and $b \neq 0$ imply $|a| \le |b|$. (5.219)

(DR6) $a|b$ implies $a|zb$ for every $z \in \mathbb{Z}$. (5.220)

(DR7) $a|b$ implies $az|bz$ for every $z \in \mathbb{Z}$. (5.221)

(DR8) $az|bz$ and $z \neq 0$ implies $a|b$ for every $z \in \mathbb{Z}$. (5.222)

(DR9) $a|b$ and $b|c$ imply $a|c$. (5.223)

(DR10) $a|b$ and $c|d$ imply $ac|bd$. (5.224)

(DR11) $a|b$ and $a|c$ imply $a|(z_1b + z_2c)$ for arbitrary $z_1, z_2 \in \mathbb{Z}$. (5.225)

(DR12) $a|b$ and $a|(b + c)$ imply $a|c$. (5.226)

5.4.1.2 Prime Numbers

1. Definition and Properties of Prime Numbers

A positive integer $p\ (p > 1)$ is called a *prime number* iff 1 and p are its only divisors in the set $\mathbb{N}$ of positive integers. Positive integers which are not prime numbers are called *composite numbers*.

For every integer, the smallest positive divisor different from 1 is a prime number. There are infinitely many prime numbers.

A positive integer $p\ (p > 1)$ is a prime number iff for arbitrary positive integers $a, b,\ p|(ab)$ implies $p|a$ or $p|b$.

2. Sieve of Eratosthenes

By the method of the "*Sieve of Eratosthenes*", every prime number smaller than a given positive integer n can be determined:

a) Write down the list of all positive integers from 2 to n.

b) Underline 2 and delete every subsequent multiple of 2.

c) If p is the first non-deleted and non-underlined number, then underline p and delete every p-th number (beginning with $2p$ and counting the numbers of the original list).

d) Repeat step c) for every $p\ (p \le \sqrt{n})$ and stop the algorithm.

[*] if and only if

Every underlined and non-deleted number is a prime number. In this way, all prime numbers $\leq n$ are obtained.
The prime numbers are called *prime elements* of the set of integers.

3. Prime Pairs

Prime numbers with a difference of 2 form *prime pairs* (twin primes).

■ $(3,5),(5,7),(11,13),(17,19),(29,31),(41,43),(59,61),(71,73),(101,103)$ are prime pairs.

4. Prime Triplets

Prime triplets consist of three prime numbers occuring among four consecutive odd numbers.

■ $(5,7,11),(7,11,13),(11,13,17),(13,17,19),(17,19,23),(37,41,43)$ are prime triplets.

5. Prime Quadruplets

If the first two and the last two of five consecutive odd numbers are prime pairs, then they are called a *prime quadruplet*.

■ $(5,7,11,13),(11,13,17,19),(101,103,107,109),(191,193,197,199)$ are prime quadruplets.

The conjecture that there exist infinitely many prime pairs, prime triplets, and prime quadruplets, is not proved still.

6. Mersenne Primes

If $2^k-1,\ \ k\in\mathbb{N}$, is a prime number, then k is also a prime number. The numbers 2^p-1 (p prime) are called Mersenne numbers. A Mersenne prime is a Mersenne number 2^p-1 which is itself a prime number.

■ 2^p-1 is a prime number for the first ten values of p: 2, 3, 5, 7, 13,17, 19, 31, 61, 89, 107, etc.

Remark: Since a few years the largest known prime is always a Mersenne prime, e.g. $2^{43112609}-1$ in 2008, $2^{57885161}-1$ in 2013. In contrary to other natural numbers the numbers of the form 2^k-1 can be tested in a relatively simple way whether they are primes: Let $p>2$ be a prime and a sequence of natural numbers is defined by $s_1=4$, $s_{i+1}:=s_i^2-2$ $(i\geq 1)$. The number 2^p-1 is a prime if and only if the term of the sequence s_{p-1} is divisible by 2^p-1.
The prime test based on this statement is called Lucas-Lehmer test .

7. Fermat Primes

If a number $2^k+1,\ \ k\in\mathbb{N}$, is an odd prime number, then k is a power of 2. The numbers $2^k+1,\ \ k\in\mathbb{N}$, are called *Fermat numbers.* If a Fermat number is a prime number, then it is called a *Fermat prime.*

■ For $k=0,1,2,3,4$ the corresponding Fermat numbers $3,5,17,257,65537$ are prime numbers. It is conjectured that there are no further Fermat primes.

8. Fundamental Theorem of Elementary Number Theory

Every positive integer $n>1$ can be represented as a product of primes. This representation is unique except for the order of the factors. Therefore n is called to have exactly one *prime factorization.*

■ $360=2\cdot 2\cdot 2\cdot 3\cdot 3\cdot 5=2^3\cdot 3^2\cdot 5$.

Remark: Analogously, the integers (except $-1,0,1$) can be represented as products of prime elements, unique apart from the order and the sign of the factors.

9. Canonical Prime Factorization

It is usual to arrange the factors of the prime factorization of a positive integer according to their size, and to combine equal factors to powers. If every non-occurring prime is assigned exponent 0, then every positive integer is uniquely determined by the sequence of the exponents of its prime factorization.

■ To $1\,533\,312=2^7\cdot 3^2\cdot 11^3$ belongs the sequence of exponents $(7,2,0,0,3,0,0,\ldots)$.

For a positive integer n, let $p_1,p_2,\ldots p_m$ be the pairwise distinct primes divisors of n, and let α_k denote the exponent of a prime number p_k in the prime factorization of n. Then

$$n=\prod_{k=1}^{m}p_k^{\alpha_k}, \tag{5.227a}$$

and this representation is called the *canonical prime factorization* of n. It is often denoted by

$$n = \prod_p p^{\nu_p(n)}, \tag{5.227b}$$

where the product applies to all prime numbers p, and where $\nu_p(n)$ is the multiplicity of p as a divisor of n. It always means a finite product because only finitely many of the exponents $\nu_p(n)$ differ from 0.

10. Positive Divisors

If a positive integer $n \geq 1$ is given by its canonical prime factorization (5.227a), then every positive divisor t of n can be written in the form

$$t = \prod_{k=1}^{m} p_k^{\tau_k} \quad \text{with } \tau_k \in \{0, 1, 2, \ldots, \alpha_k\} \text{ for } k = 1, 2, \ldots, m. \tag{5.228a}$$

The number $\tau(n)$ of all positive divisors of n is

$$\tau(n) = \prod_{k=1}^{m} (\alpha_k + 1). \tag{5.228b}$$

■ **A:** $\tau(5040) = \tau(2^4 \cdot 3^2 \cdot 5 \cdot 7) = (4+1)(2+1)(1+1)(1+1) = 60.$

■ **B:** $\tau(p_1 p_2 \cdots p_r) = 2^r$, if $p_1, p_2, \ldots, p_r$ are pairwise distinct prime numbers.

The product $P(n)$ of all positive divisors of n is given by

$$P(n) = n^{\frac{1}{2}\tau(n)}. \tag{5.228c}$$

■ **A:** $P(20) = 20^3 = 8000.$ ■ **B:** $P(p^3) = p^6$, if p is a prime number.

■ **C:** $P(pq) = p^2 q^2$, if p and q are different prime numbers.

The sum $\sigma(n)$ of all positive divisors of n is

$$\sigma(n) = \prod_{k=1}^{m} \frac{p_k^{\alpha_k+1} - 1}{p_k - 1}. \tag{5.228d}$$

■ **A:** $\sigma(120) = \sigma(2^3 \cdot 3 \cdot 5) = 15 \cdot 4 \cdot 6 = 360.$ ■ **B:** $\sigma(p) = p + 1$, if p is a prime number.

5.4.1.3 Criteria for Divisibility

1. Notation

Consider a positive integer given in decimal form:

$$n = (a_k a_{k-1} \cdots a_2 a_1 a_0)_{10} = a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_2 10^2 + a_1 10 + a_0. \tag{5.229a}$$

Then

$$Q_1(n) = a_0 + a_1 + a_2 + \cdots + a_k \tag{5.229b}$$

and

$$Q_1'(n) = a_0 - a_1 + a_2 - + \cdots + (-1)^k a_k \tag{5.229c}$$

are called the *sum of the digits* (*of first order*) and the *alternating sum of the digits* (*of first order*) of n, respectively. Furthermore,

$$Q_2(n) = (a_1 a_0)_{10} + (a_3 a_2)_{10} + (a_5 a_4)_{10} + \cdots \qquad \text{and} \tag{5.229d}$$

$$Q_2'(n) = (a_1 a_0)_{10} - (a_3 a_2)_{10} + (a_5 a_4)_{10} - + \cdots \tag{5.229e}$$

are called the *sum of the digits and the alternating sum of the digits, respectively, of second order* and

$$Q_3(n) = (a_2 a_1 a_0)_{10} + (a_5 a_4 a_3)_{10} + (a_8 a_7 a_6)_{10} + \cdots \tag{5.229f}$$

and

$$Q_3'(n) = (a_2 a_1 a_0)_{10} - (a_5 a_4 a_3)_{10} + (a_8 a_7 a_6)_{10} - + \cdots \tag{5.229g}$$

are called the *sum of the digits and alternating sum of the digits, respectively, of third order* .

■ The number 123 456 789 has the following sum of the digits: $Q_1 = 9+8+7+6+5+4+3+2+1 = 45$, $Q_1' = 9-8+7-6+5-4+3-2+1 = 5$, $Q_2 = 89+67+45+23+1 = 225$, $Q_2' = 89-67+45-23+1 = 45$, $Q_3 = 789 + 456 + 123 = 1368$ and $Q_3' = 789 - 456 + 123 = 456$.

2. Criteria for Divisibility

There are the following criteria for divisibility:

DC-1: $3|n \Leftrightarrow 3|Q_1(n),$ (5.230a) **DC-2:** $7|n \Leftrightarrow 7|Q_3'(n),$ (5.230b)

DC-3: $9|n \Leftrightarrow 9|Q_1(n),$ (5.230c) **DC-4:** $11|n \Leftrightarrow 11|Q_1'(n),$ (5.230d)

DC-5: $13|n \Leftrightarrow 13|Q_3'(n)$ (5.230e) **DC-6:** $37|n \Leftrightarrow 37|Q_3(n),$ (5.230f)

DC-7: $101|n \Leftrightarrow 101|Q_2'(n),$ (5.230g) **DC-8:** $2|n \Leftrightarrow 2|a_0,$ (5.230h)

DC-9: $5|n \Leftrightarrow 5|a_0,$ (5.230i) **DC-10:** $2^k|n \Leftrightarrow 2^k|(a_{k-1}a_{k-2}\cdots a_1a_0)_{10},$ (5.230j)

DC-11: $5^k|n \Leftrightarrow 5^k|(a_{k-1}a_{k-2}\cdots a_1a_0)_{10}.$ (5.230k)

■ **A:** $a = 123\,456\,789$ is divisible by 9 since $Q_1(a) = 45$ and $9|45$, but it is not divisible by 7 since $Q_3'(a) = 456$ and $7\not|456$.

■ **B:** 91 619 is divisible by 11 since $Q_1'(91\,619) = 22$ and $11|22$.

■ **C:** 99 994 096 is divisible by 2^4 since $2^4|4\,096$.

5.4.1.4 Greatest Common Divisor and Least Common Multiple

1. Greatest Common Divisor

For integers $a_1, a_2, \ldots, a_n$, which are not all equal to zero, the largest number in the set of common divisors of $a_1, a_2, \ldots, a_n$ is called the *greatest common divisor* of $a_1, a_2, \ldots, a_n$, and it is denoted by $\gcd(a_1, a_2, \ldots, a_n)$. If $\gcd(a_1, a_2, \ldots, a_n) = 1$, then the numbers $a_1, a_2, \ldots, a_n$ are called *coprimes.*

To determine the greatest common divisor, it is sufficient to consider the positive common divisors. If the canonical prime factorizations

$$a_i = \prod_p p^{\nu_p(a_i)} \tag{5.231a}$$

of $a_1, a_2, \ldots, a_n$ are given, then

$$\gcd(a_1, a_2, \ldots, a_n) = \prod_p p^{\left\{\min\limits_i\, [\nu_p(a_i)]\right\}}. \tag{5.231b}$$

■ For the numbers $a_1 = 15\,400 = 2^3 \cdot 5^2 \cdot 7 \cdot 11$, $a_2 = 7\,875 = 3^2 \cdot 5^3 \cdot 7$, $a_3 = 3\,850 = 2 \cdot 5^2 \cdot 7 \cdot 11$, the greatest common divisor is $\gcd(a_1, a_2, a_3) = 5^2 \cdot 7 = 175$.

2. Euclidean Algorithm

The greatest common divisor of two integers a, b can be determined by the *Euclidean algorithm* without using their prime factorization. To do this, a sequence of divisions with remainder, according to the following scheme, is performed. For $a > b$ let $a_0 = a, a_1 = b$. Then:

$$\begin{aligned}
a_0 &= q_1a_1 + a_2, && 0 < a_2 < a_1,\\
a_1 &= q_2a_2 + a_3, && 0 < a_3 < a_2,\\
&\vdots \quad \vdots \quad \vdots && \\
a_{n-2} &= q_{n-1}a_{n-1} + a_n, && 0 < a_n < a_{n-1},\\
a_{n-1} &= q_na_n.
\end{aligned} \tag{5.232a}$$

The division algorithm stops after a finite number of steps, since the sequence $a_2, a_3, \ldots$ is a strictly monotone decreasing sequence of positive integers. The last remainder a_n, different from 0 is the greatest common divisor of a_0 and a_1.

■ gcd(38, 105) = 1, as can be seen by the help of the table to the right.

$$\begin{aligned} 105 &= 2 \cdot 38 + 29 \\ 38 &= 1 \cdot 29 + 9 \\ 29 &= 3 \cdot 9 + 2 \\ 9 &= 4 \cdot 2 + 1 \\ 2 &= 2 \cdot 1 \end{aligned}$$

By the recursion formula

$$\gcd(a_1, a_2, \ldots, a_n) = \gcd(\gcd(a_1, a_2, \ldots, a_{n-1}), a_n), \qquad (5.232b)$$

the greatest common divisor of n positive integers with $n > 2$ can be determined by repeated use of the Euclidean algorithm.

■ $\gcd(150, 105, 56) = \gcd(gcd(150, 105), 56) = \gcd(15, 56) = 1.$

■ The Euclidean algorithm to determine the gcd (see also 1.1.1.4, **1.**, p. 3) of two numbers has especially many steps, if the numbers are adjacent numbers in the sequence of Fibonacci numbers (see 5.4.1.5, p. 375). The annexed calculation shows an example where all quotients are always equal to 1.

$$\begin{aligned} 55 &= 1 \cdot 34 + 21 \\ 34 &= 1 \cdot 21 + 13 \\ 21 &= 1 \cdot 13 + 8 \\ 13 &= 1 \cdot 8 + 5 \\ 8 &= 1 \cdot 5 + 3 \\ 5 &= 1 \cdot 3 + 2 \\ 3 &= 1 \cdot 2 + 1 \\ 2 &= 1 \cdot 1 + 1 \\ 1 &= 1 \cdot 1. \end{aligned}$$

3. Theorem for the Euclidean Algorithm

For two natural numbers a, b with $a > b > 0$, let $\lambda(a, b)$ denote the number of divisions with remainder in the Euclidean algorithm, and let $\kappa(b)$ denote the number of digits of b in the decimal system. Then

$$\lambda(a, b) \leq 5 \cdot \kappa(b). \qquad (5.233)$$

4. Greatest Common Divisor as a Linear Combination

It follows from the Euclidean algorithm that

$$\begin{aligned} a_2 &= a_0 - q_1 a_1 = c_0 a_0 + d_0 a_1, \\ a_3 &= a_1 - q_2 a_2 = c_1 a_0 + d_1 a_1, \\ &\vdots \quad \vdots \\ a_n &= a_{n-2} - q_{n-1} a_{n-1} = c_{n-2} a_0 + d_{n-2} a_1. \end{aligned} \qquad (5.234a)$$

Here c_{n-2} and d_{n-2} are integers. Thus the $\gcd(a_0, a_1)$ can be represented as a linear combination of a_0 and a_1 with integer coefficients:

$$\gcd(a_0, a_1) = c_{n-2} a_0 + d_{n-2} a_1. \qquad (5.234b)$$

Moreover $\gcd(a_1, a_2, \ldots, a_n)$ can be represented as a linear combination of $a_1, a_2, \ldots, a_n$, since:

$$\gcd(a_1, a_2, \ldots, a_n) = \gcd(\gcd(a_1, a_2, \ldots, a_{n-1}), a_n) = c \cdot \gcd(a_1, a_2, \ldots, a_{n-1}) + d a_n. \qquad (5.234c)$$

■ $\gcd(150, 105, 56) = \gcd(\gcd(150, 105), 56) = \gcd(15, 56) = 1$ with $15 = (-2) \cdot 150 + 3 \cdot 105$ and $1 = 15 \cdot 15 + (-4) \cdot 56)$, thus $\gcd(150, 105, 56) = (-30) \cdot 150 + 45 \cdot 105 + (-4) \cdot 56$.

5. Least Common Multiple

For integers $a_1, a_2, \ldots, a_n$, among which there is no zero, the smallest number in the set of positive common multiples of $a_1, a_2, \ldots, a_n$ is called the *least common multiple* of $a_1, a_2, \ldots, a_n$, and it is denoted by $\operatorname{lcm}(a_1, a_2, \ldots, a_n)$.

If the canonical prime factorizations (5.231a) of $a_1, a_2, \ldots, a_n$ are given, then:

$$\operatorname{lcm}(a_1, a_2, \ldots, a_n) = \prod_p p^{\left\{ \max\limits_i \, [\nu_p(a_i)] \right\}}. \qquad (5.235)$$

■ For the numbers $a_1 = 15\,400 = 2^3 \cdot 5^2 \cdot 7 \cdot 11$, $a_2 = 7\,875 = 3^2 \cdot 5^3 \cdot 7$, $a_3 = 3\,850 = 2 \cdot 5^2 \cdot 7 \cdot 11$ the least common multiple is $\operatorname{lcm}(a_1, a_2, a_3) = 2^3 \cdot 3^2 \cdot 5^3 \cdot 7 \cdot 11 = 693\,000$.

6. Relation between gcd and lcm

For arbitrary integers a, b:

$$|ab| = \gcd(a,b) \cdot \operatorname{lcm}(a,b). \tag{5.236}$$

Therefore, the $\operatorname{lcm}(a,b)$ can be determined with the help of the Euclidean algorithm without using the prime factorizations of a and b.

5.4.1.5 Fibonacci Numbers

1. Recursion Formula

The sequence

$$(F_n)_{n\in\mathbb{N}} \text{ with } F_1 = F_2 = 1 \text{ and } F_{n+2} = F_n + F_{n+1} \tag{5.237}$$

is called *Fibonacci sequence*. It starts with the elements $1, 1, 2, 3, 5, 8, 13, 21, 34, 55, 89, 144, 233, 377, \ldots$

■ The consideration of this sequence goes back to the question posed by Fibonacci in 1202: How many pairs of descendants has a pair of rabbits at the end of a year, if every pair in every month produces a new pair, which beginning with the second month itself produces new descended pairs? The answer is $F_{14} = 377$.

2. Explicit Formula

Besides the recursive definition (5.237) there is an explicit formula for the Fibonacci numbers:

$$F_n = \frac{1}{\sqrt{5}}\left(\left[\frac{1+\sqrt{5}}{2}\right]^n - \left[\frac{1-\sqrt{5}}{2}\right]^n\right). \tag{5.238}$$

Some important properties of Fibonacci numbers are the followings. For $m, n \in \mathbb{N}$:

(1) $F_{m+n} = F_{m-1}F_n + F_mF_{n+1} \quad (m > 1).$ (5.239a)

(2) $F_m | F_{mn}.$ (5.239b)

(3) $\gcd(m,n) = d$ implies $\gcd(F_m, F_n) = F_d.$ (5.239c)

(4) $\gcd(F_n, F_{n+1}) = 1.$ (5.239d)

(5) $F_m | F_k$ holds iff $m|k$ holds. (5.239e)

(6) $\sum\limits_{i=1}^{n} F_i^2 = F_nF_{n+1}.$ (5.239f)

(7) $\gcd(m,n) = 1$ implies $F_mF_n | F_{mn}.$ (5.239g)

(8) $\sum\limits_{i=1}^{n} F_i = F_{n+2} - 1.$ (5.239h)

(9) $F_nF_{n+2} - F_{n+1}^2 = (-1)^{n+1}.$ (5.239i)

(10) $F_n^2 + F_{n+1}^2 = F_{2n+1}.$ (5.239j)

(11) $F_{n+2}^2 - F_n^2 = F_{2n+2}.$ (5.239k)

5.4.2 Linear Diophantine Equations

1. Diophantine Equations

An equation $f(x_1, x_2, \ldots, x_n) = b$ is called a *Diophantine equation* in n unknowns iff $f(x_1, x_2, \ldots, x_n)$ is a polynomial in $x_1, x_2, \ldots, x_n$ with coefficients in the set $\mathbb{Z}$ of integers, b is an integer constant and only integer solutions are of interest. The name “Diophantine” reminds of the Greek mathematician Diophantus, who lived around 250 AD.

In practice, Diophantine equations occur for instance, if relations between quantities are described. Until now, only general solutions of Diophantine equations of at most second degree with two variables are known. Solutions of Diophantine equations of higher degrees are only known in special cases.

2. Linear Diophantine Equations in n Unknowns

A *linear Diophantine equation* in n unknowns is an equation of the form

$$a_1x_1 + a_2x_2 + \cdots a_nx_n = b \quad (a_i \in \mathrm{Z},\ b \in \mathrm{Z}), \tag{5.240}$$

where only integer solutions are searched for. A solution method is described in the following.

3. Conditions of Solvability

If not all the coefficients a_i are equal to zero, then the Diophantine equation (5.240) is solvable iff $\gcd(a_1, a_2, \ldots, a_n)$ is a divisor of b.

■ $114x + 315y = 3$ is solvable, since $\gcd(114, 315) = 3$.

If a linear Diophantine equation in n unknowns ($n > 1$) has a solution and Z is the domain of variables, then the equation has infinitely many solutions. Then in the set of solutions there are $n-1$ free variables. For subsets of Z, this statement is not true.

4. Solution Method for $n = 2$

Let

$$a_1x_1 + a_2x_2 = b \quad (a_1, a_2) \neq (0, 0) \tag{5.241a}$$

be a solvable Diophantine equation, i.e., $\gcd(a_1, a_2)|b$. To find a special solution of the equation, the equation is divided by $\gcd(a_1, a_2)$ and one obtains $a_1'x_1' + a_2'x_2' = b'$ with $\gcd(a_1', a_2') = 1$.
As described in 5.4.1, **4.**, p. 374, $\gcd(a_1', a_2')$ is determined to obtain finally a linear combination of a_1' and a_2': $a_1'c_1' + a_2'c_2' = 1$.
Substitution in the given equation demonstrates that the ordered pair $(c_1'b', c_2'b')$ of integers is a solution of the given Diophantine equation.

■ $114x+315y = 6$. The equation is divided by 3, since $3 = \gcd(114, 315)$. That implies $38x+105y = 2$ and $38 \cdot 47 + 105 \cdot (-17) = 1$ (see 5.4.1, **4.**, p. 374). The ordered pair $(47 \cdot 2, (-17) \cdot 2) = (94, -34)$ is a special solution of the equation $114x + 315y = 6$.

The family of solutions of (5.241a) can be obtained as follows: If (x_1^0, x_2^0) is an arbitrary special solution, which could also be obtained by trial and error, then

$$\{(x_1^0 + t \cdot a_2', x_2^0 - t \cdot a_1')|t \in \mathrm{Z}\} \tag{5.241b}$$

is the set of all solutions.

■ The set of solutions of the equation $114x + 315y = 6$ is $\{(94 + 315t, -34 - 114t)|t \in \mathrm{Z}\}$.

5. Reduction Method for $n > 2$

Suppose a solvable Diophantine equation

$$a_1x_1 + a_2x_2 + \cdots + a_nx_n = b \tag{5.242a}$$

with $(a_1, a_2, \ldots, a_n) \neq (0, 0, \ldots, 0)$ and $\gcd(a_1, a_2, \ldots, a_n) = 1$ is given. If $\gcd(a_1, a_2, \ldots, a_n) \neq 1$, then the equation should be divided by $\gcd(a_1, a_2, \ldots, a_n)$. After the transformation

$$a_1x_1 + a_2x_2 + \cdots + a_{n-1}x_{n-1} = b - a_nx_n \tag{5.242b}$$

x_n is considered as an integer constant and a linear Diophantine equation in $n-1$ unknowns is obtained, and it is solvable iff $\gcd(a_1, a_2, \ldots, a_{n-1})$ is a divisor of $b - a_nx_n$.
The condition

$$\gcd(a_1, a_2, \ldots, a_{n-1})|b - a_nx_n \tag{5.242c}$$

is satisfied iff there are integers $\underline{c}, \underline{c}_n$ such that:

$$\gcd(a_1, a_2, \ldots, a_{n-1}) \cdot \underline{c} + a_n\underline{c}_n = b. \tag{5.242d}$$

This is a linear Diophantine equation in two unknowns, and it can be solved as shown in 5.4.2,**4.**, p. 376. If its solution is determined, then it remains to solve a Diophantine equation in only $n - 1$ unknowns. This procedure can be continued until a Diophantine equation in two unknowns is obtained, which can be solved with the method given in 5.4.2, **4.**, p. 376.
Finally, the solution of the given equation is constructed from the set of solutions obtained in this way.

■ Solve the Diophantine equation

$$2x + 4y + 3z = 3. \tag{5.243a}$$

This is solvable since $\gcd(2,4,3)$ is a divisor of 3.
The Diophantine equation

$$2x + 4y = 3 - 3z \tag{5.243b}$$

in the unknowns x, y is solvable iff $\gcd(2,4)$ is a divisor of $3 - 3z$. The corresponding Diophantine equation $2z' + 3z = 3$ has the set of solutions $\{(-3+3t, 3-2t)|t \in \mathbb{Z}\}$. This implies, $z = 3 - 2t$, and now the set of solutions of the solvable Diophantine equation $2x + 4y = 3 - 3(3 - 2t)$ or

$$x + 2y = -3 + 3t \tag{5.243c}$$

is sought for every $t \in \mathbb{Z}$.
The equation (5.243c) is solvable since $\gcd(1,2) = 1|(-3+3t)$. Now $1 \cdot (-1) + 2 \cdot 1 = 1$ and $1 \cdot (3 - 3t) + 2 \cdot (-3+3t) = -3+3t$. The set of solution is $\{((3-3t)+2s, (-3+3t)-s)|s \in \mathbb{Z}\}$. That implies $x = (3-3t)+2s, y = (-3+3t)-s$, and $\{(3-3t+2s, -3+3t-s, 3-2t)|s,t \in \mathbb{Z}\}$ so obtained is the set of solutions of (5.243a).

5.4.3 Congruences and Residue Classes

1. Congruences

Let m be a positive integer m, $m > 1$. If two integers a and b have the same remainder, when divided by m, then a and b are called *congruent modulo* m, denoted by $a \equiv b \bmod m$ or $a \equiv b(m)$.

■ $3 \equiv 13 \bmod 5, \quad 38 \equiv 13 \bmod 5, \quad 3 \equiv -2 \bmod 5.$

Remark: Obviously, $a \equiv b \bmod m$ holds iff m is a divisor of the difference $a - b$. Congruence modulo m is an equivalence relation (see 5.2.4, **1.**, p. 334) in the set of integers. Note the following properties:

$$a \equiv a \bmod m \text{ for every } a \in \mathbb{Z}, \tag{5.244a}$$

$$a \equiv b \bmod m \Rightarrow \ b \equiv a \bmod m, \tag{5.244b}$$

$$a \equiv b \bmod m \wedge b \equiv c \bmod m \ \Rightarrow \ a \equiv c \bmod m. \tag{5.244c}$$

2. Calculating Rules

$$a \equiv b \bmod m \wedge c \equiv d \bmod m \Rightarrow a + c \equiv b + d \bmod m, \tag{5.245a}$$

$$a \equiv b \bmod m \wedge c \equiv d \bmod m \Rightarrow a \cdot c \equiv b \cdot d \bmod m, \tag{5.245b}$$

$$a \cdot c \equiv b \cdot c \bmod m \wedge \ \gcd(c,m) = 1 \ \Rightarrow \ a \equiv b \bmod m, \tag{5.245c}$$

$$a \cdot c \equiv b \cdot c \bmod m \wedge \ c \neq 0 \Rightarrow \ a \equiv b \bmod \frac{m}{\gcd(c,m)}. \tag{5.245d}$$

3. Residue Classes, Residue Class Ring

Since congruence modulo m is an equivalence relation in $\mathbb{Z}$, this relation induces a partition of $\mathbb{Z}$ into *residue classes modulo m*:

$$[a]_m = \{x|x \in \mathbb{Z} \wedge x \equiv a \bmod m\}. \tag{5.246}$$

The residue class " a modulo m " consists of all integers having equal remainder if divided by m. Now $[a]_m = [b]_m$ iff $a \equiv b \bmod m$.
There are exactly m residue classes modulo m, and normally they are represented by their smallest non-negative representatives:

$$[0]_m, [1]_m, \ldots, [m-1]_m. \tag{5.247}$$

In the set $\mathbb{Z}_m$ of residue classes modulo m, *residue class addition* and *residue class multiplication* are defined by

$$[a]_m \oplus [b]_m := [a+b]_m, \tag{5.248}$$

$$[a]_m \odot [b]_m := [a \cdot b]_m. \tag{5.249}$$

These residue class operations are independent of the chosen representatives, i.e.,

$$[a]_m = [a']_m \text{ and } [b]_m = [b']_m \text{ imply}$$
$$[a]_m \oplus [b]_m = [a']_m \oplus [b']_m \text{ and } [a]_m \odot [b]_m = [a']_m \odot [b']_m. \tag{5.250}$$

The residue classes modulo m form a ring with unit element, with respect to residue class addition and residue class multiplication (see 5.4.3, **1.**, p. 377), the *residue class ring modulo m.* If p is a prime number, then the residue class ring modulo p is a field (see 5.3.7, **2.**, p. 361).

4. Residue Classes Relatively Prime to *m*

A residue class $[a]_m$ with $\gcd(a, m) = 1$ is called a *residue class relatively prime to m*. If p is a prime number, then all residue classes different from $[0]_p$ are residue classes relatively prime to p.

The residue classes relatively prime to m form an Abelian group (5.3.3.1,**1.**, p. 336) with respect to residue class multiplication, the so-called *group of residue classes relatively prime to m*. The order of this group is $\varphi(m)$, where φ is the *Euler function* (see 5.4.4, **1.**, p. 381).

■ **A:** $[1]_8, [3]_8, [5]_8, [7]_8$ are residue classes relatively prime to 8.

■ **B:** $[1]_5, [2]_5, [3]_5, [4]_5$ are residue classes relatively prime to 5.

■ **C:** $\varphi(8) = \varphi(5) = 4$ is valid.

5. Primitive Residue Classes

A residue class $[a]_m$ relatively prime to m is called a *primitive residue class* if it has order $\varphi(m)$ in the group of residue classes relatively prime to m.

■ **A:** $[2]_5$ is a primitive residue class modulo 5, since $([2]_5)^2 = [4]_5,\ ([2]_5)^3 = [3]_5,\ ([2]_5)^4 = [1]_5$.

■ **B:** There is no primitive residue class modulo 8, since $[1]_8$ has order 1, and $[3]_8, [5]_8, [7]_8$ have order 2 in the group of residue classes relatively prime to m.

Remark: There is a primitive residue class modulo m, iff $m = 2, m = 4, m = p^k$ or $m = 2p^k$, where p is an odd prime number and k is a positive integer.

If there is a primitive residue class modulo m, then the group of residue classes relatively prime to m forms a cyclic group.

6. Linear Congruences

1. Definition If a, b and $m > 0$ are integers, then

$$ax \equiv b(m) \tag{5.251}$$

is called a *linear congruence* (*in the unknown x*).

2. Solutions An integer x^* satisfying $ax^* \equiv b(m)$ is a solution of this congruence. Every integer, which is congruent to x^* modulo m, is also a solution. In finding all solutions of (5.251) it is sufficient to find the integers pairwise incongruent modulo m which satisfy the congruence.

The congruence (5.251) is solvable iff $\gcd(a, m)$ is a divisor of b. In this case, the number of solutions modulo m is equal to $\gcd(a, m)$.

In particular, if $\gcd(a, m) = 1$ holds, the congruence modulo m has a unique solution.

3. Solution Method There are different solution methods for linear congruences. It is possible to transform the congruence $ax \equiv b(m)$ into the Diophantine equation $ax + my = b$, and to determine a special solution (x^0, y^0) of the Diophantine equation $a'x + m'y = b'$ with $a' = a/\gcd(a, m), m' = m/\gcd(a, m), b' = b/\gcd(a, m)$ (see 5.4.2, **1.**, p. 375).

The congruence $a'x \equiv b'(m')$ has a unique solution since $\gcd(a', m') = 1$ modulo m', and

$$x \equiv x^0(m'). \tag{5.252a}$$

The congruence $ax \equiv b(m)$ has exactly $\gcd(a, m)$ solutions modulo m:

$$x^0, x^0 + m, x^0 + 2m, \ldots, x^0 + (\gcd(a, m) - 1)m. \tag{5.252b}$$

■ $114x \equiv 6 \bmod 315$ is solvable, since $\gcd(114, 315)$ is a divisor of 6; there are three solutions modulo 315.

$38x \equiv 2 \bmod 105$ has a unique solution: $x \equiv 94 \bmod 105$ (see 5.4.2, **4.**, p. 376). 94, 199, and 304 are the solutions of $114x \equiv 3 \bmod 315$.

7. Simultaneous Linear Congruences

If finitely many congruences

$$x \equiv b_1(m_1), x \equiv b_2(m_2), \ldots, x \equiv b_t(m_t) \tag{5.253}$$

are given, then (5.253) is called a *system of simultaneous linear congruences*. A result on the set of solutions is the *Chinese remainder theorem*: Consider a given system $x \equiv b_1(m_1), x \equiv b_2(m_2), \ldots, x \equiv b_t(m_t)$, where $m_1, m_2, \ldots, m_t$ are pairwise coprime numbers. If

$$m = m_1 \cdot m_2 \cdots m_t, a_1 = \frac{m}{m_1}, a_2 = \frac{m}{m_2}, \ldots, a_t = \frac{m}{m_t} \tag{5.254a}$$

and x_j is chosen such that $a_j x_j \equiv b_j(m_j)$ for $j = 1, 2, \ldots, t$, then

$$x' = a_1 x_1 + a_2 x_2 + \cdots + a_t x_t \tag{5.254b}$$

is a solution of the system. The system has a unique solution modulo m, i.e., if x' is a solution, then x'' is a solution, too, iff $x'' \equiv x'(m)$.

■ Solve the system $x \equiv 1\,(2), \quad x \equiv 2\,(3), \quad x \equiv 4\,(5)$, where $2, 3, 5$ are pairwise coprime numbers. Then $m = 30, a_1 = 15, a_2 = 10, a_3 = 6$. The congruences $15x_1 \equiv 1\,(2),\ 10x_2 \equiv 2\,(3),\ 6x_3 \equiv 4\,(5)$ have the special solutions $x_1 = 1,\ x_2 = 2,\ x_3 = 4$. The given system has a unique solution modulo m: $x \equiv 15 \cdot 1 + 10 \cdot 2 + 6 \cdot 4\,(30)$, i.e., $x \equiv 29\,(30)$.

Remark: Systems of simultaneous linear congruences can be used to reduce the problem of solving non-linear congruences modulo m to the problem of solving congruences modulo prime number powers (see 5.4.3, **9.**, p. 380).

8. Quadratic Congruences

1. Quadratic Residues Modulo *m* One can solve every congruence $ax^2 + bx + c \equiv 0(m)$ if one can solve every congruence $x^2 \equiv a(m)$:

$$ax^2 + bx + c \equiv 0(m) \Leftrightarrow (2ax + b)^2 \equiv b^2 - 4ac(m). \tag{5.255}$$

First quadratic residues modulo m are considered: Let $m \in \mathbb{N}, m > 1$ and $a \in \mathbb{Z}$, $\gcd(a, m) = 1$. The number a is called a *quadratic residue modulo m* iff there is an $x \in \mathbb{Z}$ with $x^2 \equiv a(m)$.

If the canonical prime factorization of m is given, i.e.,

$$m = \prod_{i=1}^{\infty} p_i^{\alpha_i}, \tag{5.256}$$

then r is a quadratic residue modulo m iff r is a quadratic residue modulo $p_i^{\alpha_i}$ for $i = 1, 2, 3, \ldots$.

If a is a quadratic residue modulo a prime number p, then this is denoted by $\left(\frac{a}{p}\right) = 1$; if a is a quadratic non-residue modulo p, then it is denoted by $\left(\frac{a}{p}\right) = -1$ (Legendre symbol).

■ The numbers $1, 4, 7$ are quadratic residues modulo 9.

2. Properties of Quadratic Congruences

(E1) $p \nmid ab$ and $a \equiv b(p)$ imply $\left(\frac{a}{p}\right) = \left(\frac{b}{p}\right)$. (5.257a)

(E2) $\left(\frac{1}{p}\right) = 1.$ (5.257b)

(E3) $\left(\frac{-1}{p}\right) = (-1)^{\frac{p-1}{2}}.$ (5.257c)

(E4) $$\left(\frac{ab}{p}\right) = \left(\frac{a}{p}\right) \cdot \left(\frac{b}{p}\right) \quad \text{in particular} \quad \left(\frac{ab^2}{p}\right) = \left(\frac{a}{p}\right). \tag{5.257d}$$

(E5) $$\left(\frac{2}{p}\right) = (-1)^{\frac{p^2-1}{8}}. \tag{5.257e}$$

(E6) Quadratic reciprocity law: If p and q are distinct odd prime numbers,

then $$\left(\frac{p}{q}\right) \cdot \left(\frac{q}{p}\right) = (-1)^{\frac{p-1}{2}\frac{q-1}{2}}. \tag{5.257f}$$

■ $\left(\frac{65}{307}\right) = \left(\frac{5}{307}\right) \cdot \left(\frac{13}{307}\right) = \left(\frac{307}{5}\right) \cdot \left(\frac{307}{13}\right) = \left(\frac{2}{5}\right) \cdot \left(\frac{8}{13}\right) = (-1)^{\frac{5^2-1}{8}} \left(\frac{2^3}{13}\right) = -\left(\frac{2}{13}\right) = -(-1)^{\frac{13^2-1}{8}} = 1.$

In General: A congruence $x^2 \equiv a(2^\alpha)$, $\gcd(a, 2) = 1$, is solvable iff $a \equiv 1(4)$ for $\alpha = 2$ and $a \equiv 1(8)$ for $\alpha \geq 3$. If these conditions are satisfied, then modulo 2^α there is one solution for $\alpha = 1$, there are two solutions for $\alpha = 2$ and four solutions for $\alpha \geq 3$.

A necessary condition for solvability of congruences of the general form

$$x^2 \equiv a(m),\ m = 2^\alpha p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}, \quad \gcd(a, m) = 1, \tag{5.258a}$$

is the solvability of the congruences

$$a \equiv 1(4) \text{ for } \alpha = 2, \quad a \equiv 1(8) \text{ for } \alpha \geq 3, \quad \left(\frac{a}{p_1}\right) = 1,\ \left(\frac{a}{p_2}\right) = 1,\ \ldots,\ \left(\frac{a}{p_t}\right) = 1. \tag{5.258b}$$

If all these conditions are satisfied, then the number of solutions is equal to 2^t for $\alpha = 0$ and $\alpha = 1$, equal to 2^{t+1} for $\alpha = 2$ and equal to 2^{t+2} for $\alpha \geq 3$.

9. Polynomial Congruences

If $m_1, m_2, \ldots, m_t$ are pairwise coprime numbers, then the congruence

$$f(x) \equiv a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0 \equiv 0(m_1 m_2 \cdots m_t) \tag{5.259a}$$

is equivalent to the system

$$f(x) \equiv 0(m_1),\ f(x) \equiv 0(m_2),\ \ldots,\ f(x) \equiv 0(m_t). \tag{5.259b}$$

If k_j is the number of solutions of $f(x) \equiv 0(m_j)$ for $j = 1, 2, \ldots, t$, then $k_1 k_2 \cdots k_t$ is the number of solutions of $f(x) \equiv 0(m_1 m_2 \cdots m_t)$. This means that the solution of the congruence

$$f(x) \equiv 0\ (p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_t^{\alpha_t}), \tag{5.259c}$$

where $p_1, p_2, \ldots, p_t$ are primes, can be reduced to the solution of congruences $f(x) \equiv 0(p^\alpha)$. Moreover, these congruences can be reduced to congruences $f(x) \equiv 0(p)$ modulo prime numbers in the following way:

a) A solution of $f(x) \equiv 0(p^\alpha)$ is a solution of $f(x) \equiv 0(p)$, too.

b) A solution $x \equiv x_1(p)$ of $f(x) \equiv 0(p)$ defines a unique solution modulo p^α iff $f'(x_1)$ is not divisible by p:

Suppose $f(x_1) \equiv 0(p)$. Let $x = x_1 + pt_1$ and determine the unique solution t_1' of the linear congruence

$$\frac{f(x_1)}{p} + f'(x_1)t_1 \equiv 0(p). \tag{5.260a}$$

Substitute $t_1 = t_1' + pt_2$ into $x = x_1 + pt_1$, then $x = x_2 + p^2 t_2$ is obtained. Now, the solution t_2' of the linear congruence

$$\frac{f(x_2)}{p^2} + f'(x_2)t_2 \equiv 0(p) \tag{5.260b}$$

has to be determined modulo p^2. By substitution of $t_2 = t_2' + pt_3$ into $x = x_2 + p^2 t_2$ the result $x = x_3 + p^3 t_3$ is obtained. Continuing this process yields the solution of the congruence $f(x) \equiv 0\,(p^\alpha)$.

■ Solve the congruence $f(x) = x^4 + 7x + 4 \equiv 0\,(27)$. $f(x) = x^4 + 7x + 4 \equiv 0\,(3)$ implies $x \equiv 1\,(3)$, i.e., $x = 1 + 3t_1$. Because of $f'(x) = 4x^3 + 7$ and $3 \not| f'(1)$ now the solution of the congruence $f(1)/3 + f'(1) \cdot t_1 \equiv 4 + 11t_1 \equiv 0\,(3)$ is searched for: $t_1 \equiv 1\,(3)$, i.e., $t_1 = 1 + 3t_2$ and $x = 4 + 9t_2$. Then consider $f(4)/9 + f'(4) \cdot t_2 \equiv 0\,(3)$ and the solution $t_2 \equiv 2\,(3)$ is obtained, i.e., $t_2 = 2 + 3t_3$ and $x = 22 + 27t_3$. Therefore, 22 is the solution of $x^4 + 7x + 4 \equiv 0\,(27)$, uniquely determined modulo 27.

5.4.4 Theorems of Fermat, Euler, and Wilson

1. Euler Function

For every positive integer m with $m > 0$ one can determine the number of coprimes x with respect to m for $1 \le x \le m$. The corresponding function φ is called the Euler function. The value of the function $\varphi(m)$ is the number of residue classes relatively prime to m (s. 5.4.3, **4.**, p. 378).

For instance, $\varphi(1) = 1,\ \varphi(2) = 1,\ \varphi(3) = 2,\ \varphi(4) = 2,\ \varphi(5) = 4,\ \varphi(6) = 2,\ \varphi(7) = 6,\ \varphi(8) = 4$, etc. In general, $\varphi(p) = p - 1$ holds for every prime number p and $\varphi(p^\alpha) = p^\alpha - p^{\alpha-1}$ for every prime number power p^α. If m is an arbitrary positive integer, then $\varphi(m)$ can be determined in the following way:

$$\varphi(m) = m \prod_{p|m} \left(1 - \frac{1}{p}\right), \tag{5.261a}$$

where the product applies to all prime divisors p of m.

■ $\varphi(360) = \varphi(2^3 \cdot 3^2 \cdot 5) = 360 \cdot (1 - \frac{1}{2}) \cdot (1 - \frac{1}{3}) \cdot (1 - \frac{1}{5}) = 96$.

Furthermore

$$\sum_{d|m} \varphi(d) = m \tag{5.261b}$$

is valid. If $\gcd(m, n) = 1$ holds, then we get $\varphi(mn) = \varphi(m)\varphi(n)$.

■ $\varphi(360) = \varphi(2^3 \cdot 3^2 \cdot 5) = \varphi(2^3) \cdot \varphi(3^2) \cdot \varphi(5) = 4 \cdot 6 \cdot 4 = 96$.

2. Fermat-Euler Theorem

The *Fermat-Euler theorem* is one of the most important theorems of elementary number theory. If a and m are coprime positive numbers, then

$$a^{\varphi(m)} \equiv 1(m). \tag{5.262}$$

■ Determine the last three digits of 9^{9^9} in decimal notation. This means, determine x with $x \equiv 9^{9^9}\ (1000)$ and $0 \le x \le 999$. Now $\varphi(1000) = 400$, and according to Fermats theorem $9^{400} \equiv 1\,(1000)$. Furthermore $9^9 = (80 + 1)^4 \cdot 9 \equiv \left(\binom{4}{0} 80^0 \cdot 1^4 + \binom{4}{1} 80^1 \cdot 1^3\right) \cdot 9 = (1 + 4 \cdot 80) \cdot 9 \equiv -79 \cdot 9 \equiv 89\,(400)$. From that it follows that $9^{9^9} \equiv 9^{89} = (10 - 1)^{89} \equiv \binom{89}{0} 10^0 \cdot (-1)^{89} + \binom{89}{1} 10^1 \cdot (-1)^{88} + \binom{89}{2} 10^2 \cdot (-1)^{87} = -1 + 89 \cdot 10 - 3916 \cdot 100 \equiv -1 - 110 + 400 = 289(1000)$. The decimal notation of 9^{9^9} ends with the digits 289.

Remark: The theorem above for $m = p$, i.e., $\varphi(p) = p - 1$ was proved by Fermat; the general form was proved by Euler. This theorem forms the basis for encoding schemes (see 5.4.6). It contains a necessary criterion for the prime number property of a positive integer: If p is a prime, then $a^{p-1} \equiv 1(p)$ holds for every integer a with $p \not| a$.

3. Wilson's Theorem

There is a further prime number criterion, called the Wilson theorem:

Every prime number p satisfies $(p - 1)! \equiv -1(p)$.

The inverse proposition is also true; and therefore:

The number p is a prime number iff $(p-1)! \equiv -1(p)$.

5.4.5 Prime Number Tests

In the followings two stochastic prime tests will be presented which are useful at large numbers to test the prime property with a sufficiently small probability of mistakes. With these tests it is possible to show that a number is not a prime, without knowing its prime factors.

1. Fermat-Prime Number Test

Let n be an odd natural number and a an integer such that $\gcd(a,n) = 1$ and $a^{n-1} \equiv 1 \pmod{n}$.. Then n is called a *pseudoprime* to base a.

■ **A:** 341 is a pseudo prime to basis 2; 341 is not a pseudo prime to basis 3.

Test: Let an odd natural number $n > 1$ be given. Choose $a \in \mathbb{Z}_n \setminus \{0\}$.

• If the gcd $(a,n) > 1$, then n is not prime.

• If the gcd $(a,n) = 1$ and $\left\{\begin{array}{ll} a^{n-1} \equiv 1 & (\text{mod } n) \\ a^{n-1} \not\equiv 1 & (\text{mod } n) \end{array}\right\}$, then $n \left\{\begin{array}{c} \text{did pass} \\ \text{did not pass} \end{array}\right\}$ the test to base a. If n did not pass the test, then n is not a prime. If n did pass the test, then it may be a prime, but more tests are needed with other base, i.e. tests with further values of a.

■ **B:** $n = 15$: The test with $a = 4$ gives $4^{14} \equiv 1 \pmod{15}$. The test with $a = 7$ gives $7^{14} \equiv 4 \not\equiv 1 \pmod{15}$. Hence 15 is not a prime.

■ **C:** $n = 561$: The test with arbitrary $a \in \mathbb{Z}_{561} \setminus \{0\}$ with $\gcd(a, 561) = 1$ results in $a^{560} \equiv 1 \pmod{561}$. But $561 = 3 \cdot 11 \cdot 17$ is not a prime.

Remark: A composite number n for which $a^{n-1} \equiv 1 \pmod{n}$ for all $a \in \mathbb{Z}_n \setminus \{0\}$ with $\gcd(a,n) = 1$ is called a Carmichael number.

If n is not a prime and not a Carmichael number, then one can show that the level of error of the first kind to get a false result using k numbers with $\gcd(a,n) =1$ is at most $1/2^k$. At least for the half of the numbers in $\mathbb{Z}_n \setminus \{0\}$ with $\gcd(a,n) =1$ the relation $a^{n-1} \not\equiv 1 \pmod{n}$ holds.

2. Rabin-Miller Prim Number Test

The Rabin-Miller primality test is based on the following statement $(*)$:
Let $n > 2$ be a prime, $n - 1 = 2^t u$ (u is odd), g.c.d$(a,n) =1$. Then:

$$a^u \equiv 1 (\text{mod}\, n) \text{ or } a^{2^j u} \equiv -1 \pmod{n} \text{ for some } j \in \{0, 1, \ldots, t-1\}. \qquad (*)$$

Every odd natural number $n > 1$ can be tested about prime property in the following way:
Test: Choose $a \in \mathbb{Z}_n \setminus \{0\}$ and find the representation $n - 1 = 2^t u$ (u is odd).
• If g.c.d$(a,n) > 1$, then n is not a prime.

• If g.c.d$(a,n) =1$, then the sequence $a^u \pmod{n}$, $a^{2u}(\text{mod}n)$, …, $a^{2^{t-1}u}(\text{mod } n)$ is calculated until a value is found which satisfies $(*)$. These elements are calculated by repeated squaring mod n. If there is no such value, then n is not a prime. Otherwise n did pass the test to basis a.

■ **A**: $n = 561$, and should be tested by different values of a:

$$n - 1 = 2^4 \cdot 35, \quad a = 2\colon \begin{array}{rl} 2^{35} \equiv 263 \not\equiv \pm 1 & (\text{mod } 561), \\ 2^{70} \equiv 166 \not\equiv -1 & (\text{mod } 561), \\ 2^{140} \equiv 67 \not\equiv -1 & (\text{mod } 561), \\ 2^{280} \equiv 421 \not\equiv -1 & (\text{mod } 561). \end{array} \quad \text{561 is not a prime.}$$

If choosing k different values randomly and independently and n passes the test to basis a for each, then the error rate of the first kind that n is not a prime is $\leq 1/4^k$. In the practice $k = 25$ is chosen.

■ **B:** There is only one number $\leq 2,5 \cdot 10^{10}$ such that it passes the test to basis $a = 2, 3, 5, 7$ and it is not a prime.

3. AKS Prime Number Test

The AKS primality test is based on a polynomial algorithm to determine whether a number is prime or composite. Published by Agrawal, Kayal, and Saxena, in 2002, meanwhile it is evident that the prime property can be tested efficiently for any natural number.

The test is based on the following statements:
If $n > 1$ is a natural number and r is a prime satisfying the assumtions
- n is not divisible by primes $\le r$,
- $r^i \not\equiv 1 \pmod n$ for $i = 1, 2, \ldots, \lfloor (\log_2 n)^2 \rfloor^*$,
- $(x+a)^n \equiv x^n + a \pmod{x^r - 1, n}$ for every $1 \le a \le \sqrt{r}\log n$,

Then n is a power of a prime.

Let $n > 1$ be an odd natural number whose prime characteristic is to be tested, and $m := \lfloor (\log_2 n)^5 \rfloor$. If $n < 5690034$, then it is tested by comparing it to a list of known prime numbers whether n is a prime. For $n > 5690034$ holds $n > m$:

Test:
- Check, whether n can be divided by a natural number from the interval $[3, m]$. If yes, then n is not a prime.
- Otherwise take a prime $r < m$, such that $r^i \not\equiv 1(\bmod n)$ for $i = 1, 2, \ldots, \lfloor (\log_2 n)^2 \rfloor$. (It can be proven, that such a prime r exists.)
- Check, whether the congruence $(x+a)^n = x^n + a(\bmod x^r - 1, n)$ for $a = 1, 2, \sqrt{r}\lfloor(\log_2 n)\rfloor$ holds. If not, then n is not a prime. If yes, then n is a power of a prime. In this case it is to be tested, whether natural numbers q and $k > 1$ exist, for which $n = q^k$. If not, then n is a prime.

Different to the known and efficient stochastic algorithms, the result of the test can be trusted without even a negligible small error probability of mistakes. However in cryptography the Rabin-Miller test is preferred.

5.4.6 Codes

5.4.6.1 Control Digits

In the information theory methods are provided to recognize and to correct errors in data combinations. Some of the simplest methods are represented in the form of the following control digits.

1. International Standard Book Number ISBN-10

A simple application of the congruence of numbers is the use of control digits with the International Standard Book Number ISBN. A combination of 10 digits of the form

$$\text{ISBN } a - bcd - efghi - p. \tag{5.263a}$$

is assigned to a book. The digits have the following meaning: a is the group number (for example, $a = 3$ tells us that the book originates from Austria, Germany, or Switzerland), bcd is the publisher's number, and $efghi$ is the title number of the book by this publisher. A control digit p will be added to detect erroneous book orders and thus help reduce expenses. The control digit p is the smallest non-negative digit that fulfils the following congruence:

$$10a + 9b + 8c + 7d + 6e + 5f + 4g + 3h + 2i + p \equiv 0(11). \tag{5.263b}$$

If the control digit p is 10, a unary symbol such as X is used (see also 5.4.6, **3.**, p. 384). A presented ISBN can now be checked for a match of the control digit contained in the ISBN and the control digit determined from all the other digits. In case of no match an error is certain. The ISBN control digit method permits the detection of the following errors:

1. Single digit error and

2. interchange of two digits.

Statistical investigations showed that by this method more than 90% of all actual errors can be detected. All other observed error types have a relative frequency of less than 1%. In the majority of the cases

[*] $\lfloor x \rfloor$ is symbol for "greatest integer $\le x$".

the described method will detect the interchange of two digits or the interchange of two complete digit blocks.

2. Central Codes for Drugs and Medicines

In pharmacy, a similar numerical system with control digits is employed for identifying medicaments. In Germany, each medicament is assigned a seven digit control code:

$$abcdefp. \tag{5.264a}$$

The last digit is the control digit p. It is the smallest, non-negative number that fulfils the congruence

$$2a + 3b + 4c + 5d + 6e + 7f \equiv p(11). \tag{5.264b}$$

Here too, the single digit error or the interchange of two digits can always be detected.

3. Account Numbers

Banks and saving banks use a uniform account number system with a maximum of 10 digits (depending on the business volume). The first (at most four) digits serve the classification of the account. The remaining six digits represent the actual account number including a control digit in the last position. The individual banks and saving banks tend to apply different control digit methods, for example:

a) The digits are multiplied alternately by 2 and by 1, beginning with the rightmost digit. A control digit p will then be added to the sum of these products such that the new total is the next number divisible by 10. Given the account number $abcd\,efghi\,p$ with control digit p, then the congruence

$$2i + h + 2g + f + 2e + d + 2c + b + 2a + p \equiv 0 \pmod{10}. \tag{5.265}$$

holds.

b) As in method **a)**, however, any two-digit product is first replaced by the sum of its two digits and then the total sum will be calculated.

In case **a)** all errors caused by the interchange of adjacent digits and almost all single-digit errors will be detected.
In case **b)**, however, all errors caused by the change of one digit and almost all errors caused by the interchange of two adjacent digits will be discovered. Errors due to the interchange of non-adjacent digits and the change of two digits will often not be detected.

The reason for not using the more powerful control digit method modulo 11 is of a non-mathematical nature. The non-numerical sign X (instead of the control digit 10 (see 5.4.6, **1.**, p. 383)) would require an extension of the numerical keyboard. However, renouncing those account numbers whose control digit has the value of 10 would have barred the smooth extension of the original account number in a considerable number of cases.

4. European Article Number EAN

EAN stands for *European Article Number*. It can be found on most articles as a bar code or as a string of 13 or 8 digits. The bar code can be read by means of a scanner at the counter.
In the case of 13-digit strings the first two digits identify the country of origin, e.g., $40, 41, 42$ and 43 stand for Germany. The next five digits identify the producer, the following five digits identify a particular product. The last digit is the control digit p.

This control digit will be obtained by first multiplying all 12 digits of the string alternately by 1 and 3 starting with the left-most digit, by then totalling all values, and by finally adding a p such that the next number divisible by 10 is obtained. Given the article number $abcdefghikmn\,p$ with control digit p, then the congruence

$$a + 3b + c + 3d + e + 3f + g + 3h + i + 3k + m + 3n + p \equiv 0 \pmod{10}. \tag{5.266}$$

holds.

This control digit method always permits the detection of single digit errors in the EAN and often the detection of the interchange of two adjacent digits. The interchange of two non-adjacent digits and the change of two digits will often not be detected.

5.4.6.2 Error correcting codes

1. Model of Data Transmission and Error Correction

At transmission of messages through noisy channels the correction of errors is often possible. The message is coded first, then after transmission the usually biased codes are corrected into the right ones, so after decoding them the original message can be recovered. That case is considered now, when the length of the words of the message is k, and the length of the coded words is n, and both of them consist of only zeros and ones. Then k is the number of *information positions* and $n-k$ is the number of *redundant positions*. Every word of the message is an element of $\mathrm{GF}(2)^k$ (see 5.3.7.4 p. 363) and every word of the code is an element of $\mathrm{GF}(2)^n$. To simplify the notation the words of the message are written in the form $a_1, a_2, \ldots, a_k$, and the words of the code in the form $c_1, c_2, \ldots, c_n$. The words of the message are not transmitted, only the words of the code are.
An often used idea of error correction is to convert the transmitted word $d_1, d_2, \ldots, d_n$ first into a valid codeword $c_1, c_2, \ldots, c_n$ which differs from it in the least number of digits (decoding MLD). It depends on the properties of coding and the transmission channels that how many errors can be detected and corrected in this way.

■ At digit repeating codes the message word 0 is represented by the codeword 0000. If after transmission the receiver gets the word 0010, then he assumes that the original codeword was 0000, and it is decoded as message word 0. But if the received word is 1010, then similar assumption can not be applied, since the message word 1 is coded as 1111, so the difference is similar. At least it can be recognized that there is some error in the received word.

2. t-Error Correcting Codes

The set of all codewords is called *code* $\mathcal{C}$. The *distance* of two codewords is the number of digits (positions) in which the two words differ from each other. The *minimal distance* $d_{\min}(\mathcal{C})$ of codes is the smallest distance which occurs between the codewords of $\mathcal{C}$.

■ For $\mathcal{C}_1 = \{0000, 1111\}$, $d_{\min}(\mathcal{C})_1 = 4$. For $\mathcal{C}_2 = \{000, 011, 101, 110\}$, $d_{\min}(\mathcal{C}_2) = 2$, since there are codewords which have distance 2. For $\mathcal{C}_3 = \{00000, 01101, 10111, 11010\}$, $d_{\min}(\mathcal{C}_3) = 3$, there are codewords in $\mathcal{C}_3$ whose distance is 3.
If the minimal distance $d_{\min}(\mathcal{C})$ of a code $\mathcal{C}$ is known, then it is easy to recognize how many transmission errors can be corrected. Codes, correcting t errors, are called *t-error correcting.* A code $\mathcal{C}$ is t-error correcting if $d_{\min}(\mathcal{C}) \geq 2t+1$.

■ (Continuation) $\mathcal{C}_1$ is 1-error correcting, $\mathcal{C}_2$ is 0-error correcting (it means, that no error can be corrected), $\mathcal{C}_3$ is 1-error correcting.
For every t-error correcting code $\mathcal{C} \subseteq \mathrm{GM}(2)^n$ holds $\sum\limits_{i=0}^{t} \binom{c}{n} \cdot |\mathcal{C}| \leq 2^n$. If equality holds, then $\mathcal{C}$ is called t-perfect.

■ The digit repeating code $\mathcal{C} = \{00\ldots0, 11\ldots1\} \subseteq \mathrm{GF}(2)^{2t+1}$ is t-perfect.

3. Linear Codes

A non-empty subset $\mathcal{C} \subseteq \mathrm{GF}(2)^n$ is called *(binary) linear code*, if $\mathcal{C}$ is a sub-vector space of $\mathrm{GF}(2)^n$. If a linear code $\mathcal{C} \subseteq \mathrm{GF}(2)^n$ has dimension k, then it is called an *(n,k) linear code.*

■ (Continuation) $\mathcal{C}_1$ is a (4,1) linear code, $\mathcal{C}_2$ is a (3,2) linear code, $\mathcal{C}_3$ is a (5,2) linear code. In the case of linear codes the minimal distance (and as a consequence the number of correctible errors) is easy to determine: The minimal distance of such a code is the smallest distance of a non-zero vector from the zero vector of the vector space. The minimal distance can be found if the minimal number of ones, except with all zeros, in the codewords is given.
For every (n,k) linear code there is a *generating matrix* $\mathbf{G}$ for which $\mathcal{C} = \{aG \mid a \in \mathrm{GF}(2)^k\}$:

$$G = \begin{pmatrix} g_{11} & \cdots & g_{1n} \\ \vdots & \vdots & \vdots \\ g_{k1} & \cdots & g_{kn} \end{pmatrix}_{k\times n} = \begin{pmatrix} g_1 \\ \vdots \\ g_k \end{pmatrix}. \tag{5.267}$$

The code is uniquely defined by the generating matrix; the codeword of the message word $a_1a_2 \dots a_k$ is determined in the following way:

$$a_1a_2 \dots a_k \mapsto \underbrace{a_1g_1 + a_2g_2 + \dots + a_kg_k}_{aG} . \qquad (5.268)$$

In the case of an (n, k) linear code $\mathcal{C}$ a *check matrix* is needed for decoding:

$$H = \begin{pmatrix} h_{11} & \dots & h_{1n} \\ \vdots & \vdots & \vdots \\ h_{n-k,1} & \dots & h_{n-k,n} \end{pmatrix}_{(n-k)\times n} . \qquad (5.269)$$

The (binary) linear code $\mathcal{C}$ is 1-error correcting, if the columns of H are pairwise different and non-zero vectors. If the result of the transmission is the word $d = d_1d_2 \dots d_n$, then Hd^T is calculated. If the result is the zero vector, then d is a codeword. Otherwise if Hd^T is the i-th column of the check matrix H, then the corresponding codeword is $d + e_i$, where $e_i = (0, 0, \dots, 0, 1, 0, \dots, 0)$ and the 1 is on the i-th position.

4. Cyclic Codes

Cyclic codes are the most investigated linear codes. They provide efficient coding and decoding.

A (binary) (n, k) linear code is called *cyclic* if for every codeword $c_1c_2 \dots c_n$ the codeword obtained by a cyclic right shift of the components is also a codeword, i.e. $c_0c_1 \dots c_{n-1} \in \mathcal{C} \Rightarrow c_{n-1}c_0c_1 \dots c_{n-2} \in \mathcal{C}$

■ $\mathcal{C} = \{000, 110, 101, 011\}$ is a cyclic (3,2) linear code.

To have an efficient work with cyclic codes, the codewords are represented by polynomials of degree $\leq n-1$ with coefficients from GF(2): $\mathcal{C} = \{000, 110, 101, 011\}$ is a cyclic $(3, 2)$-linear code.

A (binary) (n, k) linear code $\mathcal{C}$ is cyclic if and only if for every $c(x)$

$$c(x) \in \mathcal{C} \Rightarrow c(x) \cdot x (\text{mod } x^n \text{ -1}) \in \mathcal{C} \qquad (5.270)$$

A cyclic (n, k) linear code can be described by a generating polynomial and a control polynomial as follows: The *generating polynomial* $g(x)$ of degree $n - k$ $(k \in \{1, 2, \dots, n-1\})$ is a divisor of $x^n - 1$. The polynomial $h(x)$ of degree k for which $g(x)h(x) = x^n - 1$ is called the *control polynomial* . Coding of $a_1a_2 \dots a_k$ in polynomial representation $a(x)$ is given by

$$a(x) \mapsto a(x) \cdot g(x) . \qquad (5.271)$$

Polynomial $d(x)$ is an element of the code, if the generator polynomial $g(x)$ is a divisor of $d(x)$, or the control polynomial $h(x)$ satisfies the relation $d(x)h(x) \equiv 0 \bmod x^n - 1$.

An important class of cyclic codes are the BCH-codes. Here a lower bound δ of the minimal distance and with it a lower bound for the number of errors can be required for which code should be corrected. Here δ is called the *design distance* of the code.

A (binary) (n, k) linear code $\mathcal{C}$ is a BCH-code with design distance δ if for the generating polynomial $g(x)$:

$$g(x) = \text{lcm}(m_{\alpha^b}(x), m_{\alpha^{b+1}}(x), \dots, m_{\alpha^{b+\delta-2}}(x)) , \qquad (5.272)$$

where α is a primitive n-th unit root and b is an integer. The polynomials $m_{\alpha^j}(x)$ are minimal polynomials of α^j.

For a BCH-code $\mathcal{C}$ with design distance δ the relation $d_{\min}(\mathcal{C}) \geq \delta$ must hold.

5.5 Cryptology

5.5.1 Problem of Cryptology

Cryptology is the science of hiding information by the transformation of data.

The idea of protecting data from unauthorized access is rather old. During the 1970s together with the introduction of *cryptosystems on the basis of public keys*, cryptology became an independent branch of science. Today, the subject of cryptological research is how to protect data from unauthorized access and against tampering.

Beside the classical military applications, the needs of the information society gain more and more in importance. Examples are the guarantee of secure message transfer via email, electronic funds transfer (home-banking), the PIN of EC-cards, etc.

Today, the fields of *cryptography* and *cryptanalysis* are subsumed under the notion of cryptology. Cryptography is concerned with the development of cryptosystems whose cryptographic strengths can be assessed by applying the methods of cryptanalysis for breaking cryptosystems.

5.5.2 Cryptosystems

An abstract cryptosystem consists of the following sets: a set M of messages, a set C of ciphertexts, sets K and K' of keys, and sets $\mathbb{E}$ and $\mathbb{D}$ of functions. A message $m \in M$ will be encrypted into a ciphertext $c \in C$ by applying a function $E \in \mathbb{E}$ together with a key $k \in K$, and will be transmitted via a communication channel. The recipient can reproduce the original message m from c if he knows an appropriate function $D \in \mathbb{D}$ and the corresponding key $k' \in K'$. There are two types of cryptosystems:

1. Symmetric Cryptosystems: The conventional symmetric cryptosystem uses the same key k for encryption of the message and for decryption of the ciphertext. The user has complete freedom in setting up his conventional cryptosystem. Encryption and decryption should, however, not become too complex. In any case, a trustworthy transmission between the two communication partners is mandatory.

2. Asymmetric Cryptosystems: The asymmetric cryptosystem (see 5.5.7.1, p. 391) uses two keys, one private key (to be kept secret) and a public key. The public key can be transmitted along the same path as the ciphertext. The security of the communication is warranted by the use of so-called *one-way functions* (see 5.5.7.2, p. 391), which makes it practically impossible for the unauthorized listener to deduce the plaintext from the ciphertext.

5.5.3 Mathematical Foundation

An alphabet $A = \{a_0, a_1, \ldots, a_{n-1}\}$ is a finite non-empty totally ordered set, whose elements a_i are called letters. $|A|$ is the length of the alphabet. A sequence of letters $w = a'_1 a'_2 \ldots a'_n$ of length $n \in \mathbb{N}$ and $a_i \in A$ is called a word of length n over the alphabet A. A^n denotes the set of all words of length n over A. Let $n, m \in \mathbb{N}$, let A, B be alphabets, and let S be a finite set.

A cryptofunction is a mapping t: $A^n \times S \to B^m$ such that the mappings t_s: $A^n \to B^m : w \to t(w, s)$ are injective for all $s \in S$. The functions t_s and t_s^{-1} are called the encryption and decryption function, respectively. w is called plaintext, $t_s(w)$ is the ciphertext.

Given a cryptofunction t, then the one-parameter family $\{t_s\}_{s\in S}$ is a cryptosystem T_S. The term *cryptosystem* will be applied if in addition to the mapping t, the structure and the size of the set of keys is significant. The set S of all the keys belonging to a cryptosystem is called the key space. Then

$$T_S = \{t_s\colon A^n \to A^n | s \in S\} \tag{5.273}$$

is called a cryptosystem on A^n.

If T_S is a cryptosystem over A^n and $n = 1$, then t_s is called a stream cipher; otherwise t_s is called a block cipher.

Cryptofunctions of a cryptosystem over A^n are suited for the encryption of plaintext of any length. The plaintext will be split into blocks of length n prior to applying the function to each individual block. The last block may need padding with filler characters to obtain a block of length n. The filler characters must not distort the plaintext.

There is a distinction between *context-free encryption*, where the ciphertext block is only a function of the corresponding plaintext block and the key, and *context sensitive encryption*, where the ciphertext block depends on other blocks of the message. Ideally, each ciphertext digit of a block depends on all digits of the corresponding plaintext block and all digits of the key. Small changes to the plaintext or to the key cause extended changes to the ciphertext (avalanche effect).

5.5.4 Security of Cryptosystems

Cryptanalysis is concerned with the development of methods for deducing from the ciphertext as much information about the plaintext as possible without knowing the key. According to A. Kerkhoff the security of a cryptosystem rests solely in the difficulty of detecting the key or, more precisely, the decryption function. The security must not be based on the assumption that the encryption algorithm is kept secret. There are different approaches to assess the security of a cryptosystem:

1. Absolutely Secure Cryptosystems: There is only one absolutely secure cryptosystem based on substitution ciphers, which is the *one-time pad*. This was proved by Shannon as part of his information theory.

2. Analytically Secure Cryptosystems: No method exists to break a cryptosystem systematically. The proof of the non-existence of such a method follows from the proof of the non-computability of a decryption function.

3. Secure Cryptosystems according to Criteria of Complexity Theory: There is no algorithm which can break a cryptosystem in polynomial time (with regard to the length of the text).

4. Practically Secure Cryptosystems: No method is known which can break the cryptosystem with available resources and with justified costs.

Cryptanalysis often applies statistical methods such as determining the frequency of letters and words. Other methods are an exhaustive search, the trial-and-error method and a structural analysis of the cryptosystem (solving of equation systems).

In order to attack a cryptosystem one can benefit from frequent flaws in encryption such as using stereotype phrases, repeated transmissions of slightly modified text, an improper and predictable selection of keys, and the use of filler characters.

5.5.4.1 Methods of Conventional Cryptography

In addition to the application of a cryptofunction it is possible to encrypt a plaintext by means of *cryptological codes*. A *code* is a bijective mapping of some subset A' of the set of all words over an alphabet A onto the subset B' of the set of all words over the alphabet B. The set of all source-target pairs of such a mapping is called a code book.

■

today evening	0815
tomorrow evening	1113

The advantage of replacing long plaintexts by short ciphertexts is contrasted with the disadvantage that the same plaintext will always be replaced by the same ciphertext. Another disadvantage of code books is the need for a complete and costly replacement of all books should the code be compromised even partially.

In the following only encryption by means of cryptofunctions will be considered. Cryptofunctions have the additional advantage that they do not require any arrangement about the contents of the messages prior to their exchange.

Transposition and *substitution* constitute conventional cryptoalgorithms. In cryptography, a transposition is a special permutation defined over geometric patterns. The substitutions will now be discussed in detail. There is a distinction between monoalphabetic and polyalphabetic substitutions according to how many alphabets are used for presenting the ciphertext. Generally, a substitution is termed polyalphabetic even if only one alphabet is used, but the encryption of the individual plaintext letter depends on its position within the plaintext.

A further, useful classification is the distinction between monographic and polygraphic substitutions. In the first case, single letters will be substituted, in the latter case, strings of letters of a fixed length > 1.

5.5.4.2 Linear Substitution Ciphers

Let $A = \{a_0, a_1, \ldots, a_{n-1}\}$ be an alphabet and $k, s \in \{0, 1, \ldots, n-1\}$ with $\gcd(k, n) = 1$. The permutation t_s^k, which maps each letter a_i to $t_s^k(a_i) = a_{ki+s}$, is called a linear substitution cipher. There exist $n\,\varphi(n)$ linear substitution ciphers on A.

Shift ciphers are linear substituting ciphers with $k = 1$. The shift cipher with $s = 3$ was already used by Julius Caesar (100 to 44 BC) and, therefore, it is called the Caesar cipher.

5.5.4.3 Vigenère Cipher

An encryption called the Vigenère cipher is based on the periodic application of a key word whose letters are pairwise distinct. The encryption of a plaintext letter is determined by the key letter that has the same position in the key as the plaintext letter in the plaintext. This requires a key that is as long as the plaintext. Shorter keys are repeated to match the length of the plaintext.

A version of the Vigenère cipher attributed to L. Carroll utilizes the so-called Vigenère tableau (see picture) for encryption and decryption. Each row represents the cipher for the key letter to its very left. The alphabet for the plaintext runs across the top. The encryption step is as follows: Given a key letter D and a plaintext letter C, then the ciphertext letter is found at the intersection of the row labeled D and the column labeled C; the ciphertext is F. Decryption is the inverse of this process.

	A	B	C	D	E	F	…
A	A	B	C	D	E	F	…
B	B	C	D	E	F	G	…
C	C	D	E	F	G	H	…
D	D	E	F	G	H	I	…
E	E	F	G	H	I	J	…
F	F	G	H	I	J	K	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋱

■ Let the key be “ HUT ”.

Plaintext:	O	N	C	E	U	P	O	N	A	T	I	M	E
Key:	H	U	T	H	U	T	H	U	T	H	U	T	H
Ciphertext:	V	H	V	L	O	I	V	H	T	A	C	F	L

Formally, the Vigenère cipher can be written in the following way: let a_i be the plaintext letter and a_j be the corresponding key letter, then $k = i + j$ determines the ciphertext letter a_k. In the above example, the first plaintext letter is $O = a_{14}$. The 15-th position of the key is taken by the letter $H = a_7$. Hence, $k = i + j = 14 + 7 = 21$ yields the ciphertext letter $a_{21} = V$.

5.5.4.4 Matrix Substitution

Let $A = \{a_0, a_1, \ldots, a_{n-1}\}$ be an alphabet and $S = (s_{ij}), s_{ij} \in \{0, 1, \ldots, m-1\}$, be a non-singular matrix of type (m, m) with gcd(detS, n)= 1. The mapping which maps the block of plaintext $a_{t(1)}$, $a_{t(2)}, \ldots, a_{t(m)}$ to the ciphertext determined by the vector (all arithmetic modulo n, vectors transposed as required)

$$\left(S \cdot \begin{pmatrix} a_{t(1)} \\ a_{t(2)} \\ \vdots \\ a_{t(m)} \end{pmatrix}\right)^T \tag{5.274}$$

is called the Hill cipher. This represents a monoalphabetic matrix substitution.

■ $S = \begin{pmatrix} 14 & 8 & 3 \\ 8 & 5 & 2 \\ 3 & 2 & 1 \end{pmatrix}$. Let the letters of the alphabet be enumerated $a_0 =$ A, $a_1 =$B, $\ldots, a_{25} =$ Z. For $m = 3$ and the plaintext AUTUMN, the strings AUT and UMN correspond to the vectors $(0, 20, 19)$ and $(20, 12, 13)$.

Then $S \cdot (0, 20, 19)^\top = (217, 138, 59)^\top \equiv (9, 8, 7)^\top(\text{mod}26)$ and $S \cdot (20, 12, 13)^\top = (415, 246, 97)^\top \equiv (25, 12, 19)^\top(\text{mod}26)$. Thus, the plaintext AUTUMN is mapped to the ciphertext JIHZMT.

5.5.5 Methods of Classical Cryptanalysis

The purpose of cryptanalytical investigations is to deduce from the ciphertext an optimum of information about the corresponding plaintext without knowing the key. These analyses are of interest not only to an unauthorized “eavesdropper” but also help assess the security of cryptosystems from the user’s point of view.

5.5.5.1 Statistical Analysis

Each natural language shows a typical frequency distribution of the individual letters, two-letter combinations, words, etc. For example, in English the letter e is used most frequently:

Letter	Relative frequency
E,	12.7 %
T, A, O, I, N, S, H, R	56.9 %
D, L	8.3 %
C, U, M, W, F, G, Y, P, B	19.9 %
V, K, J, X, Q, Z	2.2 %

Given sufficiently long ciphertexts it is possible to break a monoalphabetic, monographic substitution on the basis of the frequency distribution of letters.

5.5.5.2 Kasiski-Friedman Test

Combining the methods of Kasiski and Friedman it is possible to break the Vignère cipher. The attack benefits from the fact that the encryption algorithm applies the key periodically. If the same string of plaintext letters is encrypted with the same portion of the key then the same string of ciphertext letters will be produced. A length > 2 of the distance of such identical strings in the ciphertext must be a multiple of the key length. In the case of several reoccurring strings of ciphertext the key length is a divisor of the greatest common divisor of all distances. This reasoning is called the Kasiski test. One should, however, be aware of erroneous conclusions due to the possibility that matches may occur accidentally.

The Kasiski test permits the determination of the key length at most as a multiple of the true key length. The Friedman test yields the magnitude of the key length. Let n be the length of the ciphertext of some English plaintext encrypted by means of the Vignère method. Then the key length l is determined by

$$l = \frac{0.027n}{(n-1)\mathrm{IC} - 0.038\mathrm{n} + 0.065}. \tag{5.275a}$$

Here IC denotes the coincidence index of the ciphertext. This index can be deduced from the number n_i of occurrences of the letter a_i $(i \in \{0, 1, \ldots, 25\})$ in the ciphertext:

$$\mathrm{IC} = \frac{\sum\limits_{\mathrm{i}=1}^{26} \mathrm{n_i}(\mathrm{n_i} - 1)}{\mathrm{n}(\mathrm{n}-1)}. \tag{5.275b}$$

In order to determine the key, the ciphertext of length n is split into l columns. Since the Vignère cipher produces the contents of each column by means of a shift cipher, it suffices to determine the equivalence of E on a column base. Should V be the most frequent letter within a column, then the Vignère tableau points to the letter R

$$\begin{array}{c} \mathrm{E} \\ \vdots \\ \mathrm{R}\ldots\mathrm{V} \end{array} \tag{5.275c}$$

of the key. The methods described so far will not be successful if the Vignère cipher employs very long keys (e.g., as long as the plaintext). It is, however, possible to deduce whether the applied cipher is monoalphabetic, polyalphabetic with short period or polyalphabetic with long period.

5.5.6 One-Time Pad

The one-time pad is the only substitution cipher that is considered theoretically secure. The encryption adheres to the principle of the Vignère cipher, where the key is a random string of letters as long as the plaintext.

Usually, one-time pads are applied as binary Vignère ciphers: Plaintext and ciphertext are represented as binary numbers with addition modulo 2. In this particular case the cipher is involutory, which means that the twofold application of the cipher restores the original plaintext. A concrete implementation of the binary Vignère cipher is based on shift register circuits. These circuits combine switches and storage elements, whose states are 0 or 1, according to special rules.

5.5.7 Public Key Methods

Although the methods of conventional encryption can have efficient implementations with today's computers, and although only a single key is needed for bidirectional communication, there are a number of drawbacks:

1. The security of encryption solely depends on keeping the next key secret.

2. Prior to any communication, the key must be exchanged via a sufficiently secured channel; spontaneous communication is ruled out.

3. Furthermore, no means exist to prove to a third party that a specific message was sent by an identified sender.

5.5.7.1 Diffie-Hellman Key Exchange

The concept of encryption with public keys was developed by Diffie and Hellman in 1976. Each participant owns two keys: a public key that is published in a generally accessible register, and a private key that is solely known to the participant and kept absolutely secret. Methods with these properties are called asymmetric ciphers (see 5.5.2, p. 387).

The public key KP_i of the i-th participant controls the encryption step E_i, his private key KS_i the decryption step D_i. The following conditions must be fulfilled:

1. $D_i \circ E_i$ constitutes the identity.

2. Efficient implementations for E_i and D_i are known.

3. The private key KS_i cannot be deduced from the public key KP_i with the means available in the foreseeable future. If in addition

4. also $E_i \circ D_i$ yields the identity,

then the encryption algorithm qualifies as an electronic signature method with public keys. The electronic signature method permits the sender to attach a tamperproof signature to a message.

If A wants to send an encrypted message m to B, then A retrieves B's public key KP_B from the register, applies the encryption algorithm E_B, and calculates $E_B(m) = c$. A sends the ciphertext c via the public network to B who will regain the plaintext of the message by decrypting c using his private key KS_B in the decryption function D_B: $D_B(c) = D_B(E_B(m)) = m$. In order to prevent tampering of messages, A can electronically sign his message m to B by complying with an electronic signature method with the public key in the following way: A encrypts the message m with his private key: $D_A(m) = d$. A attaches to d his signature "A" and encrypts the total using the public key of B: $E_B(D_A(m),$ "A"$) = E_B(d,$ "A"$) = e$. The text thus signed and encrypted is sent from A to B.

The participant B decrypts the message with his private key and obtains $D_B(e) = D_B(E_B(d,$ "A"$))$ $= (d,$ "A"$)$. Based on this text B can identify A as the sender and can now decrypt d using the public key of $A : E_A(d) = E_A(D_A(m)) = m$.

5.5.7.2 One-Way Function

The encryption algorithms of a method with public key must constitute a one-way function with a "trap door". A trap door in this context is some special, additional information that must be kept secret. An injective function $f\colon X \longrightarrow Y$ is called a one-way function with a trap door, if the following conditions hold:

1. There is an efficient method to compute both f and f^{-1}.

2. The calculation of f^{-1} cannot be deduced from f without the knowledge of the secret additional information.

The efficient method to get f^{-1} from f cannot be made without the secret additional information.

5.5.7.3 RSA Codes and RSA Method

1. RSA Codes

Rivest, Shamir and Adleman (see [5.16]) developed an encryption scheme for secret messages on the basis of the Euler-Fermat theorem (see 5.4.4, **2.**, p. 381). The scheme is called the *RSA algorithm* after the initials of their last names. Part of the key required for decryption can be made public without endangering the confidentiality of the message; for this reason, the term *public key code* is used in this context as well.

In order to apply the RSA algorithm the recipient B chooses two very large prime numbers p and q, calculates $m = pq$ and selects a number r relatively prime to $\varphi(m) = (p-1)(q-1)$ and $1 < r < \varphi(m)$. B publishes the numbers m and r because they are needed for decryption.

For transmitting a secret message from sender A to recipient B the text of the message must be converted first to a string of digits that will be split into N blocks of the same length of less than 100 decimal positions. Now A calculates the remainder R of N^r divided by m.

$$N^r \equiv R(m). \tag{5.276a}$$

Sender A calculates the number R for each of the blocks N that were derived from the original text and sends the number to B. The recipient can decipher the message R if he has a solution of the linear congruence $rs \equiv 1\,(\varphi(m))$. The number N is the remainder of R^s divided by m:

$$R^s \equiv (N^r)^s \equiv N^{1+k\varphi(m)} \equiv N \cdot (N^{\varphi(m)})^k \equiv N(m). \tag{5.276b}$$

Here, the Euler-Fermat theorem (see 5.4.4, **2.**, p. 381) with $N^{\varphi(m)} \equiv 1(m)$ has been applied. Eventually, B converts the sequence of numbers into text.

■ A recipient B who expects a secret message from sender A chooses the prime numbers $p = 29$ and $q = 37$ (actually too small for practical purposes), calculates $m = 29 \cdot 37 = 1073$ (and $\varphi(1073) = \varphi(29) \cdot \varphi(37) = 1008)$), and chooses $r = 5$ (it satisfies the requirement of $\gcd(1008, 5) = 1$). B passes the values $m = 1073$ and $r = 5$ to A.

A intends to send the secret message $N = 8$ to B. A encrypts N into $R = 578$ by calculating $N^r = 8^5 \equiv 578\,(1073)$, and just sends the value $R = 578$ to B. B solves the congruence $5 \cdot s \equiv 1\,(1008)$, arrives at the solution $s = 605$, and thus determines $R^s = 578^{605} \equiv 8 = N\,(1073)$.

Remark: The security of the RSA code correlates with the time needed by an unauthorized listener to factorize m. Assuming the speed of today's computers, a user of the RSA algorithm should choose the two prime numbers p and q with at least a length of 100 decimal positions in order to impose a decryption effort of approximately 74 years on the unauthorized listener. The effort for the authorized user, however, to determine an r relatively prime to $\varphi(pq) = (p-1)(q-1)$ is comparatively small.

2. RSA Method

The RSA method is the most popular asymmetric encryption method.

1. Assumptions Let p and q be two large prime numbers with $pq \approx 10^{2048}$ and $n = pq$. The number of decimal positions of p and q should differ by a small number; yet, the difference between p and q should not be too large. Furthermore, the numbers $p-1$ and $q-1$ should contain rather big prime factors, while the greatest common divisor of $p-1$ and $q-1$ should be rather small. Let $e > 1$ be relatively prime to $(p-1)(q-1)$ and let d satisfy $d \cdot e \equiv 1\,(\mathrm{mod}(p-1)(q-1))$. Now n and e represent the public key and d the private key.

2. Encryption Algorithm

$$E\colon \{0,1,\ldots,n-1\} \to \{0,1,\ldots,n-1\} \quad E(x) := x^e \bmod n. \tag{5.277a}$$

3. Decyphering Operations

$$D\colon \{0,1,\ldots,n-1\} \to \{0,1,\ldots,n-1\} \quad D(x) := x^d \bmod n. \tag{5.277b}$$

Thus $D(E(m)) = E(D(m) = m$ for message m.

The function in this encryption method with $n > 10^{200}$ constitutes a candidate for a one-way function with trap door (see 5.5.7.2, p. 391). The required additional information is the knowledge of how to factor n. Without this knowledge it is infeasible to solve the congruence $d \cdot e \equiv 1$ (modulo $(p-1)(q-1)$).

The RSA method is considered practically secure as long as the above conditions are met. A disadvantage in comparison with other methods is the relatively large key size and the fact that RSA is 1000 times slower than DES.

5.5.8 DES Algorithm (Data Encryption Standard)

The DES method was adopted in 1976 by the National Bureau of Standards (now NIST) as the official US encryption standard. The algorithm belongs to the class of symmetric encryption methods (see 5.5.2, p. 387) and still plays a predominant role among cryptographic methods. The method is, however, no longer suited for the encryption of top secret information because today's technical means permit an attack by an exhaustive test trying all keys.

The DES algorithm combines permutations and non-linear substitutions. The algorithm requires a 56-bit key. Actually, a 64-bit key is used, however, only 56 bits can freely be chosen; the remaining eight bits serve as parity bits, one for each of the seven-bit blocks to yield odd parity.

The plaintext is split into blocks of 64 bits each. DES transforms each 64-bit plaintext block into a ciphertext block of 64 bits. First, the plaintext block will be subject to an initial permutation and is then encrypted in 16 rounds, each operating with a different subkey $K_1, K_2, \ldots, K_{16}$. The encryption completes with a final permutation that is the inverse of the initial permutation.

Decryption uses the same algorithm with the difference that the subkeys are employed in reverse order $K_{16}, K_{15}, \ldots, K_1$.

The strength of the cipher rests on the nature of the mappings that are part of each round. It can be shown that each bit of the ciphertext block depends on each bit of the corresponding plaintext and on each bit of the key.

Although the DES algorithm has been disclosed in full detail, no attack has been published so far that can break the algorithm without an exhaustive test of all 256 keys.

5.5.9 IDEA Algorithm (International Data Encryption Algorithm)

The IDEA algorithm was developed by LAI and MASSAY and patented 1991. It is a symmetric encryption method similar to the DES algorithm and constitutes a potential successor to DES. IDEA became known as part of the reputed software package PGP (Pretty Good Privacy) for the encryption of emails. In contrast to DES not only was the algorithm published but even its basic design criteria. The objective was the use of particularly simple operations (addition modulo 2, addition modulo 2^{16}, multiplication modulo 2^{16+1}).

IDEA works with keys of 128 bits length. IDEA encrypts plaintext blocks of 64 bits each. The algorithm splits a block into four subblocks of 16 bits each. From the 128-bit key 52 subkeys are derived, each 16 bits long. Each of the eight encryption rounds employs six subkeys; the remaining four subkeys are used in the final transformation which constructs the resulting 64-bit ciphertext. Decryption uses the same algorithm with the subkeys in reverse order.

IDEA is twice as fast as DES, its implementation in hardware, however, is more difficult. No successful attack against IDEA is known. Exhaustive attacks trying all 2^{56} keys are infeasible considering the length of the keys.

5.6 Universal Algebra

A *universal algebra* consists of a set, the *underlying set*, and operations on this set. Simple examples are semigroups, groups, rings, and fields discussed in sections 5.3.2, p. 336; 5.3.3, p. 336 and 5.3.7, p. 361. Universal algebras (mostly many-sorted, i.e., with several underlying sets) are handled especially in theoretical informatics. There they form the basis of algebraic specifications of abstract data types and systems and of term-rewriting systems.

5.6.1 Definition

Let Ω be a set of operation symbols divided into pairwise disjoint subsets Ω_n, $n \in \mathbb{N}$. Ω_0 contains the constants, Ω_n, $n > 0$, contain the n-ary operation symbols. The family $(\Omega_n)_{n\in\mathbb{N}}$ is called the *type* or *signature*. If A is a set, and if to every n-ary operation symbol $\omega \in \Omega_n$ an n-ary operation ω^A in A is assigned, then $A = (A, \{\omega^A|\omega \in \Omega\})$ is called an Ω *algebra* or algebra of type (or of signature) Ω.
If Ω is finite, $\Omega = \{\omega_1, \dots, \omega_k\}$, then one also writes $A = (A, \omega_1^A, \dots, \omega_k^A)$ for A.
If a ring (see 5.3.7, p. 361) is considered as an Ω algebra, then Ω is partitioned $\Omega_0 = \{\omega_1\}$, $\Omega_1 = \{\omega_2\}$, $\Omega_2 = \{\omega_3, \omega_4\}$, where to the operation symbols ω_1, ω_2, ω_3, ω_4 the constant 0, taking the inverse with respect to addition, addition and multiplication are assigned.

Let A and B be Ω algebras. B is called an Ω *subalgebra* of A, if $B \subseteq A$ holds and the operations ω^B are the restrictions of the operations ω^A $(\omega \in \Omega)$ to the subset B.

5.6.2 Congruence Relations, Factor Algebras

In constructing factor structures for universal algebras, the notion of congruence relation is needed. A congruence relation is an equivalence relation compatible with the structure: Let $A = (A, \{\omega^A|\omega \in \Omega\})$ be an Ω algebra and R be an equivalence relation in A. R is called a *congruence relation* in A, if for all $\omega \in \Omega_n$ $(n \in \mathbb{N})$ and all $a_i, b_i \in A$ with a_iRb_i $(i = 1, \dots, n)$:

$$\omega^A(a_1, \dots, a_n)\; R\; \omega^A(b_1, \dots, b_n). \tag{5.278}$$

The set of equivalence classes (factor set) with respect to a congruence relation also form an Ω algebra with respect to representative-wise calculations: Let $A = (A, \{\omega^A|\omega \in \Omega\})$ be an Ω algebra and R be a congruence relation in A. The factor set A/R (see 5.2.4, **2.**, p. 334) is an Ω algebra A/R with the following operations $\omega^{A/R}$ $(\omega \in \Omega_n, n \in \mathbb{N})$ with

$$\omega^{A/R}([a_1]_R, \dots, [a_n]_R) = [\omega^A(a_1, \dots, a_n)]_R \tag{5.279}$$

and it is called the *factor algebra* of A with respect to R.

The congruence relations of groups and rings can be defined by special substructures – normal subgroups (see 5.3.3.2, **2.** p. 338) and ideals (see 5.3.7.2, p. 362), respectively. In general, e.g., in semigroups, such a characterization of congruence relations is not possible.

5.6.3 Homomorphism

Just as with classical algebraic structures, the homomorphism theorem gives a connection between the homomorphisms and congruence relations.
Let A and B be Ω algebras. A mapping $h\colon A \to B$ is called a *homomorphism*, if for every $\omega \in \Omega_n$ and all $a_1, \dots, a_n \in A$:

$$h(\omega^A(a_1, \dots, a_n)) = \omega^B(h(a_1), \dots, h(a_n)). \tag{5.280}$$

If, in addition, h is bijective, then h is called an *isomorphism*; the algebras A and B are called *isomorphic*. The homomorphic image $h(A)$ of an Ω algebra A is an Ω subalgebra of B. Under a homomorphism h, the decomposition of A into subsets of elements with the same image corresponds to a congruence relation which is called the *kernel* of h:

$$\ker\; h = \{(a, b) \in A \times A | h(a) = h(b)\}. \tag{5.281}$$

5.6.4 Homomorphism Theorem

Let A and B be Ω algebras and h: $A \to B$ a homomorphism. h defines a congruence relation ker h in A. The factor algebra $A/\ker\ h$ is isomorphic to the homomorphic image $h(A)$.
Conversely, every congruence relation R defines a homomorphic mapping nat_R: $A \to A/R$ with $nat_R(a) = [a]_R$. **Fig. 5.19** illustrates the homomorphism theorem.

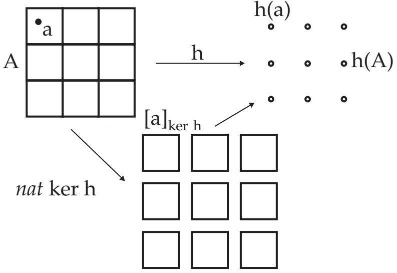


Figure 5.19

5.6.5 Varieties

A *variety* V is a class of Ω algebras, which is closed under forming direct products, subalgebras, and homomorphic images, i.e., these formations do not lead out of V. Here the direct products are defined in the following way:

Considering the operations corresponding to Ω componentwise on the Cartesian product of the underlying sets of Ω algebras, an Ω algebra, the *direct product* of these algebras is obtained. The theorem of Birkhoff (see 5.6.6, p. 395) characterizes the varieties as those classes of Ω algebras, which can be *equationally defined*.

5.6.6 Term Algebras, Free Algebras

Let $(\Omega_n)_{n\in\mathbb{N}}$ be a type (signature) and X a countable set of variables. The set $T_\Omega(X)$ of Ω terms over X is defined inductively in the following way:

1. $X \cup \Omega_0 \subseteq T_\Omega(X)$.

2. If $t_1, \ldots, t_n \in T_\Omega(X)$ and $\omega \in \Omega_n$ hold, then also $\omega t_1 \ldots t_n \in T_\Omega(X)$ holds.

The set $T_\Omega(X)$ defined in this way is an underlying set of an Ω algebra, the *term algebra* $T_\Omega(X)$ of type Ω over X, with the following operations: If $t_1, \ldots, t_n \in T_\Omega(X)$ and $\omega \in \Omega_n$ hold, then $\omega^{T_\Omega(X)}$ is defined by

$$\omega^{T_\Omega(X)}(t_1, \ldots, t_n) = \omega t_1 \ldots t_n. \tag{5.282}$$

Term algebras are the “most general” algebras in the class of all Ω algebras, i.e., no “identities” are valid in term algebras. These algebras are called *free algebras*.
An *identity* is a pair $(s(x_1, \ldots, x_n), t(x_1, \ldots, x_n))$ of Ω terms in the variables $x_1, \ldots, x_n$. An Ω algebra A *satisfies* such an equation, if for every $a_1, \ldots, a_n \in A$ holds:

$$s^A(a_1, \ldots, a_n) = t^A(a_1, \ldots, a_n). \tag{5.283}$$

A class of Ω algebras defined by identities is a class of Ω algebras satisfying a given set of identities.

Theorem of Birkhoff: The classes defined by identities are exactly the varieties.

■ Varieties are for example the classes of all semigroups, groups, Abelian groups, and rings. But, e.g., the direct product of cyclic groups is not a cyclic group, and the direct product of fields is not a field. Therefore cyclic groups or fields do not form a variety, and cannot be defined by equations.

5.7 Boolean Algebras and Switch Algebra

Calculating rules, similar to the rules established in 5.2.2, **3.**, p. 329 for set algebra and propositional calculus (5.1.1, **6.**, p. 324), can be found for other objects in mathematics too. The investigation of these rules yields the notion of Boolean algebra.

5.7.1 Definition

A set B, together with two binary operations $\sqcap$ (“conjunction”) and $\sqcup$ (“disjunction”), and a unary operation (“negation”), and two distinguished (neutral) elements 0 and 1 from B, is called a *Boolean* *algebra* $B = (B, \sqcap, \sqcup, \bar{}\,, 0, 1)$ if the following properties are valid:

(1) Associative Laws:

$$(a \sqcap b) \sqcap c = a \sqcap (b \sqcap c), \quad (5.284) \qquad (a \sqcup b) \sqcup c = a \sqcup (b \sqcup c). \quad (5.285)$$

(2) Commutative Laws:

$$a \sqcap b = b \sqcap a, \quad (5.286) \qquad a \sqcup b = b \sqcup a. \quad (5.287)$$

(3) Absorption Laws:

$$a \sqcap (a \sqcup b) = a, \quad (5.288) \qquad a \sqcup (a \sqcap b) = a. \quad (5.289)$$

(4) Distributive Laws:

$$(a \sqcup b) \sqcap c = (a \sqcap c) \sqcup (b \sqcap c), \quad (5.290) \qquad (a \sqcap b) \sqcup c = (a \sqcup c) \sqcap (b \sqcup c). \quad (5.291)$$

(5) Neutral Elements:

$$a \sqcap 1 = a, \quad (5.292) \qquad a \sqcup 0 = a, \quad (5.293)$$

$$a \sqcap 0 = 0, \quad (5.294) \qquad a \sqcup 1 = 1, \quad (5.295)$$

(6) Complement:

$$a \sqcap \overline{a} = 0, \quad (5.296) \qquad a \sqcup \overline{a} = 1. \quad (5.297)$$

A structure with the associative laws, commutative laws, and absorption laws is called a *lattice*. If the distributive laws also hold, then the lattice is called a *distributive lattice*. So a Boolean algebra is a special distributive lattice.

Remark: The notation used for Boolean algebras is not necessarily identical to the notation for the operations in propositional calculus.

5.7.2 Duality Principle

1. Dualizing

In the "axioms" of a Boolean algebra is included the following duality: Replacing $\sqcap$ by $\sqcup$, $\sqcup$ by $\sqcap$, 0 by 1, and 1 by 0 in an axiom gives always the other axiom in the same row. The axioms in a row are *dual* to each other, and the substitution process is called *dualization*. The *dual statement* follows from a statement of the Boolean algebra by dualization.

2. Duality Principle for Boolean Algebras

The dual statement of a true statement for a Boolean algebra is also a true statement for the Boolean algebra, i.e., with every proved proposition, the dual proposition is also proved.

3. Properties

One gets, e.g., the following properties for Boolean algebras from the axioms.

(E1) The Operations $\sqcap$ and $\sqcup$ are Idempotent:

$$a \sqcap a = a, \quad (5.298) \qquad a \sqcup a = a. \quad (5.299)$$

(E2) De Morgan Rules:

$$\overline{a \sqcap b} = \overline{a} \sqcup \overline{b}\,, \quad (5.300) \qquad \overline{a \sqcup b} = \overline{a} \sqcap \overline{b}\,, \quad (5.301)$$

(E3) A further Property:

$$\overline{\overline{a}} = a. \quad (5.302)$$

It is enough to prove only one of the two properties in any line above, because the other one is the dual property. The last property is self-dual.

5.7.3 Finite Boolean Algebras

All finite Boolean algebras can be described easily up to "isomorphism". Let B_1, B_2 be two Boolean algebras and $f\colon B_1 \to B_2$ a bijective mapping. f is called an *isomorphism* if

$$f(a \sqcap b) = f(a) \sqcap f(b), \quad f(a \sqcup b) = f(a) \sqcup f(b) \quad \text{and} \quad f(\overline{a}) = \overline{f(a)} \tag{5.303}$$

hold. Every finite Boolean algebra is isomorphic to the Boolean algebra of the power set of a finite set. In particular every finite Boolean algebra has 2^n elements, and every two finite Boolean algebras with the same number of elements are isomorphic.

Hereafter B denotes the Boolean algebra with two elements $\{0, 1\}$ and with the operations

$\sqcap$	0	1
0	0	0
1	0	1

$\sqcup$	0	1
0	0	1
1	1	1

	$-$
0	1
1	0

Defining the operations $\sqcap$, $\sqcup$, and $\bar{}$ componentwise on the n-times Cartesian product $B^n = \{0, 1\} \times \cdots \times \{0, 1\}$, then B^n will be a Boolean algebra with $0 = (0, \ldots, 0)$ and $1 = (1, \ldots, 1)$. B^n is called the *n times direct product* of B. Because B^n contains 2^n elements, this way one gets *all* finite Boolean algebras (out of isomorphism).

5.7.4 Boolean Algebras as Orderings

An order relation can be assigned to every Boolean algebra B: Here $a \le b$ holds if $a \sqcap b = a$ is valid (or equivalently, if $a \sqcup b = b$ holds).
So every finite Boolean algebra can be represented by a Hasse diagram (see 5.2.4, **4.**, p. 334).

■ Suppose B is the set $\{1, 2, 3, 5, 6, 10, 15, 30\}$ of the divisors of 30. Then, the least common multiple and the greatest common divisor can be defined as binary operations and the complement as unary operation. The numbers 1 and 30 correspond to the distinguished elements 0 and 1. The corresponding Hasse diagram is shown in **Fig. 5.20**.

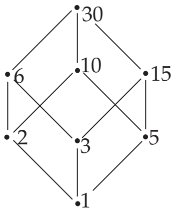


Figure 5.20

5.7.5 Boolean Functions, Boolean Expressions

1. Boolean Functions

Denoting by B the Boolean algebra with two elements as in 5.7.3, p. 397, then an *n-ary Boolean function* f is a mapping from B^n into B. There are 2^{2^n} n-ary Boolean functions. The set of all n-ary Boolean functions with the operations

$$(f \sqcap g)(b) = f(b) \sqcap g(b), \tag{5.304}$$

$$(f \sqcup g)(b) = f(b) \sqcup g(b), \tag{5.305}$$

$$\overline{f}(b) = \overline{f(b)}, \tag{5.306}$$

is a Boolean algebra. Here b always means an n tuple of the elements of $B = \{0, 1\}$, and on the right-hand side of the equations the operations are performed in B. The distinguished elements 0 and 1 correspond to the functions f_0 and f_1 with

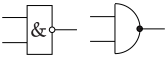
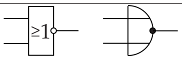
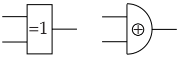
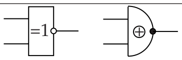
$$f_0(b) = 0, \quad f_1(b) = 1 \quad \text{for all} \quad b \in B^n. \tag{5.307}$$

■ **A:** In the case $n = 1$, i.e., for only one Boolean variable b, there are four Boolean functions:

$$\begin{array}{llll} \text{Identity} & f(b) = b, & \text{Negation} & f(b) = \overline{b}, \\ \text{Tautology} & f(b) = 1, & \text{Contradiction} & f(b) = 0. \end{array} \tag{5.308}$$

■ **B:** In the case $n = 2$, i.e., for two Boolean variables a and b, there are 16 different Boolean functions, among which the most important ones have their own names and notation. They are shown in **Table 5.6**.

Table 5.6 Some Boolean functions with two variables a and b

Name of the function	**Different notation**	**Different symbols**	**Value table for** $\begin{pmatrix}a\\b\end{pmatrix}=\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}0\\1\end{pmatrix},\begin{pmatrix}1\\0\end{pmatrix},\begin{pmatrix}1\\1\end{pmatrix}$
Sheffer or NAND	$\overline{a\cdot b}$ $a \mid b$ NAND (a,b)		$1,\quad 1,\quad 1,\quad 0$
Peirce or NOR	$\overline{a+b}$ $a \downarrow b$ NOR a,b		$1,\quad 0,\quad 0,\quad 0$
Antivalence or XOR	$\overline{a}\,b \quad + \quad a\,\overline{b}$ $a\,\mathrm{XOR}\,b$ $a \not\equiv b$ $a \oplus b$		$0,\quad 1,\quad 1,\quad 0$
Equivalence	$\overline{a}\,\overline{b}+a\,b$ $a \equiv b$ $a \leftrightarrow b$		$1,\quad 0,\quad 0,\quad 1$
Implication	$\overline{a}+b$ $a \rightarrow b$		$1,\quad 1,\quad 0,\quad 1$

2. Boolean Expressions

Boolean expressions are defined in an inductive way: Let $X = \{x, y, z, \ldots\}$ be a (countable) set of *Boolean variables* (which can take values only from $\{0,1\}$):

1. The constants 0 and 1 just as the Boolean variables from X are Boolean expressions. (5.309)

2. If S and T are Boolean expressions, so are $\overline{T}$, $(S \sqcap T)$, and $(S \sqcup T)$, as well. (5.310)

If a Boolean expression contains the variables $x_1, \ldots, x_n$, then it represents an n-ary Boolean function f_T:
Let b be a "valuation" of the Boolean variables $x_1, \ldots, x_n$, i.e., $b = (b_1, \ldots, b_n) \in B^n$.
Assigning a Boolean function to the expression T in the following way gives:

1. If $T = 0,$ then $f_T = f_0;$ if $T = 1\,,$ then $f_T = f_1.$ (5.311a)

2. If $T = x_i,$ then $f_T(b) = b_i;$ if $T = \overline{S},$ then $f_T(b) = \overline{f_S(b)}.$ (5.311b)

3. If $T = R \sqcap S,$ then $f_T(b) = f_R(b) \sqcap f_S(b).$ (5.311c)

4. If $T = R \sqcup S,$ then $f_T(b) = f_R(b) \sqcup f_S(b).$ (5.311d)

On the other hand, every Boolean function f can be represented by a Boolean expression T (see 5.7.6, p. 399).

3. Concurrent or Semantically Equivalent Boolean Expressions

The Boolean expressions S and T are called *concurrent* or *semantically equivalent* if they represent the same Boolean function. Boolean expressions are equal if and only if they can be transformed into each other according to the axioms of a Boolean algebra.
Under transformations of a Boolean expression here are considered especially two aspects:

- Transformation in a possible "simple" form (see 5.7.7, p. 399).
- Transformation in a "normal form".

5.7.6 Normal Forms

1. Elementary Conjunction, Elementary Disjunction

Let $B = (B,\sqcap,\sqcup,{}^{-},0,1)$ be a Boolean algebra and $\{x_1,\ldots,x_n\}$ a set of Boolean variables. Every conjunction or disjunction in which every variable or its negation occurs exactly once is called an *elementary conjunction* or an *elementary disjunction* respectively (in the variables $x_1,\ldots,x_n$).
Let $T(x_1,\ldots,x_n)$ be a Boolean expression. A disjunction D of elementary conjunctions with $D = T$ is called a *principal disjunctive normal form* (*PDNF*) of T. A conjunction C of elementary disjunctions with $C = T$ is called a *principal conjunctive normal form* (*PCNF*) of T.

■ **Part 1:** In order to show that every Boolean function f can be represented as a Boolean expression, the PDNF form of the function f given in the annexed table is to be constructed:

x	y	z	$f(x,y,z)$
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	0

The PDNF of the Boolean function f contains the elementary conjunctions $\overline{x}\sqcap\overline{y}\sqcap z$, $x\sqcap\overline{y}\sqcap z$, $x\sqcap y\sqcap\overline{z}$. These elementary conjunctions belong to the valuations b of the variables where the function f has the value 1. If a variable v has the value 1 in b, then v is to put in the elementary conjunction, otherwise $\overline{v}$.

■ **Part 2:** The PDNF for the example of Part 1 is:

$$(\overline{x}\sqcap\overline{y}\sqcap z)\sqcup(x\sqcap\overline{y}\sqcap z)\sqcup(x\sqcap y\sqcap\overline{z}). \tag{5.312}$$

The “dual” form for PDNF is the PCNF: The elementary disjunctions belong to the valuations b of the variables for which f has the value 0. If a variable v has the value 0 in b, then v is to put in the elementary disjunction, otherwise $\overline{v}$. So the PCNF is:

$$(x\sqcup y\sqcup z)\sqcap(x\sqcup\overline{y}\sqcup z)\sqcap(x\sqcup\overline{y}\sqcup\overline{z})\sqcap(\overline{x}\sqcup y\sqcup z)\sqcap(\overline{x}\sqcup\overline{y}\sqcup\overline{z}). \tag{5.313}$$

The PDNF and the PCNF of f are uniquely determined, if the ordering of the variables and the ordering of the valuations is given, e.g., if considering the valuations as binary numbers and arranging them in increasing order.

2. Principal Normal Forms

The principal normal form of a Boolean function f_T is considered as the principal normal form of the corresponding Boolean expression T.
Checking the equivalence of two Boolean expressions by transformations is often difficult. The principal normal forms are useful: Two Boolean expressions are semantically equivalent exactly if their corresponding uniquely determined principal normal forms are identical letter by letter.

■ **Part 3:** In the considered example (see Part 1 and 2) the expressions $(\overline{y}\sqcap z)\sqcup(x\sqcap y\sqcap\overline{z})$ and $(x\sqcup((y\sqcup z)\sqcap(\overline{y}\sqcup z)\sqcap(\overline{y}\sqcup\overline{z})))\sqcap(\overline{x}\sqcup((y\sqcup z)\sqcap(\overline{y}\sqcup z)))$ are semantically equivalent because the principal disjunctive (or conjunctive) normal forms of both are the same.

5.7.7 Switch Algebra

A typical application of Boolean algebra is the simplification of series–parallel connections (SPC). Therefore a Boolean expression is to be assigned to a SPC (transformation). This expression will be “simplified” with the transformation rules of the Boolean algebra. Finally a SPC is to be assigned to this expression (inverse transformation). The result is a simplified SPC which produces the same behavior as the initial connection system **(Fig. 5.21)**.

A SPC has two types of contact points: the so-called “make contacts” and “break contacts”, and both types have two states; namely open or closed. The usual symbolism is: When the equipment is put on, the make contacts close and the break contacts open. With Boolean variables assigned to the contacts of the switch equipment follows:
The position “off” or “on” of the equipment corresponds to the value 0 or 1 of the Boolean variables. The contacts being switched by the same equipment are denoted by the same symbol, the Boolean variable belonging to this equipment. The *contact value* of a SPC is 0 or 1, according to whether the switch is electrically non-conducting or conducting. The contact value depends on the position of the

contacts, so it is a Boolean function S (*switch function*) of the variables assigned to the switch equipment. Contacts, connections, symbols, and the corresponding Boolean expressions are represented in **Fig. 5.22**.

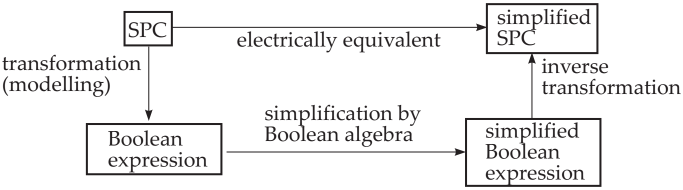

SPC
electrically equivalent
simplified SPC
transformation (modelling)
inverse transformation
simplification by Boolean algebra
Boolean expression
simplified Boolean expression

Figure 5.21

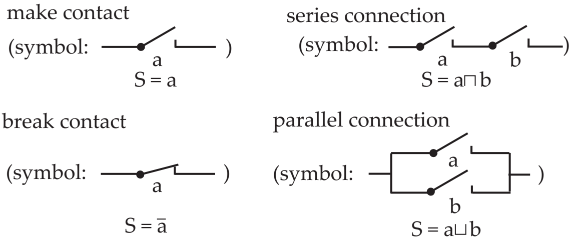


Figure 5.22

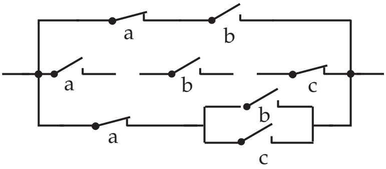


Figure 5.23

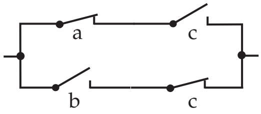


Figure 5.24

The Boolean expressions, which represent switch functions of SPC, have the special property that the negation sign can occur *only* above variables (never over subexpressions).

■ Simplification of the SPC **Fig. 5.23**. This connection corresponds to the Boolean expression

$$S = (\overline{a} \sqcap b) \sqcup (a \sqcap b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap (b \sqcup c)) \tag{5.314}$$

as switch function. According to the transformation formulas of Boolean algebra holds:

$$\begin{aligned} S &= (b \sqcap (\overline{a} \sqcup (a \sqcap \overline{c}))) \sqcup (\overline{a} \sqcap (b \sqcup c)) \\ &= (b \sqcap (\overline{a} \sqcup \overline{c})) \sqcup (\overline{a} \sqcap (b \sqcup c)) \\ &= (\overline{a} \sqcap b) \sqcup (b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap c) \\ &= (\overline{a} \sqcap b \sqcap c) \sqcup (\overline{a} \sqcap b \sqcap \overline{c}) \sqcup (b \sqcap \overline{c}) \sqcup (a \sqcap b \sqcap \overline{c}) \sqcup (\overline{a} \sqcap c) \sqcup (\overline{a} \sqcap \overline{b} \sqcap c) \\ &= (\overline{a} \sqcap c) \sqcup (b \sqcap \overline{c}). \end{aligned} \tag{5.315}$$

Here one gets $\overline{a} \sqcap c$ from $(\overline{a} \sqcap b \sqcap c) \sqcup (\overline{a} \sqcap c) \sqcup (\overline{a} \sqcap \overline{b} \sqcap c)$, and $b \sqcap \overline{c}$ from $(\overline{a} \sqcap b \sqcap \overline{c}) \sqcup (b \sqcap \overline{c}) \sqcup (a \sqcap b \sqcap \overline{c})$. The finally simplified result SPC is shown in **Fig. 5.24**.

This example shows that usually it is not so easy to get the simplest Boolean expression by transformations. In the literature one can find different methods for this procedure.

5.8 Algorithms of Graph Theory

Graph theory is a field in discrete mathematics having special importance for informatics, e.g., for representing data structures, finite automata, communication networks, derivatives in formal languages, etc. There are also applications in physics, chemistry, electrotechnics, biology and psychology. Moreover, flows can be applied in transport networks and in network analysis in operations research and in combinatorial optimization.

5.8.1 Basic Notions and Notation

1. Undirected and Directed Graphs

A *graph* G is an ordered pair (V, E) of a set V of *vertices* and a set E of *edges*. There is a mapping, defined on E, the *incidence function*, which uniquely assigns to every element of E an ordered or non-ordered pair of (not necessarily distinct) elements of V. If a non-ordered pair is assigned then G is called an *undirected graph* **(Fig. 5.25)**. If an ordered pair is assigned to every element of E, then the graph is called a *directed graph* **(Fig. 5.26)**, and the elements of E are called *arcs* or *directed edges*. All other graphs are called *mixed graphs*.
In the graphical representation, the vertices of a graph are denoted by points, the directed edges by arrows, and undirected edges by non-directed lines.

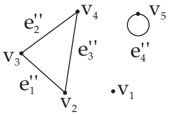


Figure 5.25

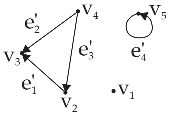


Figure 5.26

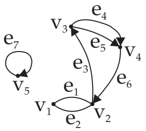


Figure 5.27

■ **A:** For the graph G in **Fig. 5.27**: $V = \{v_1, v_2, v_3, v_4, v_5\},\ E = \{e_1, e_2, e_3, e_4, e_5, e_6, e_7\}$,
$f_1(e_1) = \{v_1, v_2\},\ f_1(e_2) = \{v_1, v_2\},\ f_1(e_3) = (v_2, v_3),\ f_1(e_4) = (v_3, v_4),\ f_1(e_5) = (v_3, v_4)$,
$f_1(e_6) = (v_4, v_2),\ f_1(e_7) = (v_5, v_5)$.

■ **B:** For the graph G in **Fig. 5.26**: $V = \{v_1, v_2, v_3, v_4, v_5\}, \quad E' = \{e_1', e_2', e_3', e_4'\}$
$f_2(e_1') = (v_2, v_3),\ f_2(e_2') = (v_4, v_3),\ f_2(e_3') = (v_4, v_2),\ f_2(e_4') = (v_5, v_5)$.

■ **C:** For the graph G in **Fig. 5.25**: $V = \{v_1, v_2, v_3, v_4, v_5\}, \quad E'' = \{e_1'', e_2'', e_3'', e_4''\}$,
$f_3(e_1'') = \{v_2, v_3\},\ f_3(e_2'') = \{v_4, v_3\},\ f_3(e_3'') = \{v_4, v_2\},\ f_3(e_4'') = \{v_5, v_5\}$.

2. Adjacency

If $(v, w) \in E$, then the vertex v is called *adjacent* to the vertex w. Vertex v is called the *initial point* of (v, w), w is called the *terminal point* of (v, w), and v and w are called the *endpoints* of (v, w).
Adjacency in undirected graphs and the endpoints of undirected edges are defined analogously.

3. Simple Graphs

If several edges or arcs are assigned to the same ordered or non-ordered pairs of vertices, then they are called *multiple edges*. An edge with identical endpoints is called a *loop*. Graphs without loops and multiple edges and multiple arcs, respectively, are called *simple* graphs.

4. Degrees of Vertices

The number of edges or arcs incident to a vertex v is called the *degree* $d_G(v)$ of the vertex v. Loops are counted twice. Vertices of degree zero are called *isolated vertices*.
For every vertex v of a directed graph G, the *out-degree* $d_G^+(v)$ and *in-degree* $d_G^-(v)$ of v are distinguished as follows:

$$d_G^+(v) = |\{w|(v, w) \in E\}|, \qquad (5.316a) \qquad\qquad d_G^-(v) = |\{w|(w, v) \in E\}|. \qquad (5.316b)$$

5. Special Classes of Graphs

Finite graphs have a finite set of vertices and a finite set of edges. Otherwise the graph is called *infinite*. In *regular graphs of degree* r every vertex has degree r.

An undirected simple graph with vertex set V is called a *complete graph* if any two different vertices in V are connected by an edge. A complete graph with an n element set of vertices is denoted by K_n.

If the set of vertices of an undirected simple graph G can be partitioned into two disjoint classes X and Y such that every edge of G joins a vertex of X and a vertex of Y, then G is called a *bipartite graph*. A bipartite graph is called a *complete bipartite graph*, if every vertex of X is joined by an edge with every vertex of Y. If X has n elements and Y has m elements, then the graph is denoted by $K_{n,m}$.

■ **Fig. 5.28** shows a complete graph with five vertices.

■ **Fig. 5.29** shows a complete bipartite graph with a two-element set X and a three-element set Y.

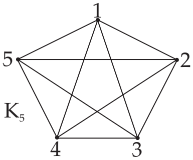


Figure 5.28

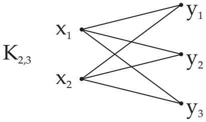


Figure 5.29

Further special classes of graphs are *plane graphs*, *trees* and *transport networks*. Their properties will be discussed in later paragraphs.

6. Representation of Graphs

Finite graphs can be visualized by assigning to every vertex a point in the plane and connecting two points by a directed or undirected curve, if the graph has the corresponding edge. There are examples in **Fig. 5.30**–**5.33**. **Fig. 5.33** shows the *Petersen graph*, which is a well-known counterexample for several graph-theoretic conjectures, which could not be proved in general.

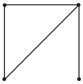


Figure 5.30

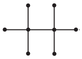


Figure 5.31

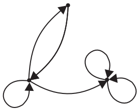


Figure 5.32

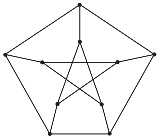


Figure 5.33

7. Isomorphism of Graphs

A graph $G_1 = (V_1, E_1)$ is called *isomorphic* to a graph $G_2 = (V_2, E_2)$ iff there are bijective mappings φ from V_1 onto V_2 and ψ from E_1 onto E_2 being compatible with the incidence function, i.e., if u, v are the endpoints of an edge or u is the initial point of an arc and v is its terminal point, then $\varphi(u)$ and $\varphi(v)$ are the endpoints of an edge and $\varphi(u)$ is the initial point and $\varphi(v)$ the terminal point of an arc, respectively. **Fig. 5.34** and **Fig. 5.35** show two isomorphic graphs. The mapping φ with $\varphi(1) = a,\ \varphi(2) = b,\ \varphi(3) = c,\ \varphi(4) = d$ is an isomorphism. In this case, every bijective mapping of $\{1, 2, 3, 4\}$ onto $\{a, b, c, d\}$ is an isomorphism, since both graphs are complete graphs with equal number of vertices.

8. Subgraphs, Factors

If $G = (V, E)$ is a graph, then the graph $G' = (V', E')$ is called a *subgraph* of G, if $V' \subseteq V$ and $E' \subseteq E$. If E' contains exactly those edges of E which connect vertices of V', then G' is called the *subgraph of* G *induced by* V' (*induced subgraph*).

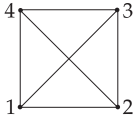


Figure 5.34

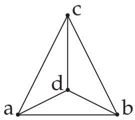


Figure 5.35

A subgraph $G' = (V', E')$ of $G = (V, E)$ with $V' = V$ is called a *partial graph* of G.
A factor F of a graph G is a regular subgraph of G containing all vertices of G.

9. Adjacency Matrix

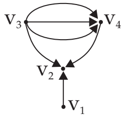
Finite graphs can be described by matrices: Let $G = (V, E)$ be a graph with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$. Let $m(v_i, v_j)$ denote the number of edges from v_i to v_j. For undirected graphs, loops are counted twice; for directed graphs loops are counted once. The matrix $\mathbf{A}$ of type (n, n) with $\mathbf{A} = (m(v_i, v_j))$ is called an *adjacency matrix*. If in addition the graph is simple, then the adjacency matrix has the following form:

$$\mathbf{A} = (a_{ij}) = \begin{cases} 1, & \text{for } (v_i, v_j) \in E, \\ 0, & \text{for } (v_i, v_j) \notin E; \end{cases} \tag{5.317}$$

i.e., in the matrix $\mathbf{A}$ there is a 1 in the i-th row and j-th column iff there is an edge from v_i to v_j. The adjacency matrix of undirected graphs is symmetric.

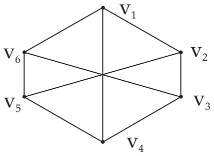
■ **A:** Beside **Fig. 5.36** there is the adjacency matrix $\mathbf{A}_1 = \mathbf{A}(G_1)$ of the directed graph G_1.

■ **B:** Beside **Fig. 5.37** there is the adjacency matrix $\mathbf{A}_2 = \mathbf{A}(G_2)$ of the undirected simple graph G_2.



$$\mathbf{A}_1 = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 3 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

Figure 5.36



$$\mathbf{A}_2 = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Figure 5.37

10. Incidence Matrix

For an undirected graph $G = (V, E)$ with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$, the matrix $\mathbf{I}$ of type (n, m) given by

$$\mathbf{I} = (b_{ij}) \text{ with } b_{ij} = \begin{cases} 0, & v_i \text{ is not incident with } e_j, \\ 1, & v_i \text{ is incident with } e_j \text{ and } e_j \text{ is not a loop}, \\ 2, & v_i \text{ is incident with } e_j \text{ and } e_j \text{ is a loop} \end{cases} \tag{5.318}$$

is called the *incidence matrix*.
For a directed graph $G = (V, E)$ with $V = \{v_1, v_2, \ldots, v_n\}$ and $E = \{e_1, e_2, \ldots, e_m\}$, the incidence matrix $\mathbf{I}$ is the matrix of type (n, m), defined by

$$\mathbf{I} = (b_{ij}) \text{ with } b_{ij} = \begin{cases} 0, & v_i \text{ is not incident with } e_j, \\ 1, & v_i \text{ is the initial point of } e_j \text{ and } e_j \text{ is not a loop}, \\ -1, & v_i \text{ is the terminal point of } e_j \text{ and } e_j \text{ is not a loop}, \\ -0, & v_i \text{ is incident to } e_j \text{ and } e_j \text{ is a loop}. \end{cases} \tag{5.319}$$

11. Weighted Graphs

If $G = (V, E)$ is a graph and f is a mapping assigning a real number to every edge, then (V, E, f) is called a *weighted graph*, and $f(e)$ is the *weight* or *length* of the edge e.

In applications, these weights of the edges represent costs resulting from the construction, maintenance or use of the connections.

5.8.2 Traverse of Undirected Graphs

5.8.2.1 Edge Sequences or Paths

1. Edge Sequences or Paths

In an undirected graph $G = (V, E)$ every sequence $F = (\{v_1, v_2\}, \{v_2, v_3\}, \ldots, \{v_s, v_{s+1}\})$ of the elements of E is called an *edge sequence* of length s.
If $v_1 = v_{s+1}$, then the sequence is called a *cycle*, otherwise it is an *open edge sequence*. An edge sequence F is called a *path* iff $v_1, v_2, \ldots, v_s$ are pairwise distinct vertices. A *closed path* is a *circuit*. A *trail* is a sequence of edges without repeated edges.

■ In the graphs in **Fig. 5.38**, $F_1 = (\{1,2\}, \{2,3\}, \{3,5\}, \{5,2\}, \{2,4\})$ is an edge sequence of length 5, $F_2 = (\{1,2\}, \{2,3\}, \{3,4\}, \{4,2\}, \{2,1\})$ is a cycle of length 5, $F_3 = (\{2,3\}, \{3,5\}, \{5,2\}, \{2,1\})$ is a path, $F_4 = (\{1,2\}, \{2,3\}, \{3,4\})$ is a path. An elementary cycle is given by $F_5 = (\{1,2\}, \{2,5\}, \{5,1\})$.

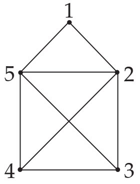


Figure 5.38

2. Connected Graphs, Components

If there is at least one path between every pair of distinct vertices v, w in a graph G, then G is called *connected*. If a graph G is not connected, it can be decomposed into *components*, i.e., into induced connected subgraphs with maximal number of vertices.

3. Distance Between Vertices

The distance $\delta(v, w)$ between two vertices v, w of an undirected graph is the length of a path with minimum number of edges connecting v and w. If such a path does not exist, then let $\delta(v, w) = \infty$.

4. Problem of Shortest Paths

Let $G = (V, E, f)$ be a weighted simple graph with $f(e) > 0$ for every $e \in E$. Determine the *shortest path* from v to w for two vertices v, w of G, i.e., a path from v to w having minimum sum of weights of edges and arcs, respectively.
There is an efficient algorithm of Dantzig to solve this problem, which is formulated for directed graphs and can be used for undirected graphs (see 5.8.6, p. 410) in a similar way.
Every graph $G = (V, E, f)$ with $V = \{v_1, v_2, \ldots, v_n\}$ has a *distance matrix* $\mathbf{D}$ of type (n, n):

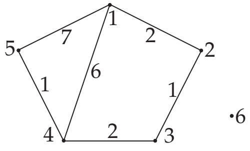
$$\mathbf{D} = (d_{ij}) \quad \text{with} \quad d_{ij} = \delta(v_i, v_j) \qquad (i, j = 1, 2, \ldots, n). \tag{5.320}$$

In the case that every edge has weight 1, i.e., the distance between v and w is equal to the minimum number of edges which have to be traversed in the graph to get from v to w, then the distance between two vertices can be determined using the adjacency matrix: Let $v_1, v_2, \ldots, v_n$ be the vertices of G. The adjacency matrix of G is $\mathbf{A} = (a_{ij})$, and the powers of the adjacency matrix with respect to the usual multiplication of matrices (see 4.1.4, **5.**, p. 272) are denoted by $\mathbf{A}^m = (a_{ij}^m)$, $m \in \mathbb{N}$.
There is a shortest path of length k from the vertex v_i to the vertex v_j $(i \neq j)$ iff:

$$a_{ij}^k \neq 0 \quad \text{and} \quad a_{ij}^s = 0 \quad (s = 1, 2, \ldots, k-1). \tag{5.321}$$

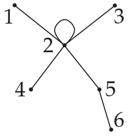
■ The weighted graph represented in **Fig. 5.39** has the distance matrix $\mathbf{D}$ beside it.

■ The graph represented in **Fig. 5.40** has the adjacency matrix $\mathbf{A}$ beside it, and for $m = 2$ or $m = 3$ the matrices $\mathbf{A}^2$ and $\mathbf{A}^3$ are obtained. Shortest paths of length 2 connect the vertices 1 and 3, 1 and 4, 1 and 5, 2 and 6, 3 and 4, 3 and 5, 4 and 5. Furthermore the shortest paths between the vertices 1 and 6, 3 and 6, and finally 4 and 6 are of length 3.



$$\mathbf{D} = \begin{pmatrix} 0 & 2 & 3 & 5 & 6 & \infty \\ 2 & 0 & 1 & 3 & 4 & \infty \\ 3 & 1 & 0 & 2 & 3 & \infty \\ 5 & 3 & 2 & 0 & 1 & \infty \\ 6 & 4 & 3 & 1 & 0 & \infty \\ \infty & \infty & \infty & \infty & \infty & 0 \end{pmatrix}$$

Figure 5.39



$$\mathbf{A} = \begin{pmatrix} 0&1&0&0&0&0 \\ 1&1&1&1&1&0 \\ 0&1&0&0&0&0 \\ 0&1&0&0&0&0 \\ 0&1&0&0&0&1 \\ 0&0&0&0&1&0 \end{pmatrix}, \quad \mathbf{A}^2 = \begin{pmatrix} 1&1&1&1&1&0 \\ 1&5&1&1&1&1 \\ 1&1&1&1&1&0 \\ 1&1&1&1&1&0 \\ 1&1&1&1&2&0 \\ 0&1&0&0&0&1 \end{pmatrix}, \quad \mathbf{A}^3 = \begin{pmatrix} 1&5&1&1&1&1 \\ 5&9&5&5&6&1 \\ 1&5&1&1&1&1 \\ 1&5&1&1&1&1 \\ 1&6&1&1&1&2 \\ 1&1&1&1&2&0 \end{pmatrix}.$$

Figure 5.40

5.8.2.2 Euler Trails

1. Euler Trail, Euler Graph

A trail containing every edge of a graph G is called an *open* or *closed Euler trail* of G.
A connected graph containing a closed Euler trail is an *Euler graph.*

■ The graph G_1 **(Fig. 5.41)** has no Euler trail. The graph G_2 **(Fig. 5.42)** has an Euler trail, but it is not an Euler graph. The graph G_3 **(Fig. 5.43)** has a closed Euler trail, but it is not an Euler graph. The graph G_4 **(Fig. 5.44)** is an Euler graph.

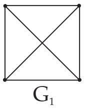


Figure 5.41

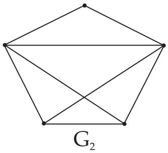


Figure 5.42

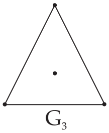


Figure 5.43

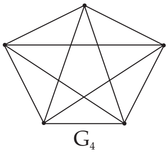


Figure 5.44

2. Theorem of Euler-Hierholzer

A finite connected graph is an Euler graph iff all vertices have positive even degrees.

3. Construction of a Closed Euler Trail

If G is an Euler graph, then one chooses an arbitrary vertex v_1 of G and constructs a trail F_1 by traversing a path, starting at v_1 and proceeding until it cannot be continued. If F_1 does not yet contain all edges of G, then one constructs another path F_2 containing the edges not in F_1, but starting at a vertex $v_2 \in F_1$ and proceeds until it cannot be continued. Then one composes a closed trail in G using F_1 and F_2: Starting to traverse F_1 at v_1 until v_2 is reached, then continuing to traverse F_2, and finishing at the edges of F_1 not used before. Repeating this method a closed Euler trail is obtained in finitely many steps.

4. Open Euler Trails

There is an open Euler trail in a graph G iff there are exactly two vertices in G with odd degrees. **Fig. 5.45** shows a graph which has no closed Euler trail, but it has an open Euler trail. The edges are consecutively enumerated with respect to an Euler trail. In **Fig. 5.46** there is a graph with a closed Euler trail.

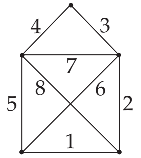


Figure 5.45

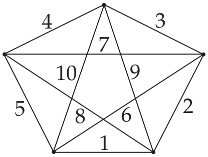


Figure 5.46

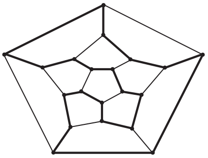


Figure 5.47

5. Chinese Postman Problem

The problem, that a postman should pass through all streets in his service area at least once and return to the initial point and use a trail as short as possible, can be formulated in graph theoretical terms as follows: Let $G = (V, E, f)$ be a weighted graph with $f(e) \geq 0$ for every edge $e \in E$. Determine an edge sequence F with minimum total length

$$L = \sum_{e \in F} f(e). \tag{5.322}$$

The name of the problem refers to the Chinese mathematician Kuan, who studied this problem first. To solve it two cases are distinguished:

1. G is an Euler graph – then every closed Euler trail is optimal – and

2. G has no closed Euler trail.

An effective algorithm solving this problem is given by Edmonds and Johnson (see [5.25]).

5.8.2.3 Hamiltonian Cycles

1. Hamiltonian Cycle

A *Hamiltonian cycle* is an elementary cycle in a graph covering all of the vertices.

■ In **Fig. 5.47**, lines in bold face show a Hamiltonian cycle.

The idea of a game to construct Hamiltonian cycles in the graph of a pentagondodecaeder, goes back to Sir W. Hamilton.

Remark: The problem of characterizing graphs with Hamiltonian cycles leads to one of the classical NP-complete problems. Therefore, an efficient algorithm to determine the Hamilton cycles cannot be given here.

2. Theorem of Dirac

If a simple graph $G = (V, E)$ has at least three vertices, and $d_G(v) \geq |V|/2$ holds for every vertex v of G, then G has a Hamiltonian cycle. This is a sufficient but not a necessary condition for the existence of Hamiltonian cycles. The following theorems with more general assumptions give only sufficient but not necessary conditions for the existence of Hamilton cycles, too.

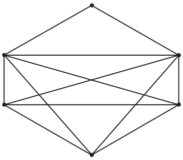


Figure 5.48

■ **Fig. 5.48** shows a graph which has a Hamiltonian cycle, but does not satisfy the assumptions of the following theorem of Ore.

3. Theorem of Ore

If a simple graph $G = (V, E)$ has at least three vertices, and $d_G(v) + d_G(w) \geq |V|$ holds for every pair of non-adjacent vertices v, w, then G contains a Hamiltonian cycle.

4. Theorem of Posa

Let $G = (V, E)$ be a simple graph with at least three vertices. There is a Hamiltonian cycle in G if the following conditions are satisfied:

1. For $1 \leq k < (|V| - 1)/2$, the number of vertices of degree not exceeding k is less than k.

2. If $|V|$ is odd, then the number of vertices of degree not exceeding $(|V| - 1)/2$ is less than or equal to $(|V| - 1)/2$.

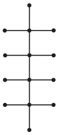


Figure 5.49

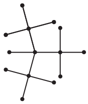


Figure 5.50

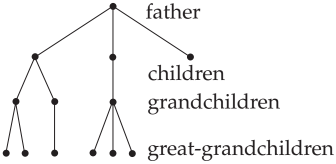


Figure 5.51

5.8.3 Trees and Spanning Trees

5.8.3.1 Trees

1. Trees

An undirected connected graph without cycles is called a *tree*. Every tree with at least two vertices has at least two vertices of degree 1. Every tree with n vertices has exactly $n - 1$ edges.
A directed graph is called a tree if G is connected and does not contain any circuit (see 5.8.6, p. 410).

■ **Fig. 5.49** and **Fig. 5.50** represent two non-isomorphic trees with 14 vertices. They demonstrate the chemical structure of butane and iso-butane.

2. Rooted Trees

A tree with a distinguished vertex is called a rooted tree, and the distinguished vertex is called the *root*. In diagrams, the root is usually on the top, and the edges are directed downwards from the root (see **Fig. 5.51**). Rooted trees are used to represent hierarchic structures, as for instance hierarchies in factories, family trees, grammatical structures.

■ **Fig. 5.51** shows the genealogy of a family in the form of a rooted tree. The root is the vertex assigned to the father.

3. Regular Binary Trees

If a tree has exactly one vertex of degree 2 and otherwise only vertices of degree 1 or 3, then it is called a *regular binary tree*.
The number of vertices of a regular binary tree is odd. Regular trees with n vertices have $(n + 1)/2$ vertices of degree 1. The *level* of a vertex is its distance from the root. The maximal level occurring in a tree is the *height* of the tree. There are several applications of regular binary rooted trees, e.g., in informatics.

4. Ordered Binary Trees

Arithmetical expressions can be represented by binary trees. Here, the numbers and variables are assigned vertices of degree 1, the operations "+","−", "·" correspond to vertices of degree > 1, and the left and right subtree, respectively, represents the first and second operand, respectively, which is, in general, also an expression. These trees are called *ordered binary trees*.
The traverse of an ordered binary tree can be performed in three different ways, which are defined in a recursive way (see also **Fig. 5.52**):

Inorder traverse: Traverse the left subtree of the root (in inorder traverse),
visit the root,
traverse the right subtree of the root (in inorder traverse).

Preorder traverse: Visit the root,
traverse the left subtree (in preorder traverse),
traverse the right subtree of the root (in preorder traverse).

Postorder traverse: Traverse the left subtree of the root (in postorder traverse),
traverse the right subtree of the root (in postorder traverse),
visit the root.

Using inorder traverse the order of the terms does not change in comparison with the given expression. The term obtained by postorder traverse is called *postfix notation* PN or *Polish notation.* Analogously, the term obtained by preorder traverse is called *prefix notation* or *reversed Polish notation.*

Prefix and postfix expressions uniquely describe the tree. This fact can be used for the implementation of trees.

■ In **Fig. 5.52** the term $a \cdot (b-c)+d$ is represented by a graph. Inorder traverse yields $a \cdot b - c + d$, preorder traverse yields $+ \cdot -bcad$, and postorder traversal yields $abc - \cdot d+$.

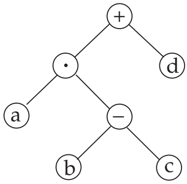


Figure 5.52

5.8.3.2 Spanning Trees

1. Spanning Trees

A tree, being a subgraph of an undirected graph G, and containing all vertices of G, is called a *spanning tree* of G. Every finite connected graph G contains a spanning tree H:

If G contains a cycle, then delete an edge of this cycle. The remaining graph G_1 is still connected and can be transformed into a connected graph G_2 by deleting a further edge of a cycle of G_1, if there exists such an edge. After finitely many steps a spanning tree of G is obtained.

■ **Fig. 5.54** shows a spanning tree H of the graph G shown in **Fig. 5.53**.

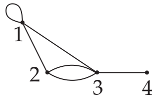


Figure 5.53

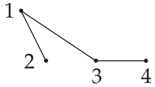


Figure 5.54

2. Theorem of Cayley

Every complete graph with n vertices ($n > 1$) has exactly n^{n-2} spanning trees.

3. Matrix Spanning Tree Theorem

Let $G = (V, E)$ be a graph with $V = \{v_1, v_2, \ldots, v_n\}$ ($n > 1$) and $E = \{e_1, e_2, \ldots, e_m\}$. Define a matrix $\mathbf{D} = (d_{ij})$ of type (n, n):

$$d_{ij} = \begin{cases} 0 \text{ for } i \neq j, \\ d_G(v_i) \text{ for } i = j, \end{cases} \tag{5.323a}$$

which is called the *degree matrix.* The difference between the degree matrix and the adjacency matrix is the admittance matrix $\mathbf{L}$ of G:

$$\mathbf{L} = \mathbf{D} - \mathbf{A}. \tag{5.323b}$$

Deleting the i-th row and the i-th column of $\mathbf{L}$ the matrix $\mathbf{L}_i$ is obtained. The determinant of $\mathbf{L}_i$ is equal to the number of spanning trees of the graph G.

■ The adjacency matrix, the degree matrix and the admittance matrix of the graph in **Fig. 5.53** are:

$$\mathbf{A} = \begin{pmatrix} 2&1&1&0 \\ 1&0&2&0 \\ 1&2&0&1 \\ 0&0&1&0 \end{pmatrix}, \qquad \mathbf{D} = \begin{pmatrix} 4&0&0&0 \\ 0&3&0&0 \\ 0&0&4&0 \\ 0&0&0&1 \end{pmatrix}, \qquad \mathbf{L} = \begin{pmatrix} 2&-1&-1&0 \\ -1&3&-2&0 \\ -1&-2&4&-1 \\ 0&0&-1&1 \end{pmatrix}.$$

Since $\det \mathbf{L}_3 = 5$, the graph has five spanning trees.

4. Minimal Spanning Trees

Let $G = (V, E, f)$ be a connected weighted graph. A spanning tree H of G is called a *minimum spanning tree* if its *total length* $f(H)$ is minimum:

$$f(H) = \sum_{e \in H} f(e). \tag{5.324}$$

Minimum spanning trees are searched for, e.g., if the edge weights represent costs, and one is interested in minimum costs. A method to find a minimum spanning tree is the *Kruskal algorithm*:

a) Choose an edge with the least weight.

b) Continue, as long as it is possible, choosing a further edge having least weight and not forming a cycle with the edges already chosen, and add such an edge to the tree.

In step **b)** the choice of the admissible edges can be made easier by the following labeling algorithm:
• Let the vertices of the graph be labeled pairwise differently.
• At every step, an edge can be added only in the case that it connects vertices with different labels.
• After adding an edge, the label of the endpoint with the larger label is changed to the value of the smaller endpoint label.

5.8.4 Matchings

1. Matchings

A set M of edges of a graph G is called a *matching* in G, iff M contains no loop and two different edges of M do not have common endpoints.
A matching M^* of G is called a *saturated matching*, if there is no matching M in G such that $M^* \subset M$.
A matching M^{**} of G is called a *maximum matching*, if there is no matching M in G such that $|M| > |M^{**}|$.
If M is a matching of G such that every vertex of G is an endpoint of an edge of M, then M is called a *perfect matching.*

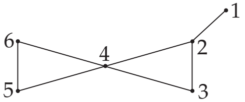


Figure 5.55

■ In the graph in **Fig. 5.55** $M_1 = \{\{2,3\},\{5,6\}\}$ is a saturated matching and $M_2 = \{\{1,2\},\{3,4\},\{5,6\}\}$ is a maximum matching which is also perfect.

Remark: In graphs with an odd number of edges there is no perfect matching.

2. Theorem of Tutte

Let $q(G-S)$ denote the number of the components of $G-S$ with an odd number of vertices. A graph $G=(V,E)$ has a perfect matching iff $|V|$ is even and for every subset S of the vertex set $q(G-S) \leq |S|$. Here $G-S$ denotes the graph obtained from G by deleting the vertices of S and the edges incident with these vertices.
Perfect machings exist for example in complete graphs with an even number of vertices, in complete bipartite graphs $K_{n,n}$ and in arbitrary regular bipartite graphs of degree $r > 0$.

3. Alternating Paths

Let G be a graph with a matching M. A path W in G is called an *alternating path* iff in W every edge e with $e \in M$ (or $e \notin M$) is followed by an edge e' with $e' \notin M$ (or $e \in M$).
An open alternating path is called an *increasing path* iff none of the endpoints of the path is incident with an edge of M.

4. Theorem of Berge

A matching M in a graph G is maximum iff there is no increasing alternating path in G.
If W is an increasing alternating path in G with corresponding set $E(W)$ of traversed edges, then $M' = (M \setminus E(W)) \cup (E(W) \setminus M)$ forms a matching in G with $|M'| = |M| + 1$.

■ In the graph of **Fig. 5.55** $(\{1,2\},\{2,3\},\{3,4\})$ is an increasing alternating path with respect to matching M_1. Matching M_2 with $|M_2| = |M_1| + 1$ is obtained as described above.

5. Determination of Maximum Matchings

Let G be a graph with a matching M.

a) First form a saturated matching M^* with $M \subseteq M^*$.

b) Chose a vertex v in G, which is not incident with an edge of M^*, and determine an increasing alternating path in G starting at v.

c) If such a path exists, then the method described above results in a matching M' with $|M'| > |M^*|$. If there is no such path, then delete vertex v and all edges incident with v in G, and repeat step **b)**.

There is an algorithm of Edmonds, which is an effective method to search for maximum matchings, but it is rather complicated to describe (see [5.24]).

5.8.5 Planar Graphs

Here, the considerations are restricted to undirected graphs, since a directed graph is planar iff the corresponding undirected graph is a planar one.

1. Planar Graph

A graph is called a *plane graph* iff G can be drawn in the plane with its edges intersecting only in vertices of G. A graph isomorphic with a plane graph is called a *planar graph.*

Fig. 5.56 shows a plane graph G_1. The graph G_2 in **Fig. 5.57** is isomorphic to G_1, it is not a plane graph but a planar graph, since it is isomorphic with G_1.

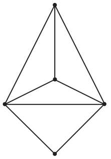


Figure 5.56

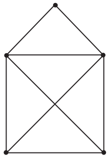


Figure 5.57

2. Non-Planar Graphs

The complete graph K_5 and the complete bipartite graph $K_{3,3}$ are non-planar graphs (see 5.8.1, **5.**, p. 402).

3. Subdivisions

A *subdivision* of a graph G is obtained if vertices of degree 2 are inserted into edges of G. Every graph is a subdivision of itself. Certain subdivisions of K_5 and $K_{3,3}$ are represented in **Fig. 5.58** and **Fig. 5.59**.

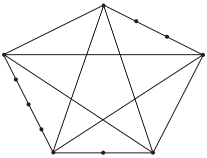


Figure 5.58

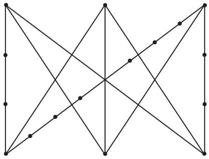


Figure 5.59

4. Kuratowski's Theorem

A graph is non-planar iff it contains a subgraph which is a subdivision either of the complete bipartite graph $K_{3,3}$ or of the complete graph K_5.

5.8.6 Paths in Directed Graphs

1. Arc Sequences

A sequence $F = (e_1, e_2, \ldots, e_s)$ of arcs in a directed graph is called a *chain* of length s, iff F does not contain any arc twice and one of the endpoints of every arc e_i for $i = 2, 3, \ldots, s-1$ is an endpoint of the arc e_{i-1} and the other one an endpoint of e_{i+1}.

A chain is called a *directed chain* iff for $i = 1, 2, \ldots, s-1$ the terminal point of the arc e_i coincides with the initial point of e_{i+1}.

Chains or directed chains traversing every vertex at most once are called *elementary chains* and *elementary directed chains*, respectively.

A closed chain is called a *cycle*. A closed directed path, with every vertex being the endpoint of exactly two arcs, is called a *circuit*.

■ **Fig. 5.60** contains examples for various kinds of arc sequences.

2. Connected and Strongly Connected Graphs

A directed graph G is called *connected* iff for any two vertices there is a chain connecting these vertices. The graph G is called *strongly connected* iff to every two vertices v, w there is is assigned a directed chain connecting these vertices.

3. Algorithm of Dantzig

Let $G = (V, E, f)$ be a weighted simple directed graph with $f(e) > 0$ for every arc e. The following algorithm yields all vertices of G, which are connected with a fixed vertex v_1 by a directed chain, together with their distances from v_1:

a) Vertex v_1 gets the label $t(v_1) = 0$. Let $S_1 = \{v_1\}$.

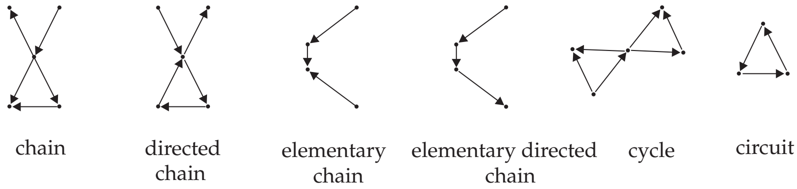


Figure 5.60

b) The set of the labeled vertices is S_m.

c) If $U_m = \{e|e = (v_i, v_j) \in E,\ v_i \in S_m,\ v_j \notin S_m\} = \emptyset$, then one finishes the algorithm.

d) Otherwise one chooses an arc $e^* = (x^*, y^*)$ with minimum $t(x^*) + f(e^*)$. One labels e^* and y^* and puts $t(y^*) = t(x^*) + f(e^*)$ and also $S_{m+1} = S_m \cup \{y^*\}$ and repeats **b)** with $m := m + 1$.

(If all arcs have weight 1, then the length of a shortest directed chain from a vertex v to a vertex w can be found using the adjacency matrix (see 5.8.2.1, **4.**, p. 404)).

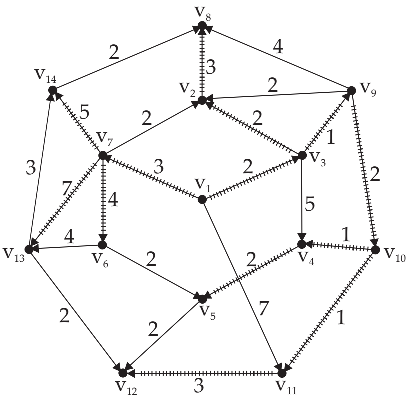


Figure 5.61

If a vertex v of G is not labeled, then there is no directed path from v_1 to v.

If v has label $t(v)$, then $t(v)$ is the length of such a directed chain. A shortest directed path from v_1 to v can be found in the tree given by the labeled arcs and vertices, the *distance tree* with respect to v_1.

■ In **Fig. 5.61**, the labeled arcs and vertices represent the distance tree with respect to v_1 in the graph. The lengths of the shortest directed chains are:

from v_1 to v_3 :	2	from v_1 to v_6 :	7
from v_1 to v_7 :	3	from v_1 to v_8 :	7
from v_1 to v_9 :	3	from v_1 to v_{14} :	8
from v_1 to v_2 :	4	from v_1 to v_5 :	8
from v_1 to v_{10} :	5	from v_1 to v_{12} :	9
from v_1 to v_4 :	6	from v_1 to v_{13} :	10
from v_1 to v_{11} :	6.		

Remark: There is also a modified algorithm to find the shortest directed chains in the case that $G = (V, E, f)$ has arcs with negative weights.

5.8.7 Transport Networks

1. Transport Network

A connected directed graph is called a *transport network* if it has two labeled vertices, called the *source* Q and *sink* S which have the following properties:

a) There is an arc u_1 from S to Q, where u_1 is the only arc with initial point S and the only arc with terminal point Q.

b) Every arc u_i different from u_1 is assigned a real number $c(u_i) \geq 0$. This number is called its *capacity*. The arc u_1 has capacity ∞.

A function φ, which assigns a real number to every arc, is called a *flow* on G, if the equality

$$\sum_{(u,v)\in G} \varphi(u, v) = \sum_{(v,w)\in G} \varphi(v, w) \tag{5.325a}$$

holds for every vertex v. The sum

$$\sum_{(Q,v)\in G} \varphi(Q,v) \tag{5.325b}$$

is called the intensity of the flow. A flow φ is called *compatible to the capacities*, if for every arc u_i of G $0 \le \varphi(u_i) \le c(u_i)$ holds.

■ For an example of a transport network see p. 412.

2. Maximum Flow Algorithm of Ford and Fulkerson

Using the maximum flow algorithm one can recognize whether a given flow φ is maximal.
Let G be a transport network and φ a flow of intensity v_1 compatible with the capacities. The algorithm given below contains the following steps for labeling the vertices, and after finishing this procedure one can realize how much the intensity of the flow could be improved depending on the chosen labeling steps.

a) One labels the source Q and sets $\varepsilon(Q) = \infty$.

b) If there is an arc $u_i = (x, y)$ with labeled x and unlabeled y and $\varphi(u_i) < c(u_i)$, then one labels y and (x, y), and sets $\varepsilon(y) = \min\{\varepsilon(x), c(u_i) - \varphi(u_i)\}$, then one repeats step **b)**, otherwise follows step **c)**.

c) If there is an arc $u_i = (x, y)$ with unlabeled x and labeled $y, \varphi(u_i) > 0$ and $u_i \ne u_1$, then one labels x and (x, y), substitutes $\varepsilon(x) = \min\{\varepsilon(y), \varphi(u_i)\}$ and returns to continue step **b)** if it is possible. Otherwise one finishs the algorithm.
If the sink S of G is labeled, then the flow in G can be improved by an amount of $\varepsilon(S)$. If the sink is not labeled, then the flow is maximal.

■ Maximum flow: For the graph in **Fig. 5.62** the weights are written next to the edges. A flow with intensity 13, compatible to these capacities, is represented in the weighted graph in **Fig. 5.63**. It is a maximum flow.

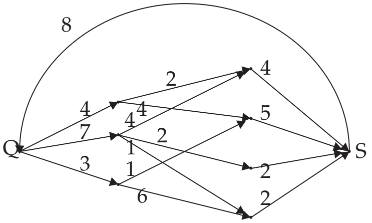


Figure 5.62

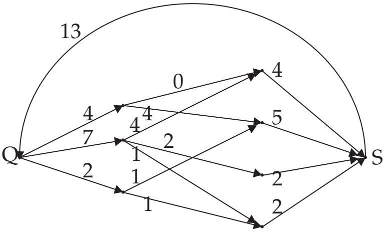


Figure 5.63

■ Transport network: A product is produced by p firms $F_1, F_2, \dots, F_p$. There are q users $V_1, V_2, \dots, V_q$. During a certain period there will be s_i units produced by F_i and t_j units required by V_j. c_{ij} units can be transported from F_i to V_j during the given period. Is it possible to satisfy all the requirements during this period? The corresponding graph is shown in **Fig. 5.64**.

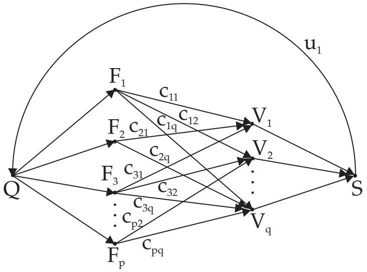


Figure 5.64

5.9 Fuzzy Logic

5.9.1 Basic Notions of Fuzzy Logic

5.9.1.1 Interpretation of Fuzzy Sets

Real situations are very often uncertain or vague in a number of ways. The word “fuzzy” also means some uncertainty, and the name of *fuzzy logic* is based on this meaning. Basically there are to distinguish two types of fuzziness: *vagueness* and *uncertainty*. There are two concepts belonging here: The theory of fuzzy sets and the theory of fuzzy measure. In the following practice-oriented introduction the notions, methods, and concepts of fuzzy sets are discussed, which are the basic mathematical tools of multi-valued logic.

1. Notions of Classical and Fuzzy Sets

The classical notion of (crisp) set is two-valued, and the classical Boolean set algebra is isomorphic to two-valued propositional logic. Let X be a fundamental set named the universe. Then for every $A \subseteq X$ there exists a function

$$f_A\colon X \to \{0, 1\}, \tag{5.326a}$$

such that it says for every $x \in X$ whether this element x belongs to the set A or not:

$$f_A(x) = 1 \Leftrightarrow x \in A \quad \text{and} \quad f_A(x) = 0 \Leftrightarrow x \notin A. \tag{5.326b}$$

The concept of fuzzy sets is based on the idea of considering the membership of an element of the set as a statement, the truth value of which is characterized by a value from the interval $[0, 1]$. For mathematical modeling of a fuzzy set A a function is necessary whose range is the interval $[0, 1]$ instead of {0,1}, i.e.:

$$\mu_A\colon X \to [0, 1]. \tag{5.327}$$

In other words: To every element $x \in X$ is to assign a number $\mu_A(x)$ from the interval $[0, 1]$, which represents the grade of membership of x in A. The mapping μ_A is called the *membership function*. The value of the function $\mu_A(x)$ at the point x is called the *grade of membership*. The fuzzy sets A, B, C, etc. over X are also called fuzzy subsets of X. The set of all fuzzy sets over X is denoted by $F(X)$.

2. Properties of Fuzzy Sets and Further Definitions

The properties below follow directly from the definition:

(E1) Crisp sets can be interpreted as fuzzy sets with grade of membership 0 and 1.

(E2) The set of the arguments x, whose grade of membership is greater than zero, i.e., $\mu_A(x) > 0$, is called the *support* of the fuzzy set A:

$$\operatorname{supp}(A) = \{x \in X \mid \mu_A(x) > 0\}. \tag{5.328}$$

The set $ker(A) = \{x \in X : \mu_A(x) = 1\}$ is called the *kernel* or *core* of A.

(E3) Two fuzzy sets A and B over the universe X are equal if the values of their membership functions are equal:

$$A = B,\ \text{if } \mu_A(x) = \mu_B(x) \text{ holds for every } x \in X. \tag{5.329}$$

(E4) Discrete representation or ordered pair representation: If the universe X is finite, i.e., $X = \{x_1, x_2, \ldots, x_n\}$ it is reasonable to define the membership function of the fuzzy set with a table of values. The tabular representation of the fuzzy set A is seen in **Table 5.7**.

Also it is possible to write

Table 5.7 Tabular representation of a fuzzy set

x_1	x_2	$\ldots$	x_n
$\mu_A(x_1)$	$\mu_A(x_2)$	$\ldots$	$\mu_A(x_n)$

$$A := \mu_A(x_1)/x_1 + \cdots + \mu_A(x_n)/x_n = \sum_{i=1}^{n} \mu_A(x_i)/x_i. \tag{5.330}$$

In (5.330) the fraction bars and addition signs have only symbolic meaning.

(E5) Ultra-fuzzy set: A fuzzy set, whose membership function itself is a fuzzy set, is called, after Zadeh, an *ultra-fuzzy set*.

3. Fuzzy Linguistics

Assigning linguistic values, e.g., “small”, “medium” or “big”, to a quantity then it is called a *linguistic quantity* or *linguistic variable.* Every linguistic value can be described by a fuzzy set, for example, by the graph of a membership function (5.9.1.2) with a given support (5.328). The number of fuzzy sets (in the case of “small”, “medium”, “big” they are three) depends on the problem.

In 5.9.1.2 the linguistic variable is denoted by x. For example, x can have linguistic values for temperature, pressure, volume, frequency, velocity, brightness, age, wearing, etc., and also medical, electrical, chemical, ecological, etc. variables.

■ By the membership function $\mu_A(x)$ of a linguistic variable, the membership degree of a fixed (crisp) value can be determined in the fuzzy set represented by $\mu_A(x)$. Namely, the modeling of a “high” quantity, e.g., the temperature, as a linguistic variable given by a trapezoidal membership function **(Fig. 5.65)** means that the given temperature α belongs to the fuzzy set “high temperature” with the degree of membership β (also degree of compatibility or degree of truth).

5.9.1.2 Membership Functions on the Real Line

The membership functions can be modeled by functions with values between 0 and 1. They represent the different grade of membership for the points of the universe being in the given set.

1. Trapezoidal Membership Functions

Trapezoidal membership functions are widespread. Piecewise (continuously differentiable) membership functions and their special cases, e.g., the triangle shape membership functions described in the following examples, are very often used. Connecting fuzzy quantities gives smoother output functions if the fuzzy quantities were represented by continuous or piecewise continuous membership functions.

■ **A:** Trapezoidal function **(Fig. 5.65)** corresponding to (5.331).

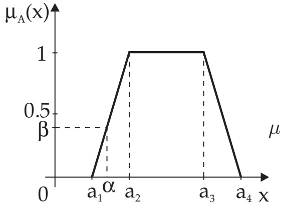


Figure 5.65

$$\mu_A(x)=\begin{cases}0 & x\le a_1\,,\\ \dfrac{x-a_1}{a_2-a_1} & a_1<x<a_2\,,\\ 1 & a_2\le x\le a_3\,,\\ \dfrac{a_4-x}{a_4-a_3} & a_3<x<a_4\,,\\ 0 & x\ge a_4.\end{cases}\qquad(5.331)$$

The graph of this function turns into a triangle function if $a_2=a_3=a$ and $a_1<a<a_4$. Choosing different values for $a_1,\dots,a_4$ gives symmetrical or asymmetrical trapezoidal functions, a symmetrical triangle function ($a_2=a_3=a$ and $|a-a_1|=|a_4-a|$) or asymmetrical triangle function ($a_2=a_3=a$ and $|a-a_1|\ne|a_4-a|$).

■ **B:** Membership function bounded to the left and to the right (**Fig. 5.66)** corresponding to (5.332):

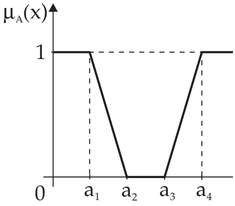


Figure 5.66

$$\mu_A(x)=\begin{cases}1 & x\le a_1,\\ \dfrac{a_2-x}{a_2-a_1} & a_1<x<a_2,\\ 0 & a_2\le x\le a_3,\\ \dfrac{x-a_3}{a_4-a_3} & a_3<x<a_4,\\ 1 & a_4\le x.\end{cases}\qquad(5.332)$$

■ **C:** Generalized trapezoidal function **(Fig. 5.67)** corresponding to (5.333).

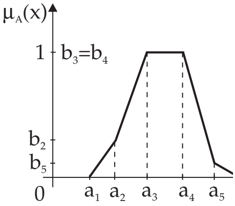


Figure 5.67

$$\mu_A(x) = \begin{cases} 0 & x \le a_1, \\ \dfrac{b_2(x-a_1)}{a_2-a_1} & a_1 < x < a_2, \\ \dfrac{(b_3-b_2)(x-a_2)}{a_3-a_2} + b_2 & a_2 \le x \le a_3, \\ b_3 = b_4 = 1 & a_3 < x < a_4, \\ \dfrac{(b_4-b_5)(a_4-x)}{a_5-a_4} + b_5 & a_4 < x \le a_5, \\ \dfrac{b_5(a_6-x)}{a_6-a_5} & a_5 < x < a_6, \\ 0 & a_6 \le x. \end{cases} \qquad (5.333)$$

2. Bell-Shaped Membership Functions

$$f(x) = \begin{cases} 0 & x \le a, \\ e^{-1/p(x)} & a < x < b, \\ 0 & x \ge b. \end{cases} \qquad (5.334)$$

■ **A:** A class of bell-shaped, differentiable membership functions is given by the function $f(x)$ from (5.334) by choosing an appropriate $p(x)$:
For $p(x) = k(x-a)(b-x)$ and, e.g., $k = 10$ or $k = 1$ or $k = 0.1$, there is a family of symmetrical curves of different width with the membership function $\mu_A(x) = f(x) \Big/ f\left(\dfrac{a+b}{2}\right)$, where $1 \Big/ f\left(\dfrac{a+b}{2}\right)$ is the normalizing factor **(Fig. 5.68)**. The exterior curve follows with the value $k = 10$ and the interior one with $k = 0.1$.
Asymmetrical membership functions in $[0,1]$ follow e.g. for $p(x) = x(1-x)(2-x)$ or for $p(x) = x(1-x)(x+1)$ **(Fig. 5.69)**, using appropriate normalizing factors. The factor $(2-x)$ in the first polynomial results in the shifting of the maximum to the left and it yields an asymmetrical curve shape. Similarly, the factor $(x+1)$ in the second polynomial results in a shifting to the right and in an asymmetric form.

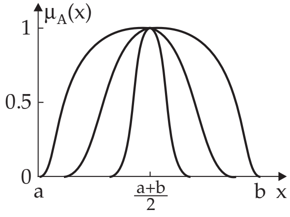


Figure 5.68

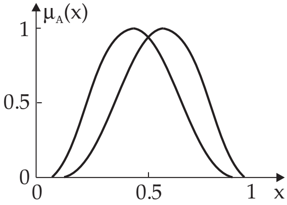


Figure 5.69

■ **B:** A more flexible class of membership functions can be got by the formula

$$F_t(x) = \frac{\displaystyle\int_a^x f\,(t(u))\;du}{\displaystyle\int_a^b f\,(t(u))\;du}, \qquad (5.335)$$

where f is defined by (5.334) with $p(x) = (x-a)(b-x)$ and t is a transformation on $[a,b]$. If t is a smooth transformation on $[a,b]$, i.e., if t is differentiable infinitely many times in the interval $[a,b]$, then F_t is also smooth, since f is smooth. Requiring t to be either increasing or decreasing and to be smooth, then the transformation t allows to change the shape of the curve of the membership function. In practice, polynomials are especially suitable for transformations. The simplest polynomial is the identity $t(x) = x$ on the interval $[a,b] = [0,1]$.

The next simplest polynomial with the given properties is $t(x) = -\frac{2}{3}cx^3 + cx^2 + \left(1 - \frac{c}{3}\right)x$ with a constant $c \in [-6,3]$. The choice $c = -6$ results in the polynomial of maximum curvature, its equation is $q(x) = 4x^3 - 6x^2 + 3x$. Choosing for q_0 the identity function, i.e., $q_0(x) = x$, then can be got recursively further polynomials q by the formula $q_i = q \circ q_{i-1}$ for $i \in \mathbb{N}$. Substituting the corresponding polynomial transformations $q_0, q_1, \ldots$ into (5.335) for t, gives a sequence of smooth functions F_{q_0}, F_{q_1} and F_{q_2} **(Fig. 5.70)**, which can be considered as membership functions $\mu_A(x)$, where F_{q_n} converges to a line. The trapezoidal membership function can be approximated by differentiable functions using the function F_{q_2}, its reflection and a horizontal line **(Fig. 5.71)**.

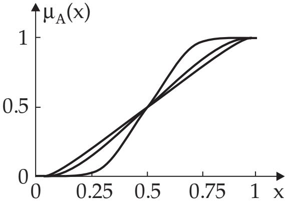


Figure 5.70

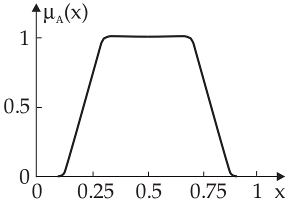


Figure 5.71

Summary: Imprecise and non-crisp information can be described by fuzzy sets and represented by membership functions $\mu(x)$.

5.9.1.3 Fuzzy Sets

1. Empty and Universal Fuzzy Sets

a) Empty fuzzy set: A set A over X is called *empty* if $\mu_A(x) = 0 \;\forall x \in X$ holds.

b) Universal fuzzy set: A set is called *universal* if $\mu_A(x) = 1 \;\forall x \in X$ holds.

2. Fuzzy Subset

If $\mu_B(x) \le \mu_A(x) \;\forall\, x \in X$, then B is called a *fuzzy subset* of A (one writes: $B \subseteq A$).

3. Tolerance Interval and Spread of a Fuzzy Set on the Real Line

If A is a fuzzy set on the real line, then the interval

$$[a,b] = \{x \in X | \mu_A(x) = 1\} \quad (a, b \text{ const, } a < b) \tag{5.336}$$

is called the *tolerance interval* of the fuzzy set A, and the interval $[c,d] = \mathrm{cl}(\mathrm{supp}A)$ (c, d const,$c < d$) is called the *spread* of A, where cl denotes the closure of the set. (The tolerance interval is sometimes also called the *peak* of set A.) The tolerance interval and the kernel coincide only if the kernel contains more then one point.

■ **A:** In **Fig. 5.65** $[a_2, a_3]$ is the tolerance interval, and $[a_1, a_4]$ is the spread.

■ **B:** $a_2 = a_3 = a$ **(Fig. 5.65)**, gives a triangle-shaped membership function μ. In that case the triangular fuzzy set has no tolerance, but its kernel is the set $\{a\}$. If additionally $a_1 = a = a_4$ holds, too, then a crisp value follows; it is called a *singleton.* A singleton A has no tolerance, but $\ker(A) = \mathrm{supp}(A) = \{a\}$.

4. Conversion of Fuzzy Sets on a Continuous and Discrete Universe

Let the universe be continuous, and let a fuzzy set be given on it by its membership function. Discretizing the universe, every discrete point together with its membership value determines a fuzzy singleton.

Conversely, a fuzzy set given on a discrete universe can be converted into a fuzzy set on the continuous universe by interpolating the membership value between the discrete points of the universe.

5. Normal and Subnormal Fuzzy Sets

If A is a fuzzy subset of X, then its *height* is defined by

$$H(A) := \max\{\mu_A(x)|x \in X\}. \tag{5.337}$$

A is called a *normal fuzzy set* if $H(A) = 1$, otherwise it is *subnormal*.
The notions and methods represented in this paragraph are limited to normal fuzzy sets, but it easy to extend them also to subnormal fuzzy sets.

6. Cut of a Fuzzy Set

The α *cut* $A^{>\alpha}$ or the *strong* α *cut* $A^{\geq\alpha}$ of a fuzzy set A are the subsets of X defined by

$$A^{>\alpha} = \{x \in X|\mu_A(x) > \alpha\}, \qquad A^{\geq\alpha} = \{x \in X|\mu_A(x) \geq \alpha\}, \quad \alpha \in (0,1]. \tag{5.338}$$

and $A^{\geq 0} = \operatorname{cl}(A^{>0})$. The α *cut* and *strong* α *cut* are also called α*-level set* and *strong* α*-level set*, respectively.

1. Properties

a) The α cuts of fuzzy sets are crisp sets.

b) The support $\operatorname{supp}(A)$ is a special α cut: $\operatorname{supp}(A) = A^{>0}$.

c) The crisp 1 cut $A^{\geq 1} = \{x \in X|\mu_A(x) = 1\}$ is called the *kernel* of A.

2. Representation Theorem

To every fuzzy subset A of X can be assigned uniquely the families of its α cuts $(A^{>\alpha})_{\alpha\in[0,1)}$ and its strong α cuts $\left(A^{\geq\alpha}\right)_{\alpha\in(0,1]}$. The α cuts and strong α cuts are monotone families of subsets from X, since:

$$\alpha < \beta \Rightarrow A^{>\alpha} \supseteq A^{>\beta} \quad \text{and} \quad A^{\geq\alpha} \supseteq A^{\geq\beta}. \tag{5.339a}$$

Conversely, if there exist the monotone families $(U_\alpha)_{\alpha\in[0,1)}$ or $(V_\alpha)_{\alpha\in(0,1]}$ of subsets from X, then there are uniquely defined fuzzy sets U and V such that $U^{>\alpha} = U_\alpha$ and $V^{\geq\alpha} = V_\alpha$ and moreover

$$\mu_U(x) = \sup\{\alpha \in [0,1))|x \in U_\alpha\}, \qquad \mu_V(x) = \sup\{\alpha \in (0,1]|x \in V_\alpha\}. \tag{5.339b}$$

7. Similarity of the Fuzzy Sets A and B

1. The fuzzy sets A, B with membership functions $\mu_A, \mu_B\colon X \to [0,1]$ are called fuzzy similar if for every $\alpha \in (0,1]$ there exist numbers α_i with $\alpha_i \in (0,1]; (i = 1,2)$ such that:

$$\operatorname{supp}(\alpha_1\mu_A)_\alpha \subseteq \operatorname{supp}(\mu_B)_\alpha, \qquad \operatorname{supp}(\alpha_2\mu_B)_\alpha \subseteq \operatorname{supp}(\mu_A)_\alpha. \tag{5.340}$$

$(\mu_C)_\alpha$ represents a fuzzy set with the membership function $(\mu_C)_\alpha = \begin{cases} \mu_C(x) & \text{if } \mu_C(x) > \alpha \\ 0 & \text{otherwise} \end{cases}$ and $(\beta\mu_C)$ represents a fuzzy set with the membership function $(\beta\mu_C) = \begin{cases} \beta \text{ if } \mu_C(x) > \beta \\ 0 \text{ otherwise.} \end{cases}$

2. Theorem: Two fuzzy sets A, B with membership functions $\mu_A, \mu_B\colon X \to [0,1]$ are fuzzy-similar if they have the same kernel:

$$\operatorname{supp}(\mu_A)_1 = \operatorname{supp}(\mu_B)_1, \tag{5.341a}$$

since the kernel is equal to the 1 cut, i.e.

$$\operatorname{supp}(\mu_A)_1 = \{x \in X|\mu_A(x) = 1\}. \tag{5.341b}$$

3. A, B with $\mu_A, \mu_B\colon X \to [0,1]$ are called strongly fuzzy-similar if they have the same support and the same kernel:

$$\mathrm{supp}(\mu_A)_1 = \mathrm{supp}(\mu_B)_1, \quad (5.342a) \qquad \mathrm{supp}(\mu_A)_0 = \mathrm{supp}(\mu_B)_0. \quad (5.342b)$$

5.9.2 Connections (Aggregations) of Fuzzy Sets

Fuzzy sets can be aggregated by operators. There are several different suggestions of how to generalize the usual set operations, such as union, intersection, and complement of fuzzy sets.

5.9.2.1 Concepts for Aggregations of Fuzzy Sets

1. Fuzzy Set Union, Fuzzy Set Intersection

The grade of membership of an arbitrary element $x \in X$ in the sets $A \cup B$ and $A \cap B$ should depend only on the grades of membership $\mu_A(x)$ and $\mu_B(x)$ of the element in the two fuzzy sets A and B. The union and intersection of fuzzy sets is defined with the help of two functions

$$s, t\colon [0,1] \times [0,1] \to [0,1], \qquad (5.343)$$

and they are defined in the following way:

$$\mu_{A\cup B}(x) := s\,(\mu_A(x), \mu_B(x))\,, \quad (5.344) \qquad \mu_{A\cap B}(x) := t\,(\mu_A(x), \mu_B(x))\,. \quad (5.345)$$

The grades of membership $\mu_A(x)$ and $\mu_B(x)$ are mapped in a new grade of membership. The functions t and s are called the t norm and t conorm; this last one is also called the s norm.

Interpretation: The functions $\mu_{A\cup B}$ and $\mu_{A\cap B}$ represent the truth values of membership, which is resulted by the aggregation of the truth values of memberships $\mu_A(x)$ and $\mu_B(x)$.

2. Definition of the *t*Norm:

The t norm is a binary operation t in $[0,1]$:

$$t\colon [0,1] \times [0,1] \to [0,1]. \qquad (5.346)$$

It is symmetric, associative, monotone increasing, it has 0 as the zero element and 1 as the neutral element. For $x, y, z, v, w \in [0,1]$ the following properties are valid:

(E1) Commutativity: $t(x,y) = t(y,x)$. (5.347a)

(E2) Associativity: $t(x, t(y,z)) = t(t(x,y), z)$. (5.347b)

(E3) Special Operations with Neutral and Zero Elements:

$t(x,1) = x$ and because of (E1): $t(1,x) = x$; $t(x,0) = t(0,x) = 0$. (5.347c)

(E4) Monotony: If $x \le v$ and $y \le w$, then $t(x,y) \le t(v,w)$ is valid. (5.347d)

3. Definition of the *s*Norm:

The s norm is a binary function in $[0,1]$:

$$s\colon [0,1] \times [0,1] \to [0,1]. \qquad (5.348)$$

It has the following properties:

(E1) Commutativity: $s(x,y) = s(y,x)$. (5.349a)

(E2) Associativity: $s(x, s(y,z)) = s(s(x,y), z)$. (5.349b)

(E3) Special Operations with Zero and Neutral Elements:

$s(x,0) = s(0,x) = x;\ s(x,1) = s(1,x) = 1$. (5.349c)

(E4) Monotony: If $x \le v$ and $y \le w$, then $s(x,y) \le s(v,w)$ is valid. (5.349d)

With the help of these properties a class T of t norms and a class S of s norms can be introduced. Detailed investigations proved that the following relations hold:

$$\min\{x,y\} \ge t(x,y)\ \forall\, t \in T,\ \forall\, x,y \in [0,1] \quad \text{and} \qquad (5.349e)$$

$$\max\{x,y\} \le s(x,y)\ \forall\, s \in S,\ \forall\, x,y \in [0,1]. \qquad (5.349f)$$

5.9.2.2 Practical Aggregation Operations of Fuzzy Sets

1. Intersection of Two Fuzzy Sets

The *intersection* $A \cap B$ of two fuzzy sets A and B is defined by the minimum operation min(.,.) on their membership functions $\mu_A(x)$ and $\mu_B(x)$. Based on the previous requirements there is:

$$C := A \cap B \text{ and } \mu_C(x) := \min\left(\mu_A(x), \mu_B(x)\right) \quad \forall\, x \in X, \quad \text{where:} \tag{5.350a}$$

$$\min(a,b) := \begin{cases} a, & \text{if } a \le b, \\ b, & \text{if } a > b. \end{cases} \tag{5.350b}$$

The intersection operation corresponds to the AND operation of two membership functions **(Fig.5.72)**. The membership function $\mu_C(x)$ is defined as the minimum value of $\mu_A(x)$ and $\mu_B(x)$.

2. Union of Two Fuzzy Sets

The *union* $A \cup B$ of two fuzzy sets is defined by the maximum operation max(.,.) on their membership functions $\mu_A(x)$ and $\mu_B(x)$:

$$C := A \cup B \text{ and } \mu_C(x) := \max\left(\mu_A(x), \mu_B(x)\right) \quad \forall\, x \in X, \quad \text{where:} \tag{5.351a}$$

$$\max(a,b) := \begin{cases} a\,, & \text{if } a \ge b, \\ b, & \text{if } a < b. \end{cases} \tag{5.351b}$$

The union corresponds to the logical OR operation. **Fig.5.73** illustrates $\mu_C(x)$ as the maximum value of the membership functions $\mu_A(x)$ and $\mu_B(x)$.

■ The t norm $t(x,y) = \min\{x,y\}$ and the s norm $s(x,y) = \max\{x,y\}$ define the intersection and the union of two fuzzy sets, respectively (see **(Fig.5.74)** and **(Fig.5.75)**).

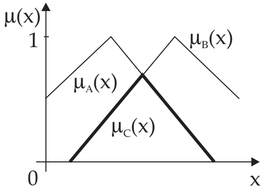


Figure 5.72

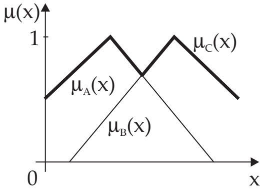


Figure 5.73

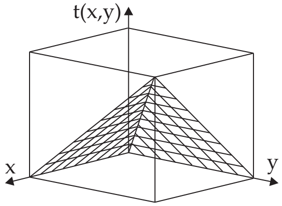


Figure 5.74

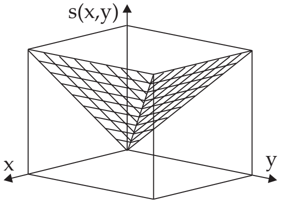


Figure 5.75

3. Further Aggregations

Further aggregations are the *bounded*, the *algebraic*, and the *drastic sum* and also the *bounded difference*, the *algebraic* and the *drastic product* (see **Table 5.8**).

The algebraic sum, e.g., is defined by

$$C := A + B \text{ and } \mu_C(x) := \mu_A(x) + \mu_B(x) - \mu_A(x) \cdot \mu_B(x) \quad \text{for every } x \in X. \tag{5.352a}$$

Similarly to the union (5.351a,b), this sum also belongs to the class of s norms. They are included in

Table 5.8 t and s norms, $p \in \mathbb{R}$

Author	***t* norm**	***s* norm**
Zadeh	intersection: $t(x,y) = \min\{x,y\}$	union: $s(x,y) = \max\{x,y\}$
Lukasiewicz	bounded difference $t_b(x,y) = \max\{0, x+y-1\}$	bounded sum $s_b(x,y) = \min\{1, x+y\}$
	algebraic product $t_a(x,y) = xy$	algebraic sum $s_a(x,y) = x+y-xy$
	drastic product $t_{dp}(x,y) = \begin{cases} \min\{x,y\}, \text{ whether } x=1 \\ \qquad\qquad \text{or } y=1 \\ 0 \;\text{ otherwise} \end{cases}$	drastic sum $s_{ds}(x,y) = \begin{cases} \max\{x,y\}, \text{ whether } x=0 \\ \qquad\qquad \text{or } y=0 \\ 1 \;\text{ otherwise} \end{cases}$
Hamacher $(p \geq 0)$	$t_h(x,y) = \dfrac{xy}{p+(1-p)(x+y-xy)}$	$s_h(x,y) = \dfrac{x+y-xy-(1-p)xy}{1-(1-p)xy}$
Einstein	$t_e(x,y) = \dfrac{xy}{1+(1-x)(1-y)}$	$s_e(x,y) = \dfrac{x+y}{1+xy}$
Frank $(p > 0, p \neq 1)$	$t_f(x,y) =$ $\log_p \left[1 + \dfrac{(p^x-1)(p^y-1)}{p-1}\right]$	$s_f(x,y) = 1-$ $\log_p \left[1 + \dfrac{(p^{1-x}-1)(p^{1-y}-1)}{p-1}\right]$
Yager $(p > 0)$	$t_{ya}(x,y) = 1-$ $\min\left(1, ((1-x)^p + (1-y)^p)^{1/p}\right)$	$s_{ya}(x,y) = \min\left(1, (x^p+y^p)^{1/p}\right)$
Schweizer $(p > 0)$	$t_s(x,y) = \max(0, x^{-p}+y^{-p}-1)^{-1/p}$	$s_s(x,y) = 1-$ $\max\left(0, (1-x)^{-p} + (1-y)^{-p} - 1\right)^{-1/p}$
Dombi $(p > 0)$	$t_{do}(x,y) =$ $\left\{1 + \left[\left(\dfrac{1-x}{x}\right)^p + \left(\dfrac{1-y}{y}\right)^p\right]^{1/p}\right\}^{-1}$	$s_{do}(x,y) = 1-$ $\left\{1 + \left[\left(\dfrac{x}{1-x}\right)^p + \left(\dfrac{y}{1-y}\right)^p\right]^{1/p}\right\}^{-1}$
Weber $(p \geq -1)$	$t_w(x,y) = \max(0, (1+p)$ $\cdot(x+y-1) - pxy)$	$s_w(x,y) = \min(1, x+y+pxy)$
Dubois $(0 \leq p \leq 1)$	$t_{du}(x,y) = \dfrac{xy}{\max(x,y,p)}$	$s_{du}(x,y) =$ $\dfrac{x+y-xy-\min(x,y,(1-p))}{\max((1-x),(1-y),p)}$

Remark: For the values of the t and s norms listed in the table, the following ordering is valid:
$t_{dp} \leq t_b \leq t_e \leq t_a \leq t_h \leq t \leq s \leq s_h \leq s_a \leq s_e \leq s_b \leq s_{ds}$.

the right-hand column of **Table 5.8**. In **Table 5.9** is given a comparision of operations in Boolean logic and fuzzy logic.

Analogously to the notion of the extended sum as a union operation, the intersection can also be extended for example by the bounded, the algebraic, and the drastic product. So, e.g., the algebraic product is defined in the following way:

$$C := A \cdot B \text{ and } \mu_C(x) := \mu_A(x) \cdot \mu_B(x) \quad \text{for every } x \in X. \tag{5.352b}$$

It also belongs to the class of t norms, similarly to the intersection (5.350a,b), and it can be found in the middle column of **Table 5.8**.

5.9.2.3 Compensatory Operators

Sometimes operators are necessary lying between the t and the s norms; they are called compensatory operators. Examples for compensatory operators are the lambda and the gamma operator.

1. Lambda Operator

$$\mu_{A\lambda B}(x) = \lambda\,[\mu_A(x)\mu_B(x)] + (1-\lambda)\,[\mu_A(x) + \mu_B(x) - \mu_A(x)\mu_B(x)] \quad \text{with} \quad \lambda \in [0,1]. \tag{5.353}$$

Case $\lambda = 0$: Equation (5.353) results in a form known as the algebraic sum (**Table 5.8**, s norms); it belongs to the OR operators.

Case $\lambda = 1$: Equation (5.353) results in the form known as the algebraic product (**Table 5.8**, t norms); it belongs to the AND operators.

2. Gamma Operator

$$\mu_{A\gamma B}(x) = [\mu_A(x)\mu_B(x)]^{1-\gamma}\,[1-(1-\mu_A(x))\,(1-\mu_B(x))]^{\gamma} \quad \text{with } \gamma \in [0,1]. \tag{5.354}$$

Case γ**= 1:** Equation (5.354) results in the representation of the algebraic sum.

Case γ**= 0:** Equation (5.354) results in the representation of the algebraic product.

The application of the gamma operator on fuzzy sets of any numbers is given by

$$\mu(x) = \left[\prod_{i=1}^{n}\mu_i(x)\right]^{1-\gamma}\left[1-\prod_{i=1}^{n}(1-\mu_i(x))\right]^{\gamma}, \tag{5.355}$$

and with weights δ_i:

$$\mu(x) = \left[\prod_{i=1}^{n}\mu_i(x)^{\delta_i}\right]^{1-\gamma}\left[1-\prod_{i=1}^{n}(1-\mu_i(x))^{\delta_i}\right]^{\gamma} \quad \text{with } x \in X, \quad \sum_{i=1}^{n}\delta_i = 1, \quad \gamma \in [0,1]. \tag{5.356}$$

5.9.2.4 Extension Principle

In the previous paragraph there are discussed the possibilities of generalizing the basic set operations for fuzzy sets. Now, the notion of mapping is extended on fuzzy domains. The basis of the concept is the *acceptance grade* of vague statements. The classical mapping Φ: $X^n \to Y$ assigns a crisp function value $\Phi(x_1, \ldots, x_n) \in Y$ to the point $(x_1, \ldots, x_n) \in X^n$. This mapping can be extended for fuzzy variables as follows: The fuzzy mapping is $\hat{\Phi}$: $F(X)^n \to F(Y)$, which assigns a fuzzy function value $\hat{\Phi}(\mu_1, \ldots, \mu_n)$ to the fuzzy vector variables $(x_1, \ldots, x_n)$ given by the membership functions $(\mu_1, \ldots, \mu_n) \in F(X)^n$.

5.9.2.5 Fuzzy Complement

A function c: $[0,1] \to [0,1]$ is called a *complement function* if the following properties are fulfilled for $\forall\, x, y \in [0,1]$:

(EK1) Boundary Conditions: $c(0) = 1$ and $c(1) = 0$. (5.357a)

(EK2) Monotony: $x < y \Rightarrow c(x) \ge c(y)$. (5.357b)

(EK3) Involutivity: $c(c(x)) = x$. (5.357c)

(EK4) Continuity: $c(x)$ should be continuous for every $x \in [0,1]$. (5.357d)

■ **A:** The most often used complement function is (continuous and involutive):

$$c(x) := 1 - x. \tag{5.358}$$

■ **B:** Other continuous and involutive complements are the *Sugeno complement* $c_\lambda(x) := (1-x)(1+\lambda x)^{-1}$ with $\lambda \in (-1, \infty)$ and the *Yager complement* $c_p(x) := (1-x^p)^{1/p}$ with $p \in (0, \infty)$.

Table 5.9 Comparison of operations in Boolean logic and in fuzzy logic

Operator	**Boolean logic**	**Fuzzy logic** $(\boldsymbol{\mu_A, \mu_B \in [0,1]})$
AND	$C = A \wedge B$	$\mu_{A\cap B} = \min(\mu_A, \mu_B)$
OR	$C = A \vee B$	$\mu_{A\cup B} = \max(\mu_A, \mu_B)$
NOT	$C = \neg A$	$\mu_A^C = 1 - \mu_A$ (μ_A^C as complement of μ_A)

5.9.3 Fuzzy-Valued Relations

5.9.3.1 Fuzzy Relations

1. Modeling Fuzzy-Valued Relations

Uncertain or fuzzy-valued relations, as e.g. "approximately equal", "practically larger than", or "practically smaller than", etc., have an important role in practical applications. A relation between numbers is interpreted as a subsets of $\mathbb{R}^2$. So, the equality "=" is defined as the set

$$\mathcal{A} = \left\{(x,y) \in \mathbb{R}^2 | x = y\right\}, \tag{5.359}$$

i.e., by a straight line $y = x$ in $\mathbb{R}^2$.
Modeling the relation "approximately equal" denoted by R_1, can be used a fuzzy subset on $\mathbb{R}^2$, the kernel of which is $\mathcal{A}$. Furthermore it is to require that the membership function should decrease and tend to zero getting far from the line $\mathcal{A}$. A linear decreasing membership function can be modeled by

$$\mu_{R_1}(x,y) = \max\{0, 1 - a|x-y|\} \quad \text{with} \quad a \in \mathbb{R},\ a > 0. \tag{5.360}$$

For modeling the relation R_2 "practically larger than", it is useful to start with the crisp relation "$\geq$". The corresponding set of values is given by

$$\left\{(x,y) \in \mathbb{R}^2 | x \leq y\right\}. \tag{5.361}$$

It describes the crisp domain above the line $x = y$.
The modifier "practically" means that a thin zone under the half-space in (5.361) is still acceptable with some grade. So, the model of R_2 is

$$\mu_{R_2}(x,y) = \left\{\begin{array}{ll} \max\{0, 1 - a|x-y|\} & \text{for } y < x \\ 1 & \text{for } y \geq x \end{array}\right\} \quad \text{with} \quad a \in \mathbb{R},\ a > 0. \tag{5.362}$$

If the value of one of the variables is fixed, e.g., $y = y_0$, then R_2 can be interpreted as a region with uncertain boundaries for the other variable.
Handling the uncertain boundaries by fuzzy relations has practical importance in fuzzy optimization, qualitative data analysis and pattern classification.

The foregoing discussion shows that the concept of fuzzy relations, i.e., fuzzy relations between several objects, can be described by fuzzy sets. In the following section the basic properties of binary relations are discussed over a universe which consists of ordered pairs.

2. Cartesian Product

Let X and Y be two universes. Their "cross product" $X \times Y$, or *Cartesian product*, is a universe G:

$$G = X \times Y = \{(x,y) | x \in X \wedge y \in Y\}. \tag{5.363}$$

Then, a fuzzy set on G is a fuzzy relation, analogously to classical set theory, if it consists of the valued pair of universes X and Y. A fuzzy relation R in G is a fuzzy subset $R \in F(G)$, where $F(G)$ denotes the set of all the fuzzy sets over $X \times Y$. R can be given by a membership function $\mu_R(x,y)$ which assigns a membership degree $\mu_R(x,y)$ from $[0,1]$ to every element of $(x,y) \in G$.

3. Properties of Fuzzy-Valued Relations

(E1) Since the fuzzy relations are special fuzzy sets, all propositions stated for fuzzy sets will also be valid for fuzzy relations.

(E2) All aggregations defined for fuzzy sets can be defined also for fuzzy relations; they yield a fuzzy relation again.

(E3) The notion of α cut defined above can be transmitted without difficulties to fuzzy relations.

(E4) The 0 cut (the closure of the support) of a fuzzy relation $R \in F(G)$ is a usual relation on G.

(E5) Denoting the membership value by $\mu_R(x, y)$, i.e., the degree by which the relation R between the pair (x, y) holds. The value $\mu_R(x, y) = 1$ means that R holds perfectly for the pair (x, y), and the value $\mu_R(x, y) = 0$ means that R does not at all hold for the pair (x, y).

(E6) Let $R \in F(G)$ be a fuzzy relation. Then the fuzzy relation $S := R^{-1}$, the inverse of R, is defined by

$$\mu_S(x, y) = \mu_R(y, x) \quad \text{for every } (x, y) \in G. \tag{5.364}$$

■ The inverse relation R_2^{-1} means “practically smaller than” (see 5.9.3.1, **1.**, p. 422); the union $R_1 \cup R_2^{-1}$ can be determined as “practically smaller or approximately equal”.

4. n-Fold Cartesian Product

Let n be the number of universal sets. Their *cross product* is an n-fold *Cartesian product*. A fuzzy set on an n-fold Cartesian product represents an n-fold fuzzy relation.

Consequences: The fuzzy sets, considered until now, are unary fuzzy relations, i.e., in the sense of the analysis they are curves above a universal set. A binary fuzzy relation can be considered as a surface over the universal set G. A binary fuzzy relation on a finite discrete support can be represented by a *fuzzy relation matrix*.

■ Colour-ripe grade relation: The well-known correspondence between the colour x and the ripe grade y of a friut is modeled in the form of a binary relation matrix with elements $\{0, 1\}$. The possible colours are $X = \{\text{green, yellow, red}\}$ and the ripe grades are $Y = \{\text{unripe, half-ripe, ripe}\}$. The relation matrix (5.365) belongs to the table:

	unripe	half-ripe	ripe
green	1	0	0
yellow	0	1	0
red	0	0	1

$$R = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{5.365}$$

Interpretation of this relation matrix: IF a fruit is green, THEN it is unripe. IF a fruit is yellow, THEN it is half-ripe. IF a fruit is red, THEN it is ripe. Green is uniquely assigned to unripe, yellow to half-ripe and red to ripe. If beyond it should be formalized that a green fruit can be considered half-ripe in a certain percentage, then the following table with discrete membership values can be arranged:

μ_R (green, unripe)	$= 1.0,$	μ_R (green, half-ripe)	$= 0.5,$
μ_R (green, ripe)	$= 0.0,$	μ_R (yellow, unripe)	$= 0.25,$
μ_R (yellow, half-ripe)	$= 1.0,$	μ_R (yellow, ripe)	$= 0.25,$
μ_R (red, unripe)	$= 0.0,$	μ_R (red, half-ripe)	$= 0.5,$
μ_R (red, ripe)	$= 1.0.$		

The relation matrix with $\mu_R \in [0, 1]$ is:

$$R = \begin{pmatrix} 1.0 & 0.5 & 0.0 \\ 0.25 & 1.0 & 0.25 \\ 0.0 & 0.5 & 1.0 \end{pmatrix}. \tag{5.366}$$

5. Rules of Calculations

The AND-type aggregation of fuzzy sets, e.g. $\mu_1 : X \to [0, 1]$ and $\mu_2 : Y \to [0, 1]$ given on different universes is formulated by the min operation as follows:

$$\mu_R(x, y) = \min(\mu_1(x), \mu_2(y)) \text{ or } (\mu_1 \times \mu_2)(x, y) = \min(\mu_1(x), \mu_2(y)) \quad \text{with} \tag{5.367a}$$

$$\mu_1 \times \mu_2 \colon G \to [0, 1], \text{ where } \quad G = X \times Y. \tag{5.367b}$$

The result of this aggregation is a fuzzy relation R on the cross product set (Cartesian product universe of fuzzy sets) G with $(x, y) \in G$. If X and Y are discrete finite sets and so $\mu_1(x), \mu_2(y)$ can be represented as vectors, then holds:

$$\mu_1 \times \mu_2 = \mu_1 \circ \mu_2^{\mathrm{T}} \quad \text{and} \quad \mu_{R^{-1}}(x, y) := \mu_R(y, x) \quad \forall\, (x, y) \in G. \tag{5.368}$$

The *aggregation operator* $\circ$ does not denote here the usual matrix product. The product is calculated here by the componentwise min operation and addition by the componentwise max operation.

The validity grade of an inverse relation R^{-1} for the pair (x, y) is always equal to the validity grade of R for the pair (y, x).

If the fuzzy relations are given on the same Cartesian product universe, then the rules of their aggregations can be given as follows: Let $R_1, R_2\colon X \times Y \to [0,1]$ be binary fuzzy relations. The evaluation rule of their AND-type aggregation uses the min operator, namely for $\forall(x, y) \in G$:

$$\mu_{R_1 \cap R_2}(x, y) = \min(\mu_{R_1}(x, y), \mu_{R_2}(x, y)). \tag{5.369}$$

A corresponding evaluation rule for the OR-type aggregation is given by the max operation:

$$\mu_{R_1 \cup R_2}(x, y) = \max(\mu_{R_1}(x, y), \mu_{R_2}(x, y)). \tag{5.370}$$

5.9.3.2 Fuzzy Product Relation $R \circ S$

1. Composition or Product Relation

Suppose $R \in F(X \times Y)$ and $S \in F(Y \times Z)$ are two relations, and it is additionally assumed that $R, S \in F(G)$ with $G \subseteq X \times Z$. Then the *composition* or the *fuzzy product relation* $R \circ S$ is:

$$\mu_{R \circ S}(x, z) := \sup_{y \in Y}\{\min(\mu_R(x, y), \mu_S(y, z))\}\ \forall\,(x, z) \in X \times Z. \tag{5.371}$$

If a matrix representation is used for a finite universal set analogously to (5.366), then the composition $R \circ S$ is motivated as follows: Let $X = \{x_1, \ldots, x_n\}, Y = \{y_1, \ldots, y_m\}, Z = \{z_1, \ldots, z_l\}$ and $R \in F(X \times Y)$, $S \in F(Y \times Z)$ and let the matrix representations R, S be in the form $R = (r_{ij})$ and $S = (s_{jk})$ for $i = 1, \ldots, n; j = 1, \ldots, m; k = 1, \ldots, l$, where

$$r_{ij} = \mu_R(x_i, y_j) \quad \text{and} \quad s_{jk} = \mu_S(y_j, z_k). \tag{5.372}$$

If the composition $T = R \circ S$ has the matrix representation t_{ik}, then

$$t_{ik} = \sup_j \min\{r_{ij}, s_{jk}\}. \tag{5.373}$$

The final result is not a usual matrix product, since instead of the summation operation there is the least upper bound (supremum) operation and instead of the product there is the minimum operator.

■ With the representations for r_{ij} and s_{jk} and with (5.371), the inverse relation $R^{-1}(r_{i,j})^{\mathrm{T}}$, can also be computed taking into consideration that R^{-1} can be represented by the transpose matrix, i.e., $R^{-1} = (r_{ij})^{\mathrm{T}}$.

Interpretation: Let R be a relation from X to Y and S be a relation from Y to Z. Then the following compositions are possible:

a) If the composition $R \circ S$ of R and S is defined as a max-min product, then the resulted fuzzy composition is called a max-min composition. The symbol sup stands for supremum and denotes the largest value, if no maximum exists.

b) If the product composition is defined as the usual matrix multiplication, then the max-prod composition is obtained.

c) For max-average composition, “multiplication” is replaced by the average.

2. Rules of Composition

The following rules are valid for the composition of fuzzy relations $R, S, T \in F(G)$:

(E1) Associative Law:

$$(R \circ S) \circ T = R \circ (S \circ T). \tag{5.374}$$

(E2) Distributive Law for Composition with Respect to the Union:

$$R \circ (S \cup T) = (R \circ S) \cup (R \circ T). \tag{5.375}$$

(E3) Distributive Law in a Weaker Form for Composition with Respect to Intersection:

$$R \circ (S \cap T) \subseteq (R \circ S) \cap (R \circ T). \tag{5.376}$$

(E4) Inverse Operations:

$$(R \circ S)^{-1} = S^{-1} \circ R^{-1}, \quad (R \cup S)^{-1} = R^{-1} \cup S^{-1} \quad \text{and} \quad (R \cap S)^{-1} = R^{-1} \cap S^{-1}. \tag{5.377}$$

(E5) Complement and Inverse:

$$\left(R^{-1}\right)^{-1} = R, \quad \left(R^{C}\right)^{-1} = \left(R^{-1}\right)^{C}. \tag{5.378}$$

(E6) Monotonic Properties:

$$R \subseteq S \Rightarrow R \circ T \subseteq S \circ T \quad \text{und } T \circ R \subseteq T \circ S. \tag{5.379}$$

■ **A:** Equation (5.371) for the product relation $R \circ S$ is defined by the min operation as we have done for intersection formation. In general, any t norm can be used instead of the min operation.

■ **B:** The α cuts with respect to the union, intersection, and complement are: $(A \cup B)^{>\alpha} = A^{>\alpha} \cup B^{>\alpha}$, $(A \cap B)^{>\alpha} = A^{>\alpha} \cap B^{>\alpha}$, $(A^C)^{>\alpha} = A^{\leq 1-\alpha} = \{x \in X | \mu_A(x) \leq 1 - \alpha\}$. Corresponding statements are valid for strong α cuts.

3. Fuzzy Logical Inferences

It is possible to make a fuzzy inference, e.g., with the IF THEN rule by the composition rule $\mu_2 = \mu_1 \circ R$. The detailed formulation for the conclusion μ_2 is given by

$$\mu_2(y) = \max_{x \in X}\Big(\min(\mu_1(x),\, \mu_R(x,y))\Big) \tag{5.380}$$

with $y \in Y,\ \mu_1\colon X \to [0,1],\ \mu_2\colon Y \to [0,1],\ R\colon G \to [0,1]$ und $G = X \times Y$.

5.9.4 Fuzzy Inference (Approximate Reasoning)

Fuzzy inference is an application of fuzzy relations with the goal of getting fuzzy logical conclusions with respect to vague information (see 5.9.6.3, p. 428). Vague information means here fuzzy information but not uncertain information. Fuzzy inference, also called *implication*, contains one or more rules, a fact and a consequence. Fuzzy inference, which is called by Zadeh, approximate reasoning, cannot be described by classical logic.

1. Fuzzy Implication, IF THEN Rule

The fuzzy implication contains one IF THEN rule in the simplest case. The IF part is called the *premise* and it represents the condition. The THEN part is the *conclusion*. Evaluation happens by $\mu_2 = \mu_1 \circ R$ and (5.380).

Interpretation: μ_2 is the fuzzy inference image of μ_1 under the fuzzy relation R, i.e., a calculation prescription for the IF THEN rule or for a group of rules.

2. Generalized Fuzzy Inference Scheme

The rule IF A_1 AND A_2 AND $A_3 \ldots$ AND A_n THEN B with $A_i\colon\ \mu_i\colon\ X_i \to [0,1]$ $(i = 1, 2, \ldots, n)$ and the membership function of the conclusion B: $\mu\colon\ Y \to [0,1]$ is described by an $(n+1)$-valued relation

$$\mathrm{R}\colon\ X_1 \times X_2 \times \cdots X_n \times Y \to [0,1]. \tag{5.381a}$$

For the actual input with crisp values $x_1', x_2', \ldots, x_n'$ the rule (5.381a) defines the actual fuzzy output by

$$\mu_{B'}(y) = \mu_R(x_1', x_2', \ldots, x_n', y) = \min(\mu_1(x_1'), \mu_2(x_2'), \ldots, \mu_n(x_n'), \mu_B(y)) \text{ where } y \in Y. \tag{5.381b}$$

Remark: The quantity $\min(\mu_1(x_1'),\ \mu_2(x_2'),\ \ldots\ \mu_n(x_n'))$ is called the *degree of fulfillment*, and the quantities $\{\mu_1(x_1'),\ \mu_2(x_2'),\ \ldots,\mu_n(x_n')\}$ represent the fuzzy-valued input quantities.

■ Forming the fuzzy relations for a connection between the quantities "medium" pressure and "high" temperature **(Fig. 5.76)**: $\tilde{\mu}_1(p,T) = \mu_1(p)\ \forall T \in X_2$ with $\mu_1 : X_1 \to [0,1]$ is a cylindrical extension **(Fig. 5.76c)** of the fuzzy set medium pressure **(Fig. 5.76a)**. Analogously, $\tilde{\mu}_2(p,T) = \mu_2(T)\ \forall p \in X_1$ with $\mu_2 :\ X_2 \to [0,1]$ is a cylindrical extension **(Fig. 5.76d)** of the fuzzy set high temperature **(Fig. 5.76b)**, where $\tilde{\mu}_1, \tilde{\mu}_2\colon G = X_1 \times X_2 \to [0,1]$.

Fig. 5.77a shows the graphic result of the formation of fuzzy relations: In **Fig. 5.77b** the result of the composition medium pressure AND high temperature with the min operator $\mu_R(p,T) = \min(\mu_1(p), \mu_2(T))$ is represented, and **(Fig. 5.77b)** shows the result of the composition OR with the max operator $\mu_R(p,T) = \max(\mu_1(p), \mu_2(T))$.

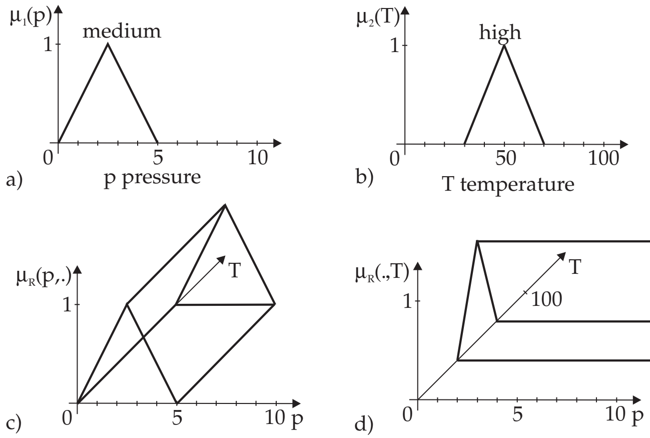


Figure 5.76

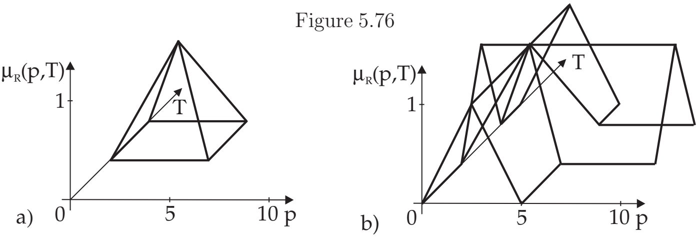


Figure 5.77

5.9.5 Defuzzification Methods

One has to get a crisp set from a fuzz-valued set in many cases. This process is called *defuzzification*. There are different methods available for this task.

1. Maximum-Criterion Method

An arbitrary value $\eta \in Y$ is selected from the domain where the fuzzy set $\mu^{\text{Output}}_{x_1,\ldots,x_n}$ has the maximal membership degree.

2. Mean-of-Maximum Method (MOM)

The output value is the mean value of the maximal membership values:

$$\sup\left(\mu^{\text{Output}}_{\mu_{x_1,\ldots,x_n}}\right) :=$$
$$\{y \in Y | \mu_{x_1,\ldots,x_n}(y) \geq \mu_{x_1,\ldots,x_n}(y^*)\ \forall\, y^* \in Y\}\ ; \quad (5.382)$$

$$\eta_{\text{MOM}} = \frac{\displaystyle\int_{y \in \sup\left(\mu^{\text{Output}}_{x_1,\ldots,x_n}\right)} y\,dy}{\displaystyle\int_{y \in \sup\left(\mu^{\text{Output}}_{x_1,\ldots,x_n}\right)} dy} . \quad (5.383)$$

i.e., the set Y is an interval, which should not be empty and it is characterized by (5.382), from which follows (5.383).

3. Center of Gravity Method (COG)

In the center of gravity method, one takes the abscissa value of the center of gravity of a surface with a fictitious homogeneous density of value 1.

$$\eta_{\text{COG}} = \frac{\displaystyle\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y) y\,dy}{\displaystyle\int_{y_{\text{inf}}}^{y_{\text{sup}}} \mu(y)\,dy} . \quad (5.384)$$

4. Parametrized Center of Gravity Method (PCOG)
The parametrized method works with the exponent $\gamma \in \mathbb{R}$. From (5.385) it follows for $\gamma = 1, \quad \eta_{\text{PCOG}} = \eta_{\text{COG}}$ and for $\gamma \to 0, \quad \eta_{\text{PCOG}} = \eta_{\text{MOM}}$.

$$\eta_{\text{PCOG}} = \frac{\displaystyle\int_{y_{\inf}}^{y_{\sup}} \mu(y)^{\gamma} y \, dy}{\displaystyle\int_{y_{\inf}}^{y_{\sup}} \mu(y)^{\gamma} \, dy} . \qquad (5.385)$$

5. Generalized Center of Gravity Method (GCOG)
The exponent γ is considered as a function of y in the PCOG method. Then (5.386) follows obviously. The GCOG method is a generalization of the PCOG method, where $\mu(y)$ can be changed by the special weight γ depending itself on y.

$$\eta_{\text{GCOG}} = \frac{\displaystyle\int_{y_{\inf}}^{y_{\sup}} \mu(y)^{\gamma(y)} y \, dy}{\displaystyle\int_{y_{\inf}}^{y_{\sup}} \mu(y)^{\gamma(y)} \, dy} . \qquad (5.386)$$

6. Center of Area (COA) Method
One calculates a line parallel to the ordinate axis so that the area under the membership function is the same on the left- and on the right-hand side of it.

$$\int_{y_{\inf}}^{\eta} \mu(y) \, dy = \int_{\eta}^{y_{\sup}} \mu(y) \, dy. \qquad (5.387)$$

7. Parametrized Center of Area (PCOA) Method

$$\int_{y_{\inf}}^{\eta_{\text{PB}}} \mu(y)^{\gamma} \, dy = \int_{\eta_{\text{PF}}}^{y_{\sup}} \mu(y)^{\gamma} \, dy. \qquad (5.388)$$

8. Method of the Largest Area (LA)
The significant subset is selected and one of the methods defined above, e.g., the method of center of gravity (COG) or center of area (COA) is used for this subset.

5.9.6 Knowledge-Based Fuzzy Systems

There are several application possibilities of multi-valued fuzzy logic, based on the unit interval, both in technical and non-technical life. The general concept consists in the fuzzification of quantities and characteristic numbers, in the aggregation them in an appropriate knowledge base with operators, and if necessary, in the defuzzification of the possibly fuzzy result set.

5.9.6.1 Method of Mamdani

The following steps are applied for a fuzzy control process:

1. Rule Base Suppose, for example, for the i-th rule

$$\mathrm{R}^i : \text{ If } e \text{ is } E^i \text{ AND } \dot{e} \text{ is } \Delta E^i \text{ THEN } u \text{ is } U^i. \qquad (5.389)$$

Here e characterizes the error, $\dot{e}$ the change of the error and u the change of the (not fuzzy valued) output value. Every quantity is defined on its domain $E, \Delta E$ and U. Let the entire domain be $E \times \Delta E \times U$. The error and the change of the error will be fuzzified on this domain, i.e., they will be represented by fuzzy sets, where linguistic description is used.

2. Fuzzifying Algorithm In general, the error e and its change $\dot{e}$ are not fuzzy-valued, so they must be fuzzified by a linguistic description. The fuzzy values will be compared with the premisses of the IF THEN rule from the rule base. From this it follows, which rules are active and how large are their weights.

3. Aggregation Module The active rules with their different weights will be combined with an algebraic operation and applied to the defuzzification.

4. Decision Module In the defuzzification process a crisp value should be given for the control quantity. With a defuzzification operation, a non-fuzzy-valued quantity is determined from the set of possible values, i.e., a crisp quantity. This quantity expresses how the control parameters of the system should be set up to keep the deviation minimal.

Fuzzy control means that the steps from **1.** to **4.** are repeated until the goal, the smallest deviation e and its change $\dot{e}$, is reached.

5.9.6.2 Method of Sugeno

The Sugeno method is also used for planning of a fuzzy control process. It differs from the Mamdani concept in the rule base and in the defuzzification method. It has the following steps:

1. Rule Base: The rule base consists of rules of the following form:

$$R^i: \text{ IF } x_1 \text{ is } A_1^i \text{ AND } \ldots \text{ AND } x_k \text{ is } A_k^i, \text{THEN } u_i = p_0^i + p_1^i x_1 + p_2^i x_2 + \cdots + p_k^i x_k. \tag{5.390}$$

The notations mean:

A_j: fuzzy sets, which can be determined by membership functions;

x_j: crisp input values as, e.g., the error e and the change of the error $\dot{e}$, which tell us something about the dynamics of the system;

p_j^i: weights of x_j $(j = 1, 2, \ldots, k)$;

u_i: the output value belonging to the i-th rule $(i = 1, 2, \ldots, n)$.

2. Fuzzifying Algorithm: A $\mu_i \in [0, 1]$ is calculated for every rule R^i.

3. Decision Module: A non-fuzzy-valued quantity is calculated from the weighted mean of u_i, where the weights are μ_i from the fuzzification:

$$u = \sum_{i=1}^{n} \mu_i u_i \left(\sum_{i=1}^{n} \mu_i \right)^{-1}. \tag{5.391}$$

Here u is a crisp value.

The defuzzification of the Mamdani method does not work here. The problem is to get the weight parameters p_j^i available. These parameters can be determined by a mechanical learning method, e.g., by an artificial neuronetwork (ANN).

5.9.6.3 Cognitive Systems

To clarify the method, the following known example will be investigated with the Mamdami method: The regulation of a pendulum that is perpendicular to its moving base **(Fig. 5.78)**. The aim of the control process is to keep a pendulum in balance so that the pendulum rod should stand vertical, i.e., the angular displacement from the vertical direction and the angular velocity should be zero. It must be done by a force F acting at the lower end of the pendulum. This force is the control quantity. The model is based on the activity of a human "control expert" (cognitive problem). The expert formulates its knowledge in linguistic rules. Linguistic rules consist, in general, of a premise, i.e., a specification of the measured values, and a conclusion which gives the appropriate control value.

For every set of values $X_1, X_2, \ldots, X_n$ for the measured values and Y for the control quantity the appropriate linguistic terms are defined as "approximately zero", "small positive", etc. Here "approximately zero" with respect to the measured value ξ_1 can have a different meaning as for the measured value ξ_2.

■ **Inverse Pendulum on a Moving Base (Fig. 5.78)**

1. Modeling For the set X_1 (values of angle) and analogously for the input quantity X_2 (values of the angular velocity) the seven linguistic terms, negative large (nl), negative medium (nm), negative small (ns), zero (z), positive small (ps), positive medium (pm) and positive large (pl) are chosen.

For the mathematical modeling, a fuzzy set must be assigned by graphs to every one of these linguistic terms **(Fig. 5.77)**, as was shown for fuzzy inference (see 5.9.4, p. 425).

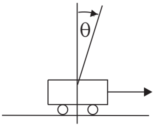


Figure 5.78

2. Choice of the Domain of Values

- Values of angles: $\Theta(-90° < \Theta < 90°)$: $X_1 := [-90°, 90°]$.
- Values of angular velocity: $\dot{\Theta}(-45°\,\text{s}^{-1} \le \dot{\Theta} \le 45°\,\text{s}^{-1})$: $X_2 := [-45°\,\text{s}^{-1}, 45°\,\text{s}^{-1}]$.
- Values of force F: $(-10\,\text{N} \le F \le 10\,\text{N})$: $Y := [-10\text{N}, 10\,\text{N}]$.

The partitioning of the input quantities X_1 and X_2 and the output quantity Y is represented graphically in **Fig. 5.79**. Usually, the initial values are actual measured values, e.g., $\Theta = 36°, \dot{\Theta} = -2.25°\,\mathrm{s}^{-1}$.

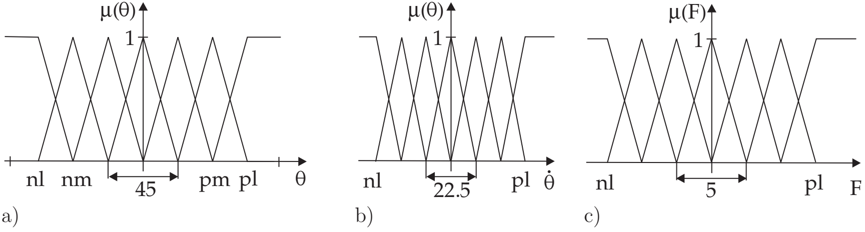


Figure 5.79

3. Choice of Rules Considering the following table, there are 49 possible rules (7 × 7) but there are only 19 important in practice, so the following two are to be discussed: **R1** and **R2**.

R1: If Θ is positive small (ps) and $\dot{\Theta}$ zero (z), then F is positive small (ps). For the *degree of fulfillment* (also called the *weight of the rules*) of the premise with $\alpha = \min\left\{\mu^{(1)}(\Theta); \mu^{(1)}(\dot{\Theta})\right\} = \min\{0.4; 0.8\} = 0.4$ one gets the output set (5.392) by an α cut, hence the output fuzzy set is positive small (ps) in the height $\alpha = 0.4$ **(Fig. 5.80c)**.

Table: Rule base with 19 practically meaningful rules

$\dot{\Theta}\backslash\Theta$	nl	nm	ns	z	ps	pm	pl
nl			ps	pl			
nm				pm			
ns	nm		ns	ps			
z	nl	nm	ns	z	ps	pm	pl
ps				ns	ps		pm
pm				nm			
pl				nl	ns		

$$\mu_{36;-2.25}^{\text{Output (R1)}}(y) = \begin{cases} \frac{2}{5}y & 0 \le y < 1, \\ 0.4 & 1 \le y \le 4, \\ 2 - \frac{2}{5}y & 4 < y \le 5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.392}$$

R2: If Θ is positive medium (pm) and $\dot{\Theta}$ is zero (z), then F is positive medium (pm). For the performance score of the premise follows $\alpha = \min\left\{\mu^{(2)}(\Theta); \mu^{(2)}\dot{\Theta}\right\} = \min\{0.6; 0.8\} = 0.6$, the output set (5.393) analogously to rule **R1 (Fig. 5.80f)**.

$$\mu_{36;-2.25}^{\text{Output (R2)}}(y) = \begin{cases} \frac{2}{5}y - 1 & 2.5 \le y < 4, \\ 0.6 & 4 \le y \le 6, \\ 3 - \frac{2}{5}y & 6 < y \le 7.5, \\ 0 & \text{otherwise.} \end{cases} \tag{5.393}$$

4. Decision Logic The evaluation of rule R$_1$ with the min operation results in the fuzzy set in **Figs. 5.80a–c**. The corresponding evaluation for the rule R$_2$ is shown in **Figs. 5.80d–f**. The control quantity is calculated finally by a defuzzification method from the fuzzy proposition set **(Fig. 5.80g)**. The result is the fuzzy set **(Fig. 5.80g)** by using the max operation and taking into account the fuzzy sets **(Fig. 5.80c)** and **(Fig. 5.80f)**.

a) Evaluation of the fuzzy set obtained in this way, which is aggregated by operators (see max-min composition 5.9.3.2, **1.**, p. 424). The decision logic yields:

$$\mu_{x_1,\dots,x_n}^{\text{Output}} : Y \to [0,1]\,; y \to \ \max_{\mathrm{r}\in\{1,\dots,\mathrm{k}\}} \left\{\min\left\{\mu_{il,r}^{(1)}(x_1), \dots, \mu_{il,r}^{(n)}(x_n), \mu_{i_r}(y)\right\}\right\}. \tag{5.394}$$

b) After taking the maximum (5.395) is obtained for the function graph of the fuzzy set.

c) For the other 17 rules results a degree of fulfillment equal to zero for the premise, i.e., it results in fuzzy sets, which are zeros themselves.

5. Defuzzification The decision logic yields no crisp value for the control quantity, but a fuzzy set. That means, by this method, one gets a mapping, which assigns a fuzzy set $\mu^{\text{Output}}_{x_1,\ldots,x_n}$ of Y to every tuple $(x_1,\ldots,x_n) \in X_1 \times X_2 \times \cdots \times X_n$ of the measured values.

$$\mu^{\text{Output}}_{36;-2.25}(y) = \begin{cases} \frac{2}{5}y & \text{for} \quad 0 \le y < 1, \\ 0.4 & \text{for} \quad 1 \le y < 3.5, \\ \frac{2}{5}y - 1 & \text{for} \quad 3.5 \le y < 4, \\ 0.6 & \text{for} \quad 4 \le y < 6, \\ 3 - \frac{2}{5}y & \text{for} \quad 6 \le y \le 7.5, \\ 0 & \text{for} \quad \text{otherwise.} \end{cases} \tag{5.395}$$

Defuzzification means that there is to determine a control quantity using defuzzification methods. The center of gravity method and the maximum criterion method result in the value for control quantity $F = 3.95$ or $F = 5.0$.

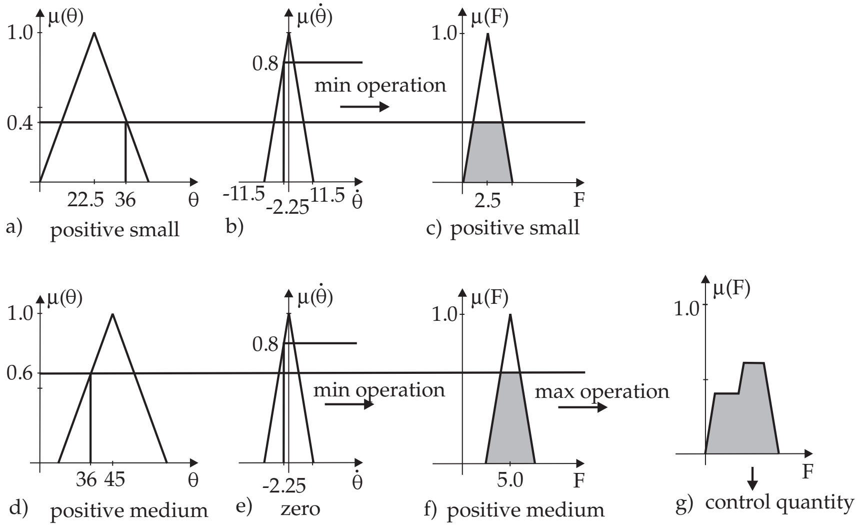


Figure 5.80

6. Remarks

1. The “knowledge-based” trajectories should lie in the rule base so that the endpoint is in the center of the smallest rule deviation.

2. By defuzzification an iteration process is introduced, which leads finally to the center of the partition space, i.e., which results in a zero control quantity.

3. Every non-linear domain of characteristics can be approximated with arbitrary accuracy by the choice of appropriate parameters on a compact domain.

5.9.6.4 Knowledge-Based Interpolation Systems

1. Interpolation Mechanism

Interpolation mechanisms can be built up with the help of fuzzy logic. Fuzzy systems are systems to process fuzzy information. With them it is possible to approximate and interpolate functions. A simple fuzzy system, by which this property can be investigated , is the Sugeno controller. It has n input variables $\xi_1, \dots, \xi_n$ and defines the value of the output variable y by rules $\mathrm{R}_1, \dots, \mathrm{R}_n$ in the form

$$\mathrm{R}_i\colon \text{ IF } \xi_1 \text{ is } A_1^{(i)} \text{ and } \cdots \text{ and } \xi_n \text{ is } A_n^{(i)}, \text{THEN is } y = f_i(\xi_1, \dots, \xi_n) \quad (i = 1, 2, \dots, n). \tag{5.396}$$

The fuzzy sets $A_j^{(1)}, \dots, A_j^{(k)}$ always partition the input sets X_j. The conclusions $f_i(\xi_1, \dots, \xi_n)$ of the rules are singletons, which can depend on the input variables $\xi_1, \dots, \xi_n$.

By a simple choice of the conclusions the expensive defuzzification can be omitted and the output value y will be calculated as a weighted sum. To do this, the controller calculates a degree of fulfillment α_i for every rule R_i with a t norm from the membership grades of the single inputs and determines the output value

$$y = \frac{\sum_{i=1}^N \alpha_i f_i(\xi_1, \dots, \xi_n)}{\sum_{i=1}^N \alpha_i}. \tag{5.397}$$

2. Restriction to the One-Dimensional Case

For fuzzy systems with only one input $x = \xi_1$, fuzzy sets represented by triangular functions are often used which are cut at the height 0.5. Such fuzzy sets satisfy the following three conditions:

1. For every rule R_i there is an input x_i, for which only one rule is fulfilled. For this input x_i, the output is calculated by f_i. By this, the output of the fuzzy system is fixed at N nodes $x_1, \dots, x_N$. Actually, the fuzzy system interpolates the nodes $x_1, \dots, x_N$. The requirement that at the node x_i only one rule R_i holds, is sufficient for an exact interpolation, but it is not necessary. For two rules R_1 and R_2, as they will be considered below, this requirement means that $\alpha_1(x_2) = \alpha_2(x_1) = 0$ holds. To fulfill the first condition, $\alpha_1(x_2) = \alpha_2(x_1) = 0$ must hold. This is a sufficient condition for an exact interpolation of the nodes.

2. There are at most two rules fulfilled between two consecutive nodes. If x_1 and x_2 are two such nodes with rules R_1 and R_2, then for inputs $x \in [x_1, x_2]$ the output y is

$$y = \frac{\alpha_1(x) f_1(x) + \alpha_2(x) f_2(x)}{\alpha_1(x) + \alpha_2(x)} = f_1(x) + g(x)\,[f_2(x) - f_1(x)] \quad \text{with } g := \frac{\alpha_2(x)}{\alpha_1(x) + \alpha_2(x)}. \tag{5.398}$$

The actual shape of the interpolation curve between x_1 and x_2 is determined by the function g. The shape depends only on the satisfaction grades α_1 and α_2, which are the values of the membership functions $\mu_{A_i^{(1)}}$ and $\mu_{A_i^{(2)}}$ at the point x, i.e., $\alpha_1 = \mu_{A^{(1)}}(x)$ and $\alpha_2 = \mu_{A^{(2)}}(x)$ are valid, or in short form $\alpha_1 = \mu_1(x)$ and $\alpha_2 = \mu_2(x)$. The shape of the curve depends only on the relation μ_1/μ_2 of the membership functions.

3. The membership functions are positive, so the output y is a convex combination of the conclusions f_i. For the given and for the general case hold (5.399) and (5.400), respectively:

$$\min(f_1, f_2) \le y \le \max(f_1, f_2), \tag{5.399}$$

$$\min_{i\epsilon\{1,2,\dots,N\}} f_i \le y \le \max_{i\epsilon\{1,2,\dots,N\}} f_i. \tag{5.400}$$

For constant conclusions, the terms f_1 and f_2 cause only a translation and stretching of the shape of the curve g. If the conclusions are dependent on the input variables, then the shape of the curve is differently perturbed in different sections. Consequently, another output function can be found.

Applying linearly dependent conclusions and membership functions with constant sum for the input x, then the output is $y = c \sum\limits_{i=1}^{N} \alpha_i(x) f_i(x)$ with α_i depending on x and a constant c, so that the interpolation functions are polynomials of second degree. These polynomials can be used for the construction of an interpolation method with polynomials of second degree.

In general, choosing polynomials of n-th degree, an interpolation polynomial of $(n+1)$-th degree is obtained as a conclusion. In this sense fuzzy systems are rule-based interpolation systems besides conventional interpolation methods interpolating locally by polynomials, e.g., with splines.