P. J. Pahl
R. Damrath

Mathematical Foundations of Computational Engineering

A Handbook

Volume II

Springer

Mathematical Foundations of Computational Engineering

Springer-Verlag Berlin Heidelberg GmbH

Engineering ONLINE LIBRARY

http://www.springer.de/engine/

Peter Jan Pahl · Rudolf Damrath

Mathematical Foundations of Computational Engineering

A Handbook

Springer

Prof. Dr. Peter Jan Pahl
Technische Universität Berlin
Institut für Allgemeine Bauingenieurmethoden
Straße des 17. Juni 135
10623 Berlin
Germany
pahl@ifb.bv.tu-berlin.de

Prof. Dr.-Ing. Rudolf Damrath
Universität Hannover
FG Angewandte Informatik im Bauingenieurwesen
Appelstraße 9A
30167 Hannover
Germany

Translation:
Dr. Felix Pahl
Schopenhauerstraße 63
14129 Berlin
Germany
fpahl@web.de

ISBN 978-3-642-63238-9

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Pahl, Peter Jan:
Mathematical foundations of computational engineering : a handbook / Peter Jan Pahl ; Rudolf Damrath. Transl. from the German by Felix Pahl. - Berlin ; Heidelberg ; New York ; Barcelona ; Hongkong ; London ; Mailand ; Paris ; Singapur ; Tokio : Springer, 2001
Dt. Ausg. u.d.T.: Pahl, Peter Jan: Mathematische Grundlagen der Ingenieurinformatik
ISBN 978-3-642-63238-9 ISBN 978-3-642-56893-0 (eBook)
DOI 10.1007/978-3-642-56893-0

http://www.springer.de

Originally published by Springer-Verlag Berlin Heidelberg New York in 2001
Softcover reprint of the hardcover 1st edition 2001

Cover Design: medio Technologies AG, Berlin
Typesetting: Digital data supplied by author
Printed on acid-free paper SPIN: 10757984 68/3020 wei - 5 4 3 2 1 0

PREFACE

Mathematics is one of the foundations of engineering. Because of the great importance of the physical behavior of engineering products, calculus usually lies at the center of the mathematical education of engineers; it is employed in the mathematical formulation of physical problems. This formulation has contributed significantly to the systematization of engineering and the mastering of engineering tasks.

Before computers were introduced into engineering, numerical solutions of the mathematical formulations of engineering problems involving irregular geometry, varying material properties, multiple influences and complex production processes were difficult to determine. Nowadays, computers amplify human mental capacities by a factor of 10^9 with respect to speed of calculation, storage capacity and speed of communication; this has created entirely new possibilities for solving mathematically formulated physical problems. New fields of science, such as computational mechanics, and widely applied new computational methods, such as the finite element method, have emerged.

While computers were being introduced, the character of engineering changed profoundly. While the key to competitiveness once lay in using better materials, developing new methods of construction and designing new engineering systems, success now depends just as much on organization and management. The reasons for these changes include a holisitic view of the market, the product, the economy and society, the importance of organization and management in global competition as well as the increased complexity of technology, the environment and the interactions among those participating in planning and production.

Given the new character of engineering, the traditional mathematical foundations no longer suffice. Branches of mathematics which are highly developed but have so far been of little importance to engineers now prove to be important tools in a computer-oriented treatment of engineering problems. These fields are, however, not readily accessible to many engineers, since frequently even fundamental concepts are not treated systematically in their education. Thus, there is no sound basis for a productive dialog between engineers and mathematicians. To make matters more difficult, many of the hitherto neglected fields are based directly on the foundations of mathematics and hence exhibit a degree of abstraction which engineers are not accustomed to.

The developments in the use of computers in engineering have shown that an inadequate education in mathematics may have grave consequences. In the areas of planning, organization and management, in particular, the potential of classical graph theory was not brought to bear on the abstraction of computer models and the systematization of the methods of solution. Numerous laws and methods which, with a sufficient background in mathematics, could have been taken from the literature, were reinvented with much effort. The foundations of topology furnish an example of this phenomenon.

Due to the rapid development of information and communication technology, with performance increasing by a factor of 100 per decade, new areas of application are constantly emerging. This makes it particularly difficult to determine that part of the abundant repertoire of mathematics which will form a solid basis for the proper utilization of computers in engineering in the coming decades. In this book, we try to compile these essentials. We have arranged the material so that it can be learned in the order of the chapters of the book. We assume that traditional mathematics for engineers is treated in addition : Essential branches of mathematics are not addressed in this book, since there is a vast literature on them.

The treatment of foundations begins with logic in Chapter 1. There are various reasons for this. For one, logic is a tool for the development of the other chapters of the book. Also, the creation of models and processes requires a systematic approach, which relies on a consistent application of the laws of logic. An example of the systematic use of logic is furnished by a correct treatment of implications and equivalences.

Set theory, treated in Chapter 2, forms the basis for the mathematical structures treated in the subsequent chapters. Set operations are of fundamental importance in all areas of computer application. Set theory leads to concepts like relation and mapping, which are fundamental for the classification and ordering of information and hence for approaches like object-oriented modelling.

Mathematics contains basic algebraic, ordinal and topological structures. All other branches of mathematics rest on these basic structures, which are treated in Chapters 3 to 5.

Algebraic structures describe operations on elements of sets. In contrast to traditional mathematics for engineers, the restriction to real numbers is lifted in order to lay a systematic foundation for general operations on values of different types, for instance logical variables, sets, vectors and matrices. These foundations are applied in all subsequent chapters.

Ordinal structures are of paramount importance for many computer-based algorithms. Reliable algorithms require a precise treatment of the properties of order relations and a systematic distinction between comparable and incomparable elements of a set. Many data structures cannot be designed or implemented without an understanding of order relations and their properties. The study of the convergence properties of iterative and sorting algorithms also relies on ordinal structures.

If a system of subsets is singled out in a set, the set acquires a topological structure. Topological structures form the basis for determining connectedness and separation of sets, convergence of sequences, nets and filters, compactness of spaces and continuity of functions. The convergence of approximation methods for solving the mathematical formulation of physical problems cannot be studied without an understanding of topological spaces. The description of geometric shapes on the computer also depends on a reliable analysis of the associated problems in topology.

Quantification in engineering relies on the natural, whole, rational, real and complex number systems and the quaternions. These number systems exhibit different algebraic, ordinal and topological structures, which are treated in Chapter 6. Knowledge of the properties of the number systems is essential for constructing reliable numerical algorithms.

Groups, treated in Chapter 7, have played an important role in the development of mathematics. Group theory deals with an operation on two elements of a set, the result of which is again an element of the set. Two of these three values are known, and the third value (an operand or the result) is to be determined. The structure of groups proves to be extraordinarily rich. It allows a systematic treatment of many fundamental mathematical problems. For example, Galois used it to prove that a circle cannot be squared with compass and straightedge alone. A systematic treatment of geometry can also be carried out on the basis of group theory. The practical applications of group theory include the systematic analysis of the topology of triangulated bodies.

In designing models and algorithms, the description of the relations between the elements of sets is of fundamental importance. This is the subject of graph theory, treated in Chapter 8. Graph theory relies on the algebra of relations. This algebra, in which graphs are described by matrices, leads to a set of theories and methods which allow the properties of graphs to be determined algebraically. Many practical problems in engineering can be solved using graphs, including problems in management and organization. Among these are the determination of paths in traffic networks, of reliability in complex systems and of the optimal order of processing steps.

Tensor theory, treated in Chapter 9, forms the basis for a reliable formulation of physical engineering problems. Tensor formulations have the special property of being independent of the chosen coordinate system. This aids the understanding of the essential characteristics of the formulated problems and thus facilitates the systematic development of algorithms. It is on this basis that complex physical processes in solids, liquids and multi-phase systems can be rendered susceptible to a universally valid implementation.

Engineers deal with events that depend on chance : The repetition of an experiment under seemingly identical conditions yields different results. Random events are studied using stochastical methods, treated in Chapter 10. These methods assign probabilities to the different outcomes of an experiment. There are typical probability distributions in engineering, for example for the reliability of a system of components and for the behavior at the nodes of a traffic network. Random processes for time-dependent random variables are of great practical importance. Their description using Markov chains forms the basis of the theory of queues, which is applied in many computer simulations of processes in engineering.

The chapters of the book are structured uniformly. Each chapter begins with an introduction, which highlights the main points of the chapter. It uses concepts and mentions properties which are defined and explained in subsequent parts of the chapter. The sections also begin with introductions, which are similarly structured. Every paragraph of the text begins with a term which appears in boldface for emphasis. This term is explained in the paragraph. The highlighted terms are intended to aid the reader in grasping the structure of the sections of the book with little effort. Proofs are included in the text, in particular where they significantly aid comprehension or form the basis for the development of algorithms for computer implementations.

The desire to write this book emerged during our long-standing cooperation at the Technische Universität Berlin in the area of "Theoretische Methoden der Bau- und Verkehrstechnik" (theoretical methods in civil engineering). While developing this field together, we realized that the topics covered in education and the information technology employed are short-lived compared to the content of other areas of engineering. Yet the application of computer science in engineering needs a stable basis. Out of this realization grew the desire to compile the mathematical foundations which are independent of the rapid developments and incessant changes in a book and thus to create a durable basis for future developments. The book differs significantly from our lecture notes, which deal with current information and communication technologies, including development environments and their applications in engineering.

Dr. Felix Pahl played an important role in shaping the book as a whole. He repeatedly proofread the chapters with great care and used his background as a physicist to make valuable suggestions for structuring the material. Particularly in the chapters on topology and group theory, Dr. Pahl contributed to most of the proofs and provided invaluable assistance in formulating them concisely. His special commitment to this book deserves our personal thanks.

The content of the present book imposes strong demands on the graphical design of the text, the figures and the formulas. With admirable intuition, Mrs. Elizabeth Maue has given the book an attractive appearance. As the book took shape over an extended period of time, during which all chapters were thoroughly revised several times, her patience has been put to a severe test. Mrs. Maue's committed participation, which resulted in the particularly appealing presentation of this book, deserves our grateful recognition.

This book took shape over the course of more than seven years. During this time our wives, Irmgard Pahl and Heidemarie Damrath, showed great understanding for our extraordinary workload. By their great patience, they gave us the freedom and support without which this book could not have been completed in its present form. We thank them with all our heart.

Berlin, May 2000

Peter Jan Pahl
Rudolf Damrath

CONTENTS

1 LOGIC

1.1 REPRESENTATION OF THOUGHT

Logic : Science structures human thought. It divides thought into individual thoughts, represents the content of these thoughts and distinguishes between true and false thoughts. Logic deals with methods that reduce misunderstandings in the representation of thoughts, and with consistency in deducing the truth of certain thoughts from the truth of given other thoughts. It follows that logic is fundamental to science.

Everyday language is not suitable as an instrument of logic, as it is ambiguous. Science therefore employs formal logic, which is expressed in a formal (artificial) language. Formal logic is an area of mathematics which makes the formulation of concepts more precise and investigates contradictions in theories. Essential tools of formal logic are symbolization, formalization and evaluation.

Symbolization is the representation of thoughts in a formal language. The core of this artificial language is a character set. The characters (symbols) in the character set must be separable; it must be possible to identify them uniquely. The character set is used to form character strings. Each thought is described by three character strings : the content, the label and the value of the thought. This formal representation of a thought is called a statement. The label identifies the statement. The value assigns the statement to one of the classes true and false.

Formalization is a set of rules which leads from given statements to other statements in a definite manner. This process is called logical deduction. Operations which describe relationships between statements are the core of formalization. A relationship between statements consists of arguments, a rule and a result. The arguments and the result are statement values. The rule prescribes the value of the result for each combination of values of the arguments. The rules of operations are stipulated and represented by a symbol. The part of formal logic which deals with operations on statements is called propositional (sentential) logic.

Evaluation is the assignment of the content of a statement to one of the classes true and false. The content of a statement is said to be true if its assertion holds according to the common judgement of a given group of people. In order to evaluate a statement, the character string of its content is resolved into its constituents, which are called language elements. A sequence of language elements is called an expression. The rules for constructing admissible expressions from the character set of the language are called the syntax of the language. The relationship between an expression which is admissible as the content of a statement and the value of this statement is called the semantics of the language. The part of formal logic that deals with the syntax and semantics of formal languages is called predicate logic.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

1.2 ELEMENTARY CONCEPTS

Concepts which carry the same meaning for all participants form the basis for communication between humans. For some fundamental concepts, agreement among the participants is presumed. Using these fundamental concepts, new concepts are defined, and these are in turn used to define further concepts. Some elementary concepts used in the formal description of thought are defined in the following.

Set : Objects of thought or of the senses which are separable and can be identified uniquely are called elements. A collection of elements with similar properties is called a set. Each property of an element is described either by its value (predicate) or by rules for determining its value. The set of all values which a property can take is called the range of the property. The elements of a set are uniquely identified using a property of the elements which takes different values for all pairs of elements. This property is called the name (label, identifier) of the element.

Sequence : A sequence is a collection of elements which are chosen successively from a given set. By virtue of the order of these choices, each element of a sequence has an additional property which it does not have in the set. The order may, for instance, be described using natural numbers, since each natural number other than zero has a predecessor. The additional property makes it possible to choose an element of the given set for the sequence repeatedly while maintaining uniqueness of the elements in the sequence.

Character string : A set of separable and uniquely identifiable symbols is called a character set. A sequence chosen from the character set is called a character string. To represent the order of the characters in the sequence, one chooses a convention, for instance horizontal arrangement from left to right. The beginning and the end of a character string are indicated according to stipulated rules, for instance by marking them with the character ”.

Value : A character string is called a value if it identifies an element of a set. The value is called a constant if the character string is the name of an element of the set. The value is called a variable if the character string is replaced by the name of an element of the set according to stipulated rules.

Operation : A rule which assigns precisely one constant to a given sequence of constants is called an operation. The given constants are called the arguments of the operation. The result is called the value of the operation. The rule for an operation is designated by a symbol.

Example 1 : Elements of a character string

Let the character set $Z = \{a,b,c\}$ be given. The character string "babac" is chosen from this set. This character string corresponds to the set $F = \{(a, 2), (a, 4), (b, 1), (b, 3), (c, 5)\}$. Each element of the set F is an ordered pair. It consists of an element of the character set Z, to which an element of the set $N = \{1, 2, 3, 4, 5\}$ of numbers is assigned as an additional property. The elements of the character string F are unique. Their order in F is arbitrary.

Example 2 : Operation on values

The operation "addition" designated by the symbol + is defined for the natural numbers. The rules of addition assign the result 8 to the arguments 3 and 5. This operation is represented by the formula $3 + 5 = 8$.

Example 3 : Russel's antinomy

Contradictions may arise if sets are defined to contain sets as elements. A contradictory definition of a set is called an antinomy. An example of such an antinomy is furnished by the collection R of all sets $M_1, M_2, ...$ which are not contained in themselves :

(1) Under the assumption that R contains itself, R is one of the sets M_i of the collection. However, the sets M_i are by definition not contained in themselves. It follows that R contains itself and at the same time does not contain itself.

(2) Under the assumption that R does not contain itself, R is by definition one of the sets M_i. It follows that R is an element of the collection and therefore contained in itself. R therefore contains itself and at the same time does not contain itself.

As R cannot simultaneously contain itself and not contain itself, the collection R is not a set. Such contradictory definitions of sets must be ruled out in set theory.

Definition : Thoughts are composed of concepts. The precise delimitation of a concept using other concepts is called the definition of the concept. The concept acquires its meaning through this delimitation. Concepts whose meaning is postulated are called fundamental concepts. All other concepts are defined.

Explicit definition : A definition is said to be explicit if the concept to be defined (definiendum) is delimited using fundamental concepts or concepts that have already been defined (definiens). Wherever it occurs, the definiendum may be replaced by the definiens. The definition of a statement value is indicated using the symbol $:\Leftrightarrow$ (equivalent by definition). The definition of terms which are not statement values is indicated using the symbol $:=$ (equal by definition).

statement value	:	definiendum	$:\Leftrightarrow$	definiens
term	:	definiendum	$:=$	definiens

Implicit definition : A definition is said to be implicit if concepts are delimited by their mutual relationships. The relationships are assumed to be true statements. Fundamental concepts such as natural number, distance and area are defined implicitly.

Recursive definition : Let a concept $G(n)$ which depends on a natural number n be given. The definition of this concept is said to be recursive if $G(0)$ is defined first, and then each $G(n)$ for $n > 0$ is defined with $G(n-1)$ as definiens. For instance, the concept $n!$ (n factorial) is defined recursively by $0! := 1$ and $n! := n(n-1)!$.

1.3 PROPOSITIONAL LOGIC

Introduction : The value of a statement can be true, false or undetermined. In a stochastic treatment, probabilities are assigned to these statement values. In a deterministic treatment, the statement has precisely one of the specified values. If no undetermined statement values are allowed, the logic is said to be two-valued.

For a two-valued logic, statements are divided into the class of true statements and the class of false statements. Thus, each statement is assigned a truth value, which is either true or false. Given statements may be connected to form a new statement. In propositional logic, connectives are defined such that the truth value of the new statement results from the truth values of the given statements in a definite manner.

Several statements may be connected to form an expression. There are expressions which are always true, regardless of the truth values of the individual statements. Such expressions are called logically valid expressions or tautologies. Among the tautologies, logical equivalences and logical implications are especially important. Logical equivalences are used to transform logical expressions into equivalent expressions. Logical implications are used to deduce new true statements from given true statements.

1.3.1 LOGICAL VARIABLES AND CONNECTIVES

Statement : A statement (proposition) is the formulation of a thought in a language. Each statement consists of a label, a content and a value. The label identifies the thought, the content defines the thought, and the value evaluates the thought. Formally, a statement is designated by a letter, its content is defined by a character string, and the result of its evaluation is expressed as a truth value.

Truth value : The determination of the truth value of a statement is a fundamental problem, and various approaches to its solution have been investigated. In a two-valued logic, it is assumed that statements can be divided into the class of true statements and the class of false statements. Accordingly, each statement is assigned either the truth value false (designated by f or 0) or the truth value true (designated by t or 1). A statement to which a truth value has been assigned is called a statement constant. A statement to which a truth value has not yet been assigned is called a statement variable. The truth value of a statement a is designated by T(a).

Example 1 : Statements and truth values

The statement a := "Every triangle has three corners." is true and therefore possesses the truth value $T(a) = t$.

The statement b := "5 is less than 2." is false and therefore possesses the truth value $T(b) = f$.

The text "Congratulations!" is not a statement, as it cannot be assigned a truth value.

Propositional connectives : Two given statements may be connected to form a new statement using words like "and", "or", "if-then" and "if and only if-then", which are called propositional (sentential) connectives. Each of these connectives corresponds to a definite rule for determining the truth value of the new statement from the truth values of the given statements. It is irrelevant whether the content of the statements being connected is related. The negation of a statement using the word "not" is also treated as a connective.

Operator and operand : The symbol that represents the rules for the truth value of connected statements is called an operator. In the context of set theory (see Chapter 2), an operator is a relation. Each connective is associated with a corresponding operator. The truth values of the connected statements are called the operands of the connective. The following operators are often used :

Name	Operator	Connective	Meaning	Rank
negation	$\neg$	$\neg\ a$	not a	5
conjunction	$\wedge$	$a \wedge b$	a and b	4
disjunction	$\vee$	$a \vee b$	a or b	3
alternation	$\oplus$	$a \oplus b$	either a or b	2
implication	$\Rightarrow$	$a \Rightarrow b$	if a then b	1
equivalence	$\Leftrightarrow$	$a \Leftrightarrow b$	a if and only if b	0

Truth tables : For each operator, the rules for determining the truth value of the new statement from the truth values of the given statements are represented by a truth table. The values of the first operand a appear on the left; if there is a second operand, its values appear at the top of the table. For each pair (a,b) of values, the table contains the value of the connective indicated at the bottom.

0	1
1	0

¬a

The negation of a statement a is designated by ¬a (not a). It is true if a is false. It is false if a is true.

	0	1
0	0	0
1	0	1

a ∧ b

The conjunction of two statements a, b is designated by a ∧ b (a and b). It is true if a is true and b is true. It is false if a is false or b is false.

	0	1
0	0	1
1	1	1

a ∨ b

The disjunction of two statements a, b is designated by a ∨ b (a or b). It is true if a is true or b is true. It is false if a is false and b is false.

	0	1
0	0	1
1	1	0

a ⊕ b

The alternation of two statements a, b is designated by a⊕b (either a or b). It is true if a and b have different truth values. It is false if a and b have the same truth value.

	0	1
0	1	1
1	0	1

a ⇒ b

The implication of two statements a, b is designated by a ⇒ b (if a then b). It is true if a is false or b is true. It is false if a is true and b is false. An implication is also called a subjunction.

	0	1
0	1	0
1	0	1

a ⇔ b

The equivalence of two statements a, b is designated by a ⇔ b (a if and only if b). It is true if a and b have the same truth value. It is false if a and b have different truth values. An equivalence is also called a bijunction.

Example 2 : Propositional connectives

Let the following statements a,b and their truth values be given :

$a :=$	”Every triangle has three corners”	$T(a) = t$
$b :=$	”Every quadrangle is red”	$T(b) = f$

The negations $\neg a$ and $\neg b$ are :

$\neg a =$	”Not every triangle has three corners”	$T(\neg a) = f$
$\neg b =$	”Not every quadrangle is red”	$T(\neg b) = t$

The conjunction $a \wedge b$ and the disjunction $a \vee b$ are :

$a \wedge b =$	”Every triangle has three corners and every quadrangle is red”	$T(a \wedge b) = f$
$a \vee b =$	”Every triangle has three corners or every quadrangle is red”	$T(a \vee b) = t$

The implication $a \Rightarrow b$ and the equivalence $a \Leftrightarrow b$ are :

$a \Rightarrow b =$	”If every triangle has three corners then every quadrangle is red”	$T(a \Rightarrow b) = f$
$a \Leftrightarrow b =$	”Every triangle has three corners if and only if every quadrangle is red”	$T(a \Leftrightarrow b) = f$

The examples demonstrate that the statements a and b are connected in a purely formal manner, irrespective of the relation of their content. The truth values of the various connectives are determined from the truth tables.

Operator basis : For a connective involving one operand (unary connective), $2^2 = 4$ different operators can be defined. One of these operators is designated by the symbol $\neg$. The remaining unary operators are replaced by logical expressions, which are shown underneath the following truth tables :

0	0	0	1	0	1	0	0
1	1	1	0	1	1	1	0
	a		$\neg a$		$a \vee (\neg a)$		$a \wedge (\neg a)$

For a connective involving two operands (binary connective), $2^{2*2} = 16$ different operators can be defined. Four of these operators are designated by the symbols $\wedge$, $\vee$, $\Rightarrow$, $\Leftrightarrow$. The remaining binary operators are replaced by logical expressions, which are shown underneath the following truth tables :

	0	1
0	0	0
1	0	1

$a \wedge b$

	0	1
0	0	1
1	1	1

$a \vee b$

	0	1
0	1	1
1	0	1

$a \Rightarrow b$

	0	1
0	1	0
1	0	1

$a \Leftrightarrow b$

	0	1
0	1	1
1	1	1

$(a \Rightarrow b) \vee (b \Rightarrow a)$

	0	1
0	0	0
1	0	0

$\neg((a \Rightarrow b) \vee (b \Rightarrow a))$

	0	1
0	1	1
1	1	0

$\neg(a \wedge b)$

	0	1
0	1	0
1	1	1

$b \Rightarrow a$

	0	1
0	1	0
1	0	0

$\neg(a \vee b)$

	0	1
0	0	0
1	1	0

$\neg(a \Rightarrow b)$

	0	1
0	0	1
1	0	0

$\neg(b \Rightarrow a)$

	0	1
0	0	1
1	1	0

$\neg(a \Leftrightarrow b)$

	0	1
0	0	0
1	1	1

$(b \Rightarrow a) \wedge (b \vee a)$

	0	1
0	1	1
1	0	0

$\neg(a \wedge b) \wedge (a \Rightarrow b)$

	0	1
0	0	1
1	0	1

$(a \vee b) \wedge (a \Rightarrow b)$

	0	1
0	1	0
1	1	0

$\neg(a \wedge b) \wedge (b \Rightarrow a)$

The truth tables show that the set $\{\neg, \wedge, \vee, \Rightarrow, \Leftrightarrow\}$ of operators generates the 20 operators of the unary and binary connectives. The question arises whether fewer generators would suffice. This is indeed the case :

(1) The set $\{\neg, \wedge, \vee\}$ generates all propositional connectives, since by the rule of elimination the connectives $a \Rightarrow b$ and $a \Leftrightarrow b$ may be replaced by the following equivalent expressions :

$$(a \Rightarrow b) \Leftrightarrow (\neg a \vee b)$$

$$(a \Leftrightarrow b) \Leftrightarrow (\neg a \vee b) \wedge (a \vee \neg b)$$

(2) The sets $\{\neg, \wedge\}$ and $\{\neg, \vee\}$ both generate all propositional connectives, since by the rule of double negation and De Morgan's laws the connectives $a \vee b$ and $a \wedge b$ may be replaced by the following equivalent expressions :

$$(a \vee b) \Leftrightarrow \neg\neg(a \vee b) \Leftrightarrow \neg(\neg a \wedge \neg b)$$

$$(a \wedge b) \Leftrightarrow \neg\neg(a \wedge b) \Leftrightarrow \neg(\neg a \vee \neg b)$$

(3) The operator basis consists of a single operator. The operator | (not and, nand) or the operator $\triangledown$ (not or, nor) may be chosen. These operators are defined as follows :

	0	1
0	1	1
1	1	0

$a \mid b \Leftrightarrow \neg (a \wedge b)$

	0	1
0	1	0
1	0	0

$(a \triangledown b) \Leftrightarrow \neg(a \vee b)$

In the following table, the generators $\{\neg, \wedge, \vee, \Rightarrow, \Leftrightarrow\}$ are expressed in terms of the basic operators. Note that the operators | and $\triangledown$ are commutative but not associative. Since the operators are not associative, expressions of the form $a \mid b \mid c$ or $a \triangledown b \triangledown c$ are not admissible without parentheses.

connective		operator $\mid$		operator $\triangledown$
$\neg a$	$\Leftrightarrow$	$a \mid a$	$\Leftrightarrow$	$a \triangledown a$
$(a \wedge b)$	$\Leftrightarrow$	$(a \mid b) \mid (a \mid b)$	$\Leftrightarrow$	$(a \triangledown a) \triangledown (b \triangledown b)$
$(a \vee b)$	$\Leftrightarrow$	$(a \mid a) \mid (b \mid b)$	$\Leftrightarrow$	$(a \triangledown b) \triangledown (a \triangledown b)$
$(a \Rightarrow b)$	$\Leftrightarrow$	$(a \mid b) \mid a$	$\Leftrightarrow$	$(b \triangledown (a \triangledown b)) \triangledown (b \triangledown (a \triangledown b))$
$(a \Leftrightarrow b)$	$\Leftrightarrow$	$((a \mid a) \mid (b \mid b)) \mid (a \mid b)$	$\Leftrightarrow$	$(a \triangledown (a \triangledown b)) \triangledown (b \triangledown (a \triangledown b))$

1.3.2 LOGICAL EXPRESSIONS

Expression : The elements of propositional logic are statement constants, statement variables, operators and the technical characters ( ). A sequence of elements is called an expression of propositional logic if it is formed according to the following syntactic rules :

(1) Every statement constant and every statement variable is an expression.

(2) If a, b are expressions, then the connectives ($\neg$ a), (a $\wedge$ b), (a $\vee$ b), (a $\Rightarrow$ b) and (a $\Leftrightarrow$ b) are also expressions.

(3) Only sequences of elements formed by rules (1) and (2) are expressions.

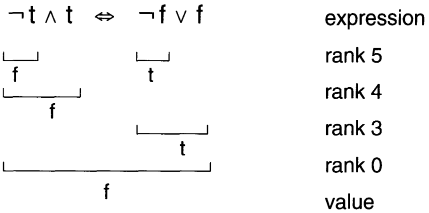
Rank of the operators : Expressions formed according to the syntactic rules contain many parentheses and are therefore difficult to read. To avoid the use of parentheses, the following rules are stipulated :

(1) Exterior parentheses may be removed.

(2) Each operator has a rank (see the table of operators). If two successive operators of different rank in a logical expression are not separated by parentheses, the operator of higher rank is applied first.

(3) If two operators of equal rank in a logical expression are not separated by parentheses, they are applied from left to right.

Example 1 : Rank of the operators



Formula : An expression of propositional logic containing one or more statement variables describes a statement formally and is called a formula. The truth value of an expression depends on the truth values of the statement variables.

Valuation : If every statement variable in an expression is assigned precisely one truth value, the collection of these assignments is called a valuation of the expression. If an expression contains n statement variables, then for n > 0 there are exactly 2^n different valuations of the expression, in which each of the statement variables takes one of the truth values t, f.

Each valuation of an expression leads to a truth value for the expression, which is determined according to the semantic rules of propositional logic. The semantic rules are determined by the truth tables defined for the connectives. Thus the truth value of an expression for a given valuation is calculated by logically evaluating the expression using the truth tables.

Example 2 : Expression and valuation

The following sequence of elements is an expression; its subexpressions (connectives) are underlined.

$$\underbrace{\underbrace{\underbrace{\neg a} \wedge b} \Leftrightarrow \underbrace{(c \Rightarrow \underbrace{b \wedge \underbrace{(a \vee d)}})}} \qquad \text{expression}$$

Let the truth values t, f, f, t be assigned to the statement variables a, b, c, d of the given expression. Using the truth tables, the expression is evaluated for this valuation as follows :

$$\underbrace{\underbrace{\underbrace{\neg t}_{f} \wedge f}_{f} \Leftrightarrow \underbrace{(f \Rightarrow \underbrace{f \wedge \underbrace{(t \vee t)}_{t}}_{f})}_{t}}_{f}$$

Logically valid, consistent and inconsistent expressions : The expressions of propositional logic are classified with respect to their valuations and their truth value as follows :

(1) An expression which is true for all valuations is said to be logically valid (a tautology).

(2) An expression which is true for at least one valuation is said to be logically consistent.

(3) An expression which is not true for any valuation is said to be logically inconsistent (a contradiction).

In propositional logic, it is possible to decide in a finite number of steps whether a given expression with a finite number of operands is logically valid, consistent or inconsistent. In fact, if an expression contains n statement variables, then for $n > 0$ there are exactly 2^n different valuations. The expression can be evaluated for each valuation using the truth tables for the connectives. Thus it takes at most 2^n evaluations to decide whether the given expression is logically valid, consistent or inconsistent.

Example 3 : Logically valid, consistent and inconsistent expressions

The expression $a \wedge \neg a$ is inconsistent, as it is false for any valuation. This is proved using the truth tables as follows :

a	$\neg a$	$a \wedge \neg a$
0	1	0
1	0	0

The expression $((a \Rightarrow b) \wedge b) \Rightarrow a$ is consistent but not logically valid, as it is true for some but not all valuations. This is proved using the truth tables as follows :

a	b	$a \Rightarrow b$	$(a \Rightarrow b) \wedge b$	$((a \Rightarrow b) \wedge b) \Rightarrow a$
0	0	1	0	1
0	1	1	1	0
1	0	0	0	1
1	1	1	1	1

The expression $(a \Rightarrow b) \Leftrightarrow (\neg a \vee b)$ is logically valid, as it is true for any valuation. This is proved using the truth tables as follows :

a	b	$a \Rightarrow b$	$\neg a$	$\neg a \vee b$	$(a \Rightarrow b) \Leftrightarrow (\neg a \vee b)$
0	0	1	1	1	1
0	1	1	1	1	1
1	0	0	0	0	1
1	1	1	0	1	1

1.3.3 LOGICAL NORMAL FORM

Logical equivalence : A logically valid expression of the form $a \Leftrightarrow b$ (a is equivalent to b) is called a logical equivalence. If $a \Leftrightarrow b$ is a logical equivalence, then a and b have the same truth value. The most important logical equivalences for expressions a, b, c with the operators $\neg$, $\wedge$, $\vee$ are :

identity	$a \wedge t$	$\Leftrightarrow$	a	$a \vee f$	$\Leftrightarrow$	a
invariance	$a \wedge f$	$\Leftrightarrow$	f	$a \vee t$	$\Leftrightarrow$	t
complementarity	$a \wedge \neg a$	$\Leftrightarrow$	f	$a \vee \neg a$	$\Leftrightarrow$	t
idempotency	$a \wedge a$	$\Leftrightarrow$	a	$a \vee a$	$\Leftrightarrow$	a
commutativity	$a \wedge b$	$\Leftrightarrow$	$b \wedge a$	$a \vee b$	$\Leftrightarrow$	$b \vee a$
associativity	$(a \wedge b) \wedge c$	$\Leftrightarrow$	$a \wedge (b \wedge c)$	$(a \vee b) \vee c$	$\Leftrightarrow$	$a \vee (b \vee c)$
distributivity	$a \wedge (b \vee c)$	$\Leftrightarrow$	$(a \wedge b) \vee (a \wedge c)$	$a \vee (b \wedge c)$	$\Leftrightarrow$	$(a \vee b) \wedge (a \vee c)$
absorption	$a \wedge (a \vee b)$	$\Leftrightarrow$	a	$a \vee (a \wedge b)$	$\Leftrightarrow$	a
double negation	$\neg \neg a$	$\Leftrightarrow$	a	a	$\Leftrightarrow$	$\neg \neg a$
De Morgan	$\neg (a \wedge b)$	$\Leftrightarrow$	$\neg a \vee \neg b$	$\neg (a \vee b)$	$\Leftrightarrow$	$\neg a \wedge \neg b$

The following logical equivalences are used in particular to reduce expressions involving the operators $\Rightarrow$, $\Leftrightarrow$ to equivalent expressions involving the operators $\neg$, $\wedge$, $\vee$.

elimination	$(a \Rightarrow b) \Leftrightarrow (\neg a \vee b)$	$(a \Leftrightarrow b) \Leftrightarrow (a \Rightarrow b) \wedge (a \Leftarrow b)$
	$(a \Leftrightarrow b) \Leftrightarrow (\neg a \vee b) \wedge (a \vee \neg b)$	$(a \Leftrightarrow b) \Leftrightarrow (a \wedge b) \vee (\neg a \wedge \neg b)$
contraposition	$(a \Rightarrow b) \Leftrightarrow (\neg b \Rightarrow \neg a)$	$(a \Leftrightarrow b) \Leftrightarrow (\neg a \Leftrightarrow \neg b)$

An expression of the form $a \wedge b \wedge \ldots \wedge c$ is called a general conjunction. An expression of the form $a \vee b \vee \ldots \vee c$ is called a general disjunction. Due to the associativity of the connectives $\wedge$ and $\vee$, no parentheses are necessary in these expressions. The expression a is admitted as a special case of a general conjunction or disjunction.

Logical transformations : A given expression of propositional logic may be transformed into a logically equivalent expression using logical equivalences : If $a \Leftrightarrow b$ is a logical equivalence and a occurs in the given expression, then a may be replaced by b, since a and b possess the same truth value due to the logical equivalence $a \Leftrightarrow b$. The aim of logical transformations is to exhibit a given expression in a more lucid and simple form. Representations which allow the truth value of the expression to be read off directly are particularly important.

Normal form : An expression of propositional logic is said to be in normal form if it contains only statement variables and negated statement variables and the operators $\wedge$ and $\vee$. A normal form may be disjunctive or conjunctive and also canonical.

1. A normal form is said to be disjunctive if it is a general disjunction of subexpressions and each subexpression is a general conjunction of statement variables or negated statement variables.
2. A normal form is said to be conjunctive if it is a general conjunction of subexpressions and each subexpression is a general disjunction of statement variables or negated statement variables.
3. A disjunctive or conjunctive normal form with n statement variables is said to be canonical if the number of subexpressions is minimal and each subexpression contains each of the n statement variables either with or without negation.

Every expression has an equivalent disjunctive normal form and an equivalent conjunctive normal form. Every consistent expression has an equivalent canonical disjunctive normal form. Every expression which is not a tautology has an equivalent canonical conjunctive normal form. The canonical disjunctive normal form and the canonical conjunctive normal form of an expression are unique up to the order of the subexpressions and of the variables inside the subexpressions.

Example 1 : Normal forms

The following expressions with the statement variables a, b, c are in normal form:

disjunctive normal form

$$(a \wedge b) \vee (\neg a \wedge b \wedge \neg c) \vee (\neg b \wedge c)$$
$$(a \wedge \neg a \wedge c) \vee (a \wedge b \wedge c \wedge \neg c)$$

conjunctive normal form

$$(a \vee b \vee c) \wedge (a \vee \neg b) \wedge (\neg a \vee \neg b \vee \neg c)$$
$$(\neg a \vee b \vee \neg b \vee c)$$

canonical disjunctive normal form

$$(a \wedge b \wedge c) \vee (\neg a \wedge \neg b \wedge c) \vee (a \wedge \neg b \wedge c)$$
$$(a \wedge b \wedge \neg c) \vee (\neg a \wedge b \wedge \neg c)$$

canonical conjunctive normal form

$$(a \vee b \vee \neg c) \wedge (a \vee \neg b \vee c) \vee (a \vee \neg b \vee \neg c)$$
$$(a \vee b \vee c) \wedge (\neg a \vee \neg b \vee c)$$

A canonical disjunctive normal form allows all valuations which yield the truth value true to be read off directly. The canonical disjunctive normal form $(a \wedge b \wedge \neg c) \vee (\neg a \wedge b \wedge c)$ is true if and only if one of the two subexpressions is true. The first subexpression $(a \wedge b \wedge \neg c)$ is true if and only if a is true and b is true and $\neg c$ is true. Thus it is true for the valuation $(T(a), T(b), T(c)) = (t, t, f)$. The second subexpression $(\neg a \wedge b \wedge c)$ is true if and only if $\neg a$ is true and b is true and c is true. Thus it is true for the valuation $(T(a), T(b), T(c)) = (f, t, t)$. The canonical disjunctive normal form $(a \wedge b \wedge \neg c) \vee (\neg a \wedge b \wedge c)$ is therefore true for the valuations $(T(a), T(b), T(c)) = (t, t, f), (f, t, t)$.

A canonical conjunctive normal form allows all valuations which yield the truth value false to be read off directly. The canonical conjunctive normal form $(a \vee b \vee c) \wedge (\neg a \vee \neg b \vee c)$ is false if and only if one of the two subexpressions is false. The first subexpression $(a \vee b \vee c)$ is false if and only if a is false and b is false and c is false. Thus it is false for the valuation $(T(a), T(b), T(c)) = (f, f, f)$. The second subexpression $(\neg a \vee \neg b \vee c)$ is false if and only if $\neg a$ is false and $\neg b$ is false and c is false. Thus it is false for the valuation $(T(a), T(b), T(c)) = (t, t, f)$. The canonical conjunctive normal form $(a \vee b \vee c) \wedge (\neg a \vee \neg b \vee c)$ is therefore false for the valuations $(T(a), T(b), T(c)) = (f, f, f), (t, t, f)$.

Example 2 : Transformation to normal form

Every logical expression may be transformed to normal form in a finite number of steps. This is demonstrated using the following expression with the statement variables a, b :

$$(a \Rightarrow b) \wedge b \Rightarrow a$$

Step 1 : The operators $\Rightarrow$ are replaced using the rule of elimination :

$$\begin{array}{ll} (a \Rightarrow b) \wedge b \Rightarrow a & \Leftrightarrow \\ (\neg a \vee b) \wedge b \Rightarrow a & \Leftrightarrow \\ \neg((\neg a \vee b) \wedge b) \vee a & \end{array}$$

Step 2 : The expression is further transformed using the rule of double negation and De Morgan's laws :

$$\begin{array}{ll} \neg((\neg a \vee b) \wedge b) \vee a & \Leftrightarrow \\ (\neg(\neg a \vee b) \vee \neg b) \vee a & \Leftrightarrow \\ ((\neg(\neg a) \wedge \neg b) \vee \neg b) \vee a & \Leftrightarrow \\ ((a \wedge \neg b) \vee \neg b) \vee a & \Leftrightarrow \\ (a \wedge \neg b) \vee (\neg b) \vee (a) & \text{disjunctive normal form} \end{array}$$

Step 3 : The disjunctive normal form is expanded using the laws of identity and complementarity and further transformed using the laws of distributivity :

$$(a \wedge \neg b) \vee (t \wedge \neg b) \vee (a \wedge t) \qquad \Leftrightarrow$$

$$(a \wedge \neg b) \vee ((a \vee \neg a) \wedge \neg b) \vee (a \wedge (b \vee \neg b)) \qquad \Leftrightarrow$$

$$(a \wedge \neg b) \vee (a \wedge \neg b) \vee (\neg a \wedge \neg b) \vee (a \wedge b) \vee (a \wedge \neg b)$$

Step 4 : The expression is reduced according to the law of idempotency by removing multiple occurrences of subexpressions :

$$(a \wedge \neg b) \vee (a \wedge \neg b) \vee (\neg a \wedge \neg b) \vee (a \wedge b) \vee (a \wedge \neg b) \qquad \Leftrightarrow$$

$$(a \wedge \neg b) \vee (\neg a \wedge \neg b) \vee (a \wedge b) \qquad \text{canonical disjunctive normal form}$$

The canonical disjunctive normal form is true for the valuations (T(a), T(b)) = (t, f), (f, f), (t, t). The logical expression

$$(a \Rightarrow b) \wedge b \Rightarrow a$$

may be transformed analogously into its canonical conjunctive normal form

$$(a \vee \neg b)$$

The canonical conjunctive normal form is false for the valuation (T(a), T(b)) = (f, t).

1.3.4 LOGICAL RULES OF INFERENCE

Logical implication : A logically valid expression of the form $a \Rightarrow b$ with expressions a, b is called a logical implication. If $a \Rightarrow b$ is a logical implication and a is true, then b is also true. If a is false, b may be true or false. The most important logical implications for statements a, b, c are :

extremality	$f \Rightarrow a$	$a \Rightarrow t$
reflexivity	$a \Rightarrow a$	$a \Leftarrow a$
contraction	$a \wedge b \Rightarrow a$	$a \vee b \Leftarrow a$
monotonicity	$(a \Rightarrow b) \Rightarrow (a \wedge c) \Rightarrow (b \wedge c)$	$(a \Rightarrow b) \Leftarrow (a \vee c) \Leftarrow (b \vee c)$
antitonicity	$(a \Rightarrow b) \Rightarrow (\neg b \Rightarrow \neg a)$	$(a \Rightarrow b) \Leftarrow (\neg b \Rightarrow \neg a)$
detachment	$(a \Rightarrow b) \wedge a \Rightarrow b$	$(a \Rightarrow b) \wedge \neg a \Rightarrow \neg b$
transitivity	$(a \Rightarrow b) \wedge (b \Rightarrow c) \Rightarrow (a \Rightarrow c)$	

The two rules of detachment are also called modus ponens and modus tollens, respectively. The rule of transitivity is also called modus barbara. These rules are frequently applied in logical deductions.

Logical implications may be obtained directly from logical equivalences: If $a \Leftrightarrow b$ is a logical equivalence, then $a \Rightarrow b$ and $a \Leftarrow b$ are logical implications. Every logical equivalence thus leads to two logical implications.

Logical deduction : Logical deduction is based on logical implications. Let the expression $a \Rightarrow b$ be a logical implication. Then $a \Rightarrow b$ is always true, independent of the truth values for a, b. If a is false, then b is either true or false, since by definition $f \Rightarrow t$ and $f \Rightarrow f$ are true. If a is true, then b is also true, since by definition $t \Rightarrow t$ is true and $t \Rightarrow f$ is false. Thus a true expression b may be deduced from a true expression a.

Rules of inference : Rules of inference are used to deduce new true statements from given true statements by logical implication.

(1) The first rule of detachment, $a \wedge (a \Rightarrow b) \Rightarrow b$, yields the following rule of inference: If the statement a is true and the implication $a \Rightarrow b$ is true, then the statement b is true.

(2) The second rule of detachment, $(a \Rightarrow b) \wedge \neg b \Rightarrow \neg a$, yields the following rule of inference : If the implication $a \Rightarrow b$ is true and the statement $\neg b$ is true, then the statement $\neg a$ is true.

(3) The rule of transitivity, $(a \Rightarrow b) \wedge (b \Rightarrow c) \Rightarrow (a \Rightarrow c)$, yields the following rule of inference : If the implication $a \Rightarrow b$ is true and the implication $b \Rightarrow c$ is true, then the implication $a \Rightarrow c$ is true.

These rules of inference are often presented in a scheme with two true premises and the true conclusion.

true premise	a	$a \Rightarrow b$	$a \Rightarrow b$
true premise	$a \Rightarrow b$	$\neg b$	$b \Rightarrow c$
true conclusion	b	$\neg a$	$a \Rightarrow c$

Example 1 : Logical deduction

Rule (1) : modus ponens

a := ”The light is red”
$a \Rightarrow b$:= ”If the light is red, then the cars stop”

---

b := ”The cars stop”

Rule (2) : modus tollens

$a \Rightarrow b$:= ” If the light is red, then the cars stop”
$\neg b$:= ”The cars do not stop”

$\neg a$:= ”The light is not red”

Rule (3) : modus barbara

$a \Rightarrow b$:= ” If the light is red, then the cars stop”
$b \Rightarrow c$:= ”If the cars stop, then a queue forms”

$a \Rightarrow c$:= ”If the light is red, then a queue forms”

The following inference is wrong, since the logical expression $(a \Rightarrow b) \wedge b \Rightarrow a$ is not logically valid and hence does not yield a logical implication (see Example 3 in Section 1.3.2).

$a \Rightarrow b$:= ”If the light is red, then the cars stop”
b := ”The cars stop”

---

a := ”The light is red”

The error in the inference may be demonstrated as follows : The cars stop not only if the light is red, but also, for example, if an accident has occurred.

1.4 PREDICATE LOGIC

Introduction : Predicate logic is an extension of propositional logic. It investigates the inner structure of statements. As in the case of a natural language, a statement is divided into its constituents, subject and predicate. Subjects are the topic of a statement, predicates describe the properties or relations of subjects.

Predicate logic allows quantified statements, which hold either for all subjects in a given set or for at least one subject in a given set. These statements are called universal statements and existential statements, respectively. Predicate logic introduces quantifiers to formulate such statements.

First-order predicate logic is an extension of propositional logic which is obtained by introducing constants and variables for subjects and predicates along with quantifiers for subject variables. Set theory is an essential basis of predicate logic. This section offers an introduction to first-order predicate logic.

Statement of predicate logic : A statement is called a statement of predicate logic if it admits analysis into subjects and predicates. The names of imaginary or real objects in a statement are called subjects. The names of properties or relations in a statement are called predicates. A predicate is said to be unary if it describes a property of one subject. A predicate is said to be binary if it describes a relationship between two subjects. A predicate is said to be n-ary (n-place) if it describes a relationship among n subjects. A truth value is associated with each statement of predicate logic.

Formula of predicate logic : A statement of predicate logic can be transformed into a formula. Every subject of the statement is replaced by a statement variable. The subject variables are usually designated by lowercase letters, for instance x, y, z. Then every predicate is replaced by a predicate variable. Each predicate variable is a truth value that depends on one or more subject variables. The predicate variables are usually designated by uppercase letters. The designation of the predicate variable is followed by a list of the subject variables on which the predicate variable depends, enclosed in parentheses, for instance K(x, z). The predicate variables are connected using operators.

Example 1 : Statements and formulas of predicate logic

The statement ”x is prime” contains the subject ”x” and the unary predicate ”is prime”. The subject ”x” is replaced by the subject variable $x \in \mathbb{N}$. The predicate ”is prime” is replaced by the predicate variable P(x), which depends on x. The corresponding formula is P(x). The statement P(3) has the value true, the statement P(8) has the value false.

The statement "x is less than y " contains the subjects "x" and "y" and the 2-place (binary) predicate "is less than". The subjects are replaced by the variables $x, y \in \mathbb{N}$, the predicate is replaced by the predicate variable $K(x, y)$. The corresponding formula is $K(x, y)$. The statement $K(2, 3)$ has the value true, the statement $K(3, 2)$ has the value false.

The statement "x is not less than y and less than z" contains the subjects "x", "y" and "z" and the predicate "is less than" as well as the operators "not" and "and". The subjects are replaced by the variables $x, y, z \in \mathbb{N}$, the predicate is replaced by $K(x, y)$ and $K(x, z)$. The corresponding formula is $\neg K(x, y) \wedge K(x, z)$.

Quantifiers : Let a formula $a(x)$ with the subject variable x and a reference set (universe) M for x with the elements $x_1, x_2, ..., x_n$ be given. A logical expression of the form $a(x_1) \wedge a(x_2) \wedge ... \wedge a(x_n)$ is called a universal statement. A universal statement is true if and only if the statement $a(x_i)$ is true for every element x_i of M. Otherwise it is false. A logical expression of the form $a(x_1) \vee a(x_2) \vee ... \vee a(x_n)$ is called an existential statement. An existential statement is true if and only if there is an element x_i for which the statement $a(x_i)$ is true. The universal quantifier $\bigwedge$ and the existential quantifier $\bigvee$ are introduced to formulate universal and existential statements :

$\bigwedge\limits_{x \in M} a(x)$	"For every element x of M the value of a(x) is true"
$\bigvee\limits_{x \in M} a(x)$	"There is an element x of M for which the value of a (x) is true"
$\bigwedge$	universal quantifier
$\bigvee$	existential quantifier
M	reference set

The reference set M may be finite or infinite. The formulation of statements using quantifiers is also applicable to formulas with several subject variables. If both universal and existential quantifiers appear in such a formulation, their order is important.

Example 2 : Universal and existential statements

The statement "In every plane triangle the sum of the interior angles is 180 degrees" is a true statement. Let D be the set of all plane triangles. The universal statement is formulated using the universal quantifier as follows :

$$\bigwedge_{d \in D} \text{(The sum of the interior angles in d is 180 degrees)}$$

The statement "Every natural number x has a natural number y as its successor" is synonymous with the statement "For every natural number x there is a natural number y which is a successor of x." Let $\mathbb{N}$ be the set of natural numbers. Then the statement is formulated using the universal quantifier and the existential quantifier as follows :

$$\bigwedge_{x \in \mathbb{N}} \bigvee_{y \in \mathbb{N}} (\text{y is successor of x})$$

The statement "For two points in the plane there is a line on which both points lie" is synonymous with the statement "For two points x, y in the plane there is a line g which contains x and contains y". Let P be the set of all points in the plane, and let G be the set of all lines in the plane. The statement is formulated using two universal quantifiers and an existential quantifier as follows :

$$\bigwedge_{x \in P} \bigwedge_{y \in P} \bigvee_{g \in G} ((\text{g contains x}) \wedge (\text{g contains y}))$$

Expression of predicate logic : The elements of predicate logic are constants, variables, operators, quantifiers and the technical characters (,). A sequence of elements is called an expression of predicate logic if it is formed according to the following rules :

(1) Every statement constant, every statement variable and every n-ary predicate with variables or constants for n subjects is an expression.

(2) If a, b are expressions, then $(\neg a)$, $(a \wedge b)$, $(a \vee b)$, $(a \Rightarrow b)$ and $(a \Leftrightarrow b)$ are also expressions.

(3) If a is an expression and x is a variable, then $(\bigwedge_{x} a)$ and $(\bigvee_{x} a)$ are also expressions.

(4) Only the sequences of elements formed according to rules (1) to (3) are expressions.

To avoid parentheses in expressions of predicate logic, a ranking of operators is defined as in propositional logic. An expression formed according to rule (1) is said to be atomic. A variable x in an expression is said to be free if it is not subject to a quantifier $\bigwedge_{x}$ or $\bigvee_{x}$. Otherwise, the variable is said to be bound. An expression without free variables is a statement of predicate logic. An expression with at least one free variable is a formula of predicate logic.

Interpretation : Every expression which is a statement of predicate logic is either true or false. The truth value depends on the meaning assigned to the subjects and predicates for a specific reference set. Such an assignment is called an interpretation. Statements of predicate logic with identical formal structure can take different truth values under different interpretations.

Example 3 : Interpretation of expressions

Let $\mathbb{N}$ be the set of natural numbers. In a first interpretation, let the binary predicate a(x, y) be assigned the meaning "x has the successor y". Since every natural number has a successor in $\mathbb{N}$, the statement $\bigwedge\limits_{x \in \mathbb{N}} \bigvee\limits_{y \in \mathbb{N}} a(x, y)$ of predicate logic is true.

In a second interpretation, let the binary predicate a(x, y) be assigned the meaning "x has the predecessor y". Since the natural number 0 has no predecessor in $\mathbb{N}$, the statement $\bigwedge\limits_{x \in \mathbb{N}} \bigvee\limits_{y \in \mathbb{N}} a(x, y)$ of predicate logic is false.

Logically valid, consistent and inconsistent expressions : The expressions of predicate logic are classified with respect to their interpretation and their truth value as follows :

(1) An expression of predicate logic which is true for every interpretation is said to be logically valid.

(2) An expression of predicate logic which is true for at least one interpretation is said to be logically consistent.

(3) An expression of predicate logic which is not true for any interpretation is said to be logically inconsistent.

Replacing every statement variable in an expression of propositional logic by an arbitrary expression of predicate logic yields a corresponding expression of predicate logic. If the expression of propositional logic is logically valid, consistent or inconsistent, then a corresponding expression of predicate logic is also logically valid, consistent or inconsistent. A logically valid expression of predicate logic derived from a logically valid expression of propositional logic is called a tautology of predicate logic.

Unlike in propositional logic, the logical validity of an arbitrary given expression of predicate logic cannot always be decided in a finite number of steps (Church's Undecidability Theorem).

Example 4 : Tautology of predicate logic

The expression $\neg(\neg a) \Leftrightarrow a$ of propositional logic is a tautology. If the statement variable a is replaced by the expression $\bigwedge\limits_{x} \bigwedge\limits_{y} a(x, y)$ of predicate logic, the corresponding expression is a tautology of predicate logic :

$$\neg(\neg \bigwedge_{x} \bigwedge_{y} a(x, y)) \Leftrightarrow \bigwedge_{x} \bigwedge_{y} a(x, y)$$

Logical equivalence for quantified expressions : A logically valid expression of predicate logic of the form $a \Leftrightarrow b$ (a is equivalent to b) is called a logical equivalence. In addition to the logical equivalences which are tautologies of predicate logic, there are further logical equivalences for quantified expressions :

relabeling	$\bigwedge_x a(x)$	$\Leftrightarrow$	$\bigwedge_y a(y)$
	$\bigvee_x a(x)$	$\Leftrightarrow$	$\bigwedge_y a(y)$
negation	$\neg \bigwedge_x a(x)$	$\Leftrightarrow$	$\bigvee_x \neg a(x)$
	$\neg \bigvee_x a(x)$	$\Leftrightarrow$	$\bigwedge_x \neg a(x)$
double negation	$\neg \bigwedge_x \neg a(x)$	$\Leftrightarrow$	$\bigvee_x a(x)$
	$\neg \bigvee_x \neg a(x)$	$\Leftrightarrow$	$\bigwedge_x a(x)$
con-/disjunction	$\bigwedge_x (a(x) \wedge b(x))$	$\Leftrightarrow$	$\bigwedge_x a(x) \wedge \bigwedge_x b(x)$
	$\bigvee_x (a(x) \vee b(x))$	$\Leftrightarrow$	$\bigvee_x a(x) \vee \bigvee_x b(x)$
commutation	$\bigwedge_x \bigwedge_y a(x, y)$	$\Leftrightarrow$	$\bigwedge_y \bigwedge_x a(x, y)$
	$\bigvee_x \bigvee_y a(x, y)$	$\Leftrightarrow$	$\bigvee_y \bigvee_x a(x, y)$

The following logical equivalences hold only if the variable x is not free in the expression a :

$$\bigwedge_x a \Leftrightarrow a \qquad\qquad \bigvee_x a \Leftrightarrow a$$

$$\bigwedge_x (a \wedge b(x)) \Leftrightarrow a \wedge \bigwedge_x b(x) \qquad\qquad \bigvee_x (a \vee b(x)) \Leftrightarrow a \vee \bigvee_x b(x)$$

$$\bigwedge_x (a \vee b(x)) \Leftrightarrow a \vee \bigwedge_x b(x) \qquad\qquad \bigvee_x (a \wedge b(x)) \Leftrightarrow a \wedge \bigvee_x b(x)$$

Prenex normal form : A given expression of predicate logic may be transformed into a logically equivalent expression using logical equivalences. An expression may be brought into prenex normal form using such transformations.

An expression of predicate logic is said to be in prenex form if it is either free of quantifiers or consists of a sequence of quantifiers followed by an expression without quantifiers. A prenex form is said to be in prenex normal form if the expression without quantifiers is a normal form of propositional logic. A prenex normal form is said to be disjunctive or conjunctive if the expression without quantifiers is a disjunctive or conjunctive normal form of propositional logic, respectively.

Every expression of predicate logic has an equivalent prenex disjunctive normal form and an equivalent prenex conjunctive normal form. The transformation of an expression of predicate logic into one of these normal forms is performed using the logical equivalences of propositional and predicate logic.

Example 5 : Prenex normal forms

The following expression of predicate logic is brought into prenex normal form, using logical equivalences of propositional and predicate logic in each step :

$$\begin{array}{lcl}
\bigvee_x a(x) \;\Rightarrow\; \bigvee_x b(x) & \Leftrightarrow & \text{(rule of elimination)} \\
\neg \bigvee_x a(x) \;\vee\; \bigvee_x b(x) & \Leftrightarrow & \text{(rule of negation)} \\
\bigwedge_x \neg a(x) \;\vee\; \bigvee_x b(x) & \Leftrightarrow & \text{(relabeling rule)} \\
\bigwedge_x \neg a(x) \;\vee\; \bigvee_y b(y) & \Leftrightarrow & \text{(prenex arrangement)} \\
\bigwedge_x \bigvee_y (\neg a(x) \;\vee\; b(y)) & \text{prenex normal form} &
\end{array}$$

In the first step, the operator $\Rightarrow$ is eliminated using the rule of elimination of propositional logic. In the second step, the rule of negation of predicate logic is applied to the first subexpression. In the third step, the relabeling rule of predicate logic is applied to the second subexpression to replace x by y. In the fourth step, the quantifiers $\bigwedge_x$ and $\bigvee_y$ are brought to the front to obtain a prenex form of the expression beginning with a sequence of quantifiers. After the fourth step, the expression is in prenex normal form, since the expression $(\neg a(x) \vee b(y))$ following the sequence of quantifiers is free of quantifiers and exhibits the normal form of propositional logic. This normal form is conjunctive.

Logical implication for quantified expressions : A logically valid expression of predicate logic of the form $a \Rightarrow b$ (a implies b) is called a logical implication. Besides the logical implications which are tautologies of predicate logic, there are logical implications for quantified expressions in a reference set M with $x, y, t \in M$:

$$\begin{array}{lrcl}
\text{elementary implication} & \bigwedge_x a(x) & \Rightarrow & a(y) \\
 & \bigvee_x a(x) & \Leftarrow & a(y) \\
\text{conjunction} & \bigvee_x (a(x) \wedge b(x)) & \Rightarrow & \bigvee_x a(x) \wedge \bigvee_x b(x) \\
\text{disjunction} & \bigwedge_x (a(x) \vee b(x)) & \Leftarrow & \bigwedge_x a(x) \vee \bigwedge_x b(x) \\
\text{subjunction} & \bigwedge_x (a(x) \Rightarrow b(x)) & \Rightarrow & (\bigwedge_x a(x) \Rightarrow \bigwedge_x b(x)) \\
 & \bigwedge_x (a(x) \Rightarrow b(x)) & \Rightarrow & (\bigvee_x a(x) \Rightarrow \bigvee_x b(x)) \\
\text{commutation} & \bigvee_x \bigwedge_y a(x, y) & \Rightarrow & \bigwedge_y \bigvee_x a(x,y)
\end{array}$$

Each logical equivalence $a \Leftrightarrow b$ yields the logical implications $a \Rightarrow b$ and $a \Leftarrow b$. The logical implications of propositional and predicate logic are used to deduce new true statements from given true statements.

1.5 PROOFS AND AXIOMS

Introduction : A field of knowledge is composed of definitions and theorems. A theorem is a statement which is proved to be true. There are different methods of proof, based on different logical rules of inference. The forms of direct and indirect proof and proof by induction are treated.

In some areas of mathematics, proofs may be formalized by introducing certain statements as axioms and certain rules of inference as rules of derivation, thereby allowing the theorems of the field to be formally derived. Which theorems can be derived depends on the axioms and rules of derivation introduced. However, according to Gödel's Incompleteness Theorem not every theorem of a theory can be formally derived on the basis of an axiomatization. The fundamental concepts of axiomatic systems are briefly explained.

Theorem : A valid statement concerning a mathematical fact is called a theorem. Theorems are often formulated as a logical implication $a \Rightarrow b$. The statement a is called the hypothesis (premise), the statement b is called the conclusion of the theorem.

Proof : The deduction of the truth of a statement from the truth of other statements is called a proof. By virtue of the proof, the statement becomes a theorem. Logical rules of inference are applied in proofs. A finite sequence of logical inferences is a proof if the following conditions are satisfied :

(1) Each inference in the sequence follows logically from the hypotheses of the theorem, theorems that have already been proved, substitution and replacement rules and rules of logic.

(2) The last inference in the sequence yields the conclusion of the theorem.

(3) The conclusion of the theorem is not used within the sequence.

Direct proof : For the hypothesis a and the conclusion b, the implication $a \Rightarrow b$ is shown to be true. Then if a is true, b is also true. This follows from the logical implication

$$(a \Rightarrow b) \wedge a \;\Rightarrow\; b$$

Example 1 : Extended Pythagorean Theorem

Hypothesis : Let a right triangle with hypotenuse c and adjacent sides a, b be given. The areas of the semicircles erected on the sides of the triangle are designated by F_a, F_b, F_c.

Conclusion : $F_a + F_b = F_c$

Proof : The direct proof consists of the following steps :

(1) The Pythagorean Theorem holds for a right triangle :

$$a^2 + b^2 = c^2$$

(2) The areas F_a, F_b of the semicircles erected over the sides a, b are determined using the formula for the area of a circle :

$$F_a = \pi\, a^2/2 \qquad\qquad F_b = \pi\, b^2/2$$

(3) From (2) and (1), the sum $F_a + F_b$ is obtained as

$$F_a + F_b = \pi a^2/2 + \pi\, b^2/2 = \pi(a^2 + b^2)\,/\,2 = \pi c^2/2$$

(4) The area F_c of the semicircle erected over the hypotenuse c is determined using the formula for the area of a circle :

$$F_c = \pi\, c^2/2$$

(5) Comparison of (3) and (4) yields the conclusion :

$$F_a + F_b = F_c$$

Indirect proof : For the hypothesis a and the conclusion b, the implication $\neg b \Rightarrow \neg a$ is shown to be true. Then if a is true, b is also true. This follows from the logical implication of the direct proof by the contraposition principle, $(a \Rightarrow b) \Leftrightarrow (\neg b \Rightarrow \neg a)$:

$$(\neg b \Rightarrow \neg a) \;\wedge\; a \Rightarrow b$$

This method of proof may also be applied by assuming a to be true and showing that the assumption $\neg b$ implies the statement $\neg a$. This results in a contradiction for the statement a. This contradiction shows that the assumption $\neg b$ must have been false, and hence that the conclusion b holds. There are further forms of indirect proof, based on the following logical implications :

$$(\neg b \Rightarrow a) \;\wedge\; \neg a \;\Rightarrow\; b$$

$$(\neg b \Rightarrow a) \;\wedge\; (\neg b \Rightarrow \neg a) \;\Rightarrow\; b$$

Example 2 : Prime numbers

Definition : A natural number $p > 1$ is said to be prime if it is divisible only by 1 and by itself.

Conclusion : There are infinitely many prime numbers.

Proof : The indirect proof consists of the following steps :

(1) The negation of the conclusion is assumed. There is only a finite number of prime numbers $p_1 < p_2 < \ldots < p_n$.

(2) The number z is formed as the product of the n prime numbers, incremented by 1 :

$$z = p_1 \cdot p_2 \cdot \ldots \cdot p_n + 1$$

(3) From (1) and (2) it follows that z is not prime, since z is greater than the greatest prime number.

(4) According to the factorization theorem, every natural number $a > 1$ which is not prime is divisible by a prime number.

(5) From (2) and (4) it follows that z is a prime number, since, due to the incrementation by 1, z is not divisible by any of the prime numbers $p_1, p_2, \ldots, p_n$.

(6) The statements (3) and (5) form a contradiction; therefore the negated conclusion (1) is false, and the conclusion is true.

Proof by induction : A statement S(n) which depends on a natural number n is proved by (mathematical) induction. S(0) is proved as the induction hypothesis. Then S(n) is deduced from S(n – 1) for an arbitrary number $n > 0$. Repeated application of this direct proof yields the conclusion S(n) for every natural number n.

$$((S(n-1) \Rightarrow S(n)) \quad \wedge \quad S(n-1)) \Rightarrow S(n)$$

Example 3 : Sums

Conclusion : The sum of the odd numbers from 1 to 2n + 1 is $(n+1)^2$.

$$S(n) \quad :\Leftrightarrow \quad \sum_{i=0}^{n} (2i+1) \quad = \quad (n+1)^2$$

Proof : The conclusion is proved by induction :

(1) Induction hypothesis : For $n = 0$, the statement is true, since the sum consists only of the number 1 and is therefore equal to $(0+1)^2 = 1$.

(2) Inference from n – 1 to n : For $n > 0$, S(n) follows from S(n – 1) :

$$\sum_{i=0}^{n}(2i+1) \quad = \quad \sum_{i=1}^{n-1}(2i+1) + 2n + 1 \quad = \quad n^2 + 2n + 1 \quad = \quad (n+1)^2$$

This result is applied for n = 1, 2, 3,... ; it follows that the conclusion S(n) is true for every natural number.

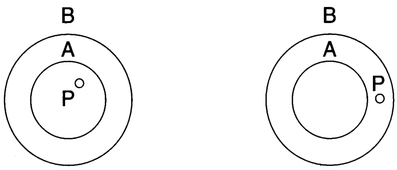
Sufficient and necessary conditions : If the conclusion b follows from the hypothesis a, the implication a ⇒ b is true. In this case, a is called a sufficient condition for b, and b is called a necessary condition for a. If the equivalence a ⇔ b is true, b is called a necessary and sufficient condition for a. These terms are motivated as follows :

(1) Since the statements a and a ⇒ b are true, it follows from the definition of the implication ⇒ that b is true. The truth of a is therefore sufficient for the truth of b. The truth of a is, however, not necessary for the truth of b. If a is false and b is true, the implication a ⇒ b remains true.

(2) Since the statement a ⇒ b is true, according to the truth table of the implication ⇒ the statement a can only be true if the statement b is true. The truth of b is therefore a necessary condition that must be satisfied if a is to be true. The truth of b is, however, not sufficient for the truth of a. If a is false and b is true, the implication a ⇒ b remains true.

(3) Since the statement a ⇔ b is true, according to the truth table for the equivalence a is true if and only if b is true. Therefore, b is a necessary and sufficient condition for a.

Example 4 : Necessary and sufficient conditions



Let the circles A and B be concentric. Let the radius of A be less than the radius of B. Let the position of a point P relative to the circles A and B be characterized by the following statements :

a := "The point P lies inside circle A"

b := "The point P lies inside circle B"

The implication a ⇒ b is true, since the point P always lies inside circle B if it lies inside circle A. The following sufficient and necessary conditions hold for the position of P :

(1) If the point P lies inside circle A, then this is a sufficient condition for the point P to lie inside circle B. It is, however, not a necessary condition. The point P can lie inside circle B without lying inside circle A.

(2) The point P can only lie inside circle A if it lies inside circle B. Lying inside circle B is therefore a necessary condition for lying inside circle A. Not every point that lies inside circle B also lies inside circle A. The condition that the point P lies inside circle B is therefore not a sufficient condition for it to lie inside circle A.

(3) If the radii of the concentric circles A and B are equal, then the equivalence $a \Leftrightarrow b$ is true. That the point P lies inside circle B is necessary and sufficient for the point P to lie inside circle A.

Axiomatization : Many areas of mathematics are axiomatized. To axiomatize a field, certain valid statements are chosen and designated as axioms. From these axioms, the theorems of the field are formally derived according to rules of inference. A set of valid statements in a field may be axiomatized in different ways. A valid statement which is an axiom in one axiomatization may be a derivable theorem in another axiomatization.

Axiomatic system : A statement that is assumed to be true in the course of an axiomatization is called an axiom. A set of axioms from which valid statements can be derived using formal rules of inference is called an axiomatic system. An axiomatic system must be consistent (free of contradictions) and should be independent :

(1) Consistency : An axiomatic system is said to be consistent (free of contradictions) if it is not possible to derive from it both a statement a and the statement $\neg a$.

(2) Independence : An axiomatic system is said to be independent if none of its axioms can be derived from the remaining axioms.

Completeness of an axiomatic system : An axiomatic system with its formal rules of inference is said to be complete for a field of mathematics if all theorems of the field can be formally derived. This means that the set of valid statements of the field and the set of statements derivable from the axiomatic system are identical.

According to Gödel's Completeness Theorem, a complete axiomatic system with formal rules of inference may be specified for propositional logic and first-order predicate logic, so that all theorems of propositional logic and first-order predicate logic are derivable. According to Gödel's Incompleteness Theorem, complete axiomatic systems with formal rules of inference cannot be specified for higher-order predicate logic.

2 SET THEORY

2.1 SETS

Introduction : The set is a fundamental concept of mathematics and computer science. Many problems in engineering deal with operations on sets. The concept of a set, the rules for forming sets (algebra of sets), the relationships between elements of sets (relations, mappings) and the corresponding classification of mathematics according to algebraic, ordinal and topological structures are treated in the following.

Formation of sets : A set M is specified either by enumerating the designations of the elements or by describing the properties of the elements. The order of enumeration of the elements is irrelevant. If two elements in the enumeration bear the same designation, they represent the same element. This element is contained in the set only once. The set without elements is called the empty set and is designated by $\emptyset$.

$M = \{a, b, c\}$	set M consists of the elements a, b, c
$M = \{x \mid E(x)\}$	set M contains every element for which the logical expression E (x) is true
$\emptyset := \{x \mid x \neq x\}$	empty set

The membership of an element a in a set M is represented using the symbols $\in$ and $\notin$:

$a \in M$	a is an element of M
$a \notin M$	a is not an element of M

Quantifier : There are statements which are true for certain elements of a set M and false for other elements of M. Such relationships between statements and elements are represented using the universal quantifier $\bigwedge$ and the existential quantifier $\bigvee$. Often the set M is not explicitly specified if it is self-evident.

universal quantifier :	$\bigwedge\limits_{x \in M} (...)$	for every x in the set M ... holds
existential quantifier :	$\bigvee\limits_{x \in M} (...)$	there is an x in the set M for which ... holds

Equal sets : Two sets A and B are said to be equal if they contain the same elements. If the sets A and B are equal, they contain the same elements. The statement $A = B$ (A equals B) possesses either the statement value true or the statement value false.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*
© Springer-Verlag Berlin Heidelberg 2001

$$(A = B) \quad :\Leftrightarrow \quad \bigwedge_{x} (x \in A \;\Leftrightarrow\; x \in B)$$

$A = B$	sets A and B are equal
$A \neq B$	sets A and B are not equal

Subset : A set A is called a subset of a set B if every element of A is also an element of B. If the set B contains at least one element not contained in A, then A is called a proper subset of B.

$$(A \subseteq B) \quad :\Leftrightarrow \quad \bigwedge_{x} (x \in A \;\Rightarrow\; x \in B)$$

$$(A \subset B) \quad :\Leftrightarrow \quad (A \subseteq B) \;\wedge\; \neg(A = B)$$

$A \subseteq B$	A is a subset of B
$A \subset B$	A is a proper subset of B

In addition to the symbols $\subseteq$ (contained in) and $\subset$ (properly contained in), the symbols $\supseteq$ (includes) and $\supset$ (properly includes) are also used.

$B \supseteq A$	set B includes set A
$B \supset A$	set B properly includes A

System of sets : A set whose elements are themselves sets is called a system of sets. A system of sets must not contain any elements which are not sets. The number of elements in different sets of a system of sets may be different.

Power set : From a given set M of n elements, 2^n subsets can be formed, including $\emptyset$ and M. The set of all subsets of M, including $\emptyset$ and M, is called the power set of M and is designated by P(M). The set M is called the reference set of the power set P(M).

Example 1 : Sets

enumeration of a set	:	M	=	$\{a, b, c, d, \ldots, x, y, z\}$
description of a set	:	M	=	$\{x \mid x \text{ is an uppercase letter}\}$
equal sets $A = B$	:	A	=	$\{a, b, c\} \quad B = \{b, c, a\}$
subset $A \subset M$	:	A	=	$\{a, d, y, z\}$
system of sets	:	S	=	$\{\{a, c, e\}, \{1, 3, 5, 9\}, \{\alpha, \beta\}\}$
power set of $\{a, b\}$	:	P	=	$\{\emptyset, \{a\}, \{b\}, \{a, b\}\}$

Family of elements : Designating the elements of a set by different names is inconvenient for sets with a large number of elements. The elements of a set X are therefore often designated by $x_1, x_2, x_3, \ldots$. The common designation by the lowercase letter x symbolizes membership in the set X, while the index $i \in \{1, 2, 3, \ldots\}$ identifies the element. The elements x_i are called a family of elements. The family of elements is designated by $\{x_i\}$.

$$X = \{ x_i \mid i \in I = \{1, 2, 3, \ldots\}\}$$

Family of sets : Designating the sets of a system M of sets by different names is often inconvenient. A family of sets is therefore formed which contains the sets A_i as elements. Each of the sets A_i may be a family of elements $\{ a_{im} \}$.

$$M = \{A_i \mid i \in I = \{1, 2, 3, \ldots\}\}$$

$$A_i = \{a_{im} \mid m \in M_i = \{1, 2, 3, \ldots\}\}$$

Example 2 : Families of elements and sets

family of elements : $B = \{b_1, b_4, b_5, b_7\} \Rightarrow B = \{b_i \mid i \in \{1, 4, 5, 7\}\}$

family of sets : $A_1 = \{a\}$ $A_2 = \{a, b\}$ $A_3 = \{b, c\}$
$M = \{A_i \mid i \in \{1, 2, 3\}\}$

2.2 ALGEBRA OF SETS

Operations on sets : A rule which, for two given sets, yields exactly one set as a result is called an operation on sets. The rule for an operation is defined using the operators $\neg$, $\wedge$, $\vee$ and $\oplus$ of propositional logic and the set membership $\in$ of the elements. Each operation is designated by a symbol. For the sets A and B, the following operations are defined :

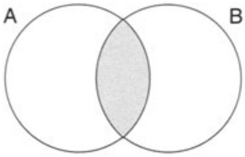
intersection : $A \cap B := \{x \mid x \in A \;\wedge\; x \in B\}$

union : $A \cup B := \{x \mid x \in A \;\vee\; x \in B\}$

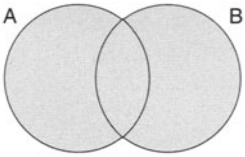
difference : $A - B := \{x \mid x \in A \;\wedge\; x \notin B\}$

symmetric difference : $A \oplus B := \{x \mid x \in A \;\oplus\; x \in B\}$

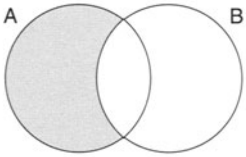
Set diagram : A set is schematically represented by a region in a plane. Every element of the set is represented by a point in this region. The following set diagrams represent the operations on sets.



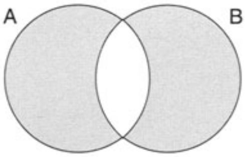
intersection A∩B



union A∪B



difference A – B



symmetric difference A⊕B

Example 1 : Operations on sets

Applying the operations to the sets $A = \{a, c, d\}$ and $B = \{c, x, z\}$ leads to the following results :

intersection : $A \cap B = \{c\}$

union : $A \cup B = \{a, c, d, x, z\}$

difference : $A - B = \{a, d\}$

difference : $B - A = \{x, z\}$

sym. difference : $A \oplus B = \{a, d, x, z\}$

Complement of a set : Let the set A be a subset of the set M. The difference M – A is called the complement of A with respect to M and is designated by $\overline{A}$.

complement : $\overline{A} := M - A$

operations : $A \cap \overline{A} = \emptyset \qquad A \cup \overline{A} = M$

Generalized operations : The set of all elements contained in each set A_i of an indexed system of sets is called the generalized intersection $\cap A_i$ of the system of sets. The set of all elements contained in at least one set A_i of an indexed system of sets is called the generalized union $\cup A_i$ of the system of sets.

$$\bigcap_{i \in I} A_i := \{x \mid \bigwedge_{i \in I} (x \in A_i)\}$$

$$\bigcup_{i \in I} A_i := \{x \mid \bigvee_{i \in I} (x \in A_i)\}$$

Disjoint sets : The sets A and B are said to be disjoint if they have no elements in common. A system of sets is said to be disjoint if its elements A_i are pairwise disjoint.

disjunction : $A \cap B = \emptyset$

Partition : A subdivision of a set M into a disjoint system of subsets T_i is called a partition of M. Every element of M is contained in exactly one of these subsets, none of the subsets is empty, and the union of the subsets is the set M.

$$M = T_1 \cup T_2 \cup ... \cup T_n$$

$$\bigwedge_i \bigwedge_m (i = m \ \vee \ T_i \cap T_m = \emptyset)$$

Example 2 : Disjoint systems of sets

indexed sets	:	$A_1 = \{a, b\}$	$A_2 = \{b, c, d\}$	$A_3 = \{d, e, f\}$
disjoint system of sets	:	$M = \{A_1, A_3\}$		
non-disjoint system	:	$M = \{A_1, A_2, A_3\}$		

Set-valued expressions : Set-valued expressions are the fundamental objects of the algebra of sets. A sequence of symbols is called a set-valued expression if it is formed according to the following syntactic rules :

(1) Every set value is an expression.

(2) If A and B are expressions, then $\overline{A}$, $(A \cap B)$, $(A \cup B)$ and $(A - B)$ are also expressions.

(3) Only sequences of symbols formed by rules (1) and (2) are expressions.

Valuation of a set-valued expression : A set value is either a set constant or a set variable. A set constant is the representation of a specific set, enclosed in curly brackets. A set variable is a character string which is replaced by a set constant. A valuation for an expression is obtained by assigning a set constant to each set variable in the expression. The operations contained in the expression are performed for this valuation. The resulting value of the expression is a set.

Rules of calculation : If the set-valued expressions Y and Z are equal for all valuations, the logical expression $Y = Z$ is called a rule of calculation. The rules of calculation follow from the definitions of the operations $\cap, \cup$ and $-$ together with the truth tables of the operators $\neg$, $\wedge$ and $\vee$. The following expressions involving the set variables A, B and C are rules of calculation of the algebra of sets.

idempotency	$A \cap A = A$	$A \cup A = A$
commutativity	$A \cap B = B \cap A$	$A \cup B = B \cup A$
associativity	$(A \cap B) \cap C = A \cap (B \cap C)$	$(A \cup B) \cup C = A \cup (B \cup C)$
distributivity	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
absorption	$A \cap (A \cup B) = A$	$A \cup (A \cap B) = A$

If A, B and C are subsets of M and all complements are formed with respect to M, the following expressions are also rules of calculation.

identity	$A \cap M = A$	$A \cup \emptyset = A$
invariance	$A \cap \emptyset = \emptyset$	$A \cup M = M$
reflexivity	$A \subseteq A$	$A \supseteq A$
extremality	$\emptyset \subseteq A$	$M \supseteq A$
contraction	$A \cap B \subseteq A$	$A \cup B \supseteq A$
monotonicity	$A \subseteq B \Rightarrow A \cap C \subseteq B \cap C$	$A \supseteq B \Rightarrow A \cup C \supseteq B \cup C$
complementarity	$A \cap \overline{A} = \emptyset$	$A \cup \overline{A} = M$
double complement	$\overline{\overline{A}} = A$	$\overline{\overline{A}} = A$
antitonicity	$A \subseteq B \Rightarrow \overline{A} \supseteq \overline{B}$	$A \supseteq B \Rightarrow \overline{A} \subseteq \overline{B}$
De Morgan	$(\overline{A \cap B}) = \overline{A} \cup \overline{B}$	$(\overline{A \cup B}) = \overline{A} \cap \overline{B}$

2.3 RELATIONS

Introduction : There may be relationships between the elements of sets. The order in which the sets are considered may be relevant in these relationships. Such relationships are treated in the following, using the concepts of ordered pair, direct product, relation and class.

Ordered pair : In a set, the order of elements is irrelevant, so that $\{a, b\} = \{b, a\}$. Two elements a and b whose order is relevant are called an ordered pair. An ordered pair is enclosed in parentheses. The elements a and b may be contained in different sets. Two ordered pairs (a, b) and (c, d) are equal if and only if $a = c$ and $b = d$.

ordered pair : $(a, b) := \{\{a\}, \{a, b\}\}$
a first component of the ordered pair (a, b)
b second component of the ordered pair (a, b)

equal pairs : $(a, b) = (c, d) \quad \Leftrightarrow \quad (a = c) \wedge (b = d)$

Cartesian product : Let the sets A and B be given. The set of all ordered pairs (a,b) that can be formed using elements $a \in A$ and $b \in B$ is called the cartesian product (direct product) of the sets A and B. The cartesian product is designated by $A \times B$ (A times B).

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

Relation : Let the sets A and B be given, together with an operation on the elements $a \in A$ and $b \in B$ whose value is a logical constant. The value of the operation for the ordered pair (a, b) in the product $A \times B$ is designated by aRb (a is related to b) and is either true or false.

The subset R of pairs (a, b) for which aRb is true is called a relation on A and B. Thus the relation is a set containing the pairs of elements for which the relationship specified by the operation holds. The order of the elements a and b in the operation is relevant to the result of the operation.

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

Relation in M : The subset $R \subseteq M \times M$ of the cartesian product of a set with itself for which aRb is true is called a relation in M. The relationships between the statement values aRb and bRa of the pairs (a, b) and (b, a) determine the properties of the relation. These properties are defined in the following for $a, b, c \in M$.

$$R := \{(a, b) \in M \times M \mid aRb\}$$

R is reflexive :⇔ $\bigwedge\limits_{a} (aRa)$

R is antireflexive :⇔ $\bigwedge\limits_{a} (\neg aRa)$

R is symmetric :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} (aRb \Rightarrow bRa)$

R is asymmetric :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} (aRb \Rightarrow \neg bRa)$

R is antisymmetric :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} (aRb \wedge bRa \Rightarrow a = b)$

R is linear :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} (aRb \vee bRa)$

R is connex :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} (a \neq b \Rightarrow aRb \vee bRa)$

R is transitive :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} \bigwedge\limits_{c} (aRb \wedge bRc \Rightarrow aRc)$

Example 1 : The strict order relation for natural numbers

The strict order relation $a < b$ in the set $\mathbb{N}$ of natural numbers is antireflexive, asymmetric, connex and transitive.

Totality of a relation on A and B : The subset $R \subseteq A \times B$ for which aRb is true is a relation on the sets A and B. The subset of A for which there exists $b \in B$ such that aRb is true is called the domain of R. The subset of B for which there exists $a \in A$ such that aRb is true is called the codomain of R. The relation is said to be left-total if its domain is A. The relation is said to be right-total if its range is B. A relation which is left- and right-total is said to be bitotal.

R is left-total :⇔ $\bigwedge\limits_{a} \bigvee\limits_{b} (aRb)$

R is right-total :⇔ $\bigwedge\limits_{b} \bigvee\limits_{a} (aRb)$

R is bitotal :⇔ R is left-total $\wedge$ R is right-total

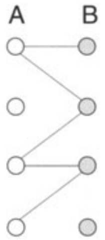
Uniqueness of a relation on A and B : A relation on A and B is said to be left-unique if the statements aRb and cRb are true only for $a = c$. The relation is said to be right-unique if the statements aRb and aRc are true only for $b = c$. A relation which is left-unique and right-unique is said to be bi-unique.

R is left-unique :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} \bigwedge\limits_{c} (aRb \wedge cRb \Rightarrow a = c)$

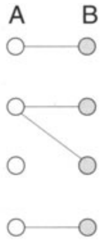
R is right-unique :⇔ $\bigwedge\limits_{a} \bigwedge\limits_{b} \bigwedge\limits_{c} (aRb \wedge aRc \Rightarrow b = c)$

R is bi-unique :⇔ R is left-unique $\wedge$ R is right-unique

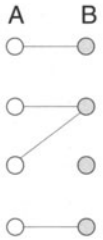
Relational diagram : A relational diagram shows three sets : the sets A and B as well as the relation R. The elements of A and B are represented by different symbols, for instance empty and filled circles. The elements of R are represented by line segments. For $R \subseteq A \times B$ the elements $a \in A$ and $b \in B$ for which aRb is true are joined by line segments. The following relational diagrams illustrate the uniqueness of R.



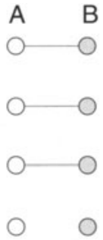
R general
m : n relationship



R left-unique
1 : n relationship



R right-unique
m : 1 relationship



R bi-unique
1 : 1 relationship

n-ary relation : A relation on two sets is called a binary relation. The concept of a relation is extended to describe relationships among n sets. An arrangement of n elements whose order is relevant is called an n-tuple. The n-tuple is defined recursively using the ordered pair.

$$(x_1,\ x_2, \ldots, x_n) \quad := \quad ((x_1,\ x_2, \ldots, x_{n-1}),\ x_n)$$

The set of all n-tuples which can be formed using elements x_i of the sets M_i with $i \in \{1, 2, \ldots, n\}$ is called the n-ary (n-fold) product $M_1 \times M_2 \times \ldots \times M_n$. If the sets M_i are equal, the n-ary product is written as M^n.

Let an operation on the n-tuple $(x_1,\ x_2, \ldots, x_n)$ be defined whose result is a logical constant, designated by $Rx_1 x_2 \ldots x_n$. The n-ary (n-place) relation $R \subseteq M_1 \times \ldots \times M_n$ is the subset of n-tuples $(x_1,\ x_2, \ldots, x_n)$ for which $R x_1\, x_2 \ldots x_n$ is true.

$$R := \{(x_1,\ x_2, \ldots, x_n) \ \in \ M_1 \times M_2 \times \ldots \times M_n \ \mid \ Rx_1\, x_2 \ldots x_n\ \}$$

Example 2 : The natural numbers a and b and their sum $c = a + b$ form a ternary (3-place) relation on $\mathbb{N}^3$:

$$R \ = \ \{(a, b, c) \in \mathbb{N}^3 \ \mid \ c \ = \ a + b\}$$

2.4 TYPES OF RELATIONS

Every relation is a subset of a direct product. Relations often have additional properties. Relations with common properties belong to a type of relations. Some types of relations are defined in the following. They lead to the concept of a class, which is of central importance in the formation of models.

Identity relation : The set of all ordered pairs (a, a) in the product $A \times A$ is called the identity relation I_A in the set A.

$$I_A \quad := \quad \{(a, a) \mid a \in A\}$$

Dual relation : The set R^{-1} is called the dual (inverse) relation of the relation R if the order of the elements in the ordered pairs (a, b) of R is exchanged in R^{-1}.

$$R^{-1} \quad := \quad \{(b, a) \mid (a, b) \in R\}$$

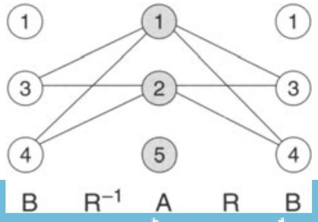
Composition : Let a relation R on the sets A and B and a relation S on the sets B and C be given. The set of ordered pairs $(a, c) \in A \times C$ for which there is a common element in B is called the composition of R and S. The order of R and S is relevant, as b is the second element of R and the first element of S. The composition is designated by $S \circ R$.

$$S \circ R \; := \; \{(a, c) \in A \times C \mid \bigvee_{b \in B} (aRb \;\wedge\; bSc)\}$$

Example 1 : Dual relation and composition

Let the sets $A = \{1, 2, 5\}$ and $B = \{1, 3, 4\}$ be given. Let the relation R be the set of all pairs (a, b) with $a \in A$ and $b \in B$ for which $a < b$ is true. The composition of this relation with its dual relation does not yield an identity!

product	: $A \times B$	=	{(1,1), (1,3), (1,4), (2,1), (2,3), (2,4), (5,1), (5,3), (5,4)}
relation	: R	=	{(1,3), (1,4), (2,3), (2,4)}
dual relation	: R^{-1}	=	{(3,1), (4,1), (3,2), (4,2)}
composition :	$R \circ R^{-1}$	=	{(3,3), (3,4), (4,3), (4,4)}



Equivalence relation : A relation $E \subseteq M \times M$ is called an equivalence relation in the set M if it is reflexive, symmetric and transitive. The elements x and y of the set M are said to be equivalent if the set E contains the pair (x, y) ; this relationship is designated by $x \sim y$ or xEy.

E is reflexive : $x \sim x$

E is symmetric : $x \sim y \quad \Rightarrow \quad y \sim x$

E is transitive : $x \sim y \quad \wedge \quad y \sim z \quad \Rightarrow \quad x \sim z$

Equivalence class : A subset of a set M is called an equivalence class in M if the elements of the subset are pairwise equivalent. An equivalence class is designated by choosing an arbitrary element a of the class and enclosing it in square brackets [a]. The selected element a is called a representative of its class.

$$[a] \quad := \quad \{ x \in M \quad \mid \ (a, x) \in E \}$$

Partitioning by equivalence : The equivalence classes in a set M for a given equivalence relation E form a partition of M :

(1) Every element x of the set M is contained in at least one equivalence class, since (x, x) is an element of the reflexive relation E.

(2) None of the equivalence classes [x] is empty, since $(x, x) \in E$ and hence at least x itself is an element of [x].

(3) Every element z of the set M is contained in exactly one equivalence class. In fact, if z is an element of the classes [x] and [y], then since E is symmetric and transitive $z \sim x$ and $z \sim y$ imply $x \sim z$ and $x \sim y$; hence $[x] = [y]$.

Quotient set : The set of equivalence classes of a set M for an equivalence relation E is called a quotient set and is designated by M/E (M modulo E). A subset $R \subseteq M$ is called a system of representatives of the quotient set M/E if it contains exactly one representative from each class of M/E.

$$M/E \ := \ \{ [x] \ \mid \ x \in M \}$$

Example 2 : Parallel lines

Let a set M of lines in a plane be given. Let the lines be parallel either to the x-axis or to the y-axis. Let two lines be equivalent if they are parallel. The relation ”Line a is parallel to line b” has the properties of an equivalence relation :

Reflexivity : Every line is parallel to itself.

Symmetry : If line a is parallel to line b, then b is also parallel to a.

Transitivity : If line a is parallel to line b and line b is parallel to line c, then a is also parallel to c.

The equivalence relation is derived from the abstraction "direction of a line". It partitions the set M into two equivalence classes. One class contains all lines which are parallel to the x-axis. The other class contains all lines which are parallel to the y-axis. A system of representatives contains one line parallel to the x-axis and one line parallel to the y-axis.

Closure : A relation $H \subseteq A \times B$ is called a closure of the relation $R \subseteq A \times B$ if elements of the set R enter into the rule for forming the set H. The closure H of a relation R is designated by $<R>_x$. The symbol x stands for the properties of the relation R which enter into the rule for H.

Symmetric closure : The relation $R \subseteq M \times M$ in the set M is generally not symmetric. The symmetric closure $<R>_s$ of R is formed by letting every element $(x, y) \in R$ and its dual element (y, x) be contained in the closure. The symmetric closure $<R>_s$ is an extension of R.

$$<R>_s \; := \; \{(x, y) \mid (x, y) \in R \;\vee\; (y, x) \in R\}$$

Connection : Consider the relation $R \subseteq M \times M$ in the set M. An n-tuple $(x_1, x_2, \ldots, x_n) \in M^n$ is called a connection of the elements a and b by R in M if all ordered pairs (x_i, x_{i+1}) are contained in the relation R and $x_1 = a$, $x_n = b$. The number $n-1$ of ordered pairs is called the length of the connection. For given elements a, b in M, there may be several connections with equal or different lengths. The statement "The elements a and b are connected by R" is designated by aV_Rb.

$$V_R \; := \; \{(x_1, x_2, \ldots, x_n) \mid \bigwedge_{i \in \{1,\ldots,n-1\}} ((x_i, x_{i+1}) \in R)\}$$

$$aV_Rb \; :\Leftrightarrow \; \bigvee_{(x_1,\ldots,x_n) \in V_R} (x_1 = a \;\wedge\; x_n = b)$$

Transitive closure : The relation $R \subseteq M \times M$ in the set M contains only binary connections. The transitive closure $<R>_t$ of R contains all ordered pairs $(a, b) \in M^2$ which are connected by R. The transitive closure $<R>_t$ is an extension of R.

$$<R>_t \; := \; \{(a, b) \in M^2 \mid aV_Rb\}$$

Reflexive transitive closure : If a relation $R \subseteq M \times M$ is not reflexive, then its transitive closure $<R>_t$ is not reflexive either. The reflexive transitive closure $<R>_{rt}$ of R is formed by adding all ordered pairs $(a, a) \in M^2$ to the transitive closure.

$$<R>_{rt} \; := \; \{(a, b) \in M^2 \mid (a, b) \in \; <R>_t \;\vee\; a = b\}$$

Reflexive symmetric transitive closure : The reflexive transitive closure of the symmetric closure of a relation $R \subseteq M \times M$ is called the reflexive symmetric transitive closure $<R>_{rst}$ of R. This closure is an equivalence relation and can therefore be used to classify the elements of M.

$$<R>_{rst} \; := \; \{(a, b) \in M^2 \mid (a, b) \in \; <<R>_s>_t \;\vee\; a = b\}$$

Example 3 : Component problems

Let M be the set of parts and partially assembled units which occur in the assembly of a steel construction. These parts and partial assemblies are called components of the assembly. The statement ”Component x_1 is directly necessary for the assembly of component x_2” leads to the component relation $R \subseteq M \times M$. Let the set R be given. The value of the statement ”Component x_1 is directly or indirectly needed for the assembly of component x_2” is to be determined. The statement is true if (x_1, x_2) is an element of the transitive closure $<R>_t$ of the component relation.

Example 4 : Train connections

Let M be a set of railway stations. The statement ”A train goes non-stop from station x_1 to station x_2” leads to a traffic relation $R \subseteq M \times M$. Let the set R be given. The value of the statement ”There is a train connection from station x_1 to station x_2” is to be determined. This statement is true if (x_1, x_2) is an element of the transitive closure $<R>_t$ of the traffic relation.

2.5 MAPPINGS

Introduction : Relations generally do not establish unique relationships between the elements of sets. However, in many applications it is convenient to assign to each element of a set A exactly one element of a set Z. The same element of Z may be assigned to different elements of A. Relations of this type are called mappings. If A and Z are sets of numbers, the term "function" is often used instead of the term "mapping".

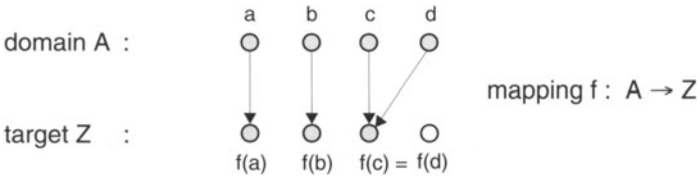
Mapping : A relation $f \subseteq A \times Z$ is called a mapping if it is left-total and right-unique. The following notation is used :

$f: A \to Z$	f is a mapping from A to Z
A	domain of f
Z	target of f

Image of an element : If the mapping f assigns the element $z \in Z$ to the element $a \in A$, then z is called the image of a under the mapping f. The element a is called a preimage (inverse image) of z. The following notation is used :

$$f : a \to z \qquad \text{or} \qquad f(a) = z$$

Arrow diagram : Mappings are depicted using arrow diagrams. Every element of the domain is the starting point of an arrow. The arrow points to the image in the target.



Fiber : Every element a of the domain A of a mapping $f : A \to Z$ has a unique image f(a) in Z. An element z of the target Z may have zero, one or several preimages. The set of all preimages of an element z in the target is called the fiber of f over z and is designated by $f^{-1}(z)$.

$$f^{-1}(z) := \{x \in M \mid f(x) = z\}$$

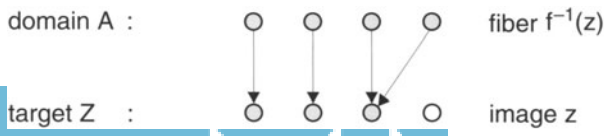
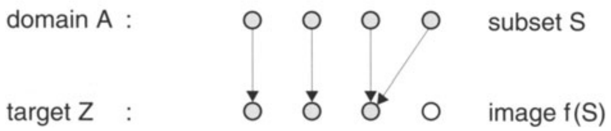
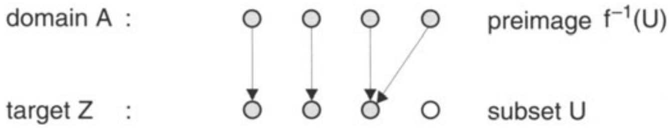


Image of a subset : Let S be a subset of the domain A. The set of images of the elements of S under the mapping $f : A \rightarrow Z$ is called the image of the subset S and is designated by $f(S)$.

$$f(S) := \{ z \in Z \mid z = f(x) \wedge x \in S \}$$



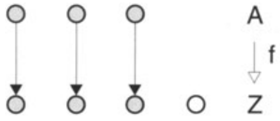
Preimage of a subset : Let U be a subset of the target Z. The union of the fibers of the elements of U under the mapping $f : A \rightarrow Z$ is called the preimage of the subset U and is designated by $f^{-1}(U)$.



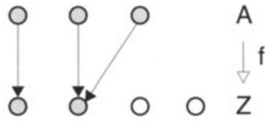
2.6 TYPES OF MAPPINGS

Introduction : All mappings are left-total and right-unique relations. Mappings often have additional properties. Mappings with common additional properties belong to a type of mappings. Types of mappings are often defined according to the effect of successive mappings. Some types of mappings are described in the following.

Injective mapping : A mapping $f: A \rightarrow Z$ is said to be injective (an injection) if two different elements $a \neq b$ of the set A always possess two different images $f(a) \neq f(b)$. An injection is a left-total, bi-unique relation. From $f(a) = f(b)$ it follows that $a = b$.

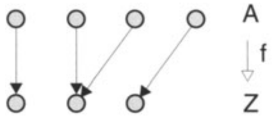


injection

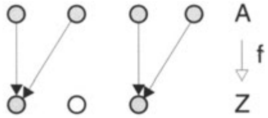


not an injection

Surjective mapping : A mapping $f : A \rightarrow Z$ is said to be surjective (a surjection) if each element of the target Z is the image of at least one element of A. A surjection is a bitotal, right-unique relation. An element $z \in Z$ may be the image of more than one element in A.

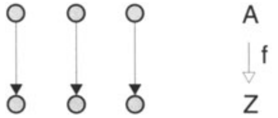


surjection

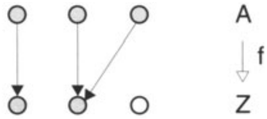


not a surjection

Bijective mapping : A mapping $f: A \rightarrow Z$ is said to be bijective (a bijection) if every element of Z is the image of exactly one element of A. A bijection is a bitotal, bi-unique relation. The number of elements in A and Z is the same.



bijection



not a bijection

Permutation : A bijective mapping $p: A \rightarrow A$ of a set to itself is called a permutation. The permutations of the set $\{a, b, c\}$ are the following sets:

$P_1 = \{(a,a), (b,b), (c,c)\}$ $\quad$ $P_4 = \{(a,a), (b,c), (c,b)\}$

$P_2 = \{(a,c), (b,a), (c,b)\}$ $\quad$ $P_5 = \{(a,b), (b,a), (c,c)\}$

$P_3 = \{(a,b), (b,c), (c,a)\}$ $\quad$ $P_6 = \{(a,c), (b,b), (c,a)\}$

Identity mapping : A permutation $1_A : A \to A$ is called an identity mapping if each element $a \in A$ is its own image $1_A(a)$:

$$1_A : A \to A \qquad \text{with} \qquad 1_A(a) = a$$

Equal mappings : The mappings $f : A \to Z$ and $g : A \to Z$ are said to be equal if the images f(a) and g(a) are equal for every element $a \in A$:

$$(f = g) \quad :\Leftrightarrow \quad \bigwedge_{a \in A} (f(a) = g(a))$$

Constant mapping : The mapping $f : A \to Z$ is called a constant mapping to the value c if all elements of the set A have the same image c in the set Z :

$$f : A \to Z \text{ is constant} \quad :\Leftrightarrow \quad \bigvee_{c \in Z} \bigwedge_{a \in A} (f(a) = c)$$

Composition : A mapping $g \circ f : A \to C$ is called the composition of the mappings f and g if first the mapping $f : A \to B$ is applied and then the mapping $g : B \to C$ is applied. For every element $a \in A$, first the image $b = f(a)$ is determined. For the element $b \in B$, the image $c = g(b)$ in C is then determined. Composition of mappings is associative, but generally not commutative.

$$g \circ f : A \to C \qquad \text{with} \qquad c = g(f(a))$$

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

Commutative diagram : The relationships between the mappings in a composition are represented in arrow diagrams. A diagram of mappings is said to be commutative if all compositions with the same domain and target are equal.

Example 1 : Commutative diagram

$$\begin{tikzcd}
A \arrow[r, "f_3"] \arrow[dd, "f_1"'] & D \arrow[dd, "f_4"'] \arrow[dr, "f_5"] & \\
& & E \\
B \arrow[r, "f_2"'] & C \arrow[ur, "f_6"'] &
\end{tikzcd}$$

Applying the mapping $f_2 : B \to C$ after the mapping $f_1 : A \to B$ and applying the mapping $f_4 : D \to C$ after the mapping $f_3 : A \to D$ leads to the same mapping. The mapping $f_5 : D \to E$ and the composition of the mapping $f_6 : C \to E$ with the mapping $f_4 : D \to C$ also coincide.

$$f_2 \circ f_1 = f_4 \circ f_3$$

$$f_6 \circ f_4 = f_5$$

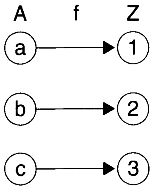
Establishing the type of a mapping : A mapping f : A → Z is injective if and only if there is a mapping g : Z → A such that the composition $g \circ f$ is the identity mapping 1_A. The mapping f is surjective if and only if there is a mapping g : Z → A such that the composition $f \circ g$ is the identity mapping 1_Z. If f is injective and surjective, then f is bijective.

$$f: A \to Z \text{ is injective} \quad \Leftrightarrow \quad \bigvee_g (g: Z \to A \ \wedge \ g \circ f = 1_A)$$

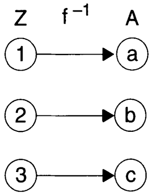
$$f: A \to Z \text{ is surjective} \quad \Leftrightarrow \quad \bigvee_g (g: Z \to A \ \wedge \ f \circ g = 1_Z)$$

$$f: A \to Z \text{ is bijective} \quad \Leftrightarrow \quad \bigvee_g (g: Z \to A \ \wedge \ g \circ f = 1_A \ \wedge \ f \circ g = 1_Z)$$

Inverse mapping : For every bijective mapping f : A → Z there is an inverse mapping $f^{-1}: Z \to A$. From $f(a) = z$ it follows that $f^{-1}(z) = a$. If a mapping f is not bijective, it does not possess an inverse mapping.



mapping



inverse mapping

Example 2 : Composition of mappings and inverse mappings

Let the permutations f and g of a set {1, 2, 3, 4} be given :

$$f = \{(1,2), (2,3), (3,4), (4,1)\}$$
$$g = \{(1,3), (2,4), (3,2), (4,1)\}$$

The two mappings are bijective and possess the following inverses :

$$f^{-1} = \{(1,4), (2,1), (3,2), (4,3)\}$$
$$g^{-1} = \{(1,4), (2,3), (3,1), (4,2)\}$$

The compositions $g \circ f$ and $f \circ g$ are different :

$$\begin{array}{lllll} g(f(1)) = g(2) = 4 & \qquad & f(g(1)) = f(3) = 4 \\ g(f(2)) = g(3) = 2 & & f(g(2)) = f(4) = 1 \\ g(f(3)) = g(4) = 1 & & f(g(3)) = f(2) = 3 \\ g(f(4)) = g(1) = 3 & & f(g(4)) = f(1) = 2 \end{array}$$

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

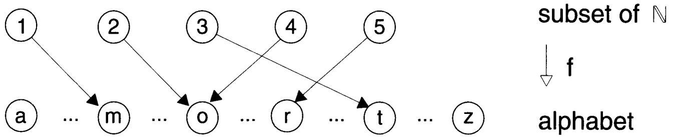
The composition of f and f^{-1} is the identity mapping :

$$f^{-1} \circ f = f \circ f^{-1} = \{(1,1), (2,2), (3,3), (4,4)\}$$

Sequence : A mapping of elements from the set $\mathbb{N}$ of natural numbers to a set M is called a sequence of elements from M. If an element of M is assigned to every natural number, the sequence is said to be infinite. If a segment $\{1, 2, ..., n\}$ of $\mathbb{N}$ is mapped to M, the sequence is said to be finite. If a sequence maps the natural number n to the element a_n, the sequence is designated by $<a_n>$. The members of a sequence are distinguished by their indices.

Example 3 : Sequence of characters

The character string ”motor” may be viewed as a sequence of characters, namely a mapping from the natural numbers {1, 2, 3, 4, 5} to the set of characters of the lowercase alphabet.



Canonical mapping : The surjection from a set M to its quotient set M / E for a given equivalence relation E is called a canonical mapping of M. The image of the element $a \in M$ is the equivalence class [a].

$$k : M \rightarrow M/E \qquad \text{with} \qquad k(a) = [a]$$

Example 4 : Canonical mapping of marbles

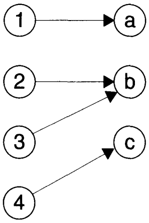
Let a set $M = \{a, b, c, d\}$ of marbles be given. Let the marbles a and d be white, and let the marbles b and c be black. The equivalence relation ”of the same color” partitions the set M into the color classes {a, d} and {b, c}. The colors a and c are chosen as representatives of the classes. The canonical mapping $k : M \rightarrow M/\text{color}$ maps the marbles to the color classes as follows :

$$k(a) = [a] \qquad k(b) = [c]$$
$$k(d) = [a] \qquad k(c) = [c]$$

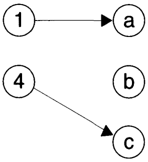
Restriction of a mapping : A mapping $g : S \rightarrow B$ is called a restriction of the mapping $f : A \rightarrow B$ if S is a subset of A and the images f(s) and g(s) of every element s of S coincide. The restriction of f to the set S is designated by f I S (f restricted to S).

$$f \mid S = \{s \in A \mid s \in S \quad \wedge \quad f(s) = g(s)\}$$

Example 5 : Restriction of a surjective mapping $f: A \to B$ to $S = \{1, 4\}$



mapping $f: A \to B$



mapping $f \mid \{1, 4\}: S \to B$

Continuation of a mapping : A mapping $h: M \to B$ is called a continuation of the mapping $f: A \to B$ if f is the restriction of h to the set A.

$$h: M \to B \text{ is a continuation of } f: A \to B \quad :\Leftrightarrow \quad f = h \mid A$$

Projection of a product set : Let the index set of the cartesian product $M_1 \times ... \times M_n$ be $I = \{1, ..., n\}$. Let the index set $K = \{i_1, ..., i_m\}$ be a subset of I, so that $i_k \in I$ and $m \le n$. The surjection p_k from the product $M_1 \times ... \times M_n$ to the product $M_{i_1} \times ... \times M_{i_m}$ is called the projection of the product set $M_1 \times ... \times M_n$ with respect to the index set K.

$$p_k: \quad M_1 \times M_2 \times ... \times M_n \quad \to \quad M_{i_1} \times M_{i_2} \times ... \times M_{i_m}$$

Function : The concept of a function is defined differently in mathematics and computer science. A mapping $f: A \to Z$ is called a mathematical function if the elements of the sets A and Z are numbers. The set A is called the domain of the function f. The set Z is called the target (codomain) of the function f. The mapping rule $f(a)$ is called the function term. The equation $z = f(a)$ is called the functional equation. The image of a given element $x \in A$ is called the function value $f(x)$ for x. The function value is thus a valuation of the function term.

Example 6 : Real function of one real variable

If the sets A and Z are subsets of the set $\mathbb{R}$ of real numbers, then $f: A \to Z$ is called a real function of one real variable.

Functions of several variables : A mapping $f: A \to Z$ is called a mathematical function of n variables if $A \subseteq \mathbb{R}^n$ and $Z \subseteq \mathbb{R}^m$ are product sets. For $m = 1$, the function f is called a scalar function. For $m > 1$, the function is called a vector function.

2.7 CARDINALITY AND COUNTABILITY

Introduction : The number of elements in a finite set is described by a natural number. This concept cannot be extended directly to infinite sets (for instance to the set of natural numbers). Infinite sets are collections which are never completed by successively adding their elements, and which can therefore never be completely enumerated. The concept which corresponds to the number of elements of a finite set for general (finite or infinite) sets is the cardinal number (cardinality) of a set. Different kinds of infinite sets may possess different cardinal numbers. Cardinal numbers and operations on cardinal numbers are treated in the following.

Equipotency : Two sets A and B are said to be equipotent (equinumerous, equipollent) if there is a bijective mapping $f : A \to B$. Equipotency is designated by $A \sim B$ (A is equipotent with B). Due to the properties of bijective mappings, the equipotency relation $\sim$ is an equivalence relation :

$\sim$ is reflexive : The mapping $f : A \to A$ with $f(x) = x$ is bijective
$A \sim A$

$\sim$ is symmetric : $f : A \to B$ is bijective $\Rightarrow$ $f^{-1} : B \to A$ is bijective
$A \sim B \Rightarrow B \sim A$

$\sim$ is transitive : $f : A \to B$ is bijective $\wedge$ $g : B \to C$ is bijective $\Rightarrow$
$h : A \to C$ with $h = g \circ f$ is bijective
$A \sim B \wedge B \sim C \Rightarrow A \sim C$

Example 1 : Equipotent sets

The finite sets $A = \{1, 3, 6, 7, 8\}$ and $B = \{a, b, d, i, m\}$ are equipotent, since there is a bijective mapping $f : A \to B$. The infinite sets $\mathbb{N} = \{0, 1, 2, 3, ...\}$ and $Z = \{0, 4, 8, 12, ...\}$ are also equipotent, since there is a bijective mapping $g : \mathbb{N} \to Z$ with $f(x) = 4x$. The sets A and $\mathbb{N}$ have different cardinality, since A is finite and hence there is no bijective mapping from A to $\mathbb{N}$.

Cardinal numbers : In a given system $M = \{A, B, ...\}$ of sets, the quotient set $M/\sim$ with respect to the equivalence relation $\sim$ (equipotent) is formed. An element of the quotient set $M/\sim$ is called a cardinal number (cardinality) of the given system of sets. The canonical mapping card : $M \to M/\sim$ assigns a cardinality card (A) to each set $A \in M$. The cardinality of A is alternatively designated by [A] or by $|A|$.

Finite and infinite sets : Using the concept of cardinality, the concepts of a finite set and an infinite set are defined without reference to the natural numbers. A set M is said to be infinite if there is a proper subset $A \subset M$ which is equipotent with M, that is $A \sim M$ or card (A) = card (M). Otherwise the set M is said to be finite. The cardinal number of a finite set is said to be natural, the cardinal number of an infinite set is said to be transfinite.

Example 2 : Finite and infinite sets

The set $\mathbb{N}$ in Example 1 is infinite, as it contains the equipotent proper subset Z. By contrast, the sets $\emptyset$, $\{a\}$ and $\{3, 6, 7\}$ are finite.

Operations on cardinal numbers : Operations in the set of cardinal numbers of a given system of sets are defined as follows (without proof of compatibility) :

(1) The sum of the cardinal numbers of disjoint sets A and B is the cardinal number of the union of A and B :

$$A \cap B = \emptyset \quad \Rightarrow \quad \text{card}\,(A) + \text{card}\,(B) = \text{card}\,(A \cup B)$$

(2) The product of the cardinal numbers of the sets A and B is the cardinal number of the cartesian product of A and B :

$$\text{card}\,(A) \cdot \text{card}\,(B) = \text{card}\,(A \times B)$$

(3) The power card (B) of a cardinal number card (A) is the cardinal number of the set of all mappings from B to A :

$$\text{card}\,(A)^{\text{card}\,(B)} = \text{card}\,(A^B)$$

$$A^B := \{f \mid f : B \to A\}$$

Countable set : A set M is said to be countable if there is an injection $f : M \to \mathbb{N}$ from the set M to the set $\mathbb{N} = \{0, 1, 2, \ldots\}$ of natural numbers. Otherwise the set is said to be uncountable. A countable set may be finite or infinite.

$$M \text{ is countable} \quad :\Leftrightarrow \quad \bigvee_f \ (f : M \to \mathbb{N} \text{ is injective})$$

Properties of countable sets

(A1) The cartesian product $\mathbb{N} \times \mathbb{N}$ of the set $\mathbb{N} = \{0, 1, 2, \ldots\}$ of natural numbers is countable.

(A2) For every injection $f : A \to B$ with $A \neq \emptyset$ there is a surjection $g : B \to A$ such that $g \circ f = 1_A$.

(A3) If the set A is countable and the mapping $f : A \to B$ is surjective, then the set B is countable.

(A4) Every subset of a countable set is countable.

(A5) If the sets A and B are countable, then their cartesian product $A \times B$ is countable.

(A6) A countable union of countable sets is countable.

Proof : Properties of countable sets

(A1) The product $\mathbb{N} \times \mathbb{N}$ is countable if the mapping $f : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ defined by $f(a, m) = (2a + 1)2^m - 1$ is an injection. By definition, the mapping f contains exactly one element $k \in \mathbb{N}$ for every element $(a, m) \in \mathbb{N} \times \mathbb{N}$. For every $k \in \mathbb{N}$ there is an element $(a, m) \in \mathbb{N} \times \mathbb{N}$:

(a) If k is even, then $m = 0$ and $2a = k$.

(b) If k is odd, then $k + 1 = (2a + 1)2^m$ is even. Divide $k + 1$ by 2 until an odd number v results. This determines $m > 0$ and $2a = v - 1$.

The numbers $k_1 = (2a + 1)2^m - 1$ and $k_2 = (2b + 1)2^n - 1$ are equal if and only if $a = b \ \wedge \ m = n$:

$$k_1 = k_2 \quad \Rightarrow \quad (2a + 1)2^m = (2b + 1)2^n \quad \Rightarrow \quad a = b \ \wedge \ m = n$$

Since every element $k \in \mathbb{N}$ has a unique preimage $(a, m) \in \mathbb{N} \times \mathbb{N}$, the mapping f is injective. It is even bijective !

(A2) Let the injection $f : A \to B$ be given. To construct the surjection $g : B \to A$, an arbitrary element $a \in A$ is chosen. Then g(y) for an arbitrary element $y \in B$ is defined as follows :

$$\begin{aligned} y \in f(A) &: \quad g(y) = f^{-1}(y) \\ y \notin f(A) &: \quad g(y) = a \end{aligned}$$

For an arbitrary element $x \in A$, this yields :

$$\begin{aligned} y := f(x) \quad \Rightarrow \quad g \circ f(x) &= g(y) \qquad \text{with} \quad y \in f(A) \\ &= f^{-1}(y) \\ &= x \end{aligned}$$

(A3) Let the surjective mapping $f : A \to B$ of a countable set A be given. Since A is countable, there exists an injection $g : A \to \mathbb{N}$. For every element $y \in B$, let $S_y := \{x \in A \mid x \in f^{-1}(y)\}$. The least element of the image $g(S_y)$ is determined and designated by m_y. Then the mapping $h : B \to \mathbb{N}$ with $h(y) = m_y$ is an injection. Hence the set B is countable.

(A4) By definition, there is an injection $g : A \to \mathbb{N}$ for every countable set A. For a subset $B \subseteq A$, this induces a restricted injection $g_B : B \to \mathbb{N}$. The set B is countable by virtue of the injection g_B.

(A5) For the countable sets A and B there are injections $f : A \to \mathbb{N}$ with $f(a) = n_1$ and $g : B \to \mathbb{N}$ with $f(b) = n_2$. Hence there is an injection $h : A \times B \to \mathbb{N} \times \mathbb{N}$ with $h(a,b) = (n_1,n_2)$. By property (A1), there is an injection $i : \mathbb{N} \times \mathbb{N} \to \mathbb{N}$. The composition $i \circ h : A \times B \to \mathbb{N}$ is an injection. Hence $A \times B$ is countable.

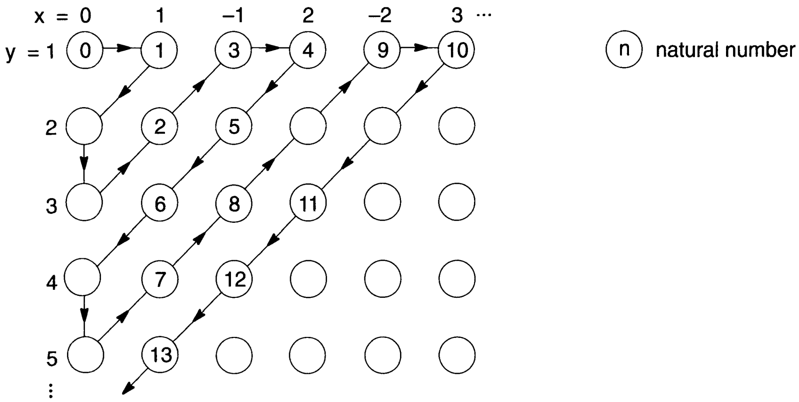
(A6) Let the set $A \neq \emptyset$ be countable. Let a countable set X_a be defined for every $a \in A$. Since A and every X_a are countable, there are surjections $f_a : \mathbb{N} \rightarrow X_a$ and $g : \mathbb{N} \rightarrow A$. The mapping h is defined as follows :

$$h : \mathbb{N} \times \mathbb{N} \rightarrow \bigcup \{X_a \mid a \in A\} \quad \text{with} \quad h(n,k) = f_{g(n)}(k)$$

Thus h is surjective. Since $\mathbb{N} \times \mathbb{N}$ is countable, it follows from (A3) that the union $\bigcup X_a$ is countable.

Example 3 : Countability of the rational numbers

The set $\mathbb{Q}$ of rational numbers is countably infinite, since a bijective mapping $f : \mathbb{N} \rightarrow \mathbb{Q}$ may be obtained using the following scheme. The vertices represent all fractions $\frac{x}{y}$. The natural numbers {0, 1, 2,...} are assigned as illustrated to the vertices which correspond to the normal form of a rational number. The mapping $f : \mathbb{N} \rightarrow \mathbb{Q}$ with $f(n) = \frac{x}{y}$ is bijective. Hence $\mathbb{Q}$ and $\mathbb{N}$ are equipotent.



Example 4 : Uncountability of the open unit interval

The uncountability of the open unit interval $J = \{x \in \mathbb{R} \mid 0 < x < 1\}$ is proved indirectly by proceeding from the assumption that J is countable. In this case there is a bijective mapping $f : \mathbb{N} \rightarrow J$ from the natural numbers n to the infinite decimal fractions x_n.

$$\begin{array}{lllllll} f(1) & = & x_1 & = & 0.z_{11} \; z_{12} \; z_{13} \cdots & & z_{im} \in \{0,1,2,3,4,5,6,7,8,9\} \\ f(2) & = & x_2 & = & 0.z_{21} \; z_{22} \; z_{23} \cdots & & \\ f(3) & = & x_3 & = & 0.z_{31} \; z_{32} \; z_{33} \cdots & & \\ \vdots & & & & & & \end{array}$$

A number $y = 0.b_1 b_2 b_3 \ldots$ in the interval J is formed such that every digit b_i differs from the digit z_{ii} of the number x_i. A possible choice is $b_i := 1$ for $z_{ii} = 0$ and $b_i := 0$ for $z_{ii} \neq 0$. Then y is not contained in the enumeration. This is a contradiction, since y lies in the interval J. Hence the open unit interval is uncountable. The method of proof used here is called the diagonal method.

Example 5 : Uncountability of the real numbers

The set $\mathbb{R}$ of real numbers is uncountable, since there is a bijective mapping from the open unit interval J to the real numbers. The sets J and $\mathbb{R}$ are therefore equipotent. In Example 4, the set J is shown to be uncountable.

$$f : J \rightarrow \mathbb{R} \quad \text{with} \quad f(x) = \frac{x - 0.5}{x(x - 1)}$$

2.8 STRUCTURES

Introduction : Mathematics comprises numerous fields, whose boundaries have historical origins. The sets and relations of different fields have common properties. If mathematics is classified according to these properties, basic algebraic, ordinal and topological structures become apparent. From these basic structures and additional axioms, the mixed structures and the fields of mathematics are derived. These connections are highly relevant to engineering, since the computer provides only basic structures for the solution of engineering problems.

Unstructured set : A set is said to be unstructured if there are no relations defined in it. All sets are initially unstructured when they are defined. In particular, the order in which elements are enumerated in the definition of the set is irrelevant.

Domain : The ordered pair $(M ; R)$ of a set $M = \{a_1, \ldots, a_m\}$ of elements and a set $R = \{R_1, \ldots, R_n\}$ of relations is called a domain (structured set). The relations in R provide the set M with structure. The compatibility of the relations in R is ensured by rules for the domain. The rules of compatibility determine the theory of the domain.

Basic structures : A family of domains with similar relations is called a basic structure. The following basic structures have emerged in mathematics :

algebraic structure : the relations describe relationships between elements of the given sets

ordinal structure : the relations describe relationships between elements and subsets of the given sets

topological structure : the relations describe relationships between subsets of the given sets

Mixed structures : A domain equipped with relations from more than one basic structure possesses a mixed structure (multiple structure). The compatibility of the relations of a mixed structure is ensured by rules for the domain.

Derived structures : New domains are derived from a domain $(M ; R)$ by equipping the sets $S \subseteq M$, M^n and M/E derived from M with similar relations.

substructure : certain relations in R are restricted to subsets $S \subseteq M$.

product structure : a cartesian product $M_1 \times \ldots \times M_n$ of similarly structured sets is equipped with the common structure R.

quotient structure : the quotient set M / E for an equivalence relation E is equipped with the structure of M.

Example 1 : Basic algebraic structure of the natural numbers

The domain $(\mathbb{N} ; + , \cdot)$ consists of the set $\mathbb{N}$ of natural numbers, together with the relations addition (symbol $+$) and multiplication (symbol $\cdot$). The relations $+$ and $\cdot$ are made compatible by the distributive law.

$$a \cdot (b + c) = (a \cdot b) + (a \cdot c)$$

Example 2 : Mixed structure of the rational numbers

The domain $(\mathbb{Q} ; + , \cdot, \leq, d)$ is a mixed structure on the set $\mathbb{Q}$ of rational numbers. The domain $(\mathbb{Q} ; + , \cdot)$ is the basic algebraic structure of the rational numbers. The domain $(\mathbb{Q} ; \leq)$ is the basic ordinal structure of the rational numbers. The domain $(\mathbb{Q} ; d)$ with the distance metric d is the basic topological structure of the rational numbers.

Example 3 : Addition of vectors

The domain $(\mathbb{Z} ; +)$ equips the set $\mathbb{Z}$ of integers with the relation of addition. An n-tuple $(z_1, z_2, ..., z_n)$ in the product $\mathbb{Z}^n$ is called a vector. The relation of addition (symbol $+$) is defined for a vector by applying the relation $+$ of the domain $(\mathbb{Z} ; +)$ to each component of $\mathbb{Z}^n$. The domain $(\mathbb{Z}^n ; +)$ is thus derived from the domain $(\mathbb{Z} ; +)$.

$$+ := \{ (a, b, c) \mid a_i = b_i + c_i \quad \wedge \quad i \in \{1, 2, ..., n\}\}$$

Structurally compatible mapping : For a mapping $f : A \rightarrow Z$, the sets A and Z may be components of domains, for instance $(A ; +)$ and $(Z ; +)$. The mapping f is said to be compatible with the structure of the domains if the image $f(a_1 + a_2)$ of the sum of two elements a_1 and a_2 from A is equal to the sum of their images $f(a_1)$ and $f(a_2)$ in the target Z. A structurally compatible mapping preserves essential properties of the domain A in the target Z. The target Z may therefore be used to study the structure of the domain A. This bears a particular advantage if the number of elements in the image f (A) is significantly less than the number of elements in A.

$$f(a_1 + a_2) = f(a_1) + f(a_2)$$

Morphism : A mapping $f : A \rightarrow Z$ from the set of a domain $(A ; R_a)$ to the set of a domain $(Z ; R_z)$ is called a morphism if the rule f (a) of the mapping is structurally compatible. If the ordered pair (a, b) of the elements $a, b \in A$ is contained in the relation R_a, then for a structurally compatible mapping the ordered pair (f(a), f(b)) of the images $f(a), f(b) \in Z$ is contained in the relation R_z.

Isomorphism : The concept of isomorphism allows the structures of different domains to be compared. The domains $(A\ ;\ R_a)$ and $(Z\ ;\ R_z)$ are said to be isomorphic if there is a morphism $f : A \rightarrow Z$ whose inverse $f^{-1} : Z \rightarrow A$ is also a morphism. Since the mapping f is bijective, the correspondence between the elements of A and Z is one-to-one.

Automorphism : An isomorphism $f : A \rightarrow A$ is called an automorphism, since the domain $(A\ ;\ R_a)$ is mapped to itself.

Example 4 : Isotonic mapping as an example of a morphism

Let two sets M_1 and M_2 be given, and let both sets be equipped with a relation $<$. A mapping $f : M_1 \rightarrow M_2$ is said to be isotonic if $x_1 < x_2$ implies $f(x_1) < f(x_2)$. The isotonic mapping from the domain $(M_1\ ;\ <)$ to the domain $(M_2\ ;\ <)$ is a morphism which preserves the property ”ordered set”.

3 ALGEBRAIC STRUCTURES

3.1 INTRODUCTION

A set is equipped with an algebraic structure by defining operations on elements of the set. The operands may also belong to different sets. The type of the sets considered determines the branch of algebra, for example :

- boolean algebra for truth values
- algebra of numbers for sets of numbers
- algebra of sets for subsets
- vector algebra for vector spaces

Computers can perform the basic algebraic operations for truth values and different types of numbers directly. Algebraic structures are therefore very important in engineering applications of computers. The properties of other structures which cannot be handled directly by a computer are studied on the computer using properties of related algebraic structures.

This chapter provides an overview of algebraic structures. The definitions of inner and outer operations and the properties of these operations are of fundamental importance. Semigroups and groups are domains typical for sets with one operation; semirings, rings, fields and lattices are domains typical for sets with two operations. Outer operations lead to vector spaces and matrix algebra. Important subjects such as group theory and graph theory are treated in separate chapters.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

3.2 INNER OPERATIONS

Inner operation : The elementary algebraic structure of a set is provided by the inner operation on the elements of the set. To every ordered pair (a, b) of the direct product $M \times M$, an inner operation assigns exactly one element c of the set M. An inner operation in a set M is therefore a mapping f from $M \times M$ to M. This mapping is by definition left-total and right-unique.

$$f : M \times M \rightarrow M \quad \text{with} \quad f(a, b) = c \quad \text{and} \quad a, b, c \in M$$

An operator symbol is often used instead of a letter to designate an inner operation. In this case, the mapping is represented as follows :

$$\circ : M \times M \rightarrow M \quad \text{with} \quad a \circ b = c \quad \text{and} \quad a, b, c \in M$$

Associative operation : An operation $\circ$ in a set M is said to be associative if the result of two successive operations does not depend on the order in which the operations are performed.

$$(a \circ b) \circ c = a \circ (b \circ c) \qquad a, b, c \in M$$

Commutative operation : An operation $\circ$ in a set M is said to be commutative if the order of the elements a and b in the operation $a \circ b$ does not influence the result of the operation.

$$a \circ b = b \circ a \qquad a, b \in M$$

Identity element : An element e of the set M is called the identity element of the domain $(M ; \circ)$ if every element a of M remains invariant when operated on with e. There is at most one identity element for an inner operation. For if e_1 and e_2 are two identity elements, then $e_1 \circ e_2 = e_1 = e_2$ shows that they are equal.

$$e \text{ is an identity element} \quad :\Leftrightarrow \quad \bigwedge_{a \in M} (a \circ e = e \circ a = a)$$

Inverse : An element a^{-1} of a set M is called the inverse of the element a in the domain $(M ; \circ)$ if the inner operation yields the identity element e of $(M ; \circ)$ when applied to a and a^{-1}. If a certain element a possesses an inverse a^{-1}, then this is unique in M. In fact, if x and y are two inverses of a, then it follows from $x = e \circ x = y \circ a \circ x = y \circ e = y$ that they are equal.

$$a^{-1} \text{ is the inverse of a} \quad :\Leftrightarrow \quad a \circ a^{-1} = a^{-1} \circ a = e$$

Rules of calculation for invertible elements : In a domain $(M ; \circ)$ with the identity element e, the following rules of calculation hold for the invertible elements of $(M ; \circ)$:

(1) The identity element e is invertible.
$e^{-1} = e$

(2) If the element $a \in M$ is invertible, then its inverse a^{-1} is also invertible.
$(a^{-1})^{-1} = a$

(3) If the elements $a, b \in M$ are invertible, then $a \circ b$ is also invertible.
$(a \circ b)^{-1} = b^{-1} \circ a^{-1}$

Powers of an element : Let $(M ; \circ)$ be a domain with an associative operation and an identity element e. If n is an integer and a is an invertible element, then the n-fold product of a with itself is called the n-th power of a.

$$\begin{array}{lll} n > 0 : & a^n := a \circ a \circ \ldots \circ a & n \text{ times} \\ n = 0 : & a^n := e & \\ n < 0 : & a^n := a^{-1} \circ a^{-1} \circ \ldots \circ a^{-1} & |n| \text{ times} \end{array}$$

Rules of calculation for powers : Let $(M ; \circ)$ be a domain with an associative and commutative operation and an identity element e. Let the elements a and b of M be invertible. Then the following rules of calculation hold for the integers m and n :

(1) $a^m \circ a^n = a^{m+n}$ (2) $(a^m)^n = a^{mn}$

(3) $a^m \circ b^n = b^n \circ a^m$ (4) $(a \circ b)^n = a^n \circ b^n$

Idempotency of an element : Let $(M ; \circ)$ be a domain with an associative operation. An element $a \in M$ is said to be idempotent if operating on a with itself again yields a. This definition implies that every power a^n of an idempotent element a with $n > 0$ is itself idempotent.

$$a \text{ is idempotent} \quad :\Leftrightarrow \quad a^2 = a \circ a = a$$

Nilpotency of an element : Let $(M ; \circ)$ be a domain with an associative operation and an identity element e. An element a is said to be nilpotent if there is a positive integer m such that $a^m = e$. If an element is nilpotent, the least positive integer n with $a^n = e$ is called the degree of nilpotency of a.

$$a \text{ is nilpotent of degree } n \quad :\Leftrightarrow \quad (a^n = e \mid \bigwedge_{k=1}^{n-1} a^k \neq e)$$

An element a is said to be self-inverse if it is its own inverse, that is if $a = a^{-1}$. A self-inverse element is nilpotent of degree $n = 2$.

$$a \text{ is self-inverse} \quad :\Leftrightarrow \quad a^2 = a \circ a = e$$

Notation as a product or sum : Inner operations are represented either as a product using the symbol $\circ$ or as a sum using the symbol +. In the product notation, the identity element e is designated by 1, in the sum notation it is designated by 0. The inverse of the element a is designated by a^{-1} in the product notation and by $-a$ in the sum notation. The n-fold application of the operation to an element is designated by a^n in the product notation and by na in the sum notation. The meaning of the expressions is independent of notation.

rules of calculation (inverse)	$1^{-1} = 1$	$-0 = 0$
	$(a^{-1})^{-1} = a$	$-(-a) = a$
	$(a \circ b)^{-1} = b^{-1} \circ a^{-1}$	$-(a+b) = (-b) + (-a)$
rules of calculation (n-fold)	$a^m \circ a^n = a^{m+n}$	$ma + na = (m+n)a$
	$(a^m)^n = a^{mn}$	$m(na) = (mn)a$
	$a^m \circ b^n = b^n \circ a^m$	$ma + nb = nb + ma$
	$(a \circ b)^n = a^n \circ b^n$	$n(a+b) = na + nb$
idempotent	$a \circ a = a$	$a + a = a$
self-inverse	$a \circ a = 1$	$a + a = 0$

3.3 SETS WITH ONE OPERATION

Introduction : The simplest algebraic structures consist of a set and one inner operation. An algebraic structure is defined by stipulating certain properties for the operation by definition. In the following these properties are called the defining properties of the structure. The rules for the structure are derived from the defining properties. The defining properties and the rules form the theory of the algebraic structure.

This section is an introduction to the theory of semigroups and groups. The sum notation with the symbol $+$ for the inner operation is used. The product notation with the symbol $\circ$ is also used in the literature. As explained in Section 3.2, the two notations are equivalent. The algebraic rules for semigroups and groups are formulated and deduced from the defining properties. Their applicability is demonstrated for numbers, sets, relations and geometric shapes.

Semigroup : A domain $(M\,;+)$ with the inner operation $+$ in the set M is called a semigroup if :

(1) The operation $+$ is associative for all elements of M.

A semigroup $(M\,;+)$ is said to be commutative if the operation $+$ is commutative for all elements of M. A semigroup $(M\,;+)$ is said to be idempotent if all elements of M are idempotent. A commutative and idempotent semigroup is also called a semilattice.

Semigroup with identity element : A domain $(M\,;+)$ with the inner operation $+$ in the set M is called a semigroup with identity element (monoid) if in addition to (1) :

(2) The set M contains an identity element for the operation $+$.

Group : A domain $(M\,;+)$ with the inner operation $+$ in the set M is called a group if in addition to (1) and (2) :

(3) The set M contains an inverse $-a$ for every element a of M.

A group $(M\,;+)$ is said to be commutative (abelian) if the operation $+$ is commutative for all elements of M. A group $(M\,;+)$ is said to be self-inverse if all elements of M are self-inverse.

Rules for idempotent semigroups : Idempotent semigroups with an identity element are particularly important in applications. The defining properties imply the following algebraic rules :

(1) If $(M ; +)$ is an idempotent semigroup with the identity element n, then there is no inverse for any $a \neq n$.

(2) If $(M ; +)$ is an idempotent semigroup with the identity element n, then $a + b = n$ implies $a = b = n$.

$$a + b = n \quad \Rightarrow \quad a = n \quad \wedge \quad b = n$$

(3) If $(M ; +)$ is an idempotent semigroup with more than one element, then $(M ; +)$ cannot be a group.

Proof : Rules for idempotent semigroups

(1) By definition, a semigroup $(M ; +)$ is idempotent if $a + a = a$ for all elements $a \in M$. If a possesses an inverse $(-a)$, then $a = a + n$ and $n = (a + (-a))$ together with the idempotency $a + a = a$ and the associative law imply that $a = a + n = a + (a + (-a)) = (a + a) + (-a) = a + (-a) = n$. Hence the identity element is the only element which has an inverse.

(2) Since $a + b = n$, it follows that a is the inverse of b and b is the inverse of a. However, by (1) there is no inverse for $a \neq n$. Hence $a + b = n$ holds only for $a = n$ and $b = n$.

(3) By definition, in a group $(M ; +)$ every element $a \in M$ has an inverse $(-a)$. However, by (1) an idempotent semigroup $(M ; +)$ contains no inverse for any element $a \neq n$. Hence an idempotent semigroup with more than one element cannot be a group.

Rules for groups : The defining properties for groups imply the following algebraic rules :

(1) If $(M ; +)$ is a group, the following cancellation laws hold :

$$a + b = a + c \quad \Rightarrow \quad b = c$$
$$b + a = c + a \quad \Rightarrow \quad b = c$$

(2) If $(M ; +)$ is a group, every equation has a unique solution :

$$a + x = b \qquad x = (-a) + b$$
$$x + a = b \qquad x = b + (-a)$$

(3) If $(M ; +)$ is a commutative group, the equations $a + x = b$ and $x + a = b$ have the same solution.

(4) If $(M ; +)$ is a self-inverse group, then it is commutative.

Proof : Rules for groups

(1) Operating on the equation $a+b=a+c$ with $(-a)$ from the left yields $(-a)+(a+b)=((-a)+a)+b=b$ for the left-hand side of the equation and $(-a)+(a+c)=((-a)+a)+c=c$ for the right-hand side, so that $b=c$. This establishes the validity of the first cancellation law. The validity of the second cancellation law is proved analogously.

(2) Operating on the equation $a+x=b$ with $(-a)$ from the left yields $(-a)+(a+x)=((-a)+a)+x=x$ for the left-hand side and $(-a)+b$ for the right-hand side, so that $x=(-a)+b$. The proof that the equation $x+a=b$ has the unique solution $x=b+(-a)$ is carried out analogously.

(3) If the group $(M ; +)$ is commutative, the equations $a+x=b$ and $x+a=b$ are equivalent, since $a+x=x+a$. As the solution is unique, they have the same solution.

(4) If the group $(M ; +)$ is self-inverse, then by definition $a+a=n$ for all elements $a \in M$. Operating on the equation $(a+b)+(a+b)=n$ with a from the left and with b from the right yields $a+(a+b)+(a+b)+b = (a+a)+b+a+(b+b) = n+b+a+n = b+a$ for the left-hand side and $a+n+b=a+b$ for the right-hand side, so that the commutative law $b+a=a+b$ holds. Hence a self-inverse group $(M ; +)$ is commutative.

Example 1 : Domains of numbers

(1) The domain $(\mathbb{N}' ; +)$ for the addition of positive natural numbers is a commutative semigroup.

(2) The domain $(\mathbb{N} ; +)$ for the addition of natural numbers including 0 is a commutative semigroup with the identity element 0.

(3) The domain $(\mathbb{Z} ; +)$ for the addition of integers is a commutative group with the identity element 0,

(4) The domain $(\mathbb{Q} ; \circ)$ for the multiplication of rational numbers is a commutative semigroup with the identity element 1. The element 0 is the only element of $\mathbb{Q}$ without an inverse. The domain $(\mathbb{Q}' ; \circ)$ for the multiplication of non-zero rational numbers is a commutative group.

(5) The domain $(\mathbb{N} ; \min)$ for the minimum $\min\{a, b\}$ of natural numbers is an idempotent and commutative semigroup, and therefore a semilattice.

- The operation min is associative :

  $\min\{a, \min\{b, c\}\} = \min\{\min\{a, b\}, c\}$

- The operation min is commutative :

  $\min\{a, b\} = \min\{b, a\}$

- The operation min is idempotent :

  $\min\{a, a\} = a$

Example 2 : Domain of sets

Let a reference set M be given. Every subset $A \subseteq M$ is an element of the power set P(M). The inner operations union $\cup$, intersection $\cap$ and symmetric difference $\oplus$ are defined in the power set P(M) (see Section 2.2.2). The properties of the domains $(P(M); \cup)$, $(P(M); \cap)$ and $(P(M); \oplus)$ are compiled in the following table for elements $A, B, C \in P(M)$.

Property	$(P(M) ; \cup)$	$(P(M) ; \cap)$	$(P(M) ; \oplus)$
associative	$A \cup (B \cup C) =$ $(A \cup B) \cup C$	$A \cap (B \cap C) =$ $(A \cap B) \cap C$	$A \oplus (B \oplus C) =$ $(A \oplus B) \oplus C$
commutative	$A \cup B = B \cup A$	$A \cap B = B \cap C$	$A \oplus B = B \oplus A$
identity element	$A \cup \emptyset = A$	$A \cap M = A$	$A \oplus \emptyset = A$
idempotent	$A \cup A = A$	$A \cap A = A$	
self-inverse			$A \oplus A = \emptyset$

The algebraic structure of the various domains may be read off directly :

(1) The domain $(P(M); \cup)$ is a commutative and idempotent semigroup with the empty set $\emptyset$ acting as an identity element, and hence a semilattice with the identity element $\emptyset$.

(2) The domain $(P(M); \cap)$ is a commutative and idempotent semigroup with the reference set M acting as an identity element, and hence a semilattice with the identity element M.

(3) The domain $(P(M); \oplus)$ is a commutative group with the empty set $\emptyset$ acting as an identity element. It has the special property that every element $A \in P(M)$ is self-inverse. The domain $(P(M); \oplus)$ is therefore a self-inverse group with the identity element $\emptyset$.

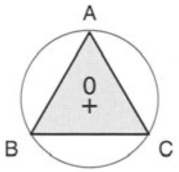
Example 3 : Domain of relations

Let a reference set M be given. Every relation A is a set of ordered pairs (a, b) with $a, b \in M$. It is a subset of the direct product $M \times M$, and hence an element of the power set $P(M \times M)$. In addition to the union $\cup$ and the intersection $\cap$, the composition $\circ$ is defined as an inner operation in the power set $P(M \times M)$ (see Section 2.4). As in Example 2, the domains $(P(M \times M) ; \cup)$ and $(P(M \times M) ; \cap)$ are semilattices. The domain $(P(M \times M) ; \circ)$ is a semigroup with the identity relation I as an identity element, since the following properties hold for elements $A, B, C \in P(M \times M)$:

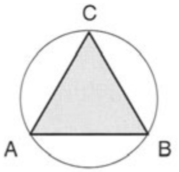
Property	$(P(M \times M) ; \circ)$
associative	$A \circ (B \circ C) = (A \circ B) \circ C$
identity element	$A \circ I = A = I \circ A$

Example 4 : Covering rotations of an equilateral triangle

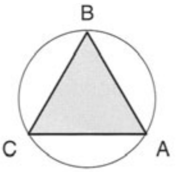
An equilateral triangle ABC covers itself when rotated through 0 degrees, 120 degrees or 240 degrees about its center 0. The rotation a_0 through 0 degrees leaves the triangle in its original position. The rotation a_1 through 120 degrees takes A to B, B to C and C to A. The rotation a_2 through 240 degrees takes A to C, B to A and C to B. Rotations which differ by 360 degrees are considered to be identical.



rotation a_0 : 0 degrees



rotation a_1 : 120 degrees



rotation a_2 : 240 degrees

Let the set G contain the distinguishable covering rotations $\{a_0, a_1, a_2\}$. The composition $a_i \circ a_m$ (rotation a_i after rotation a_m) is chosen as an inner operation $\circ : G \times G \to G$. The result $a_n = a_i \circ a_m$ is the rotation a_n which leads to the same position of the triangle ABC as performing the rotation a_i and then the rotation a_m. The results of the operation are arranged in the following multiplication table for $a_i \circ a_m$.

$\circ$	a_0	a_1	a_2
a_0	a_0	a_1	a_2
a_1	a_1	a_2	a_0
a_2	a_2	a_0	a_1

The domain (G ; $\circ$) is a group. The identity element is a_0. The inverse elements are $a_0^{-1} = a_0$, $a_1^{-1} = a_2$ and $a_2^{-1} = a_1$. The associative law holds in (G ; $\circ$). The multiplication table is symmetric, so that $a_i \circ a_k = a_k \circ a_i$ holds. Hence the group is commutative.

Let the equation $a_1 \circ x = a_0$ be given. The solution $x = a_1^{-1} \circ a_0$ is determined using the multiplication table as follows :

$$x = a_1^{-1} \circ a_0 = a_2 \circ a_0 = a_2$$

3.4 SETS WITH TWO OPERATIONS

3.4.1 INTRODUCTION

An algebraic structure for a set with two inner operations may be decomposed into two algebraic substructures for the set with one inner operation each, together with the compatibility properties of the two operations. According to this principle, an algebraic structure with two operations is defined by defining the two algebraic substructures and the compatibility properties consistently. The rules for the algebraic structure are derived from these defining properties. The defining properties and the rules form the theory of the algebraic structure.

Algebraic structures for additive and multiplicative domains $(M ; +, \circ)$ are treated as a generalization of number theory. The subdomains $(M ; +)$ and $(M ; \circ)$ for the addition $+$ and the multiplication $\circ$ have different mathematical properties. The compatibility of the addition $+$ and the multiplication $\circ$ is ensured by the distributive law. Different properties of the subdomains $(M ; +)$ and $(M ; \circ)$ lead to different algebraic structures. Semirings, rings and fields are among the important structures.

Algebraic structures for dual domains $(M ; \sqcup, \sqcap)$ with the disjunction $\sqcup$ and the conjunction $\sqcap$ are treated as a generalization of the theory of truth values and of set theory. The subdomains $(M ; \sqcup)$ and $(M ; \sqcap)$ have the same mathematical properties, so that the operations $\sqcup$ and $\sqcap$ are interchangeable. This interchangeability is the basis of duality. Compatibility is ensured by the adjunctive, distributive and complementary laws in dual form. Lattices and boolean lattices are examples of domains with dual structure.

3.4.2 ADDITIVE AND MULTIPLICATIVE DOMAINS

Introduction : Domains $(M\,;+,\circ)$ with the inner operations $+$ (addition) and $\circ$ (multiplication) in the set M are treated as generalizations of the algebraic structure of numbers. The additive domain $(M\,;+)$ and the multiplicative domain $(M\,;\circ)$ possess different mathematical properties, so that the operations $+$ and $\circ$ are not interchangeable. The compatibility of the two operations is guaranteed by the distributive laws. It is assumed that the set M contains more than one element.

Rank of the operations : The expression $a+b\circ c$ in a domain $(M\,;+,\circ)$ is ambiguous; its value depends on the order in which the operations $+$ and $\circ$ are performed. The order of execution is therefore determined according to the rank of the operations ($\circ$ before $+$). A different order may be prescribed using parentheses, as in the expression $(a+b)\circ c$: In this case, the addition is performed before the multiplication.

Distributive laws : The distributive laws ensure the compatibility of the operations $+$ and $\circ$ in the domain $(M\,;+,\circ)$. The operation $\circ$ is said to be distributive with respect to the operation $+$ if for all $a, b, c \in M$:

$$\begin{aligned} a\circ(b+c) &= a\circ b + a\circ c \\ (a+b)\circ c &= a\circ c + b\circ c \end{aligned}$$

Identity elements : If there is an identity element for the addition $+$, it is called the zero element and designated by 0. If there is an identity element for the multiplication $\circ$, it is called the unit element and designated by 1.

$$\begin{array}{lllll} \text{zero element} & a+0 & = & 0+a & = & a \\ \text{unit element} & a\circ 1 & = & 1\circ a & = & a \end{array}$$

Zero sums : A domain $(M\,;+,\circ)$ with the zero element 0 is said to be without zero sums if $a+b=0$ implies $a=b=0$. This is equivalent to the property that there is no additive inverse in M for any element $a \neq 0$. If the addition $+$ is idempotent, then there is no inverse for any $a \neq 0$, so that the domain $(M\,;+,\circ)$ is without zero sums.

$$(M\,;+,\circ) \text{ is without zero sums } :\Leftrightarrow \bigwedge_a \bigwedge_b (a+b=0 \Rightarrow a=0 \wedge b=0)$$

Zero divisors : A domain $(M\,;+,\circ)$ with the zero element 0 is said to be without zero divisors if $a\circ b=0$ implies $a=0$ or $b=0$.

$$(M\,;+,\circ) \text{ is without zero divisors } :\Leftrightarrow \bigwedge_a \bigwedge_b (a\circ b=0 \Rightarrow a=0 \vee b=0)$$

Semiring : A domain $(M\,;+,\circ)$ with the inner operations $+$ and $\circ$ in the set M is called a semiring if :

(1) The domain $(M\,;+)$ is a commutative semigroup.

(2) The domain $(M\,;\circ)$ is a semigroup.

(3) The multiplication $\circ$ is distributive with respect to the addition $+$.

The semiring $(M\,;+,\circ)$ is called a semiring with zero element if there is a zero element 0 which acts as an identity element under the addition $+$. The semiring $(M\,;+,\circ)$ is called a semiring with unit element if there is a unit element 1 which acts as an identity element under the multiplication $\circ$. The semiring $(M\,;+,\circ)$ is said to be commutative if the multiplication $\circ$ is commutative. Semirings with zero element may be without zero sums or without zero divisors.

Ring : A domain $(M\,;+,\circ)$ with the inner operations $+$ and $\circ$ in the set M is called a ring if :

(1) The domain $(M\,;+)$ is a commutative group.

(2) The domain $(M\,;\circ)$ is a semigroup.

(3) The multiplication $\circ$ is distributive with respect to the addition $+$.

A ring $(M\,;+,\circ)$ differs from a semiring in that the additive domain $(M\,;+)$ in a semiring is a semigroup, while in a ring it is required to be a group. A ring contains a zero element 0 which acts as an identity element under addition.

A ring $(M\,;+,\circ)$ is called a ring with unit element if there is a unit element 1 which acts as an identity element under the multiplication $\circ$. A ring $(M\,;+,\circ)$ is said to be commutative if the multiplication $\circ$ is commutative. A ring may be without zero divisors. A commutative ring without zero divisors is called an integral ring (integral domain).

Rules for rings : The defining properties of a ring $(M\,;+,\circ)$ imply the rules for the additive group $(M\,;+)$, as well as additional rules for multiplication :

(1) If $(M\,;+,\circ)$ is a ring, then the zero element 0 is invariant with respect to multiplication.

$$a \circ 0 = 0 \circ a = 0$$

(2) If $(M\,;+,\circ)$ is a ring, then multiplication with additive inverses follows the rules known as sign rules in number theory.

$$a \circ (-b) = (-a) \circ b = -(a \circ b)$$

$$(-a) \circ (-b) = a \circ b$$

(3) If $(M ; +, \circ)$ is a ring without zero divisors, then the following cancellation laws hold for multiplication :

$$a \circ b = a \circ c \quad \Rightarrow \quad b = c \qquad \text{for} \quad a \neq 0$$

$$b \circ a = c \circ a \quad \Rightarrow \quad b = c \qquad \text{for} \quad a \neq 0$$

Proof : Rules for rings

(1) By the first distributive law, $a \circ b + a \circ c = a \circ (b + c)$. For $b = c = 0$, this yields $a \circ 0 + a \circ 0 = a \circ 0$. Adding $-(a \circ 0)$ on both sides yields $-(a \circ 0) + (a \circ 0 + a \circ 0) = (-(a \circ 0) + a \circ 0) + a \circ 0 = a \circ 0$ for the left-hand side and $-(a \circ 0) + a \circ 0 = 0$ for the right-hand side, so that the multiplicative invariance $a \circ 0 = 0$ of the zero element holds. The multiplicative invariance $0 \circ a = 0$ of the zero element is proved using the second distributive law.

(2) By the first distributive law, $a \circ b + a \circ c = a \circ (b + c)$. On substituting $c = (-b)$, it follows that $a \circ b + a \circ (-b) = a \circ (b + (-b)) = a \circ 0 = 0$. But $a \circ b + a \circ (-b) = 0$ implies that $a \circ (-b)$ is the additive inverse of $a \circ b$, and hence $a \circ (-b) = -(a \circ b)$. The rule $(-a) \circ b = -(a \circ b)$ is proved analogously using the second distributive law. For a and $-b$, the two rules yield $(-a) \circ (-b) = -((-a) \circ b) = -(-(a \circ b)) = a \circ b$.

(3) Applying the first distributive law to $0 = a \circ 0 = a \circ (c + (-c))$ yields $a \circ c + a \circ (-c) = 0$. Then $a \circ c = a \circ b$ implies $a \circ b + a \circ (-c) = a \circ (b + (-c)) = 0$. If $a \neq 0$, then in a ring without zero divisors it follows that $(b + (-c)) = 0$. Thus $-c$ is the additive inverse of b, and hence $b = c$. This yields the first cancellation law $a \circ b = a \circ c \Rightarrow b = c$ for $a \neq 0$. The second cancellation law is proved analogously.

Boolean ring : A ring $(M ; +, \circ)$ is said to be boolean if every element of M is idempotent with respect to multiplication. Idempotency with respect to multiplication leads to the following properties :

(1) The domain $(M ; +)$ is a self-inverse group.

(2) The domain $(M ; \circ)$ is a semilattice.

Proof : Properties of a boolean ring

By definition, the multiplicative semigroup $(M ; \circ)$ of the ring $(M ; +, \circ)$ is idempotent, so that $a^2 = a \circ a = a$ holds for all $a \in M$.

(1) The multiplicative idempotency of $(a + a)$ implies $(a + a) = (a + a)^2 = (a + a) \circ (a + a) = a^2 + a^2 + a^2 + a^2 = (a + a) + (a + a)$, so that $(a + a) = (a + a) + (a + a)$. Adding $-(a + a)$ to both sides yields $-(a + a) + (a + a) = 0$ on the left-hand side and $-(a + a) + ((a + a) + (a + a)) = (-(a + a) + (a + a)) + (a + a) = a + a$ on the right-hand side, and hence $a + a = 0$. Thus a is its own additive inverse. The group $(M ; +)$ is therefore self-inverse.

(2) The multiplicative idempotency for $(a+b)$ implies $(a+b)=(a+b)^2 = (a+b)\circ(a+b) = a^2 + a\circ b + b\circ a + b^2 = a+b+a\circ b+b\circ a$, so that $a+b=a+b+a\circ b+b\circ a$. Adding $-(a+b)$ to both sides yields $-(a+b)+(a+b) = 0$ for the left-hand side and $-(a+b)+(a+b)+a\circ b+b\circ a = a\circ b+b\circ a$ for the right-hand side, so that $a\circ b+b\circ a = 0$. Adding $b\circ a$ on both sides yields $a\circ b+b\circ a+b\circ a = a\circ b+0 = a\circ b$ for the left-hand side (due to the self-inverse property of addition) and $0+b\circ a = b\circ a$ for the right-hand side, so that the multiplicative commutativity $a\circ b = b\circ a$ holds. Thus the multiplicative semigroup $(M ; \circ)$ of the ring $(M ; +, \circ)$ is idempotent and commutative, and hence a semilattice.

Field : A domain $(M ; +, \circ)$ with the inner operations $+$ and $\circ$ in the set M is called a field if :

(1) The domain $(M ; +)$ is a commutative group.

(2) The domain $(M - \{0\} ; \circ)$ is a group.

(3) The multiplication $\circ$ is distributive with respect to the addition $+$.

A field $(M ; +, \circ)$ differs from a ring in that the multiplicative domain $(M - \{0\} ; \circ)$ in a field is a group, while the multiplicative domain $(M ; \circ)$ in a ring is a semigroup. A field $(M ; +, \circ)$ contains a zero element 0, which acts as an identity element under addition, and a unit element 1, which acts as an identity element under multiplication. The definition of a field cannot require that $(M ; \circ)$ be a group, since the zero element 0, due to its multiplicative invariance $a\circ 0 = 0\circ a = 0$, has no multiplicative inverse a such that $a\circ 0 = 0\circ a = 1$. A field $(M ; +, \circ)$ is said to be commutative if the multiplication $\circ$ is commutative.

Rules for fields : The defining properties of a field $(M ; +, \circ)$ imply the algebraic rules for the additive group $(M ; +)$, for the multiplicative group $(M - \{0\} ; \circ)$ and for rings, as well as the following additional rules :

(1) A field $(M ; +, \circ)$ has no zero divisors, so that $a \circ b = 0$ implies $a = 0$ or $b = 0$.

$$a \circ b = 0 \quad \Rightarrow \quad a = 0 \quad \vee \quad b = 0$$

(2) If the field $(M ; +, \circ)$ is commutative, then the rules of calculation for fractions apply :

$$(a \circ b^{-1}) \circ (c \circ d^{-1}) = (a \circ c) \circ (b \circ d)^{-1} \qquad b, d \neq 0$$

$$(a \circ b^{-1}) + (c \circ d^{-1}) = (a \circ d + c \circ b) \circ (b \circ d)^{-1} \qquad b, d \neq 0$$

Proof : Rules for fields

(1) Multiplying $a \circ b = 0$ with $a \neq 0$ by a^{-1} from the left yields $a^{-1}\circ a\circ b = 1\circ b = b = a^{-1}\circ 0 = 0$, so that $b = 0$. Multiplying $a \circ b = 0$ with $b \neq 0$ by b^{-1} from the right yields $a\circ b\circ b^{-1} = a\circ 1 = a = 0 \circ b^{-1} = 0$, so that $a = 0$.

(2) The rules are proved using the associative, commutative and distributive laws for addition and multiplication.

$$(a \circ c) \circ (b \circ d)^{-1} = (a \circ c) \circ (d^{-1} \circ b^{-1}) = a \circ (c \circ d^{-1}) \circ b^{-1} = (a \circ b^{-1}) \circ (c \circ d^{-1})$$

$$(a \circ d + c \circ b) \circ (b \circ d)^{-1} = (a \circ d + c \circ b) \circ (d^{-1} \circ b^{-1}) =$$

$$(a \circ d) \circ (d^{-1} \circ b^{-1}) + (c \circ b) \circ (b^{-1} \circ d^{-1}) =$$

$$a \circ (d \circ d^{-1}) \circ b^{-1} + c \circ (b \circ b^{-1}) \circ d^{-1} =$$

$$a \circ 1 \circ b^{-1} + c \circ 1 \circ d^{-1} = a \circ b^{-1} + c \circ d^{-1}$$

Summary : Additive and multiplicative domains (M ; +, ∘)

Property	Semiring	Ring	Integral ring	Boolean ring	Field
operation +					
associative	yes	yes	yes	yes	yes
commutative	yes	yes	yes	yes	yes
zero element 0		yes	yes	yes	yes
inverse		yes	yes	yes	yes
idempotent				no	
self-inverse				yes	
without zero sums		no	no	no	no
operation ∘					
associative	yes	yes	yes	yes	yes
commutative			yes	yes	
unit element 1					yes
inverse except for 0					yes
idempotent				yes	
self-inverse				no	
without zero divisors			yes		yes
operations + and ∘					
distributive	yes	yes	yes	yes	yes

Note that the operation + or ∘ cannot be both idempotent and self-inverse, and that the existence of additive inverses implies the existence of zero sums.

Example 1 : Domains of numbers

(1) The domain ($\mathbb{N}'$; + , ∘) for the addition and multiplication of positive natural numbers is a commutative semiring with the unit element 1.

(2) The domain ($\mathbb{N}$; + , ∘) for the addition and multiplication of natural numbers including zero is a commutative semiring without zero sums and without zero divisors, with the zero element 0 and the unit element 1.

(3) The domain ($\mathbb{Z}$; + , ∘) for the addition and multiplication of integers is a commutative ring without zero divisors with the zero element 0. It is therefore an integral ring.

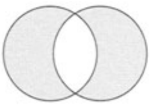
(4) The domain ($\mathbb{Q}$; + , ∘) for the addition and multiplication of rational numbers is a commutative field with the zero element 0 and the unit element 1.

Example 2 : Ring of sets

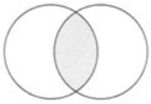
The domains (P(M) ; ⊕) for the symmetric difference and (P(M) ; ∩) for the intersection of sets are treated in Example 2 of Section 3.3. The domain (P(M) ; ⊕) is a commutative group. The domain (P(M) ; ∩) is a commutative semigroup. The intersection ∩ is distributive with respect to the symmetric difference ⊕. The domain (P(M) ; ⊕, ∩) is therefore a commutative ring with the empty set ∅ acting as the zero element and the reference set M acting as the unit element. Since idempotency holds for the intersection $A \cap A = A$, the ring is boolean. The following properties hold for $A, B, C \in P(M)$:

Property	Symmetric difference ⊕	Intersection ∩
associative	$A \oplus (B \oplus C) = (A \oplus B) \oplus C$	$A \cap (B \cap C) = (A \cap B) \cap C$
commutative	$A \oplus B = B \oplus A$	$A \cap B = B \cap A$
distributive	$A \cap (B \oplus C) = (A \cap B) \oplus (A \cap C)$	$(A \oplus B) \cap C = (A \cap C) \oplus (B \cap C)$
zero element	$A \oplus \emptyset = A$	$A \cap \emptyset = \emptyset$
unit element		$A \cap M = A$
inverse	$A \oplus A = \emptyset$	
idempotent		$A \cap A = A$

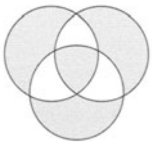
$A \oplus B$



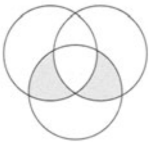
$A \cap B$



$A \oplus B \oplus C$



$(A \oplus B) \cap C$



Example 3 : Semiring of relations

The domains $(P(M \times M)\,;\,\cup)$ for the union and $(P(M \times M)\,;\,\circ)$ for the composition of relations are treated in Example 3 of Section 3.3. The domain $(P(M \times M)\,;\,\cup)$ is a commutative semigroup. The domain $(P(M \times M)\,;\,\circ)$ is a semigroup. The composition $\circ$ is distributive with respect to the union $\cup$. The domain $(P(M \times M)\,;\,\cup\,,\circ)$ is therefore a semiring with the empty set $\emptyset$ acting as the zero element and the identity relation I acting as the unit element. The following properties hold for $A, B, C \in P(M \times M)$:

Property	Union $\cup$	Composition $\circ$
associative	$A \cup (B \cup C) = (A \cup B) \cup C$	$A \circ (B \circ C) = (A \circ B) \circ C$
commutative	$A \cup B = B \cup A$	
distributive	$A \circ (B \cup C) = (A \circ B) \cup (A \circ C)$	$(A \cup B) \circ C = (A \circ C) \cup (B \circ C)$
zero element	$A \cup \emptyset = A$	$A \circ \emptyset = \emptyset = \emptyset \circ A$
unit element		$A \circ \mathrm{I} = A = \mathrm{I} \circ A$
idempotent	$A \cup A = A$	

The semiring $(P(M \times M)\,;\,\cup\,,\circ)$ of relations has the additional property that the union $\cup$ is idempotent and that the zero element $\emptyset$ is invariant with respect to the composition $\circ$.

3.4.3 DUAL DOMAINS

Introduction : Domains $(M ; \sqcup, \sqcap)$ with the inner operations $\sqcup$ (disjunction) and $\sqcap$ (conjunction) in the set M are treated as generalizations of the algebraic structure of truth values and sets. The disjunctive domain $(M ; \sqcup)$ and the conjunctive domain $(M ; \sqcap)$ have the same properties. The laws for the compatibility of the operations are formulated such that the operations $\sqcup$ and $\sqcap$ are interchangeable. Domains with these properties are called dual domains. It is assumed that the set M contains more than one element.

Adjunctive laws : The operations $\sqcup$ and $\sqcap$ of the domain $(M ; \sqcup, \sqcap)$ are said to be adjunctive if for all $a, b \in M$:

$$a \sqcap (a \sqcup b) = a$$
$$a \sqcup (a \sqcap b) = a$$

Distributive laws : The operations $\sqcup$ and $\sqcap$ of the domain $(M ; \sqcup, \sqcap)$ are said to be mutually distributive if for all $a, b, c \in M$:

$$a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c)$$
$$a \sqcup (b \sqcap c) = (a \sqcup b) \sqcap (a \sqcup c)$$

Identity elements : If there are identity elements for disjunction and conjunction, they are called the zero element and the unit element and are designated by 0 and 1, respectively.

zero element: $a \sqcup 0 = 0 \sqcup a = a$

unit element : $a \sqcap 1 = 1 \sqcap a = a$

Lattice : A domain $(M ; \sqcup, \sqcap)$ with the inner operations $\sqcup$ and $\sqcap$ in the set M is called a lattice if :

(1) The domain $(M ; \sqcup)$ is a commutative semigroup.

(2) The domain $(M ; \sqcap)$ is a commutative semigroup.

(3) The operations $\sqcup$ and $\sqcap$ are adjunctive.

A lattice $(M ; \sqcup, \sqcap)$ is called a lattice with zero and unit element if there is a zero element that acts as the identity element with respect to disjunction and a unit element that acts as the identity element with respect to conjunction. A lattice $(M ; \sqcup, \sqcap)$ is said to be distributive if the operations are mutually distributive.

Rules for lattices : The defining properties of a lattice $(M ; \sqcup, \sqcap)$ imply the following algebraic rules :

(1) If $(M ; \sqcup, \sqcap)$ is a lattice, then the elements are idempotent with respect to both operations.

$$a \sqcup a = a \qquad\qquad a \sqcap a = a$$

(2) If $(M ; \sqcup, \sqcap)$ is a lattice, then the following consistency rule holds :

$$a \sqcup b = b \quad \Leftrightarrow \quad a \sqcap b = a$$

(3) If $(M ; \sqcup, \sqcap)$ is a lattice with zero and unit element, then the zero element 0 is invariant under the operation $\sqcap$ and the unit element 1 is invariant under the operation $\sqcup$.

$$a \sqcap 0 = 0 \sqcap a = 0 \qquad\qquad a \sqcup 1 = 1 \sqcup a = 1$$

(4) If $(M ; \sqcup, \sqcap)$ is a distributive lattice, then the following uniqueness rule holds : The elements a and b are equal if their operations with an element c yield identical results.

$$(a \sqcup c = b \sqcup c) \quad \wedge \quad (a \sqcap c = b \sqcap c) \quad \Rightarrow \quad a = b$$

Proof : Rules for lattices

(1) The first adjunctive law $a \sqcap (a \sqcup b) = a$ with $b = a$ yields the equation $a \sqcap (a \sqcup a) = a$. The second adjunctive law $a \sqcup (a \sqcap b) = a$ with $b = a \sqcup a$ yields the equation $a \sqcup (a \sqcap (a \sqcup a)) = a$. Substituting the first equation into the second equation yields the idempotency rule $a \sqcup a = a$ for the operation $\sqcup$. Substituting the idempotency rule $a \sqcup a = a$ into the first equation yields the idempotency rule $a \sqcap a = a$.

(2) Substituting $a \sqcup b = b$ into the first adjunctive law $a \sqcap (a \sqcup b) = a$ yields $a \sqcap b = a$, so that $a \sqcup b = b \;\Rightarrow\; a \sqcap b = a$ holds. The second adjunctive law $b \sqcup (b \sqcap a) = b$ is equivalent to $(a \sqcap b) \sqcup b = b$ by commutativity; together with $a \sqcap b = a$, this yields $a \sqcup b = b$, so that $a \sqcap b = a \;\Rightarrow\; a \sqcup b = b$ holds. Together, $a \sqcup b = b \;\Rightarrow\; a \sqcap b = a$ and $a \sqcap b = a \;\Rightarrow\; a \sqcup b = b$ imply the consistency rule $a \sqcup b = b \;\Leftrightarrow\; a \sqcap b = a$.

(3) The second adjunctive law $b \sqcup (b \sqcap a) = b$ with $b = 0$ yields the invariance $0 \sqcup (0 \sqcap a) = 0 \sqcap a = 0$. By commutativity, this implies $0 \sqcap a = a \sqcap 0 = 0$. The first adjunctive law $b \sqcup (b \sqcap a) = b$ with $b = 1$ yields the invariance $1 \sqcap (1 \sqcup a) = 1 \sqcup a = 1$. By commutativity, this implies $1 \sqcup a = a \sqcup 1 = 1$.

(4) The equality $a = b$ is proved from the conditions $a \sqcup c = b \sqcup c$ and $a \sqcap c = b \sqcap c$ using the adjunctive, commutative and distributive laws :

$$\begin{aligned} a &= a \sqcap (a \sqcup c) = a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c) = \\ &(b \sqcap a) \sqcup (b \sqcap c) = b \sqcap (a \sqcup c) = b \sqcap (b \sqcup c) = b \end{aligned}$$

Boolean lattice : A domain $(M ; \sqcup, \sqcap, ^{-})$ with the binary operations $\sqcup$ and $\sqcap$ and the unary operation $^{-}$ (complement) in the set M is called a boolean lattice if :

(1) The domain $(M ; \sqcup)$ is commutative and possesses the identity element 0.

$a \sqcup b = b \sqcup a \qquad\qquad a \sqcup 0 = a$

(2) The domain $(M ; \sqcap)$ is commutative and possesses the identity element 1.

$a \sqcap b = b \sqcap a \qquad\qquad a \sqcap 1 = a$

(3) The operations $\sqcup$ and $\sqcap$ are mutually distributive.

$a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c) \qquad\qquad a \sqcup (b \sqcap c) = (a \sqcup b) \sqcap (a \sqcup c)$

(4) The complement $\bar{a}$ of a satisfies the following conditions :

$a \sqcup \bar{a} = 1 \qquad\qquad a \sqcap \bar{a} = 0$

Note that, in contrast to the definition of a lattice, the definition of a boolean lattice does not include the associative and adjunctive laws. These laws are derived from the defining properties (1) to (4). They are due to the operation $^{-}$ (complement).

Rules for boolean lattices : The defining properties of the boolean lattice $(M ; \sqcup, \sqcap, ^{-})$ imply the following rules :

(1) The boolean lattice has all properties of a distributive lattice with zero and unit element.

(2) The double complement of an element is the element itself.

$\bar{\bar{a}} = a$

(3) The zero element and the unit element are complements of each other.

$\bar{0} = 1 \qquad\qquad \bar{1} = 0$

(4) De Morgan's rules hold for the complements.

$\overline{(a \sqcup b)} = \bar{a} \sqcap \bar{b} \qquad\qquad \overline{(a \sqcap b)} = \bar{a} \sqcup \bar{b}$

Proof : Rules for boolean lattices

(1) The boolean lattice is a distributive lattice with zero and unit element if and only if the operations $\sqcup$ and $\sqcap$ are adjunctive and associative. Adjunctivity and associativity are proved by using the complementary properties to prove idempotency, the invariance of the identity elements and the uniqueness rule. To improve the legibility of the proof, the defining properties of the identity elements and the complementary, commutative and distributive laws are designated by N, K, C and D, respectively. These designations serve as a reference to the property being used in the transformation of an expression.

identity	:	$a \sqcup 0 = 0 \sqcup a = a$	$a \sqcap 1 = 1 \sqcap a = a$	(N)
complement	:	$a \sqcup \bar{a} = \bar{a} \sqcup a = 1$	$a \sqcap \bar{a} = \bar{a} \sqcap a = 0$	(K)
commutative	:	$a \sqcup b = b \sqcup a$	$a \sqcap b = b \sqcap a$	(C)
distributive	:	$a \sqcup (b \sqcap c) = (a \sqcup b) \sqcap (a \sqcup c)$	$a \sqcap (b \sqcup c) = (a \sqcap b) \sqcup (a \sqcap c)$	(D)

All properties derived from the dual defining properties are themselves dual. By virtue of this duality, a proof of one of the two dual properties suffices. The corresponding dual property is obtained by interchanging the operations $\sqcup$ and $\sqcap$, including their identity elements.

- The operations $\sqcup$ and $\sqcap$ are idempotent.

  $$a \sqcup a = a \qquad\qquad a \sqcap a = a \qquad\qquad \text{(id)}$$

  Proof : Idempotency of the operation $\sqcup$

  $$a \sqcup a \overset{N}{=} (a \sqcup a) \sqcap 1 \overset{K}{=} (a \sqcup a) \sqcap (a \sqcup \bar{a}) \overset{D}{=} a \sqcup (a \sqcap \bar{a}) \overset{K}{=} a \sqcup 0 \overset{N}{=} a$$

- The zero element 0 is invariant with respect to the operation $\sqcap$, and the unit element 1 is invariant with respect to the operation $\sqcup$.

  $$a \sqcap 0 = 0 \qquad\qquad a \sqcup 1 = 1 \qquad\qquad \text{(in)}$$

  Proof : Invariance of the zero element

  $$a \sqcap 0 \overset{N}{=} (a \sqcap 0) \sqcup 0 \overset{K}{=} (a \sqcap 0) \sqcup (a \sqcap \bar{a}) \overset{D}{=} a \sqcap (0 \sqcup \bar{a}) \overset{N}{=} a \sqcap \bar{a} \overset{K}{=} 0$$

- The elements a and b are equal if their operations $\sqcup$ and $\sqcap$ with an element c and its complement $\bar{c}$ yield identical results.

  $$(a \sqcup c = b \sqcup c) \quad \wedge \quad (a \sqcup \bar{c} = b \sqcup \bar{c}) \quad \Rightarrow \quad a = b$$
  $$(a \sqcap c = b \sqcap c) \quad \wedge \quad (a \sqcap \bar{c} = b \sqcap \bar{c}) \quad \Rightarrow \quad a = b \qquad\qquad \text{(un)}$$

  Proof : Uniqueness rule with the operation $\sqcup$

  $$(a \sqcup c) \sqcap (a \sqcup \bar{c}) = (b \sqcup c) \sqcap (b \sqcup \bar{c}) \overset{D}{\Rightarrow}$$
  $$a \sqcup (c \sqcap \bar{c}) = b \sqcup (c \sqcap \bar{c}) \overset{K}{\Rightarrow} a \sqcup 0 = b \sqcup 0 \overset{N}{\Rightarrow} a = b$$

- The operations $\sqcup$ and $\sqcap$ are adjunctive.

  $$a \sqcap (a \sqcup b) = a \qquad\qquad a \sqcup (a \sqcap b) = a \qquad\qquad \text{(ad)}$$

  Proof : First adjunctive law

  $$a \sqcap (a \sqcup b) \overset{N}{=} (a \sqcup 0) \sqcap (a \sqcup b) \overset{D}{=} a \sqcup (0 \sqcap b) \overset{C}{=} a \sqcup (b \sqcap 0) \overset{in}{=} a \sqcup 0 \overset{N}{=} a$$

- The operations $\sqcup$ and $\sqcap$ are associative.

  $$a \sqcup (b \sqcup c) = (a \sqcup b) \sqcup c \qquad\qquad a \sqcap (b \sqcap c) = (a \sqcap b) \sqcap c \quad \text{(as)}$$

Proof : Associative law for the operation $\sqcup$

The proof is performed in three steps. In the first step, it is proved that $a \sqcap ((a \sqcup b) \sqcup c) = a \sqcap (a \sqcup (b \sqcup c))$. In the second step, it is proved that $\bar{a} \sqcap ((a \sqcup b) \sqcup c) = \bar{a} \sqcap (a \sqcup (b \sqcup c))$. In the third step, the uniqueness rule (un) is used to prove the associative law $(a \sqcup b) \sqcup c = a \sqcup (b \sqcup c)$.

$$a \sqcap ((a \sqcup b) \sqcup c) \overset{D}{=} (a \sqcap (a \sqcup b)) \sqcup (a \sqcap c) \overset{ad}{=} a \sqcup (a \sqcap c) \overset{ad}{=} a \overset{ad}{=} a \sqcap (a \sqcup (b \sqcup c))$$

$$\bar{a} \sqcap ((a \sqcup b) \sqcup c) \overset{D}{=} (\bar{a} \sqcap (a \sqcup b)) \sqcup (\bar{a} \sqcap c) \overset{D}{=} ((\bar{a} \sqcap a) \sqcup (\bar{a} \sqcap b)) \sqcup (\bar{a} \sqcap c) \overset{K}{=}$$
$$(0 \sqcup (\bar{a} \sqcap b)) \sqcup (\bar{a} \sqcap c) \overset{N}{=} (\bar{a} \sqcap b) \sqcup (\bar{a} \sqcap c) \overset{D}{=} \bar{a} \sqcap (b \sqcup c) \overset{N}{=} 0 \sqcup (\bar{a} \sqcap (b \sqcup c)) \overset{K}{=}$$
$$(\bar{a} \sqcap a) \sqcup (\bar{a} \sqcap (b \sqcup c)) \overset{D}{=} \bar{a} \sqcap (a \sqcup (b \sqcup c))$$

$$a \sqcap ((a \sqcup b) \sqcup c) = a \sqcap (a \sqcup (b \sqcup c)) \quad \wedge$$
$$\bar{a} \sqcap ((a \sqcup b) \sqcup c) = \bar{a} \sqcap (a \sqcup (b \sqcup c)) \quad \overset{un}{\Rightarrow} \quad (a \sqcup b) \sqcup c = a \sqcup (b \sqcup c)$$

(2) The double complement $\bar{\bar{a}}$ is the element a itself.

$$\bar{\bar{a}} = a \qquad \text{(k2)}$$

Proof : By definition, the complement $\bar{a}$ of a satisfies $a \sqcup \bar{a} = 1$ and $a \sqcap \bar{a} = 0$. As the complement of $\bar{a}$, the double complement $\bar{\bar{a}}$ must therefore satisfy $\bar{\bar{a}} \sqcup \bar{a} = 1$ and $\bar{\bar{a}} \sqcap \bar{a} = 0$. Hence $a \sqcup \bar{a} = \bar{\bar{a}} \sqcup \bar{a}$ and $a \sqcap \bar{a} = \bar{\bar{a}} \sqcap \bar{a}$. Together with the uniqueness rule (4) for a distributive lattice, this yields $a = \bar{\bar{a}}$:

$$(a \sqcup \bar{a} = \bar{\bar{a}} \sqcup \bar{a}) \;\wedge\; (a \sqcup \bar{a} = \bar{\bar{a}} \sqcap \bar{a}) \quad \Rightarrow \quad a = \bar{\bar{a}}$$

(3) The zero element and the unit element are complements of each other.

$$\bar{0} = 1 \qquad \bar{1} = 0 \qquad \text{(nk)}$$

Proof : Complement of the zero element

By definition, the complement $\bar{0}$ of 0 satisfies $\bar{0} \sqcup 0 = 1$ and $\bar{0} \sqcap 0 = 0$. By the invariance (in) of the zero and unit element and by commutativity, $1 \sqcup 0 = 1$ and $1 \sqcap 0 = 0$. Hence $\bar{0} \sqcup 0 = 1 \sqcup 0$ and $\bar{0} \sqcap 0 = 1 \sqcap 0$. Together with the uniqueness rule (4) for a distributive lattice, this yields $\bar{0} = 1$:

$$(\bar{0} \sqcup 0 = 1 \sqcup 0) \;\wedge\; (\bar{0} \sqcap 0 = 1 \sqcap 0) \quad \Rightarrow \quad \bar{0} = 1$$

(4) De Morgan's rules hold for the complements.

$$\overline{(a \sqcup b)} = \bar{a} \sqcap \bar{b} \qquad \overline{(a \sqcap b)} = \bar{a} \sqcup \bar{b} \qquad \text{(M)}$$

Proof : First of De Morgan's rules

The proof is performed in three steps. In the first step, $(a \sqcup b) \sqcup (\overline{a} \sqcap \overline{b}) = 1$ is shown to hold. In the second step, $(a \sqcup b) \sqcap (\overline{a} \sqcap \overline{b}) = 0$ is shown to hold. These two results together form the complementary law for $(a \sqcup b)$ and $(\overline{a} \sqcap \overline{b})$, so that in the third step it may be inferred that the complement $\overline{(a \sqcup b)}$ of $(a \sqcup b)$ is $(\overline{a} \sqcap \overline{b})$.

$$(a \sqcup b) \sqcup (\overline{a} \sqcap \overline{b}) \overset{D}{=} ((a \sqcup b) \sqcup \overline{a}) \sqcap ((a \sqcup b) \sqcup \overline{b}) \overset{C}{=}$$
$$(\overline{a} \sqcup (a \sqcup b)) \sqcap ((a \sqcup b) \sqcup \overline{b}) \overset{as}{=} ((\overline{a} \sqcup a) \sqcup b) \sqcap (a \sqcup (b \sqcup \overline{b})) \overset{K}{=}$$
$$(1 \sqcup b) \sqcap (a \sqcup 1) \overset{in}{=} 1 \sqcap 1 \overset{N}{=} 1$$

$$(a \sqcup b) \sqcap (\overline{a} \sqcap \overline{b}) \overset{D}{=} (a \sqcap (\overline{a} \sqcap \overline{b})) \sqcup (b \sqcap (\overline{a} \sqcap \overline{b})) \overset{C}{=}$$
$$(a \sqcap (\overline{a} \sqcap \overline{b})) \sqcup ((\overline{a} \sqcap \overline{b}) \sqcap b) \overset{as}{=} ((a \sqcap \overline{a}) \sqcap \overline{b}) \sqcup (\overline{a} \sqcap (\overline{b} \sqcap b)) \overset{K}{=}$$
$$(0 \sqcap \overline{b}) \sqcup (\overline{a} \sqcup 0) \overset{in}{=} 0 \sqcup 0 \overset{N}{=} 0$$

$$((a \sqcup b) \sqcup (\overline{a} \sqcap \overline{b}) = 1) \ \wedge \ ((a \sqcup b) \sqcap (\overline{a} \sqcap \overline{b})) \ = 0) \ \overset{K}{\Rightarrow} \ \overline{(a \sqcup b)} = \overline{a} \sqcap \overline{b}$$

Boolean lattice and boolean ring : Every boolean lattice $(M ; \sqcup, \sqcap)$ may be used to construct a boolean ring $(M ; +, \circ)$ with unit element by the following transformation :

$$a \circ b := a \sqcap b$$
$$a + b := (a \sqcap \overline{b}) \sqcup (\overline{a} \sqcap b)$$

Conversely, every boolean ring $(M ; +, \circ)$ with unit element may be used to construct a boolean lattice $(M ; \sqcup, \sqcap)$ by the following transformation:

$$a \sqcap b := a \circ b$$
$$a \sqcup b := (a + b) + (a \circ b)$$

Proof : Transformation of a boolean lattice into a boolean ring

The specified transformation rule for a boolean lattice $(M ; \sqcup, \sqcap)$ yields a boolean ring, which by definition has the following properties :

(1) The domain $(M ; \circ)$ is an idempotent semigroup.

(2) The domain $(M ; +)$ is a commutative group.

(3) The multiplication $\circ$ is distributive with respect to the addition $+$.

The same designations as in the proof of the rules for boolean lattices are used in the proof of these properties.

(1) By definition, the product $a \circ b$ is identical with the operation $a \sqcap b$. The operation $\sqcap$ of the boolean lattice is associative and idempotent. Hence $(M\,;\,\circ)$ is an idempotent semigroup. It possesses an identity element, namely the unit element 1.

(2) The sum $a + b$ is defined by $(a \sqcap \bar{b}) \sqcup (\bar{a} \sqcap b)$. In the following, the domain $(M\,;\,+)$ is shown to possess all defining properties of a commutative group.

- The zero element 0 is the identity element of the operation $\sqcup$ and of the addition $+$:

  $$0 + b := (0 \sqcap \bar{b}) \sqcup (\bar{0} \sqcap b) \overset{\text{in}}{=} 0 \sqcup (\bar{0} \sqcap b) \overset{\text{nk}}{=} 0 \sqcup (1 \sqcap b) \overset{\text{in}}{=} 0 \sqcup b$$

- Every element a is its own inverse, so that $a + a = 0$.

  $$a + a := (a \sqcap \bar{a}) \sqcup (\bar{a} \sqcap a) \overset{\text{K}}{=} 0 \sqcup 0 \overset{\text{N}}{=} 0$$

- Addition is commutative, so that $a + b = b + a$.

  $$a + b := (a \sqcap \bar{b}) \sqcup (\bar{a} \sqcap b) \overset{\text{C}}{=} (\bar{a} \sqcap b) \sqcup (a \sqcap \bar{b}) \overset{\text{C}}{=} (b \sqcap \bar{a}) \sqcup (\bar{b} \sqcap a) = b + a$$

- Addition is associative, so that $a + (b + c) = (a + b) + c$. In the proof of associativity, the complement of the sum $a + b$ is determined according to De Morgan's rules.

  $$\overline{(a+b)} = \overline{(a \sqcap \bar{b}) \sqcup (\bar{a} \sqcap b)} \overset{\text{M}}{=} \overline{(a \sqcap \bar{b})} \sqcap \overline{(\bar{a} \sqcap b)} \overset{\text{M}}{=} (\bar{a} \sqcup \bar{\bar{b}}) \sqcap (\bar{\bar{a}} \sqcup \bar{b}) \overset{\text{K2}}{=}$$
  $$(\bar{a} \sqcup b) \sqcap (a \sqcup \bar{b}) \overset{\text{D}}{=} (\bar{a} \sqcap (a \sqcup \bar{b})) \sqcup (b \sqcap (a \sqcup \bar{b})) \overset{\text{D}}{=} ((\bar{a} \sqcap a) \sqcup (\bar{a} \sqcap \bar{b})) \sqcup$$
  $$((b \sqcap a) \sqcup (b \sqcap \bar{b})) \overset{\text{K}}{=} (0 \sqcup (\bar{a} \sqcap \bar{b})) \sqcup ((b \sqcap a) \sqcup 0) \overset{\text{N}}{=} (\bar{a} \sqcap \bar{b}) \sqcup (b \sqcap a) \overset{\text{C}}{=}$$
  $$(\bar{a} \sqcap \bar{b}) \sqcup (a \sqcap b)$$

  $$a + (b + c) = (a \sqcap \overline{(b+c)}) \sqcup (\bar{a} \sqcap (b + c)) = (a \sqcap ((\bar{b} \sqcap \bar{c}) \sqcup (b \sqcap c)) \sqcup$$
  $$(\bar{a} \sqcap ((b \sqcap \bar{c}) \sqcup (\bar{b} \sqcap c)) \overset{\text{D}}{=} (a \sqcap \bar{b} \sqcap \bar{c}) \sqcup (a \sqcap b \sqcap c) \sqcup (\bar{a} \sqcap b \sqcap \bar{c}) \sqcup (\bar{a} \sqcap \bar{b} \sqcap c)$$

  $$(a + b) + c = ((a + b) \sqcap \bar{c}) \sqcup (\overline{(a+b)} \sqcup c)) = (((a \sqcap \bar{b}) \sqcup (\bar{a} \sqcap b)) \sqcap \bar{c}) \sqcup$$
  $$(((\bar{a} \sqcap \bar{b}) \sqcup (a \sqcap b)) \sqcap c) \overset{\text{D}}{=} (a \sqcap \bar{b} \sqcap \bar{c}) \sqcup (\bar{a} \sqcap b \sqcap \bar{c}) \sqcup (\bar{a} \sqcap \bar{b} \sqcap c) \sqcup (a \sqcap b \sqcap c)$$

  The expressions for $a + (b + c)$ and $(a + b) + c$ contain the same terms connected by $\sqcup$. The order of these terms is irrelevant, since the operation $\sqcup$ is commutative and associative. Hence the associative law $a + (b + c) = (a + b) + c$ holds.

(3) The multiplication $\circ$ is distributive with respect to addition, so that the first distributive law $a \circ (b+c) = a \circ b + a \circ c$ holds.

$$a \circ (b+c) \;=\; a \sqcap (b+c) \;=\; a \sqcap ((b \sqcap \bar{c}) \sqcup (\bar{b} \sqcap c)) \overset{D}{=} (a \sqcap b \sqcap \bar{c}) \sqcup (a \sqcap \bar{b} \sqcap c)$$

$$a \circ b + a \circ c \;=\; ((a \circ b) \sqcap \overline{(a \circ c)}) \sqcup (\overline{(a \circ b)} \sqcap (a \circ c)) \;=$$

$$((a \sqcap b) \sqcap \overline{(a \sqcap c)}) \sqcup (\overline{(a \sqcap b)} \sqcap (a \sqcap c)) \overset{M}{=} ((a \sqcap b) \sqcap (\bar{a} \sqcup \bar{c})) \sqcup ((\bar{a} \sqcup \bar{b}) \sqcap (a \sqcap c)) \overset{D}{=}$$

$$(a \sqcap b \sqcap \bar{a}) \sqcup (a \sqcap b \sqcap \bar{c}) \sqcup (\bar{a} \sqcap a \sqcap c) \sqcup (\bar{b} \sqcap a \sqcap c) \overset{C}{=}$$

$$(a \sqcap \bar{a} \sqcap b) \sqcup (a \sqcap b \sqcap \bar{c}) \sqcup (\bar{a} \sqcap a \sqcap c) \sqcup (a \sqcap \bar{b} \sqcap c) \overset{N}{=}$$

$$(0 \sqcap b) \sqcup (a \sqcap b \sqcap \bar{c}) \sqcup (0 \sqcap c) \sqcup (a \sqcap \bar{b} \sqcap c) \overset{in}{=} (a \sqcap b \sqcap \bar{c}) \sqcup (a \sqcap \bar{b} \sqcap c)$$

The expressions for $a \circ (b+c)$ and $a \circ b + a \circ c$ contain the same terms connected by $\sqcup$ in the same order. Hence the first distributive law holds. The validity of the second distributive law is proved analogously.

Example 1 : Lattice of numbers

The domain ($\mathbb{Q}$; min, max) for the minimum $\min\{a,b\}$ and the maximum $\max\{a,b\}$ of rational numbers is a distributive lattice, since :

- The operations min and max are associative :

$$\min\{a, \min\{b,c\}\} = \min\{\min\{a,b\}, c\}$$
$$\max\{a, \max\{b,c\}\} = \max\{\max\{a,b\}, c\}$$

- The operations min and max are commutative :

$$\min\{a,b\} = \min\{b,a\}$$
$$\max\{a,b\} = \max\{b,a\}$$

- The operations min and max are adjunctive :

$$\max\{a, \min\{a,b\}\} = a$$
$$\min\{a, \max\{a,b\}\} = a$$

- The operations min and max are mutually distributive :

$$\max\{a, \min\{b,c\}\} = \min\{\max\{a,b\}, \max\{a,c\}\}$$
$$\min\{a, \max\{b,c\}\} = \max\{\min\{a,b\}, \min\{a,c\}\}$$

The domain ([0.0, 1.0] ; min, max) for the minimum $\min\{a,b\}$ and the maximum $\max\{a,b\}$ of the real numbers in the closed interval $0.0 \le a,b \le 1.0$ is a distributive lattice with the zero element 1.0 and the unit element 0.0.

- zero element 1.0 : $\min\{a, 1.0\} = \min\{1.0, a\} = a$
- unit element 0.0 : $\max\{a, 0.0\} = \max\{0.0, a\} = a$

Example 2 : Logical lattice

The domain $(T; \vee, \wedge)$ with the set $T = \{0,1\} = \{\text{false, true}\}$ of truth values and the logical operations $\vee$ and $\wedge$ is a boolean lattice. The identity elements are $n = 0$ and $e = 1$. The complement $\bar{a}$ of a is $\neg a$. The following laws hold for elements a, b, $c \in T$ (see Section 2.1.3) :

Property	$\vee$ (logical or)	$\wedge$ (logical and)
associative	$(a \vee b) \vee c = a \vee (b \vee c)$	$(a \wedge b) \wedge c = a \wedge (b \wedge c)$
commutative	$a \vee b = b \vee a$	$a \wedge b = b \wedge a$
adjunctive	$a \vee (a \wedge b) = a$	$a \wedge (a \vee b) = a$
identity	$a \vee 0 = a$	$a \wedge 0 = 0$
	$a \vee 1 = 1$	$a \wedge 1 = a$
complementary	$a \vee (\neg a) = 1$	$a \wedge (\neg a) = 0$
distributive	$a \vee (b \wedge c) = (a \vee b) \wedge (a \vee c)$	$a \wedge (b \vee c) = (a \wedge b) \vee (a \vee c)$
	$(a \wedge b) \vee c = (a \vee c) \wedge (b \vee c)$	$(a \vee b) \wedge c = (a \wedge c) \vee (b \wedge c)$

Example 3 : Lattice of sets

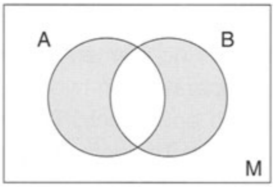
The domain $(P(M) ; \cup, \cap)$ with the power set $P(M)$ of a non-empty set M and the operations $\cup$ (union) and $\cap$ (intersection) on sets is a boolean lattice. The identity elements in $P(M)$ are the empty set $n = \emptyset$ and the reference set $e = M$. The complement $\overline{A}$ of an element A of $P(M)$ is the difference $M - A$. The following rules hold for elements $A, B, C \in P(M)$ (see Example 2 in Section 3.3) :

Property	$\cup$ (union)	$\cap$ (intersection)
associative	$(A \cup B) \cup C = A \cup (B \cup C)$	$(A \cap B) \cap C = A \cap (B \cap C)$
commutative	$A \cup B = B \cup A$	$A \cap B = B \cap A$
adjunctive	$A \cup (A \cap B) = A$	$A \cap (A \cup B) = A$
identity	$A \cup \emptyset = A$	$A \cap \emptyset = \emptyset$
	$A \cup M = M$	$A \cap M = A$
complementary	$A \cup \overline{A} = M$	$A \cap \overline{A} = \emptyset$
distributive	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$	$A \cap (B \cup C) = (A \cap B) \cup (B \cap C)$
	$(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$	$(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$

Example 4 : Lattice of sets and ring of sets

The lattice (P(M) ; ∪, ∩) of sets with the operations ∪ (union) and ∩ (intersection) on sets is treated in Example 3. The lattice (P(M) ; ∪, ∩) is boolean and may therefore be transformed into a boolean ring (P(M) ; +, ∘) according to the following prescription :

$$\begin{aligned} A \circ B &:= A \cap B \\ A + B &:= (A \cap \overline{B}) \cup (\overline{A} \cap B) \\ &= (A \cap (M-B)) \cup ((M-A) \cap B) \\ &= (A-(A \cap B)) \cup (B-(A \cap B)) \\ &= (A-B) \cup (B-A) \\ &= A \oplus B \end{aligned}$$



A ⊕ B

The boolean ring (P(M) ; +, ∘) is identical with the ring (P(M) ; ⊕, ∩) of sets with the operations ⊕ (symmetric difference) and ∩ (intersection). This ring of sets is treated in Example 2 of Section 3.4.2.

3.5 VECTOR SPACES

3.5.1 GENERAL VECTOR SPACES

Introduction : The theory of vector spaces developed out of the theory of systems of linear equations. The characteristic property of vector spaces is an operation on elements taken from different sets. Such an operation is called an outer operation on two sets. A vector space is defined as a domain with two sets and two operations. The rules for vector spaces are derived from the defining properties. The defining properties and the rules together form the theory of vector spaces. This section gives an introduction to the theory of vector spaces.

Outer operation : An outer operation f on the sets A and V assigns exactly one element $\mathbf{y}$ of the set V to every pair $(a, \mathbf{x})$ in the direct product $A \times V$. Thus an outer operation is a mapping f from $A \times V$ to V. To distinguish between the two sets, elements in A appear in normal type, while the elements of V appear in boldface.

$$f : A \times V \to V \quad \text{with} \quad f(a, \mathbf{x}) = \mathbf{y} \quad \text{and} \quad \mathbf{x}, \mathbf{y} \in V \,; \quad a \in A$$

An operator symbol is often used instead of a letter to designate an outer operation. In this case, the mapping is represented as follows :

$$\circ : A \times V \to V \quad \text{with} \quad a \circ \mathbf{x} = \mathbf{y} \quad \text{and} \quad \mathbf{x}, \mathbf{y} \in V \,; \quad a \in A$$

Due to this representation, the set A is called the operator set. In the context of vector spaces, the set V is called the set of vectors. In the outer operation $a \circ \mathbf{x}$, the order of the elements a and $\mathbf{x}$ may be changed; the meaning of the expression is unambiguous, since $a \in A$ and $\mathbf{x} \in V$. Often the operator symbol $\circ$ in the expression $a \circ \mathbf{x}$ is dropped, and one writes $a\,\mathbf{x}$ or $\mathbf{x}\,a$ instead.

Vector space : The domain $(V \,; +)$ is called a vector space over the domain $(A \,; +, \circ)$ and is designated by $(A, V \,; +, \circ)$ if :

(1) The domain $(V \,; +)$ is a commutative group.

(2) The domain $(A \,; +, \circ)$ is a commutative field.

(3) The operation $\circ$ in $(A, V \,; +, \circ)$ is an outer operation with the following properties for all elements $a, b \in A$ and $\mathbf{x}, \mathbf{y} \in V$:

associative : $(a \circ b) \circ \mathbf{x} = a \circ (b \circ \mathbf{x})$

distributive : $(a + b) \circ \mathbf{x} = a \circ \mathbf{x} + b \circ \mathbf{x}$
$a \circ (\mathbf{x} + \mathbf{y}) = a \circ \mathbf{x} + a \circ \mathbf{y}$

identitive : $1_A \circ \mathbf{x} = \mathbf{x}$

In the definition of vector spaces, the symbols $+$ and $\circ$ for addition and multiplication designate different operations in different domains.

(1) The symbol $+$ in the domain $(V\,;+)$ represents an inner operation in the set V of vectors. The symbol $+$ in the domain $(A\,;+,\circ)$ represents an inner operation in the operator set A. The meaning of the symbol $+$ in expressions like $a+b$ or $\mathbf{x}+\mathbf{y}$ is determined by the membership of the operands in the sets A and V.

(2) The symbol $\circ$ in the domain $(A\,;+,\circ)$ represents an inner operation in the operator set A. The symbol $\circ$ in the domain $(A, V\,;+,\circ)$ represents an outer operation on the operator set A and the set V of vectors. The meaning of the symbol $\circ$ in expressions like $a\circ b$ or $a\circ\mathbf{x}$ is determined by the membership of the operands in the sets A and V.

The identity element of the addition $+$ in the domain $(V\,;+)$ is called the zero vector and is designated by $\mathbf{0}$. The identity element of the addition $+$ in the domain $(A\,;+,\circ)$ is called the zero element and is designated by 0_A or 0. The identity element of the multiplication $\circ$ in the domain $(A\,;+,\circ)$ is called the unit element and is designated by 1_A or 1.

Rules for vector spaces : Besides the rules for the commutative group $(V\,;+)$ and the rules for the commutative field $(A\,;+,\circ)$, the defining properties of a vector space imply additional rules for the outer operation of $(A, V\,;+,\circ)$:

(1) The outer operation $\circ$ with the identity elements $\mathbf{0}$ and 0 of addition yields the zero vector.

$$0\circ\mathbf{x} = \mathbf{x}\circ 0 = \mathbf{0} \qquad\qquad a\circ\mathbf{0} = \mathbf{0}\circ a = \mathbf{0}$$

(2) The outer operation $\circ$ applied to additive inverses obeys rules which are called sign rules in the algebra of real vectors.

$$a\circ(-\mathbf{x}) = (-a)\circ\mathbf{x} = -(a\circ\mathbf{x})$$
$$(-a)\circ(-\mathbf{x}) = a\circ\mathbf{x}$$

(3) The vector space has no zero divisors with respect to the outer operation $\circ$, so that $a\circ\mathbf{x}=\mathbf{0}$ implies $a=0$ or $\mathbf{x}=\mathbf{0}$.

$$a\circ\mathbf{x} = \mathbf{0} \quad\Rightarrow\quad a = 0 \;\vee\; \mathbf{x} = \mathbf{0}$$

Proof : Rules for vector spaces

(1) By the first distributive law, $a\circ\mathbf{x}+b\circ\mathbf{x}=(a+b)\circ\mathbf{x}$. Substituting $a=b=0$ yields $0\circ\mathbf{x}+0\circ\mathbf{x}=0\circ\mathbf{x}$. Adding $-(0\circ\mathbf{x})$ on both side yields $-(0\circ\mathbf{x})+(0\circ\mathbf{x}+0+\mathbf{x})=(-0\circ\mathbf{x}+0\circ\mathbf{x})+0\circ\mathbf{x}=0\circ\mathbf{x}$ on the left-hand side and $-(0\circ\mathbf{x})+0\circ\mathbf{x}=\mathbf{0}$ on the right-hand side, so that $0\circ\mathbf{x}=\mathbf{0}$ holds. The invariance $a\circ\mathbf{0}=\mathbf{0}$ is proved similarly using the second distributive law.

(2) By the first distributive law, $a \circ \mathbf{x} + b \circ \mathbf{x} = (a+b) \circ \mathbf{x}$. Substituting $b = -a$ yields $a \circ \mathbf{x} + (-a) \circ \mathbf{x} = (a + (-a)) \circ \mathbf{x} = 0 \circ \mathbf{x} = \mathbf{0}$. Further $a \circ \mathbf{x} + (-a) \circ \mathbf{x} = \mathbf{0}$ implies that $(-a) \circ \mathbf{x}$ is the additive inverse of $a \circ \mathbf{x}$, that is $(-a) \circ \mathbf{x} = -(a \circ \mathbf{x})$. The validity of $a \circ (-\mathbf{x}) = -(a \circ \mathbf{x})$ is proved similarly using the second distributive law. For $-a$ and $-\mathbf{x}$, the two rules yield $(-a) \circ (-\mathbf{x}) = -((-a) \circ \mathbf{x}) = -(-(a \circ \mathbf{x})) = a \circ \mathbf{x}$.

(3) Multiplying $a \circ \mathbf{x} = \mathbf{0}$ for $a \neq 0$ by a^{-1} yields $a^{-1} \circ (a \circ \mathbf{x}) = (a^{-1} \circ a) \circ \mathbf{x} = 1 \circ \mathbf{x} = \mathbf{x}$ on the left-hand side and $a^{-1} \circ \mathbf{0} = \mathbf{0}$ on the right-hand side, so that $\mathbf{x} = \mathbf{0}$. If $a = 0$, then $a \circ \mathbf{x} = \mathbf{0}$ by (1). Hence $a \circ \mathbf{x} = \mathbf{0}$ implies $a = 0$ or $\mathbf{x} = \mathbf{0}$.

Linear combination : Vectors $\mathbf{x}_1, \mathbf{x}_2, ...$ are chosen from the set of vectors of a vector space $(A, V ; +, \circ)$ and multiplied by elements $a_1, a_2, ...$ of the operator set A. The sum of the products $a_i \circ \mathbf{x}_i$ is called a linear combination of the vectors $\mathbf{x}_i$. The elements a_i are called the coefficients of the linear combination. The order of the products in the sum is irrelevant, since vector addition is commutative. Every linear combination yields a vector $\mathbf{x}$ of the set V of vectors. The notation for linear combinations is simplified by introducing the symbol $\sum$ for the summation and dropping the symbol $\circ$ in products.

$$\mathbf{x} = a_1 \circ \mathbf{x}_1 + a_2 \circ \mathbf{x}_2 + ... = \sum_{i \in I} a_i \, \mathbf{x}_i \qquad \mathbf{x}, \mathbf{x}_i \in V \ ; \quad a_i \in A$$

Rules for linear combinations : The following rules follow from the defining properties of a vector space and the definition of a linear combination :

(1) Scaling a linear combination
A linear combination is scaled by a factor $p \in A$ by multiplying each coefficient of the linear combination with p.

$$p\mathbf{x} = p \sum_{i \in I} a_i \, \mathbf{x}_i = \sum_{i \in I} (p \circ a_i) \, \mathbf{x}_i$$

(2) Sum of two linear combinations
Let two linear combinations with the coefficients a_i and b_i for the vectors $\mathbf{x}_i$ be given. The sum of the two linear combinations is formed by adding the coefficients a_i and b_i for each vector $\mathbf{x}_i$.

$$\mathbf{x} + \mathbf{y} = \sum_{i \in I} a_i \, \mathbf{x}_i + \sum_{i \in I} b_i \mathbf{x}_i = \sum_{i \in I} (a_i + b_i) \, \mathbf{x}_i$$

(3) Difference of two linear combinations
Let two linear combinations with the coefficients a_i and b_i for the vectors $\mathbf{x}_i$ be given. The difference of the two linear combinations is formed by subtracting the coefficients a_i and b_i for each vector $\mathbf{x}_i$. The difference $a_i - b_i$ is the sum of a_i and the additive inverse $(-b_i)$ of b_i.

$$\mathbf{x} - \mathbf{y} = \sum_{i \in I} a_i \, \mathbf{x}_i - \sum_{i \in I} b_i \mathbf{x}_i = \sum_{i \in I} (a_i - b_i) \, \mathbf{x}_i = \sum_{i \in I} (a_i + (-b_i)) \, \mathbf{x}_i$$

Linear closure : The set of all vectors which can be formed as a linear combination of a given subset $X = \{\mathbf{x}_1, \mathbf{x}_2, ...\}$ of the set V of vectors of the vector space $(A, V ; +, \circ)$ is called the linear closure (span) of X and is designated by L(X).

$$L(X) := \{\mathbf{x} \in V \mid \mathbf{x} = \sum_{i \in I} a_i \, \mathbf{x}_i \ \wedge \ \mathbf{x}_i \in X\}$$

Generating set : A subset E of the set V of vectors is said to generate (span) the vector space $(A, V ; +, \circ)$ if V is the linear closure of E.

$$E \text{ generates } (A, V ; +, \circ) \quad :\Leftrightarrow \quad E \subseteq V \quad \wedge \quad V = L(E)$$

The set V of vectors is always a generating set, since $L(V) = V$. A generating set E is said to be finite if E is a finite subset of the set V of vectors. A vector space is called a vector space of finite type if it has a finite generating set. The generating set of a vector space is not unique.

Linear independence : A subset X of the set V of vectors of a vector space $(A, V ; +, \circ)$ is said to be linearly independent if the choice $a_i = 0$ for the coefficients is the only way to represent the zero vector **0** as a linear combination of the vectors $\mathbf{x}_i$. Otherwise, X is said to be linearly dependent.

$$\sum_{i \in I} a_i \, \mathbf{x}_i = \mathbf{0} \quad \Rightarrow \quad \bigwedge_{i \in I} (a_i = 0) \qquad\qquad a_i \in A, \ \mathbf{x}_i \in V$$

Basis of a vector space : A subset B of the set V of vectors is called a basis of the vector space $(A, V ; +, \circ)$ if B is a linearly independent generating set. The vectors $\mathbf{x}_i$ of a basis are called basis vectors. A basis is said to be finite if the number of its basis vectors is finite. Otherwise, it is said to be infinite.

Construction of a basis : Every vector space $(A, V ; +, \circ)$ has a basis. A basis B may be constructed step by step from a generating set $E \subseteq V$ of the vector space :

(1) Before the first step, the set B is empty.

(2) In every step i, a vector $\mathbf{x}_i \neq \mathbf{0}$ of the generating set E is added to the set B such that $B = \{\mathbf{x}_1, ..., \mathbf{x}_{i-1}, \mathbf{x}_i\}$ is linearly independent.

(3) If the set B cannot be further enlarged after a finite number of steps, then B is a finite basis of the vector space. Otherwise, B is an infinite basis of the vector space.

The choice of the vector added to the basis in a step is not unique. Hence the constructed basis is not unique.

Properties of bases : A vector space $(A, V\ ; +, \circ)$ has the following properties with respect to its bases :

(1) Every vector of a vector space has a unique representation as a linear combination of the basis vectors of a basis.

(2) Finite bases of a vector space contain the same number of basis vectors.

Proof : Properties of bases

(1) Let a vector **x** of the vector space with the basis B be represented by two linear combinations of the basis vectors $\mathbf{x}_i \in B$. Then the difference of the two linear combinations is the zero vector **0**.

$$\mathbf{x} = \sum_{i \in I} a_i\, \mathbf{x}_i \qquad a_i \in A$$

$$\mathbf{x} = \sum_{i \in I} b_i\, \mathbf{x}_i \qquad b_i \in A$$

$$\mathbf{0} = \sum_{i \in I} (a_i - b_i)\ \mathbf{x}_i = \sum_{i \in I} c_i\, \mathbf{x}_i \qquad c_i \in A$$

All coefficients $c_i = a_i - b_i$ are 0, since the basis vectors are by definition linearly independent. Hence $a_i = b_i$, and the two linear combinations for the vector **x** are identical.

(2) Let $B_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_s\}$ be a finite basis with s basis vectors, and let $B_2 = \{\mathbf{y}_1, \ldots, \mathbf{y}_s, \mathbf{y}_{s+1}\}$ be a finite basis with s + 1 basis vectors. Then by (1) each basis vector $\mathbf{y}_i \in B_2$ has a unique representation as a linear combination of the basis vectors $\mathbf{x}_i \in B_1$. This leads to the following system of equations :

$$\begin{aligned}
\mathbf{y}_1 &= a_{11}\, \mathbf{x}_1 + \ldots + a_{1s}\, \mathbf{x}_s \\
&\vdots \\
\mathbf{y}_s &= a_{s1}\, \mathbf{x}_1 + \ldots + a_{ss}\, \mathbf{x}_s \\
\mathbf{y}_{s+1} &= a_{s+1,1}\, \mathbf{x}_1 + \ldots + a_{s+1,s}\, \mathbf{x}_s
\end{aligned}$$

In the i-th equation, at least one coefficient a_{im} is non-zero. The i-th equation is used to eliminate $\mathbf{x}_m$ in all other equations. Then the system of equations is reduced by $\mathbf{x}_m$ and the equation i.

None of the left-hand sides of the reduced system of equations is **0**, since the basis vectors $\mathbf{y}_i$ are linearly independent. There is no reduced equation in which all coefficients on the right-hand side are zero, since the basis vectors $\mathbf{x}_m$ are linearly independent. The right-hand side takes the value **0** only when all basis vectors $\mathbf{x}_m$ have been eliminated. The elimination is repeated until all basis vectors $\mathbf{x}_m$ have been eliminated.

After the elimination of all basis vectors $\mathbf{x}_m$, the reduced system consists of a single equation. Let this be the p-th equation of the original system. Then the left-hand side of this equation is a linear combination of the basis vectors $\mathbf{y}_i$ in which $\mathbf{y}_p$ still has the coefficient 1. This linear combination is non-zero by definition. This contradicts the fact that the right-hand side of the equation is $\mathbf{0}$. Hence the number of basis vectors in B_2 is not greater than in B_1.

An analogous argument shows that the number of basis vectors in B_1 is not greater than in B_2. Hence the number of basis vectors in B_1 and B_2 is equal.

Rank of a vector space : A vector space is said to be of infinite rank if a basis of the vector space is infinite. A vector space is said to be of rank m if a basis is finite and contains m basis vectors.

Vector subspace : Let $(A, V \,;\, +, \circ)$ and $(A, W \,;\, +, \circ)$ be vector spaces over the same commutative field $(A \,;\, +, \circ)$. The vector space $(A, W \,;\, +, \circ)$ is called a (vector) subspace of $(A, V \,;\, +, \circ)$ if W is a subset of V.

Basis and rank of a subspace : Let the vector space $(A, W \,;\, +, \circ)$ of rank m_W be a subspace of the vector space $(A, V \,;\, +, \circ)$ of rank m_V. From $W \subseteq V$ it follows directly that $m_W \le m_V$. The method for the construction of a basis shows that the basis B_W can be extended stepwise to a basis B_V of V. Hence there are bases B_W and B_V for which $B_W \subseteq B_V$.

Example 1 : Vector space of complex numbers

The additive domain $(\mathbb{C} \,;\, +)$ of complex numbers is a vector space of rank 2 over the commutative field $(\mathbb{R} \,;\, +, \circ)$ of real numbers, as demonstrated in the following.

(1) A complex number $\mathbf{x} \in \mathbb{C}$ is represented in the form $x_r + i\, x_i$ with the real part $x_r \in \mathbb{R}$ and the imaginary part $x_i \in \mathbb{R}$. The real part and the imaginary part are real numbers. Two complex numbers are added by adding the two real parts and the two imaginary parts separately.

complex number	$\mathbf{x} := (x_r + i\, x_i) \in \mathbb{C}$	$x_r, x_i \in \mathbb{R}$
addition	$(x_r + i\, x_i) + (y_r + i\, y_i) = (x_r + y_r) + i\,(x_i + y_i)$	

The addition of complex numbers is associative and commutative. The complex number $(0 + i\,0)$ acts as the identity element under addition. For every complex number $x_r + i\, x_i$ there is a complex number $((-x_r) + i(-x_i))$ that is its additive inverse. The additive domain $(\mathbb{C} \,;\, +)$ with these properties is a commutative group according to the definition in Section 3.3.2.

(2) According to Section 3.4.2, the additive and multiplicative domain $(\mathbb{R} \,;\, +, \circ)$ of real numbers is a commutative field with the zero element 0 and the unit element 1.

(3) A complex number is scaled by a real number $a \in \mathbb{R}$ by separately multiplying the real part and the imaginary part by a.

scaling $\quad a \circ (x_r + i\, x_i) := ((a\, x_r) + i(a\, x_i)) \qquad a \in \mathbb{R}$

The scaling operation is the outer operation $\circ$ on the sets $\mathbb{R}$ and $\mathbb{C}$. It satisfies the defining properties of a vector space. That is, for $a, b \in \mathbb{R}$ and $(x_r + i\, x_i)$, $(y_r + i\, y_i) \in \mathbb{C}$, the following properties hold :

associative $\quad (a \circ b) \circ (x_r + i\, x_i) = a \circ (b \circ (x_r + i\, x_i))$

distributive $\quad (a + b) \circ (x_r + i\, x_i) = a \circ (x_r + i\, x_i) + b \circ (x_r + i\, x_i)$

$\quad a \circ ((x_r + i\, x_i) + (y_r + i\, y_i)) = a \circ (x_r + i\, x_i) + a \circ (y_r + i\, y_i)$

identitive $\quad 1 \circ (x_r + i\, x_i) = (x_r + i\, x_i)$

(4) With properties (1), (2), (3) the domain $(\mathbb{R}, \mathbb{C}\, ; +, \circ)$ has the defining properties of a vector space. Every complex number is a vector of this vector space.

(5) The vector space $(\mathbb{R}, \mathbb{C}\, ; +, \circ)$ of complex numbers has rank 2. Every basis of the vector space contains exactly two non-zero complex numbers, which are linearly independent. The following examples show two different bases.

first basis : $\{1, i\}$

second basis : $\{(1 + i), (1 - i)\}$

Example 2 : Vector space of real polynomial functions

The additive domain $(P_n(x)\, ; +)$ of real polynomial functions of degree n is a vector space of rank n + 1 over the commutative field $(\mathbb{R}\, ; +, \circ)$ of real numbers. This is demonstrated in the following.

(1) A real polynomial function $p(x) \in P_n(x)$ of degree n depending on the variable x is defined as follows :

real polynomial function : $\quad p(x) = \sum_{k=0}^{n} c_k\, x^k \qquad x, p(x) \in \mathbb{R}$

coefficient of the k-th term : $\quad c_k \in \mathbb{R}$

degree of the polynomial : $\quad n \in \mathbb{N}$

To simplify the notation, the index $k = 0,...,n$ on the summation symbols is dropped in the following. Two real polynomial functions are added by adding the coefficients for each term x^k separately.

addition $\quad \sum c_k\, x^k + \sum d_k\, x^k = \sum (c_k + d_k)\, x^k$

The addition of real polynomial functions is associative and commutative. The zero polynomial $\sum 0\, x^k$ acts as the identity element under addition. For every polynomial function $\sum c_k\, x^k$ there is a polynomial function $\sum (-c_k)\, x^k$ which is its additive inverse. According to the definition in Section 3.3, the additive domain $(P_n(x)\, ; +)$ is a commutative group.

(2) According to Section 3.4.2, the additive and multiplicative domain ($\mathbb{R}$; $+$, $\circ$) of real numbers is a commutative field with the zero element 0 and the unit element 1.

(3) A real polynomial function is scaled by a real number $a \in \mathbb{R}$ by multiplying each coefficient by a.

scaling $\quad a \circ \sum c_k x^k = \sum (ac_k) x^k \qquad a \in \mathbb{R}$

The scaling operation is the outer operation $\circ$ on the sets $\mathbb{R}$ and $P_n(x)$. It satisfies the defining properties of a vector space. That is, for $a, b \in \mathbb{R}$ and $\sum c_k x^k$, $\sum d_k x^k \in P_n(x)$ the following properties hold :

associative $\quad (a \circ b) \circ \sum c_k x^k = a \circ (b \circ \sum c_k x^k)$

distributive $\quad (a + b) \circ \sum c_k x^k = a \circ \sum c_k x^k + b \circ \sum c_k x^k$

$\quad a \circ (\sum c_k x^k + \sum d_k x^k) = a \circ \sum c_k x^k + a \circ \sum d_k x^k$

identitive $\quad 1 \circ \sum c_k x^k = \sum c_k x^k$

(4) With properties (1), (2), (3) the domain ($\mathbb{R}$, $P_n(x)$; $+$, $\circ$) has the defining properties of a vector space. Every real polynomial function of degree n is a vector of this vector space.

(5) The vector space ($\mathbb{R}$, $P_n(x)$; $+$, $\circ$) of real polynomial functions has rank $n + 1$. It possesses a basis consisting of the $n + 1$ basis functions x^k for $k = 0,...,n$. Every real polynomial function of degree n is a linear combination of these basis functions with real coefficients.

basis : $\quad \{x^0, x^1, ..., x^n\}$

3.5.2 REAL VECTOR SPACES

Introduction : A real vector space is a special case of a general vector space. The characteristic property of a real vector is that it is represented as an n-tuple of real numbers. Addition and multiplication of real numbers are applied to real vectors elementwise. This leads to the theory of real vector spaces.

Real vector : A vector is called a real vector of dimension n if it is an n-tuple $(x_1, x_2, \ldots, x_n)$ with the elements $x_i \in \mathbb{R}$. The set of all real vectors of dimension n is the n-fold cartesian product $\mathbb{R}^n$.

Real vectors of dimension n are usually not represented in the n-tuple notation. Instead, the elements x_i of a vector **x** are arranged in a column, the index of the element being used as a row index. This representation is called a column vector.

$$\mathbf{x} = \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array} \in \mathbb{R}^n$$

Real vector addition : Two real vectors of equal dimension are added by adding the elements with the same index. The additive structure $(\mathbb{R} \, ; +)$ of the real numbers thus carries over elementwise to real vectors.

$$\mathbf{x} + \mathbf{y} = \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array} + \begin{array}{|c|}\hline y_1 \\ \hline y_2 \\ \hline \vdots \\ \hline y_n \\ \hline \end{array} := \begin{array}{|c|}\hline x_1 + y_1 \\ \hline x_2 + y_2 \\ \hline \vdots \\ \hline x_n + y_n \\ \hline \end{array}$$

The addition $+$ is an inner operation in the set $\mathbb{R}^n$ of n-dimensional real vectors. By Section 3.3, The additive domain $(\mathbb{R}^n \, ; +)$ satisfies all defining properties of a commutative group :

(1) The addition is associative, since for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$:

$$(\mathbf{x} + \mathbf{y}) + \mathbf{z} = \mathbf{x} + (\mathbf{y} + \mathbf{z})$$

(2) The addition is commutative, since for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\mathbf{x} + \mathbf{y} = \mathbf{y} + \mathbf{x}$$

(3) The zero vector **0**, whose elements are all 0, acts as an identity element under addition, since for $\mathbf{0}, \mathbf{x} \in \mathbb{R}^n$:

$$\mathbf{x} + \mathbf{0} \;=\; \mathbf{0} + \mathbf{x} \;=\; \mathbf{x}$$

(4) For every real vector **x** there exists a vector $-\mathbf{x}$ which is its additive inverse, so that $\mathbf{x} + (-\mathbf{x}) = \mathbf{0}$. The vector $-\mathbf{x}$ contains the elements of the vector **x** with opposite signs.

Real vector scaling : A real vector is scaled by a real number $a \in \mathbb{R}$ by multiplying each element of the vector by a. The multiplicative structure $(\mathbb{R}\ ;\ \circ)$ of the real numbers thus carries over elementwise to real vectors.

$$\mathbf{x} \circ a \;=\; a \circ \mathbf{x} \;=\; a \circ \begin{array}{|c|} \hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array} \;:=\; \begin{array}{|c|} \hline a\,x_1 \\ \hline a\,x_2 \\ \hline \vdots \\ \hline a\,x_n \\ \hline \end{array}$$

The scaling operation is an outer operation on the set $\mathbb{R}$ of real numbers and the set $\mathbb{R}^n$ of n-dimensional real vectors. It has the following properties :

(1) The scaling operation is associative, since for $a, b \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$:

$$(a \circ b) \circ \mathbf{x} \;=\; a \circ (b \circ \mathbf{x})$$

(2) The scaling operation is distributive, since for $a, b \in \mathbb{R}$ and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$(a + b) \circ \mathbf{x} \;=\; a \circ \mathbf{x} \;+\; b \circ \mathbf{x}$$
$$a \circ (\mathbf{x} + \mathbf{y}) \;=\; a \circ \mathbf{x} \;+\; a \circ \mathbf{y}$$

(3) The scaling operation is identitive, since for $1 \in \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$:

$$1 \circ \mathbf{x} \;=\; \mathbf{x}$$

Complete real vector space : The additive domain $(\mathbb{R}^n\ ; +)$ of n-dimensional real vectors is a vector space over the commutative field $(\mathbb{R}\ ; +, \circ)$ of real numbers. This vector space is called the complete n-dimensional real vector space and is designated by $(\mathbb{R}, \mathbb{R}^n\ ; +, \circ)$. It has the defining properties of a vector space specified in Section 3.5.1 :

(1) The domain $(\mathbb{R}^n\ ; +)$ is the infinite set of n-dimensional real vectors with the vector addition $+$ as an inner operation. It is a commutative group.

(2) The domain $(\mathbb{R}\ ; +, \circ)$ is the infinite set of real numbers with the addition $+$ and the multiplication $\circ$ as inner operations. It is a commutative field.

(3) The vector scaling operation $\circ$ is an outer operation on the infinite sets of real numbers and of n-dimensional real vectors. It is associative, distributive and identitive. Thus it has the defining properties of the operation $\circ$ of a vector space.

Real unit vectors and canonical basis : A vector of the complete n-dimensional real vector space is called a canonical unit vector and is designated by $\mathbf{e}_i$ if it contains the unit element 1 of the commutative field $(\mathbb{R}\ ; +, \circ)$ of real numbers at position i and the zero element 0 at all other positions. The set $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n\}$ of canonical unit vectors is called the canonical basis (standard basis) of the complete n-dimensional real vector space. Every n-dimensional real vector $\mathbf{x} \in \mathbb{R}^n$ is a linear combination of the n canonical unit vectors with the elements x_i as coefficients.

$$\mathbf{x} = \sum_{i=1}^{n} x_i \, \mathbf{e}_i = x_1 \begin{array}{|c|}\hline 1 \\ \hline 0 \\ \hline \vdots \\ \hline 0 \\ \hline \end{array} + x_2 \begin{array}{|c|}\hline 0 \\ \hline 1 \\ \hline \vdots \\ \hline 0 \\ \hline \end{array} + \ldots + x_n \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline \vdots \\ \hline 1 \\ \hline \end{array} = \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array}$$

General basis and rank : The canonical basis with the n unit vectors is a special basis of the complete n-dimensional real vector space. According to the general rules in Section 3.5.1, all finite bases of a vector space contain the same number of basis vectors. Thus every basis of the complete n-dimensional real vector space contains exactly n basis vectors. Hence the complete real n-dimensional vector space has rank n.

Coordinates of a real vector in a basis : Let a basis $B = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ and a vector $\mathbf{x}$ of a complete n-dimensional real vector space be given. The vector $\mathbf{x}$ has a unique representation as a linear combination of the basis vectors of B. The coefficients of this linear combination are called the coordinates of the vector $\mathbf{x}$ in the basis B.

$$\mathbf{x} = \sum_{i=1}^{n} a_i \, \mathbf{v}_i = a_1 \begin{array}{|c|}\hline v_{11} \\ \hline v_{21} \\ \hline \vdots \\ \hline v_{n1} \\ \hline \end{array} + a_2 \begin{array}{|c|}\hline v_{12} \\ \hline v_{22} \\ \hline \vdots \\ \hline v_{n2} \\ \hline \end{array} + \ldots + a_n \begin{array}{|c|}\hline v_{1n} \\ \hline v_{2n} \\ \hline \vdots \\ \hline v_{nn} \\ \hline \end{array} = \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array}$$

v_{ki} element k of the basis vector $\mathbf{v}_i$

a_i coordinate of the vector $\mathbf{x}$ for the basis vector $\mathbf{v}_i$

To determine the coordinates a_i of the vector $\mathbf{x}$ in the basis B, the above linear combination is written in rows. With $a_i \mathbf{v}_i = \mathbf{v}_i a_i$, this leads to a system of n linear equations for the unknown coefficients $a_1, a_2, \ldots, a_n$.

$$\begin{array}{lcl} v_{11}\, a_1 + v_{12}\, a_2 + \ldots + v_{1n}\, a_n & = & x_1 \\ v_{21}\, a_1 + v_{22}\, a_2 + \ldots + v_{2n}\, a_n & = & x_2 \\ \vdots & & \\ v_{n1}\, a_1 + v_{n2}\, a_2 + \ldots + v_{nn}\, a_n & = & x_n \end{array}$$

Due to the linear independence of the n basis vectors, this system of equations possesses a unique solution $a_1, a_2, \ldots, a_n$ for the coordinates of the vector $\mathbf{x}$ in the basis B. The coordinates a_i are assembled in a vector $\mathbf{a}$. The vector $\mathbf{a}$ is called the coordinate vector of the vector $\mathbf{x}$ in the basis B.

Example 1 : Real three-dimensional vector space

vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^3$

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix}$$

vector addition $\mathbf{x} + \mathbf{y}$

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \\ 5 \end{bmatrix}$$

vector subtraction $\mathbf{x} + (-\mathbf{y})$

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} + \begin{bmatrix} 1 \\ -1 \\ -4 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \\ -3 \end{bmatrix}$$

vector scaling $2 \circ \mathbf{x} = \mathbf{x} \circ 2$

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \circ 2 = \begin{bmatrix} 2 \\ 4 \\ 2 \end{bmatrix}$$

vector scaling $(-2) \circ \mathbf{y} = \mathbf{y} \circ (-2)$

$$\begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix} \circ (-2) = \begin{bmatrix} 2 \\ -2 \\ -8 \end{bmatrix}$$

canonical basis and linear combination $\mathbf{x} = x_1 \mathbf{e}_1 + x_2 \mathbf{e}_2 + x_3 \mathbf{e}_3$

$$\begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} = 1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Example 2 : Coordinates of a vector

The vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ shown below form a basis of the vector space $\mathbb{R}^3$. The coordinates $\{a_1, a_2, a_3\}$ of the vector $\mathbf{x}$ in this basis are determined by solving the system of equations $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + a_3\mathbf{v}_3 = \mathbf{x}$.

$$\mathbf{v}_1 = \begin{array}{|c|}\hline 4 \\ \hline 1 \\ \hline -1 \\ \hline \end{array} \qquad \mathbf{v}_2 = \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline 1 \\ \hline \end{array} \qquad \mathbf{v}_3 = \begin{array}{|c|}\hline -1 \\ \hline 1 \\ \hline 4 \\ \hline \end{array} \qquad \mathbf{x} = \begin{array}{|c|}\hline 12 \\ \hline 7 \\ \hline 17 \\ \hline \end{array}$$

$$\left.\begin{array}{rcrcrcr} 4a_1 & + & a_2 & - & a_3 & = & 12 \\ a_1 & + & 2a_2 & + & a_3 & = & 7 \\ -a_1 & + & a_2 & + & 4a_3 & = & 17 \end{array}\right] \quad \begin{array}{lcr} a_1 & = & 5 \\ a_2 & = & -2 \\ a_3 & = & 6 \end{array}$$

3.6 LINEAR MAPPINGS

Introduction : The study of the properties of vector spaces requires a definition of the concept of "vector spaces with identical structure". Two vector spaces are identically structured (isomorphic) if there exists a bijective mapping between them which preserves structure (is homomorphic) in both directions. Homomorphic mappings of vector spaces are called linear mappings.

Matrices are introduced to define the mapping rules of linear mappings. Matrices of the same dimension with the inner operation of matrix addition and the outer operation of matrix scaling form a vector space. The composition of linear mappings leads to the concept of matrix multiplication.

Linear mapping : Let $(A, V ; +, \circ)$ and $(A, W ; +, \circ)$ be vector spaces over the same commutative field $(A ; +, \circ)$. The vector sets V and W with their inner and outer operations may be different. A mapping $f : V \rightarrow W$ with $f(v) = w$ is called a linear (homomorphic) mapping from the vector space $(A, V ; +, \circ)$ to the vector space $(A, W ; +, \circ)$ if for all vectors $\mathbf{x}, \mathbf{y} \in V$ and for all $a \in A$ the order in which the operation and the mapping are applied may be changed without changing the result :

inner operation $\quad f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y})$

outer operation $\quad f(a \circ \mathbf{x}) = a \circ f(\mathbf{x})$

The defining properties of a linear mapping imply that the image of a linear combination with coefficients $a_i \in A$ and vectors $\mathbf{x}_i \in V$ is equal to the linear combination of the images $f(\mathbf{x}_i)$ with the same coefficients a_i.

linear combination $\quad f(\sum\limits_{i \in I} a_i \mathbf{x}_i) = \sum\limits_{i \in I} a_i f(\mathbf{x}_i)$

Image space of a linear mapping : A linear mapping $f : V \rightarrow W$ maps a vector space $(A, V ; +, \circ)$ to a vector space $(A, W ; +, \circ)$. Every vector $\mathbf{x} \in V$ is assigned a unique image $f(\mathbf{x}) \in W$. The set of the images $f(\mathbf{x})$ for all $\mathbf{x} \in V$ is called the image of V under f and is designated by f(V). The image of V is a subset of W, since there may be vectors $\mathbf{z} \in W$ which have no preimage in V. The domain $(A, f(V) ; +, \circ)$ is a vector space and is called the image space of $(A, V ; +, \circ)$ with respect to the linear mapping f (the vector space induced by f). Since $f(V) \subseteq W$, the image space $(A, f(V) ; +, \circ)$ is a subspace of the vector space $(A, W ; +, \circ)$.

image $\quad f(V) := \{ f(\mathbf{x}) \mid \mathbf{x} \in V \}$

image space $\quad (A, f(V) ; +, \circ)$

Linear mapping rule : Let B be a basis of a vector space (A, V ; +, ∘) of rank m with the basis vectors $\mathbf{v}_1,...,\mathbf{v}_m$. Then every vector $\mathbf{x} \in$ V has a unique representation as a linear combination of the basis vectors.

$$B = \{ \mathbf{v}_1,...,\mathbf{v}_m \} \qquad \mathbf{x} = \sum_{i=1}^{m} a_i \, \mathbf{v}_i$$

A linear mapping f : V → W maps the basis B to an image f(B) consisting of the images $f(\mathbf{v}_1),...,f(\mathbf{v}_m)$ of the basis vectors. Since the mapping f is linear, the image $f(\mathbf{x})$ is determined as a linear combination of the images of the basis vectors.

$$f(B) = \{ f(\mathbf{v}_1),...,f(\mathbf{v}_m) \} \qquad f(\mathbf{x}) = \sum_{i=1}^{m} a_i \, f(\mathbf{v}_i)$$

Hence a linear mapping f : V → W is uniquely determined by the images of the basis vectors of a basis B ⊆ V. The image f(B) spans the image space (A, f(V); +, ∘), since every element $f(\mathbf{x})$ of the image f(V) is a linear combination of the vectors of f(B). The image f(B) is, however, not necessarily a basis, since the vectors of f(B) may be linearly dependent.

Image basis of a linear mapping : Let (A, V ; +, ∘) be a vector space, and let f : V → W be a linear mapping which induces the image space (A, f(V); +, ∘). A basis F ⊆ f(V) of the image space is called an image basis of the linear mapping. According to the definition in Section 3.5.1, the rank of the image space is equal to the number of vectors in the image basis.

Construction of an image basis : Let the image f(B) of a basis B of a vector space (A, V ; +, ∘) under the linear mapping f : V → W be given. Then f(B) spans the image space f(V). An image basis F may be constructed from the generating set f(B) using the method described in Section 3.5.1.

Defect of a linear mapping : Let a vector space (A, V ; +, ∘) of rank m be mapped to an image space of rank r by a linear mapping f : V → W. Then the construction of the image basis yields $r \le m$. The difference $d = m - r$ is called the defect of the linear mapping f.

Types of linear mappings : Let a vector space (A, V ; +, ∘) of rank m and a vector space (A, W ; +, ∘) of rank n be given, along with a linear mapping f : V → W which induces an image space (A, f(V); +, ∘) of rank r. A linear mapping is injective, surjective or bijective, respectively, if and only if :

(1) f is injective ⇔ $r = m$

(2) f is surjective ⇔ $r = n$

(3) f is bijective ⇔ $r = m = n$

Proof : Types of linear mappings

The vector space (A, V ; +, ∘) has a basis B which contains exactly m linearly independent basis vectors $\mathbf{v}_1,...,\mathbf{v}_m \in V$. Every basis vector $\mathbf{v}_i \in V$ is mapped to an image $f(\mathbf{v}_i) \in W$. Since the image space is of rank r, exactly r images of the basis vectors are linearly independent. They form an image basis F of the image space (A, f(V) ; +, ∘). Let the basis vectors $\mathbf{v}_1,...,\mathbf{v}_r$ be chosen such that their images $f(\mathbf{v}_1),...,f(\mathbf{v}_r)$ form an image basis.

basis	$B = \{\mathbf{v}_1,...,\mathbf{v}_r,...,\mathbf{v}_m\} \subseteq V$
image basis	$F = \{f(\mathbf{v}_1),...,f(\mathbf{v}_r)\} \subseteq f(V) \subseteq W$

(1) Let the mapping f be injective. Then two different vectors $\mathbf{x} \neq \mathbf{y}$ of V have two different images $f(\mathbf{x}) \neq f(\mathbf{y})$ in W. The zero vector $\mathbf{0}_V \in V$ is mapped to the zero vector $\mathbf{0}_W \in W$. Every basis vector $\mathbf{v}_i \neq \mathbf{0}_V$ is mapped to an image $f(\mathbf{v}_i) \neq \mathbf{0}_W$. Every linear combination $\sum a_i \mathbf{v}_i$ with coefficients $a_i \neq 0$ differs from the zero vector $\mathbf{0}_V$. Its image is the linear combination $\sum a_i f(\mathbf{v}_i) \neq \mathbf{0}_W$. Therefore the images $f(\mathbf{v}_i)$ of all basis vectors $\mathbf{v}_i$ are linearly independent, and hence the rank r is equal to the rank m.

Conversely, let the rank r be equal to the rank m. Then the images $f(\mathbf{v}_i) \in W$ of all basis vectors $\mathbf{v}_i \in V$ are linearly independent. This implies that for two different vectors $\mathbf{x} = \sum a_i \mathbf{v}_i$ and $\mathbf{y} = \sum b_i \mathbf{v}_i$ of V the images $\sum a_i f(\mathbf{v}_i)$ and $\sum b_i f(\mathbf{v}_i)$ are also different. Hence the mapping f is injective.

(2) Let the mapping f be surjective. Then every vector $\mathbf{z} \in W$ has a preimage $\mathbf{x} \in V$ with $\mathbf{z} = f(\mathbf{x})$, and hence f(V) = W. Every image basis F is a basis of the image space (A, f(V) ; +, ∘), and thus a basis of the vector space (A, W ; +, ∘). Hence the rank r is equal to the rank n.

Conversely, let the rank r be equal to the rank m. Then every basis of the vector space (A, W ; +, ∘) is an image basis of the image space (A, f(V) ; +, ∘). Two vector spaces with the same basis are identical. Thus W = f(V), and hence the mapping f is surjective.

(3) A mapping is bijective if it is both injective and surjective. By (1) and (2), this is the case if and only the rank r is equal to the rank m and to the rank n.

Isomorphic vector spaces : Two vector spaces (A, V ; +, ∘) and (A, W ; +, ∘) are said to be isomorphic if there is a bijective linear mapping $f : V \to W$. Isomorphic vector spaces have the same algebraic structure. They differ only in the meaning of the set of vectors and of the vector operations. Two vector spaces of finite rank over the same domain (A ; +, ∘) are isomorphic if and only if they are of equal rank. Every vector space ($\mathbb{R}$, V ; +, ∘) of rank n over the field ($\mathbb{R}$; +, ∘) of real numbers is isomorphic to the complete n-dimensional real vector space ($\mathbb{R}$, $\mathbb{R}^n$; +, ∘).

Example 1 : Complete two-dimensional real vector space

The vector space $(\mathbb{R}, \mathbb{C}\,; +, \circ)$ for the addition of complex numbers over the field $(\mathbb{R}\,; +, \circ)$ of real numbers is treated in Example 1 of Section 3.5.1. It is isomorphic to the complete two-dimensional real vector space $(\mathbb{R}, \mathbb{R}^2\,; +, \circ)$, since there is a bijective linear mapping $f : \mathbb{C} \rightarrow \mathbb{R}^2$.

(1) Every complex number $x_r + i\,x_i \in \mathbb{C}$ corresponds to a pair $(x_r, x_i) \in \mathbb{R}^2$ with the real part $x_r \in \mathbb{R}$ and the imaginary part $x_i \in \mathbb{R}$. This pair corresponds to a two-dimensional real vector. This one-to-one correspondence is a bijective mapping $f : \mathbb{C} \rightarrow \mathbb{R}^2$.

$$f : \mathbb{C} \rightarrow \mathbb{R}^2$$

$$f(\mathbf{x}) = \mathbf{r} \qquad \mathbf{x} := x_r + i\,x_i \in \mathbb{C} \qquad \mathbf{r} = \begin{array}{|c|}\hline x_r \\ \hline x_i \\ \hline \end{array} \in \mathbb{R}^2$$

(2) The mapping f has the defining properties of a linear mapping :

$$f(\mathbf{x} + \mathbf{y}) = f(\mathbf{x}) + f(\mathbf{y}) = \begin{array}{|c|}\hline x_r \\ \hline x_i \\ \hline \end{array} + \begin{array}{|c|}\hline y_r \\ \hline y_i \\ \hline \end{array} = \begin{array}{|c|}\hline x_r + y_r \\ \hline x_i + y_i \\ \hline \end{array}$$

$$f(a \circ x) = a \circ f(\mathbf{x}) = \begin{array}{|c|}\hline x_r \\ \hline x_i \\ \hline \end{array} \circ a = \begin{array}{|c|}\hline a\ x_r \\ \hline a\ x_i \\ \hline \end{array}$$

(3) Since the mapping f is bijective and linear, every basis $B \subseteq \mathbb{C}$ corresponds to a basis $f(B) \in \mathbb{R}^2$. This is demonstrated for two examples :

$$B = \{1, i\} \subseteq \mathbb{C} \qquad f(B) = \left\{ \begin{array}{|c|}\hline 1 \\ \hline 0 \\ \hline \end{array}, \begin{array}{|c|}\hline 0 \\ \hline 1 \\ \hline \end{array} \right\} \subseteq \mathbb{R}^2$$

$$B = \{1 + i, 1 - i\} \subseteq \mathbb{C} \qquad f(B) = \left\{ \begin{array}{|c|}\hline 1 \\ \hline 1 \\ \hline \end{array}, \begin{array}{|c|}\hline 1 \\ \hline -1 \\ \hline \end{array} \right\} \subseteq \mathbb{R}^2$$

Matrix of a linear mapping : Let a vector space $(A, V\,; +, \circ)$ of rank m, a vector space $(A, W\,; +, \circ)$ of rank n and a linear mapping $f : V \rightarrow W$ be given. Every basis vector $\mathbf{v}_i \in V$ of a basis $B_V \subseteq V$ is mapped to an image $f(\mathbf{v}_i) \in W$. Then every image vector $f(\mathbf{v}_i)$ has a unique representation as a linear combination of basis vectors $\mathbf{w}_k \in W$ of a basis $B_W \subseteq W$ with coefficients $a_{ik} \in A$:

$$f(\mathbf{v}_i) = \sum_{k=1}^{n} a_{ik}\, \mathbf{w}_k \qquad i = 1,...,m$$

The coefficients a_{ik} are arranged in a rectangular scheme, using the first index as a row index and the second index as a column index. This scheme is called a matrix scheme. The coefficient a_{ik} appears in row i and column k of the matrix scheme.

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|l}
\hline
a_{11} & \cdots & a_{1k} & \cdots & a_{1n} & \leftarrow 1 \\ \hline
\vdots & & \vdots & & \vdots & \\ \hline
a_{i1} & \cdots & a_{ik} & \cdots & a_{in} & \leftarrow i \\ \hline
\vdots & & \vdots & & \vdots & \\ \hline
a_{m1} & \cdots & a_{mk} & \cdots & a_{mn} & \leftarrow m \\ \hline
\multicolumn{1}{c}{\uparrow} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{\uparrow} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{\uparrow} & \\
\multicolumn{1}{c}{1} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{k} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{n} &
\end{array}$$

row dimension	:	m	column dimension	:	n
row index	:	$i = 1, 2, \ldots, m$	column index	:	$k = 1, 2, \ldots, n$
row i	:	$(a_{i1}, \ldots, a_{in})$	column k	:	$(a_{1k}, \ldots, a_{mk})$

An $m \cdot n$-tuple with elements $a_{ik} \in A$ is called a matrix and is designated by an uppercase letter in boldface. The set of all matrices with row dimension m and column dimension n is the $m \cdot n$-fold direct product $A^m \times A^n$, which is designated by $A_{m,n}$. A matrix **A** is said to be quadratic if its row dimension is equal to its column dimension. A matrix is said to be real if its elements are real numbers.

Matrix addition : Two matrices $\mathbf{A}, \mathbf{B} \in A_{m,n}$ are added by adding elements that have the same indices. The additive structure $(A\,;+)$ thus carries over to matrices elementwise.

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \qquad \text{with } c_{ik} := a_{ik} + b_{ik} \qquad i = 1,\ldots,m\,;\, k = 1,\ldots,n$$

The addition $+$ is an inner operation in the set $A_{m,n}$ of matrices. The additive domain possesses the defining properties of a commutative group specified in Section 3.3 :

(1) The addition is associative, since for $\mathbf{A}, \mathbf{B}, \mathbf{C} \in A_{m,n}$:
$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C})$

(2) The addition is commutative, since for $\mathbf{A}, \mathbf{B} \in A_{m,n}$:
$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A}$

(3) The zero matrix **0**, whose elements are all 0, acts as an identity element under addition, since for $\mathbf{0}, \mathbf{A} \in A_{m,n}$:
$\mathbf{A} + \mathbf{0} = \mathbf{0} + \mathbf{A} = \mathbf{A}$

(4) For every matrix **A** there exists an additive inverse $-\mathbf{A}$ such that $\mathbf{A} + (-\mathbf{A}) = \mathbf{0}$. If the elements of the matrix **A** are $a_{ik} \in A$, then the elements of the matrix $-\mathbf{A}$ are the additive inverses $-a_{ik} \in A$.

Matrix scaling : A matrix $\mathbf{A} \in A_{m,n}$ is scaled by a number $a \in A$ by multiplying each element of the matrix by a. The algebraic structure (A ; ∘) thus carries over to matrices elementwise.

$$\mathbf{B} = a \circ \mathbf{A} = \mathbf{A} \circ a \quad \text{with } b_{ik} = a\, a_{ik} \qquad i = 1,...,m \,;\, k = 1,...,n$$

The scaling operation is an outer operation on the set A and the set $A_{m,n}$ of matrices. It has the following properties :

(1) The scaling operation is associative, since for $a, b \in A$ and $\mathbf{A} \in A_{m,n}$:
$$(a \circ b) \circ \mathbf{A} = a \circ (b \circ \mathbf{A})$$

(2) The scaling operation is distributive, since for $a, b \in A$ and $\mathbf{A}, \mathbf{B} \in A_{m,n}$:
$$(a + b) \circ \mathbf{A} = a \circ \mathbf{A} + b \circ \mathbf{A}$$
$$a \circ (\mathbf{A} + \mathbf{B}) = a \circ \mathbf{A} + a \circ \mathbf{B}$$

(3) The scaling operation is identitive, since for $1 \in A$ and $\mathbf{A} \in A_{m,n}$:
$$1 \circ \mathbf{A} = \mathbf{A}$$

Vector space of matrices : The additive domain $(A_{m,n}\,;+)$ of matrices with row dimension m and column dimension n over the commutative field (A ; + , ∘) is designated by $(A, A_{m,n}\,; +, \circ)$. It is a vector space of rank m • n which has the defining properties required in Section 3.5.1 :

(1) The domain $(A_{m,n}\,;+)$ is a commutative group.

(2) The domain (A; + , ∘) is a commutative field.

(3) The outer operation ∘ on the sets A and $A_{m,n}$ is the scaling operation. It is associative, distributive and identitive.

For each $j = 1,...,m$ and each $q = 1,...,n$ the canonical basis of the vector space $(A, A_{m,n}\,; +, \circ)$ contains the matrix $\mathbf{A}_{jq}$, whose elements a_{ik} are 1 for $i = j$ and $k = q$ and 0 otherwise. Thus the vector space has rank m • n.

Example 2 : Linear mapping

Let a vector space $(\mathbb{R}, V\,; +, \circ)$ of rank 2 with the basis vectors $\mathbf{v}_1, \mathbf{v}_2 \in V \subset \mathbb{R}^3$ and a vector space $(\mathbb{R}, W\,; +, \circ)$ of rank 2 with the basis vectors $\mathbf{w}_1, \mathbf{w}_2 \in W \subset \mathbb{R}^4$ be given. The two vector spaces are isomorphic. The bijective mapping $f : V \rightarrow W$ is defined by a quadratic matrix $\mathbf{A} \in \mathbb{R}_{2,2}$.

$$\mathbf{v}_1 = \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} \qquad \mathbf{v}_2 = \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} \qquad \mathbf{A} = \begin{bmatrix} 3 & -2 \\ 2 & 1 \end{bmatrix}$$

$$\mathbf{w}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{w}_2 = \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix}$$

To determine the image $f(\mathbf{x}) \in V$ of a given vector $\mathbf{x} \in V$, the vector $\mathbf{x}$ is first expressed as a linear combination of the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ with the coefficients a_1 and a_2. Then the images of the basis vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ are calculated using the matrix $\mathbf{A}$. The image $f(\mathbf{x})$ is the linear combination of the images $f(\mathbf{v}_1)$ and $f(\mathbf{v}_2)$ with the coefficients a_1 and a_2.

$$\mathbf{v}_1 a_1 + \mathbf{v}_2 a_2 = \mathbf{x}$$

$$\begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix} a_1 + \begin{bmatrix} -2 \\ 1 \\ 1 \end{bmatrix} a_2 = \begin{bmatrix} -5 \\ 2 \\ 6 \end{bmatrix} \quad \rightarrow \quad \begin{array}{l} a_1 = 1 \\ a_2 = 3 \end{array}$$

$$f(\mathbf{v}_1) = 3 \circ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} - 2 \circ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 5 \\ 3 \\ -2 \end{bmatrix}$$

$$f(\mathbf{v}_2) = 2 \circ \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix} + 1 \circ \begin{bmatrix} 1 \\ -1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \\ 2 \\ 1 \end{bmatrix}$$

$$f(\mathbf{x}) = 1 \circ \begin{bmatrix} 1 \\ 5 \\ 3 \\ -2 \end{bmatrix} + 3 \circ \begin{bmatrix} 3 \\ 1 \\ 2 \\ 1 \end{bmatrix} = \begin{bmatrix} 10 \\ 8 \\ 9 \\ 1 \end{bmatrix}$$

Composition of linear mappings : Let the vector spaces $(A, U; +, \circ)$, $(A, V; +, \circ)$ and $(A, W; +, \circ)$ with ranks m, n, s and basis vectors $\mathbf{u}_i \in U$, $\mathbf{v}_k \in V$, $\mathbf{w}_r \in W$ be given, as well as mappings $f : U \rightarrow V$ and $g : V \rightarrow W$. Every image $f(\mathbf{u}_i) \in V$ has a unique representation as a linear combination of the basis vectors $\mathbf{v}_k \in V$ with coefficients $a_{ik} \in A$. Every image $g(\mathbf{v}_k) \in W$ has a unique representation as a linear combination of the basis vectors $\mathbf{w}_r \in W$ with coefficients b_{kr}.

$$f(\mathbf{u}_i) = \sum_{k=1}^{n} a_{ik}\,\mathbf{v}_k \qquad i = 1,...,m$$

$$g(\mathbf{v}_k) = \sum_{r=1}^{s} b_{kr}\,\mathbf{w}_r \qquad k = 1,...,n$$

The composition of the linear mappings g and f is a linear mapping $g \circ f : U \rightarrow W$. Every image $g(f(\mathbf{u}_i)) \in W$ has a unique representation as a linear combination of the basis vectors $\mathbf{w}_r \in W$ with coefficients $c_{ir} \in A$.

$$g(f(\mathbf{u}_i)) = \sum_{k=1}^{n} a_{ik}\, g(\mathbf{v}_k) = \sum_{k=1}^{n} \sum_{r=1}^{s} a_{ik}\, b_{kr}\, \mathbf{w}_r$$

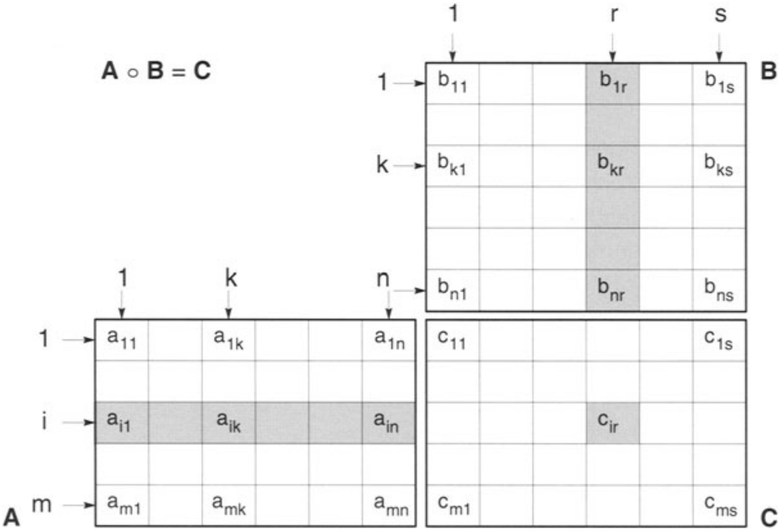
$$g(f(\mathbf{u}_i)) = \sum_{r=1}^{s} c_{ir}\, \mathbf{w}_r \qquad c_{ir} = \sum_{k=1}^{n} a_{ik}\, b_{kr}$$

Matrix multiplication : The matrix **C** for a composition $g \circ f : U \rightarrow W$ is called the product of the matrices **A** and **B** for the mappings $f : U \rightarrow V$ and $g : V \rightarrow W$. The product of **A** and **B** is designated by $\mathbf{A} \circ \mathbf{B}$. Often the symbol $\circ$ is dropped and the product is designated by **AB**. The product of **A** and **B** is only defined if the column dimension of **A** and the row dimension of **B** coincide.

$$\mathbf{C} = \mathbf{A} \circ \mathbf{B} \qquad \mathbf{A} \in A_{m,n} \quad \mathbf{B} \in A_{n,s} \quad \mathbf{C} \in A_{m,s}$$

$$c_{ir} := \sum_{k=1}^{n} a_{ik}\, b_{kr} \qquad i = 1,...,m\,; \quad r = 1,...,s$$

Matrix multiplication allows a convenient graphical representation. The following example shows the calculation of the coefficient c_{ir} of the matrix **C** using row i of the matrix **A** and column r of the matrix **B**. Coefficients a_{ik} and b_{kr} with the same index k are multiplied. The products are summed.



In the general case, the multiplication of two matrices is an outer operation on the sets $A_{m,n}$ and $A_{n,r}$ with the target $A_{m,r}$. In the case of quadratic matrices with $m = n = r$, it is an inner operation in the set $A_{m,m}$. The multiplicative domain $(A_{m,m} ; \circ)$ has the defining properties of a semigroup with identity element specified in Section 3.3.

(1) The multiplication is associative, since for **A**, **B**, **C** :

$$(\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C} = \mathbf{A} \circ (\mathbf{B} \circ \mathbf{C})$$

(2) The identity matrix **I**, whose elements a_{ij} are 1 for $i = j$ and 0 for $i \neq j$, acts as the identity element under multiplication, since for $\mathbf{I}, \mathbf{A} \in A_{m,m}$:

$$\mathbf{A} \circ \mathbf{I} = \mathbf{I} \circ \mathbf{A} = \mathbf{A}$$

Ring of matrices : The additive and multiplicative domain $(A_{m,m} ; +, \circ)$ of quadratic matrices is a ring with the defining properties specified in Section 3.4.2 :

(1) The domain $(A_{m,m} ; +)$ is a commutative group.

(2) The domain $(A_{m,m} ; \circ)$ is a semigroup.

(3) The multiplication $\circ$ is distributive with respect to the addition $+$, since for $\mathbf{A}, \mathbf{B}, \mathbf{C} \in A_{m,m}$:

$$\mathbf{A} \circ (\mathbf{B} + \mathbf{C}) = \mathbf{A} \circ \mathbf{B} + \mathbf{A} \circ \mathbf{C}$$

$$(\mathbf{A} + \mathbf{B}) \circ \mathbf{C} = \mathbf{A} \circ \mathbf{C} + \mathbf{B} \circ \mathbf{C}$$

The ring ($A_{m,m}$; $+$, $\circ$) is a ring with identity element, since the identity matrix $\mathbf{I}$ acts as an identity element under the multiplication $\circ$. However, the ring contains zero divisors, since $\mathbf{A} \circ \mathbf{B} = \mathbf{0}$ does not imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$.

Example 3 : Composition of linear mappings

Let the real vector spaces of dimensions 2,3,4 with the sets $\mathbb{R}^2$, $\mathbb{R}^3$, $\mathbb{R}^4$ of vectors be given. Let the matrix of the linear mapping $f : \mathbb{R}^2 \to \mathbb{R}^3$ be $\mathbf{A} \in \mathbb{R}_{2,3}$, and let the matrix of the linear mapping $g : \mathbb{R}^3 \to \mathbb{R}^4$ be $\mathbf{B} \in \mathbb{R}_{3,4}$. The matrix $\mathbf{C} \in \mathbb{R}_{2,4}$ of the composition $g \circ f : \mathbb{R}^2 \to \mathbb{R}^4$ is calculated as the product $\mathbf{C} = \mathbf{A} \circ \mathbf{B}$.

$$\begin{array}{c|c|c|c||c|c|c|c|l}
\cline{5-8}
 & \multicolumn{3}{c||}{} & 1 & -1 & 1 & 2 & \mathbf{B} \\
\cline{5-8}
 & \multicolumn{3}{c||}{} & -1 & 2 & 1 & 1 & \\
\cline{5-8}
 & \multicolumn{3}{c||}{} & 0 & 1 & -1 & 0 & \\
\hline
 & 4 & -2 & 2 & 6 & -6 & 0 & 6 & \\
\cline{2-8}
\mathbf{A} & 1 & 3 & 0 & -2 & 5 & 4 & 5 & \mathbf{C} = \mathbf{A} \circ \mathbf{B} \\
\cline{2-8}
\end{array}$$

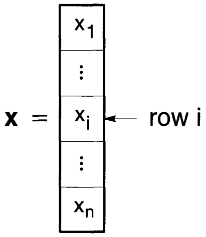
3.7 VECTOR AND MATRIX ALGEBRA

Introduction : Vector and matrix algebra has many applications in the formulation and solution of geometric, physical and technical problems. It developed out of the theory of systems of linear equations. The basic theory is treated in a general form in the preceding sections on vector spaces and linear mappings. This section builds on that basis and presents the elementary definitions, operations and rules for real and complex vectors and matrices which are important in practical applications.

3.7.1 DEFINITIONS

Scalar : A quantity is said to be scalar if it is described by exactly one value. Real and complex scalars are considered in the following.

Vector : An n-tuple of elements is called a vector of dimension n and is designated by a lowercase letter in boldface. A vector is said to be real if all its elements are real. A vector is said to be complex if all its elements are complex. The elements x_i of a vector **x** are arranged in a column by using the index of the element as a row index. This representation is called a column vector.



Special vectors : A vector is called a zero vector and is designated by **0** if all of its elements are 0. A vector is called a canonical unit vector and is designated by $\mathbf{e}_k$ if the element x_k is 1 and all other elements are 0. Canonical unit vectors are represented using the Kronecker symbol δ_{ik}, which has the value 0 for $i \neq k$ and the value 1 for $i = k$.

zero vector **0** : $x_i = 0$ $\qquad i = 1,\ldots,n$

unit vector $\mathbf{e}_k$: $x_i = \delta_{ik}$

Kronecker symbol : $\delta_{ik} = 1 \quad \text{for} \quad k = i$

$\delta_{ik} = 0 \quad \text{for} \quad k \neq i$

Matrix : An m • n-tuple of elements is called a matrix with the dimensions m, n and is designated by an uppercase letter in boldface. A matrix is said to be real if all of its elements are real. A matrix is said to be complex if all of its elements are complex. The elements a_{ik} of a matrix **A** with the dimensions m, n are arranged in a rectangular scheme with m rows and n columns by using the first index as a row index and the second index as a column index. The elements a_{ik} are called diagonal elements for i = k and non-diagonal elements for i ≠ k. A matrix is said to be quadratic if the row dimension and the column dimension coincide.

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|l} \hline a_{11} & \cdots & a_{1k} & \cdots & a_{1n} & \\ \hline \vdots & & \vdots & & \vdots & \\ \hline a_{i1} & \cdots & a_{ik} & \cdots & a_{in} & \leftarrow \text{row i} \\ \hline \vdots & & \vdots & & \vdots & \\ \hline a_{m1} & \cdots & a_{mk} & \cdots & a_{mn} & \\ \hline \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{\uparrow \text{column k}} & \multicolumn{1}{c}{} & \multicolumn{1}{c}{} & \end{array}$$

m	row dimension
n	column dimension
m = n	quadratic matrix
a_{kk}	diagonal element
a_{ik}	non-diagonal element for i ≠ k

Special quadratic matrices : Some designations for matrices with special patterns of elements are defined in the following. In the graphical schemes, zero elements are represented by empty squares and elements with an arbitrary value are represented by shaded squares.

zero matrix :

$$\mathbf{0} = \begin{array}{|c|c|c|c|c|} \hline 0 & \cdots & 0 & \cdots & 0 \\ \hline \vdots & & \vdots & & \vdots \\ \hline 0 & \cdots & 0 & \cdots & 0 \\ \hline \vdots & & \vdots & & \vdots \\ \hline 0 & \cdots & 0 & \cdots & 0 \\ \hline \end{array} \qquad x_{ik} := 0$$

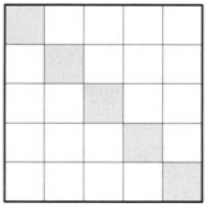
A matrix is called a zero matrix if all of its elements are 0.

identity matrix :

$$\mathbf{I} = \begin{array}{|c|c|c|c|c|}\hline 1 & & & & \\ \hline & \ddots & & & \\ \hline & & 1 & & \\ \hline & & & \ddots & \\ \hline & & & & 1 \\ \hline \end{array} \qquad x_{ik} := \delta_{ik}$$

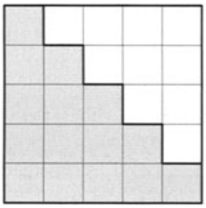
A matrix is called an identity matrix if all of its diagonal elements are 1 and all of its non-diagonal elements are 0.

diagonal matrix :

$\mathbf{D} =$  $x_{ik} := 0 \quad \text{for} \quad i \neq k$

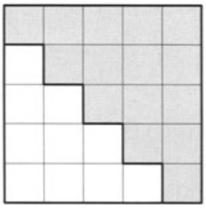
A matrix is called a diagonal matrix if all of its non-diagonal elements are 0.

lower triangular matrix :

$\mathbf{L} =$  $x_{ik} := 0 \quad \text{for} \quad i < k$

A matrix is called a lower triangular matrix if all elements above its diagonal are 0.

upper triangular matrix :

$\mathbf{R} =$  $x_{ik} := 0 \quad \text{for} \quad i > k$

A matrix is called an upper triangular matrix if all elements below its diagonal are 0.

3.7.2 ELEMENTARY VECTOR OPERATIONS

Equality : Two vectors **u** and **v** are equal if they have the same dimension n and elements with the same indices are equal.

$$\mathbf{u} = \mathbf{v} \quad :\Leftrightarrow \quad \bigwedge_{i=1}^{n} (u_i = v_i)$$

Addition and subtraction : Two vectors **u** and **v** can only be added or subtracted if they have the same dimension n. The addition + and the subtraction − are carried out by adding and subtracting, respectively, the elements with identical indices.

$$\begin{array}{lll} \mathbf{w} = \mathbf{u} + \mathbf{v} & \qquad w_i := u_i + v_i & \qquad i = 1,...,n \\ \mathbf{w} = \mathbf{u} - \mathbf{v} & \qquad w_i := u_i - v_i & \qquad i = 1,...,n \end{array}$$

Scaling : A vector **u** is scaled by a scalar c by multiplying each element of the vector by c.

$$\mathbf{v} = c\mathbf{u} \qquad v_i := cu_i \qquad i = 1,...,n$$

Rules for vector operations : The rules for the elementary vector operations follow from the theory of vector spaces :

(1) Vector addition is associative and commutative.

$$(\mathbf{u}+\mathbf{v}) + \mathbf{w} = \mathbf{u}+(\mathbf{v}+\mathbf{w})$$
$$\mathbf{u}+\mathbf{v} = \mathbf{v}+\mathbf{u}$$

(2) Vector scaling is associative, and it is distributive with respect to addition.

$$\begin{array}{lcl} (ab)\mathbf{u} & = & a(b\mathbf{u}) \\ (a+b)\mathbf{u} & = & a\mathbf{u}+b\mathbf{u} \\ a(\mathbf{u}+\mathbf{v}) & = & a\mathbf{u}+a\mathbf{v} \end{array}$$

(3) For the zero vector **0** and the scalars 0 and 1 :

$$\begin{array}{lll} \mathbf{u}+\mathbf{0} = \mathbf{u} & \qquad 1\mathbf{u} = \mathbf{u} & \\ \mathbf{u}-\mathbf{u} = \mathbf{0} & \qquad 0\mathbf{u} = \mathbf{0} & \qquad a\mathbf{0} = \mathbf{0} \end{array}$$

From $a\mathbf{u} = \mathbf{0}$ it follows that $a = 0$ or $\mathbf{u} = \mathbf{0}$.

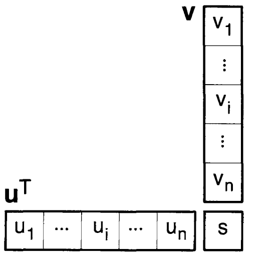
Scalar product : The scalar product of two vectors **u** and **v** can only be formed if the two vectors have the same dimension n. The scalar product is designated by $\mathbf{u} \circ \mathbf{v}$ or by $\mathbf{u}^T\mathbf{v}$. The result of a scalar product is a scalar s, which is calculated as follows :

$$s = \mathbf{u} \circ \mathbf{v} := \mathbf{u}^T\mathbf{v} \qquad s := \sum_{i=1}^{n} u_i v_i$$

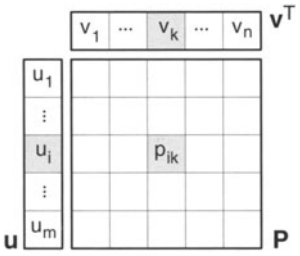
Dyadic product : The dyadic product of a vector **u** of dimension m and a vector **v** of dimension n is a matrix **P** of row dimension m and column dimension n. It is designated by $\mathbf{u}\,\mathbf{v}^T$ and is calculated as follows :

$$\mathbf{P} = \mathbf{u}\,\mathbf{v}^T \qquad p_{ik} := u_i\, v_k \qquad i = 1,...,m \;;\; k = 1,...,n$$

Graphical representation : The scalar product $\mathbf{u}^T\mathbf{v}$ and the dyadic product $\mathbf{u}\,\mathbf{v}^T$, which are special cases of the general matrix product, may be represented graphically. The vectors $\mathbf{u}^T$ and $\mathbf{v}^T$ are called transposed vectors of **u** and **v**. The elements of a transposed vector are arranged in a row by using the index of the element as a column index. Transposed vectors are therefore also called row vectors. The following example shows the graphical representation of the calculation of the scalar product and the dyadic product.



scalar product $s = \mathbf{u}^T\mathbf{v}$



dyadic product $\mathbf{P} = \mathbf{u}\,\mathbf{v}^T$

Rules for scalar products : The definition of the scalar product implies the following rules for real vectors :

(1) The scalar product is commutative.

$$\mathbf{u} \circ \mathbf{v} = \mathbf{v} \circ \mathbf{u}$$

(2) The scalar product is distributive with respect to vector addition.

$$\mathbf{u} \circ (\mathbf{v} + \mathbf{w}) = \mathbf{u} \circ \mathbf{v} + \mathbf{u} \circ \mathbf{w}$$

(3) The scalar product is associative with respect to vector scaling.

$$(c\mathbf{u}) \circ \mathbf{v} = c(\mathbf{u} \circ \mathbf{v})$$

(4) The scalar product $\mathbf{u} \circ \mathbf{u}$ is positive for $\mathbf{u} \neq \mathbf{0}$.

$$\mathbf{u} \circ \mathbf{u} > 0 \quad \text{for} \quad \mathbf{u} \neq \mathbf{0}$$

Orthogonal and orthonormal vectors : Two real non-zero vectors **u** and **v** are said to be orthogonal if the scalar product $\mathbf{u} \circ \mathbf{v}$ is zero. They are said to be orthonormal if the scalar product $\mathbf{u} \circ \mathbf{v}$ is zero and the scalar products $\mathbf{u} \circ \mathbf{u}$ and $\mathbf{v} \circ \mathbf{v}$ are one.

u and **v** are orthogonal $\;:\Leftrightarrow\; \mathbf{u} \circ \mathbf{v} = 0 \;\wedge\; \mathbf{u} \neq \mathbf{0} \;\wedge\; \mathbf{v} \neq \mathbf{0}$

u and **v** are orthonormal $\;:\Leftrightarrow\; \mathbf{u} \circ \mathbf{v} = 0 \;\wedge\; \mathbf{u} \circ \mathbf{u} = 1 \;\wedge\; \mathbf{v} \circ \mathbf{v} = 1$

Examples : Scalar products

Let real vectors **x**, **y** of the three-dimensional space $\mathbb{R}^3$ be given :

$$\mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} -1 \\ 1 \\ 4 \end{bmatrix}$$

The scalar products $\mathbf{x} \circ \mathbf{x}$, $\mathbf{y} \circ \mathbf{y}$, $\mathbf{x} \circ \mathbf{y}$ are calculated as follows :

$$\begin{aligned}
\mathbf{x} \circ \mathbf{x} &= \mathbf{x}^T\mathbf{x} = 1 \cdot 1 + 2 \cdot 2 + 1 \cdot 1 = 6 \\
\mathbf{y} \circ \mathbf{y} &= \mathbf{y}^T\mathbf{y} = (-1) \cdot (-1) + 1 \cdot 1 + 4 \cdot 4 = 18 \\
\mathbf{x} \circ \mathbf{y} &= \mathbf{x}^T\mathbf{y} = 1 \cdot (-1) + 2 \cdot 1 + 1 \cdot 4 = 5
\end{aligned}$$

The positive square root of the scalar product $\mathbf{x} \circ \mathbf{x}$ is called the length of the vector **x**. It is a norm (see Section 3.7.4) and is designated by $\|\mathbf{x}\|$. The lengths of the vectors **x** and **y** are calculated as follows :

$$\begin{aligned}
\|\mathbf{x}\| &= \sqrt{\mathbf{x} \circ \mathbf{x}} = \sqrt{6} = 2.449 \\
\|\mathbf{y}\| &= \sqrt{\mathbf{y} \circ \mathbf{y}} = \sqrt{18} = 4.243
\end{aligned}$$

A vector is normalized with respect to its length by scaling it with the reciprocal of its length. The normalized vectors $\mathbf{x}_N$, $\mathbf{y}_N$ are calculated as follows :

$$\mathbf{x}_N = \mathbf{x} / \|\mathbf{x}\| = \frac{1}{\sqrt{6}}\mathbf{x} = \begin{bmatrix} 0.408 \\ 0.816 \\ 0.408 \end{bmatrix} \qquad \mathbf{x}_N \circ \mathbf{x}_N = 1$$

$$\mathbf{y}_N = \mathbf{y} / \|\mathbf{y}\| = \frac{1}{\sqrt{18}}\mathbf{y} = \begin{bmatrix} -0.236 \\ 0.236 \\ 0.943 \end{bmatrix} \qquad \mathbf{y}_N \circ \mathbf{y}_N = 1$$

The angle γ between the vectors **x** and **y** is calculated from the scalar products according to the cosine rule of trigonometry :

$$\cos\gamma = \frac{\mathbf{x} \circ \mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|} = \frac{\mathbf{x} \circ \mathbf{y}}{\sqrt{(\mathbf{x} \circ \mathbf{x})(\mathbf{y} \circ \mathbf{y})}} \qquad 0 \le \gamma \le \pi$$

$$\cos\gamma = \frac{5}{\sqrt{6 \cdot 18}} = 0.481$$

$$\gamma = \text{arc cos } 0.481 = 1.069$$

From two vectors **x** and **y**, a linear combination $\mathbf{z} \neq \mathbf{0}$ is to be formed which is orthogonal to **x**, so that $\mathbf{x} \circ \mathbf{z} = 0$. It is assumed that $\mathbf{x} \circ \mathbf{x} \neq 0$.

$$\mathbf{z} \quad = \quad a\mathbf{x} + b\mathbf{y}$$

$$\mathbf{x} \circ \mathbf{z} = \mathbf{x} \circ (a\mathbf{x} + b\mathbf{y}) \quad = \quad a(\mathbf{x} \circ \mathbf{x}) + b(\mathbf{x} \circ \mathbf{y}) \quad = \quad 0$$

$$a \quad = \quad -b\,\frac{\mathbf{x} \circ \mathbf{y}}{\mathbf{x} \circ \mathbf{x}}$$

$$\mathbf{z} \quad = \quad -b\,\frac{\mathbf{x} \circ \mathbf{y}}{\mathbf{x} \circ \mathbf{x}}\,\mathbf{x} + b\mathbf{y} \quad = \quad b\left(\mathbf{y} - \frac{\mathbf{x} \circ \mathbf{y}}{\mathbf{x} \circ \mathbf{x}}\,\mathbf{x}\right)$$

The vector **z** is not uniquely determined, since b may be an arbitrary non-zero real scalar. For b = 1 :

$$\mathbf{z} = \mathbf{y} - \frac{5}{6}\mathbf{x} \quad = \quad \begin{array}{|c|}\hline -1 \\ \hline 1 \\ \hline 4 \\ \hline \end{array} \; - \; \frac{5}{6} \begin{array}{|c|}\hline 1 \\ \hline 2 \\ \hline 1 \\ \hline \end{array} \quad = \quad \frac{1}{6} \begin{array}{|c|}\hline -11 \\ \hline -4 \\ \hline 19 \\ \hline \end{array}$$

$$\mathbf{x} \circ \mathbf{z} = 0$$

3.7.3 ELEMENTARY MATRIX OPERATIONS

Equality : Two matrices **A** and **B** are equal if they have the same row dimension m and the same column dimension n and the elements with identical indices are equal.

$$\mathbf{A} = \mathbf{B} \quad :\Leftrightarrow \quad \bigwedge_{i=1}^{m} \bigwedge_{k=1}^{n} (a_{ik} = b_{ik})$$

Addition and subtraction : Two matrices **A** and **B** can only be added or subtracted if they have the same row dimension m and the same column dimension n. The addition + and the subtraction − are carried out by adding and subtracting, respectively, the elements with identical indices.

$$\mathbf{C} = \mathbf{A} + \mathbf{B} \qquad c_{ik} := a_{ik} + b_{ik} \qquad i=1,...,m \; ; \; k=1,...,n$$

$$\mathbf{C} = \mathbf{A} - \mathbf{B} \qquad c_{ik} := a_{ik} - b_{ik} \qquad i=1,...,m \; ; \; k=1,...,n$$

Scaling : A matrix **A** is scaled by a scalar c by multiplying each element of the matrix **A** by c.

$$\mathbf{B} = c\mathbf{A} \qquad b_{ik} = ca_{ik} \qquad i=1,...,m \; ; \; k=1,...,n$$

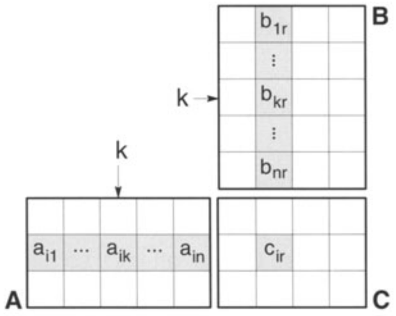
Multiplication : A matrix **A** can be multiplied by a matrix **B** from the right only if the column dimension of **A** and the row dimension of **B** coincide. The matrix product **AB,** a matrix with the row dimension of **A** and the column dimension of **B**, is calculated as follows :

$$\mathbf{C} = \mathbf{AB} \qquad c_{ir} = \sum_{k=1}^{n} a_{ik}\, b_{kr} \qquad i=1,...,m \; ; \; r=1,...,s$$

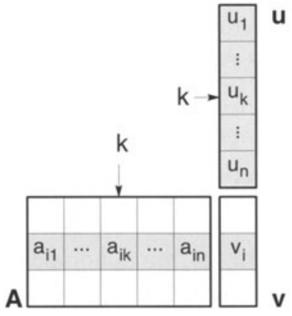
The product of a matrix and a vector is a special case of matrix multiplication. A matrix **A** can only be multiplied by a vector **u** from the right if the column dimension of **A** and the dimension of **u** coincide. The product of a matrix and a vector, a vector whose dimension is the the row dimension of **A**, is calculated as follows :

$$\mathbf{v} = \mathbf{Au} \qquad v_i := \sum_{k=1}^{n} a_{ik}\, u_k \qquad i=1,...,m$$

Graphical representation : The product of two matrices and the product of a matrix and a vector are graphically represented as follows :



matrix product $\mathbf{C} = \mathbf{AB}$



product of matrix and vector $\mathbf{v} = \mathbf{Au}$

Rules for matrix operations : The rules for the elementary matrix operations follow from the theory of vector spaces :

(1) Matrix addition is associative and commutative.

$$(\mathbf{A}+\mathbf{B}) + \mathbf{C} = \mathbf{A}+(\mathbf{B}+\mathbf{C})$$
$$\mathbf{A}+\mathbf{B} = \mathbf{B}+\mathbf{A}$$

(2) Matrix scaling is associative, and it is distributive with respect to addition.

$$(ab)\mathbf{A} = a(b\mathbf{A})$$
$$(a+b)\mathbf{A} = a\mathbf{A}+b\mathbf{A}$$
$$a(\mathbf{A}+\mathbf{B}) = a\mathbf{A}+a\mathbf{B}$$

(3) Matrix multiplication and the multiplication of a matrix and a vector are associative, and they are distributive with respect to addition. Matrix multiplication for quadratic matrices is not commutative in general.

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \qquad (\mathbf{AB})\mathbf{u} = \mathbf{A}(\mathbf{Bu})$$
$$(\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC} \qquad (\mathbf{A}+\mathbf{B})\mathbf{u} = \mathbf{Au}+\mathbf{Bu}$$
$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \qquad \mathbf{A}(\mathbf{u}+\mathbf{v}) = \mathbf{Au}+\mathbf{Av}$$
$$\mathbf{AB} \neq \mathbf{BA}$$

(4) For the zero matrix $\mathbf{0}$ and the identity matrix $\mathbf{I}$ and the scalars 0 and 1 :

$$\mathbf{A}+\mathbf{0} = \mathbf{A} \qquad \mathbf{IA} = \mathbf{A} \qquad \mathbf{AI} = \mathbf{A} \qquad 1\mathbf{A} = \mathbf{A} \qquad \mathbf{Iu} = \mathbf{u}$$
$$\mathbf{A}-\mathbf{A} = \mathbf{0} \qquad \mathbf{A0} = \mathbf{0} \qquad \mathbf{0A} = \mathbf{0} \qquad 0\mathbf{A} = \mathbf{0}$$

From $c\mathbf{A} = \mathbf{0}$ it follows that $c = 0$ or $\mathbf{A} = \mathbf{0}$. However, from $\mathbf{AB} = \mathbf{0}$ it does not follow that $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$.

Inverse : A matrix $\mathbf{X}$ is called a right inverse of the matrix $\mathbf{A}$ if $\mathbf{AX} = \mathbf{I}$. A matrix $\mathbf{Y}$ is called a left inverse of the matrix $\mathbf{A}$ if $\mathbf{YA} = \mathbf{I}$. If the matrix $\mathbf{A}$ has row dimension m and column dimension n, then a left or right inverse of $\mathbf{A}$ has row dimension n and column dimension m. The following rules hold for the existence of inverses of a matrix $\mathbf{A}$:

(1) If $m < n$, the matrix $\mathbf{A}$ may possess a right inverse, but not a left inverse. If a right inverse exists, it is not unique.

(2) If $m > n$, the matrix $\mathbf{A}$ may possess a left inverse, but not a right inverse. If a left inverse exists, it is not unique.

(3) If $m = n$, the quadratic matrix $\mathbf{A}$ may possess both a left inverse and a right inverse. If a left and right inverse exist, they are identical and unique. The inverse of a quadratic matrix $\mathbf{A}$ is designated by $\mathbf{A}^{-1}$.

Regular and singular matrices : A quadratic matrix $\mathbf{A}$ is said to be regular if it has an inverse $\mathbf{A}^{-1}$. It is said to be singular if it does not have an inverse.

Rules for inverses : The definition of the inverse of a regular matrix implies the following rules :

(1) The inverse of the inverse of a regular matrix $\mathbf{A}$ is the matrix $\mathbf{A}$.

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A}$$

(2) The inverse of a scaled matrix $c\mathbf{A}$ is equal to the inverse of $\mathbf{A}$ scaled by a factor c^{-1}.

$$(c\mathbf{A})^{-1} = c^{-1}\mathbf{A}^{-1}$$

(3) The inverse of a product $\mathbf{AB}$ of two regular matrices $\mathbf{A}$ and $\mathbf{B}$ is equal to the product of the inverses in reverse order.

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$$

Transposition : A matrix $\mathbf{A}$ of row dimension m and column dimension n is transposed by interchanging its rows and columns in the matrix scheme. The transpose of $\mathbf{A}$ is designated by $\mathbf{A}^T$; it is a matrix of row dimension n and column dimension m.

$$\mathbf{B} = \mathbf{A}^T \qquad b_{ki} := a_{ik} \qquad i = 1,\ldots,m \quad ; \quad k = 1,\ldots,n$$

Rules for transposes : The definition of the transpose of a matrix implies the following rules :

(1) The transpose of the transpose of a matrix **A** is the matrix **A**.

$$(\mathbf{A}^T)^T = \mathbf{A}$$

(2) The transpose of the sum of two matrices **A** and **B** is the sum of the two transposes.

$$(\mathbf{A}+\mathbf{B})^T = \mathbf{A}^T + \mathbf{B}^T$$

(3) The transpose of a scaled matrix c**A** is equal to the transpose of **A** scaled by a factor c.

$$(c\mathbf{A})^T = c\mathbf{A}^T$$

(4) The transpose of the product of two matrices **A** and **B** is the product of the two transposes in reverse order.

$$(\mathbf{A}\mathbf{B})^T = \mathbf{B}^T\mathbf{A}^T$$

(5) The transpose of the inverse of a regular matrix **A** is the inverse of the transpose. It is sometimes designated by $\mathbf{A}^{-T}$.

$$(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1} = \mathbf{A}^{-T}$$

Symmetric and antisymmetric matrices : A quadratic matrix **A** is said to be symmetric if the matrix and its transpose coincide. It is said to be antisymmetric if the matrix and its negative transpose coincide. The diagonal elements of an antisymmetric matrix are zero.

$$\begin{array}{lll} \mathbf{A} \text{ is symmetric} & :\Leftrightarrow & \mathbf{A} = \mathbf{A}^T \\ \mathbf{A} \text{ is antisymmetric} & :\Leftrightarrow & \mathbf{A} = -\mathbf{A}^T \end{array}$$

Rules for symmetric and antisymmetric matrices : The definition of the symmetry and antisymmetry of matrices implies the following rules :

(1) The sum of two symmetric matrices is a symmetric matrix. The sum of two antisymmetric matrices is an antisymmetric matrix.

(2) Scaling a symmetric matrix yields a symmetric matrix. Scaling an antisymmetric matrix yields an antisymmetric matrix.

(3) Every quadratic matrix **A** has a unique representation as the sum of a symmetric matrix $\mathbf{A}_S$ and an antisymmetric matrix $\mathbf{A}_A$.

$$\mathbf{A} = \mathbf{A}_S + \mathbf{A}_A \qquad \mathbf{A}_S = \tfrac{1}{2}(\mathbf{A}+\mathbf{A}^T) \qquad \mathbf{A}_A = \tfrac{1}{2}(\mathbf{A}-\mathbf{A}^T)$$

(4) The inverse of a regular symmetric matrix is symmetric.

Orthonormal matrix : A real quadratic matrix **A** is said to be orthonormal if its column vectors are pairwise orthonormal. This is equivalent to the condition that the transpose $\mathbf{A}^T$ and the inverse $\mathbf{A}^{-1}$ coincide.

$$\mathbf{A} \text{ is orthonormal} \quad :\Leftrightarrow \quad \mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I} \quad \Leftrightarrow \quad \mathbf{A}^T = \mathbf{A}^{-1}$$

Rules for orthonormal matrices : The definition of orthonormality of matrices implies the following rules :

(1) The product of two orthonormal matrices is an orthonormal matrix.

(2) The inverse of an orthonormal matrix is orthonormal.

Nilpotent matrix : A quadratic matrix **A** is said to be nilpotent if the product **A A** is equal to the zero matrix **0**. The dyadic product $\mathbf{x}\mathbf{y}^T$ of two vectors **x, y** is a nilpotent matrix if and only if the vectors **x** and **y** are orthogonal, so that $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = 0$:

$$\begin{aligned}
\mathbf{A} &= \mathbf{x}\mathbf{y}^T \\
\mathbf{A}\mathbf{A} &= \mathbf{x}\mathbf{y}^T\mathbf{x}\mathbf{y}^T = \mathbf{x}(\mathbf{y}^T\mathbf{x})\mathbf{y}^T = (\mathbf{y}^T\mathbf{x})\mathbf{x}\mathbf{y}^T = 0 \cdot \mathbf{x}\mathbf{y}^T = \mathbf{0}
\end{aligned}$$

Idempotent matrix : A quadratic matrix **A** is said to be idempotent if the product **AA** is equal to **A**. The matrix $\mathbf{A} = \mathbf{I} - \mathbf{x}\mathbf{y}^T$ is idempotent if the vectors **x** and **y** are not orthogonal and $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = 1$:

$$\begin{aligned}
\mathbf{A} &= \mathbf{I} - \mathbf{x}\mathbf{y}^T \\
\mathbf{A}\mathbf{A} &= (\mathbf{I} - \mathbf{x}\mathbf{y}^T)(\mathbf{I} - \mathbf{x}\mathbf{y}^T) = \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T + \mathbf{x}\mathbf{y}^T\mathbf{x}\mathbf{y}^T \\
&= \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T + \mathbf{x}(\mathbf{y}^T\mathbf{x})\mathbf{y}^T = \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T + \mathbf{x}\mathbf{y}^T = \mathbf{I} - \mathbf{x}\mathbf{y}^T = \mathbf{A}
\end{aligned}$$

Self-inverse matrix : A quadratic matrix **A** is said to be self-inverse if the inverse $\mathbf{A}^{-1}$ coincides with **A**, so that $\mathbf{A}\mathbf{A} = \mathbf{I}$. The matrix $\mathbf{A} = \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T$ is self-inverse if the vectors **x** and **y** are not orthogonal and $\mathbf{x}^T\mathbf{y} = \mathbf{y}^T\mathbf{x} = 1$:

$$\begin{aligned}
\mathbf{A} &= \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T \\
\mathbf{A}\mathbf{A} &= (\mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T)(\mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T) = \mathbf{I} - 4\,\mathbf{x}\mathbf{y}^T + 4\,\mathbf{x}\mathbf{y}^T\mathbf{x}\mathbf{y}^T \\
&= \mathbf{I} - 4\,\mathbf{x}\mathbf{y}^T + 4\,\mathbf{x}(\mathbf{y}^T\mathbf{x})\mathbf{y}^T = \mathbf{I} - 4\,\mathbf{x}\mathbf{y}^T + 4\,\mathbf{x}\mathbf{y}^T = \mathbf{I}
\end{aligned}$$

A self-inverse matrix **A** with $\mathbf{A}\mathbf{A} = \mathbf{I}$ is orthonormal if **A** is symmetric, so that $\mathbf{A} = \mathbf{A}^T$. The self-inverse matrix $\mathbf{A} = \mathbf{I} - 2\,\mathbf{x}\mathbf{y}^T$ with $\mathbf{x}^T\mathbf{y} = 1$ is symmetric if $\mathbf{y} = \mathbf{x}$ and hence $\mathbf{x}^T\mathbf{x} = 1$:

$$\begin{aligned}
\mathbf{A} &= \mathbf{I} - 2\,\mathbf{x}\mathbf{x}^T \\
\mathbf{A}^T &= (\mathbf{I} - 2\,\mathbf{x}\mathbf{x}^T)^T = \mathbf{I}^T - 2\,(\mathbf{x}^T)^T\mathbf{x}^T = \mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^T = \mathbf{A}
\end{aligned}$$

Example 1 : Matrix products

The product $\mathbf{AA}^T$ is symmetric.

$$\begin{array}{cccc|cc|l}
 & & & & 1 & -1 & \mathbf{A}^T \\
 & & & & 2 & 1 & \\
 & & & & 1 & 0 & \\
\hline
 & 1 & 2 & 1 & 6 & 1 & \\
\mathbf{A} & -1 & 1 & 0 & 1 & 2 & \mathbf{AA}^T
\end{array}$$

The product $\mathbf{L\,M}$ of two lower triangular matrices is a lower triangular matrix.

$$\begin{array}{cccc|ccc|l}
 & & & & -2 & 0 & 0 & \mathbf{M} \\
 & & & & 1 & -1 & 0 & \\
 & & & & -1 & 2 & 1 & \\
\hline
 & -1 & 0 & 0 & 2 & 0 & 0 & \\
 & 1 & 1 & 0 & -1 & -1 & 0 & \\
\mathbf{L} & 2 & 1 & 1 & -4 & 1 & 1 & \mathbf{L\,M}
\end{array}$$

The product $\mathbf{A\,B}$ of the following matrices $\mathbf{A}$ and $\mathbf{B}$ is a zero matrix $\mathbf{0}$, although neither $\mathbf{A}$ nor $\mathbf{B}$ is a zero matrix.

$$\begin{array}{cccc|cc|l}
 & & & & 1 & 2 & \mathbf{B} \\
 & & & & 2 & 4 & \\
 & & & & 1 & 2 & \\
\hline
 & -1 & 0 & 1 & 0 & 0 & \\
\mathbf{A} & 1 & -1 & 1 & 0 & 0 & \mathbf{AB} = \mathbf{0}
\end{array}$$

Example 2 : Inverse matrices

Let the following matrix **A** be given. The rectangular matrix **B** is a right inverse of **A** for arbitrary values of a, b ∈ ℝ, since **A B** = **I** :

$$\begin{array}{cccc|cc|l}
 & & & & -a & 1-b & \mathbf{B} \\
 & & & & 1 & -1 & \\
 & & & & a & b & a,b\in\mathbb{R} \\
\hline
 & 1 & 1 & 1 & 1 & 0 & \\
\mathbf{A} & 1 & 0 & 1 & 0 & 1 & \mathbf{AB} = \mathbf{I}
\end{array}$$

Let the following symmetric matrix **A** be given. It is regular and has a unique symmetric inverse $\mathbf{A}^{-1}$. The products $\mathbf{A}\,\mathbf{A}^{-1}$ and $\mathbf{A}^{-1}\,\mathbf{A}$ are equal to the identity matrix **I**.

$$\begin{array}{cccc|ccc|l}
 & & & & 0.375 & 0 & -0.250 & \mathbf{A}^{-1} \\
 & & & & 0 & 0.500 & 0.500 & \\
 & & & & -0.250 & 0.500 & 1.000 & \\
\hline
 & 4 & -2 & 2 & 1 & 0 & 0 & \\
 & -2 & 5 & -3 & 0 & 1 & 0 & \\
\mathbf{A} & 2 & -3 & 3 & 0 & 0 & 1 & \mathbf{A}\,\mathbf{A}^{-1} = \mathbf{I}
\end{array}$$

Example 3 : Orthonormal matrices

Let an orthonormal matrix **R** containing sine and cosine functions of an angle α be given. The products $\mathbf{R}^T\mathbf{R}$ and $\mathbf{R}\mathbf{R}^T$ are equal to the identity matrix since $\sin^2\alpha + \cos^2\alpha = 1$.

$$\begin{array}{ccc|cc|l}
 & & & \cos\alpha & -\sin\alpha & \mathbf{R} \\
 & & & \sin\alpha & \cos\alpha & \\
\hline
 & \cos\alpha & \sin\alpha & 1 & 0 & \\
\mathbf{R}^T & -\sin\alpha & \cos\alpha & 0 & 1 & \mathbf{R}^T\mathbf{R} = \mathbf{I}
\end{array}$$

$$\begin{array}{ccc|cc|l}
 & & & \cos\alpha & \sin\alpha & \mathbf{R}^T \\
 & & & -\sin\alpha & \cos\alpha & \\
\hline
 & \cos\alpha & -\sin\alpha & 1 & 0 & \\
\mathbf{R} & \sin\alpha & \cos\alpha & 0 & 1 & \mathbf{R}\,\mathbf{R}^T = \mathbf{I}
\end{array}$$

3.7.4 DERIVED SCALARS

Bilinear form : A bilinear form is a scalar quantity b which is calculated from the vectors **u** and **v** and a matrix **A** as follows :

$$b := \mathbf{u}^T \mathbf{A}\, \mathbf{v} = \sum_{i=1}^{m} \sum_{k=1}^{n} u_i\, a_{ik}\, v_k$$

Quadratic form : A quadratic form is a scalar quantity q which is calculated from a vector **u** and a quadratic matrix **A** as follows :

$$q := \mathbf{u}^T \mathbf{A}\, \mathbf{u} = \sum_{i=1}^{m} \sum_{k=1}^{m} u_i\, a_{ik}\, u_k$$

Positive definite matrix : A quadratic real matrix **A** is said to be positive definite if the quadratic form q is invariably positive for arbitrary non-zero real vectors **u**. It is said to be positive semidefinite if q is positive or zero.

$$\mathbf{A} \text{ is positive definite} \quad :\Leftrightarrow \quad \bigwedge_{\mathbf{u} \neq \mathbf{0}} \mathbf{u}^T \mathbf{A}\, \mathbf{u} > 0$$

$$\mathbf{A} \text{ is positive semidefinite} \quad :\Leftrightarrow \quad \bigwedge_{\mathbf{u} \neq \mathbf{0}} \mathbf{u}^T \mathbf{A}\, \mathbf{u} \geq 0$$

Trace : The trace of a quadratic matrix **A** is the sum of its diagonal elements :

$$\text{tr}\, \mathbf{A} := \sum_{k=1}^{m} a_{kk}$$

Determinant : The determinant of a quadratic matrix **A** is a scalar which is calculated according to the following recursive rule :

$$\det \mathbf{A} := \sum_{k=1}^{n} (-1)^{i+k}\, a_{ik} \det \mathbf{A}_{ik} \qquad i \in \{1,...,n\}$$

n dimension of the quadratic matrix **A**

$\mathbf{A}_{ik}$ the quadratic matrix of dimension n – 1 which is obtained from **A** by deleting row i and column k

The determinants of quadratic matrices of dimension n = 1, 2, 3 are calculated as follows :

$$\det \begin{array}{|c|}\hline a_{11} \\ \hline \end{array} = a_{11}$$

$$\det \begin{array}{|c|c|}\hline a_{11} & a_{12} \\ \hline a_{21} & a_{22} \\ \hline \end{array} = a_{11}\, a_{22} - a_{12}\, a_{21}$$

$$\det \begin{array}{|c|c|c|}\hline a_{11} & a_{12} & a_{13} \\ \hline a_{21} & a_{22} & a_{23} \\ \hline a_{31} & a_{32} & a_{33} \\ \hline \end{array} \begin{array}{l} = a_{11}\, a_{22}\, a_{33} + a_{12}\, a_{23}\, a_{31} + a_{13}\, a_{21}\, a_{32} \\ - a_{13}\, a_{22}\, a_{31} - a_{11}\, a_{23}\, a_{32} - a_{12}\, a_{21}\, a_{33} \end{array}$$

Rules for determinants : The following rules hold for determinants :

(1) The determinants of a matrix **A** and of its transpose $\mathbf{A}^T$ are equal.

$$\det \mathbf{A} = \det \mathbf{A}^T$$

(2) The determinant of the inverse $\mathbf{A}^{-1}$ of a regular matrix **A** is equal to the reciprocal of the determinant of **A**.

$$\det \mathbf{A}^{-1} = 1 / \det \mathbf{A}$$

(3) The determinant of a product **AB** of two quadratic matrices **A** and **B** is equal to the product of the determinants of **A** and of **B**.

$$\det (\mathbf{AB}) = \det \mathbf{A} * \det \mathbf{B}$$

(4) If the matrix **A** is a lower triangular matrix **L**, an upper triangular matrix **R** or a diagonal matrix **D**, then its determinant is the product of its diagonal elements.

$$\det \mathbf{A} = \prod_{k=1}^{n} a_{kk} \qquad\qquad \mathbf{A} = \mathbf{L}, \mathbf{R}, \mathbf{D}$$

Norms : Norms are scalars which are defined for a vector **v** and a quadratic matrix **A** as follows :

$$\| \mathbf{v} \|_\infty := \max_k | v_k | \qquad\qquad \| \mathbf{A} \|_\infty := \max_i \sum_{k=1}^{n} | a_{ik} |$$

$$\| \mathbf{v} \|_1 := \sum_{k=1}^{n} | v_k | \qquad\qquad \| \mathbf{A} \|_1 := \max_k \sum_{i=1}^{n} | a_{ik} |$$

$$\| \mathbf{v} \|_2 := \left[\sum_{k=1}^{n} | v_k |^2 \right]^{1/2} \qquad\qquad \| \mathbf{A} \|_2 := \left[\sum_{i=1}^{n} \sum_{k=1}^{n} | a_{ik} |^2 \right]^{1/2}$$

Rules for norms : The following rules hold for norms with the same index :

$$\| c\mathbf{v} \| = | c | \| \mathbf{v} \| \qquad\qquad \| c\mathbf{A} \| = | c | \| \mathbf{A} \|$$

$$\| \mathbf{u} + \mathbf{v} \| \leq \| \mathbf{u} \| + \| \mathbf{v} \| \qquad\qquad \| \mathbf{A} + \mathbf{B} \| \leq \| \mathbf{A} \| + \| \mathbf{B} \|$$

$$\| \mathbf{A} \mathbf{v} \| \leq \| \mathbf{A} \| \| \mathbf{v} \| \qquad\qquad \| \mathbf{AB} \| \leq \| \mathbf{A} \| \| \mathbf{B} \|$$

Example 1 : Determinants

Let a quadratic matrix **A** be given; its determinant is calculated according to the recursive rule with i = 1 as follows.

$$\det \mathbf{A} = \det \begin{array}{|c|c|c|}\hline 4 & -2 & 2 \\ \hline -2 & 5 & -3 \\ \hline 2 & -3 & 3 \\ \hline \end{array} = \sum_{k=1}^{3} (-1)^{k+1}\, a_{1k} \det \mathbf{A}_{1k}$$

$$\det \mathbf{A}_{11} = \det \begin{array}{|c|c|}\hline 5 & -3 \\ \hline -3 & 3 \\ \hline \end{array} = 5 * 3 - (-3) * (-3) = 6$$

$$\det \mathbf{A}_{12} = \det \begin{array}{|c|c|}\hline -2 & -3 \\ \hline 2 & 3 \\ \hline \end{array} = (-2) * 3 - (-3) * 2 = 0$$

$$\det \mathbf{A}_{13} = \det \begin{array}{|c|c|}\hline -2 & 5 \\ \hline 2 & -3 \\ \hline \end{array} = (-2) * (-3) - 5 * 2 = -4$$

$$\begin{aligned} \det \mathbf{A} &= a_{11} \det \mathbf{A}_{11} - a_{12} \det \mathbf{A}_{12} + a_{13} \det \mathbf{A}_{13} \\ &= 4 * 6 - (-2) * 0 + 2 * (-4) = 24 - 8 = 16 \end{aligned}$$

The determinant of the inverse $\mathbf{A}^{-1}$ is calculated from the determinant of **A** according to the rules for determinants :

$$\det \mathbf{A}^{-1} = 1/\det \mathbf{A} = \frac{1}{16}$$

Example 2 : Quadratic form and norms

Let a quadratic matrix **A** and a vector **u** be given. The calculation of the vector $\mathbf{v} = \mathbf{A}\,\mathbf{u}$ and the quadratic form $\mathbf{u}^T\mathbf{v} = \mathbf{u}^T\mathbf{A}\mathbf{u}$ is represented graphically :

				2	**u**
				–1	
				1	
	4	–2	2	12	
	–2	5	–3	–12	
A	2	–3	3	10	$\mathbf{Au} = \mathbf{v}$
$\mathbf{u}^T$	2	–1	1	46	$\mathbf{u}^T\mathbf{v} = \mathbf{u}^T\mathbf{A}\,\mathbf{u}$

The norms $\|\mathbf{u}\|_i$, $\|\mathbf{v}\|_i$ and $\|\mathbf{A}\|_i$ are calculated as follows :

$$\|\mathbf{u}\|_\infty = \max\{2, 1, 1\} = 2 \qquad \|\mathbf{v}\|_\infty = 12$$

$$\|\mathbf{u}\|_1 = 2+1+1 = 4 \qquad \|\mathbf{v}\|_1 = 34$$

$$\|\mathbf{u}\|_2 = [2^2 + (-1)^2 + 1^2]^{0.5} = \sqrt{6} \qquad \|\mathbf{v}\|_2 = \sqrt{388}$$

$$\|\mathbf{A}\|_\infty = \max\{(4+2+2), (2+5+3), (2+3+3)\} = 10$$

$$\|\mathbf{A}\|_1 = \max\{(4+2+2), (2+5+3), (2+3+3)\} = 10$$

$$\|\mathbf{A}\|_2 = [4^2 + (-2)^2 + 2^2 + (-2)^2 + 5^2 + (-3)^2 + 2^2 + (-3)^2 + 3^2]^{0.5} = \sqrt{84}$$

The norms $\|\mathbf{A}\|_\infty$ and $\|\mathbf{A}\|_1$ are equal, since the matrix **A** is symmetric. The vector and matrix norms are compatible :

$$\|\mathbf{Au}\|_\infty \le \|\mathbf{A}\|_\infty\ \|\mathbf{u}\|_\infty \quad : \quad 12 \le 10 * 2$$

$$\|\mathbf{Au}\|_1 \le \|\mathbf{A}\|_1\ \|\mathbf{u}\|_1 \quad : \quad 34 \le 10 * 4$$

$$\|\mathbf{Au}\|_2 \le \|\mathbf{A}\|_2\ \|\mathbf{u}\|_2 \quad : \quad \sqrt{388} \le \sqrt{84 * 6}$$

3.7.5 COMPLEX VECTORS AND MATRICES

Introduction : The definitions, operations and rules in Sections 3.7.1 to 3.7.4 are valid for complex vectors and matrices, unless an explicit restriction was made. The specific characteristics of complex vectors and matrices arise in connection with their conjugates. Some of the specific properties of complex vectors and matrices are named after the mathematician Hermite.

Complex quantities : A complex number is represented by a real part and an imaginary part. This form of representation is transferred to vectors and matrices.

complex scalar	$c = a + i\,b$	a, b real scalars
complex vector	$\mathbf{c} = \mathbf{a} + i\,\mathbf{b}$	$\mathbf{a}, \mathbf{b}$ real vectors
complex matrix	$\mathbf{C} = \mathbf{A} + i\,\mathbf{B}$	$\mathbf{A}, \mathbf{B}$ real matrices
number i	$i^2 = -1$	

Conjugate quantities : A complex quantity is transformed into the corresponding conjugate quantity by changing the sign of the imaginary part. Conjugate quantities are designated by a horizontal line over the symbol.

conjugate scalar	$\bar{c} = a - i\,b$
conjugate vector	$\bar{\mathbf{c}} = \mathbf{a} - i\,\mathbf{b}$
conjugate matrix	$\bar{\mathbf{C}} = \mathbf{A} - i\,\mathbf{B}$

Hermitian transpose : The transpose of the conjugate of a complex matrix $\mathbf{C}$ is called the hermitian transpose and is designated by $(\bar{\mathbf{C}})^T$ or $\mathbf{C}^H$.

$$\mathbf{C} = \mathbf{A} + i\,\mathbf{B} \qquad\qquad \mathbf{C}^H := (\bar{\mathbf{C}})^T = \mathbf{A}^T - i\,\mathbf{B}^T$$

The hermitian transpose of a complex vector $\mathbf{c}$ is defined as a special case of a complex matrix as follows :

$$\mathbf{c} = \mathbf{a} + i\,\mathbf{b} \qquad\qquad \mathbf{c}^H := (\bar{\mathbf{c}})^T = \mathbf{a}^T - i\,\mathbf{b}^T$$

Hermitian scalar product : The scalar product $s = \bar{\mathbf{u}} \circ \mathbf{v}$ of two complex vectors $\mathbf{u}$ and $\mathbf{v}$ is called the hermitian scalar product and is designated by $(\bar{\mathbf{u}})^T\mathbf{v}$ or $\mathbf{u}^H\,\mathbf{v}$. It is calculated as follows :

$$s = \bar{\mathbf{u}} \circ \mathbf{v} := \mathbf{u}^H\,\mathbf{v} \qquad \mathbf{u} = \mathbf{a} + i\,\mathbf{b} \qquad \mathbf{v} = \mathbf{c} + i\,\mathbf{d}$$

$$s = \sum_{k=1}^{n} (a_k - i\,b_k)(c_k + i\,d_k)$$

By definition, the hermitian scalar product may be reduced to the real vectors $\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}$ as follows :

$$s = (\mathbf{a} - i\,\mathbf{b}) \circ (\mathbf{c} + i\,\mathbf{d}) = (\mathbf{a} \circ \mathbf{c} + \mathbf{b} \circ \mathbf{d}) + i\,(\mathbf{a} \circ \mathbf{d} - \mathbf{b} \circ \mathbf{c})$$

The conjugate of the hermitian scalar product is given by :

$$\bar{s} = \overline{\bar{\mathbf{u}} \circ \mathbf{v}} = \mathbf{u} \circ \bar{\mathbf{v}} = \bar{\mathbf{v}} \circ \mathbf{u} = \mathbf{v}^H \mathbf{u}$$

Rules for hermitian scalar products : The definition of the hermitian scalar product implies the following rules for complex vectors, which reduce to the rules for scalar products in Section 3.7.2 in the special case of real vectors :

(1) The hermitian scalar product is commutative.

$$\bar{\mathbf{u}} \circ \mathbf{v} = \mathbf{v} \circ \bar{\mathbf{u}}$$

(2) The hermitian scalar product is distributive with respect to vector addition.

$$\bar{\mathbf{u}} \circ (\mathbf{v} + \mathbf{w}) = \bar{\mathbf{u}} \circ \mathbf{v} + \bar{\mathbf{u}} \circ \mathbf{w}$$

(3) The hermitian scalar product is associative with respect to vector scaling.

$$(c\bar{\mathbf{u}}) \circ \mathbf{v} = c\,(\bar{\mathbf{u}} \circ \mathbf{v})$$

(4) The hermitian scalar product $\bar{\mathbf{u}} \circ \mathbf{u}$ is real and positive for $\mathbf{u} \neq \mathbf{0}$.

$$\bar{\mathbf{u}} \circ \mathbf{u} > 0 \quad \text{for} \quad \mathbf{u} \neq \mathbf{0}$$

Unitary vectors : Two complex non-zero vectors **u** and **v** are said to be unitary if the hermitian scalar product $\bar{\mathbf{u}} \circ \mathbf{v}$ is zero and the hermitian scalar products $\bar{\mathbf{u}} \circ \mathbf{u}$ and $\bar{\mathbf{v}} \circ \mathbf{v}$ are one. If $\bar{\mathbf{u}} \circ \mathbf{v} = 0$, then conjugation yields $\bar{\mathbf{v}} \circ \mathbf{u} = 0$.

$$\begin{aligned} \mathbf{u} \text{ and } \mathbf{v} \text{ are unitary} \quad &:\Leftrightarrow \quad \bar{\mathbf{u}} \circ \mathbf{v} = 0 \;\wedge\; \bar{\mathbf{u}} \circ \mathbf{u} = 1 \;\wedge\; \bar{\mathbf{v}} \circ \mathbf{v} = 1 \\ &\;\Leftrightarrow \quad \bar{\mathbf{v}} \circ \mathbf{u} = 0 \;\wedge\; \bar{\mathbf{u}} \circ \mathbf{u} = 1 \;\wedge\; \bar{\mathbf{v}} \circ \mathbf{v} = 1 \end{aligned}$$

Hermitian and antihermitian matrices : A complex quadratic matrix is said to be hermitian if the matrix and its hermitian transpose coincide. It is said to be antihermitian if the matrix and its negative hermitian transpose coincide.

$$\begin{aligned} \mathbf{A} \text{ is hermitian} \quad &:\Leftrightarrow \quad \mathbf{A} = \mathbf{A}^H \\ \mathbf{A} \text{ is antihermitian} \quad &:\Leftrightarrow \quad \mathbf{A} = -\mathbf{A}^H \end{aligned}$$

The rules for hermitian and antihermitian matrices are analogous to the rules for symmetric and antisymmetric matrices in Section 3.7.3.

Unitary matrices : A complex quadratic matrix is said to be unitary if its column vectors are pairwise unitary. This is equivalent to the condition that the hermitian transpose $\mathbf{A}^H$ and the inverse $\mathbf{A}^{-1}$ coincide.

$$\mathbf{A} \text{ is unitary} \quad :\Leftrightarrow \quad \mathbf{A}^H \mathbf{A} = \mathbf{A}\,\mathbf{A}^H = \mathbf{I} \quad \Leftrightarrow \quad \mathbf{A}^H = \mathbf{A}^{-1}$$

The rules for unitary matrices are analogous to the ones for orthonormal matrices in Section 3.7.3.

Example 1 : Scalar products

Let complex vectors **x**, **y** of the complex two-dimensional space $\mathbb{C}^2$ be given :

$$\mathbf{x} = \begin{array}{|c|}\hline 1+i \\ \hline 1-i \\ \hline\end{array} \qquad\qquad \mathbf{y} = \begin{array}{|c|}\hline 2+i \\ \hline -i \\ \hline\end{array}$$

The conjugate vectors $\bar{\mathbf{x}}$, $\bar{\mathbf{y}}$ are :

$$\bar{\mathbf{x}} = \begin{array}{|c|}\hline 1-i \\ \hline 1+i \\ \hline\end{array} \qquad\qquad \bar{\mathbf{y}} = \begin{array}{|c|}\hline 2-i \\ \hline i \\ \hline\end{array}$$

The hermitian scalar products $\bar{\mathbf{x}} \circ \mathbf{x}$, $\bar{\mathbf{y}} \circ \mathbf{y}$, $\bar{\mathbf{x}} \circ \mathbf{y}$, $\bar{\mathbf{y}} \circ \mathbf{x}$ are calculated as follows :

$$\begin{array}{lllll}
\bar{\mathbf{x}} \circ \mathbf{x} = \mathbf{x}^H \mathbf{x} & = (1-i)(1+i) + (1+i)(1-i) & = (1-i^2) + (1-i)^2 & = 4 \\
\bar{\mathbf{y}} \circ \mathbf{y} = \mathbf{y}^H \mathbf{y} & = (2-i)(2+i) + i(-i) & = (2-i^2) - i^2 & = 4 \\
\bar{\mathbf{x}} \circ \mathbf{y} = \mathbf{x}^H \mathbf{y} & = (1-i)(2+i) + (1+i)(-i) & = (2-i-i^2) + (-i-i^2) & \\
 & = (3-i) + (1-i) & = 4 - 2i & \\
\bar{\mathbf{y}} \circ \mathbf{x} = \mathbf{y}^H \mathbf{x} & = (2-i)(1+i) + i(1-i) & = (2+i-i^2) + (i-i^2) & \\
 & = (3+i) + (1+i) & = 4 + 2i &
\end{array}$$

The hermitian scalar products $\bar{\mathbf{x}} \circ \mathbf{y}$ and $\bar{\mathbf{y}} \circ \mathbf{x}$ are conjugate scalars.

Example 2 : Hermitian and unitary matrix

Let the following matrix **A** be given :

$$\mathbf{A} = \mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^H \qquad\qquad \mathbf{x}^H\mathbf{x} = 1$$

The matrix **A** is hermitian, since $\mathbf{A} = \mathbf{A}^H$:

$$\mathbf{A}^H = (\mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^H)^H = \mathbf{I}^H - 2\,(\mathbf{x}^H)^H\,\mathbf{x}^H = \mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^H = \mathbf{A}$$

The matrix **A** is unitary, since $\mathbf{A}\,\mathbf{A}^H = \mathbf{A}^H\mathbf{A} = \mathbf{A}\,\mathbf{A} = \mathbf{I}$:

$$\begin{aligned}\mathbf{A}\,\mathbf{A} &= (\mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^H)\,(\mathbf{I} - 2\,\mathbf{x}\,\mathbf{x}^H) = \mathbf{I} - 4\,\mathbf{x}\,\mathbf{x}^H + 4\,\mathbf{x}\,(\mathbf{x}^H\mathbf{x})\,\mathbf{x} \\ &= \mathbf{I} - 4\,\mathbf{x}\,\mathbf{x}^H + 4\,\mathbf{x}\,\mathbf{x}^H = \mathbf{I}\end{aligned}$$

Since $\mathbf{A}\,\mathbf{A} = \mathbf{I}$, the matrix **A** is also self-inverse, so that $\mathbf{A} = \mathbf{A}^{-1}$. The properties of the matrix **A** are demonstrated in the following numerical example for a given vector **x** :

$$\mathbf{x} = \begin{array}{|c|}\hline 0.80 \\ \hline 0.60\,i \\ \hline\end{array} \qquad\qquad \begin{array}{r|c|cc|l} & & 0.80 & -0.60\,i & \mathbf{x}^H \\ \hline & 0.80 & 0.64 & -0.48\,i & \\ \mathbf{x} & 0.60\,i & 0.48\,i & 0.36 & \mathbf{x}\,\mathbf{x}^H \end{array}$$

$$\mathbf{A} = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline\end{array} - 2 \begin{array}{|c|c|}\hline 0.64 & -0.48\,i \\ \hline 0.48\,i & 0.36 \\ \hline\end{array} = \begin{array}{|c|c|}\hline -0.28 & 0.96\,i \\ \hline -0.96\,i & 0.28 \\ \hline\end{array} = \mathbf{A}^H$$

$$\begin{array}{r|cc|cc|l} & & & -0.28 & 0.96\,i & \mathbf{A} \\ & & & -0.96\,i & 0.28 & \\ \hline & -0.28 & 0.96\,i & 1 & 0 & \\ \mathbf{A} & -0.96\,i & 0.28 & 0 & 1 & \mathbf{A}\,\mathbf{A} = \mathbf{I} \end{array}$$

4 ORDINAL STRUCTURES

4.1 INTRODUCTION

The elements of an unstructured set are not ordered and therefore cannot be compared. However, many applications of sets require a comparison between their elements. For this purpose, order structures are defined in the following. The properties of ordered sets are used, for instance, to construct sorting algorithms for the elements of a set.

To order a set, a relation is defined in the set. Such a relation may allow comparison of all elements (total order) or only some elements (partial order) of the set. Order relations, which are reflexive, antisymmetric and transitive, correspond to the relation $\leq$ (less than or equal to) in the set of integers. Strict order relations, which are antireflexive, asymmetric and transitive, correspond to the relation $<$ (less than) in the set of integers. Ordered sets are graphically represented in order diagrams.

Ordered sets may possess extreme elements. An element a of a set is minimal if every element of the set is not less than a. The element a is the least element of the set if a is less than every other element of the set. A set may contain several minimal elements, but at most one least element. Analogously, there may be maximal elements and a greatest element of a set.

For a subset A of a set M, bounds in M are defined. In contrast to minimal and maximal elements of A and a least or greatest element of A, all of which are contained in A, bounds of A may be elements of M which are not contained in A. An element a of M, which need not be an element of A, is an upper bound of A in M if every element of A is less than or equal to a. The least upper bound of A in M is of particular importance. Lower bounds are defined analogously. These concepts are required, for example, for the definition of real numbers.

Ordered sets may have various extremality properties. Well-ordered sets, which are totally ordered and in which every subset contains a least element, are especially important. Directed sets and lattices are further examples of ordered sets with extremality properties.

The comparison of properties of different ordered sets relies on the concept of similarly structured sets. Two ordered sets are similarly structured if there is a mapping between them which is not only bijective but also isotonic in both directions. Similarly structured sets form an order type. A well-ordered order type is called an ordinal number.

The order type of a set determines whether a certain mapping of this set onto itself possesses fixed points. These fixed points are used to determine the properties of ordered sets. Examples are furnished by Zorn's Lemma, Zermelo's Theorem and the Axiom of Choice. The chapter concludes with a treatment of the comparison of cardinal numbers.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

4.2 ORDERED SETS

Introduction : To order a set, an order relation is defined in the set. The definition of an order relation does not require that any two elements of the set be comparable : In general, the set is only partially ordered. The total ordering of a set is a special case : In this case, any two elements of the set are comparable. As in the set of natural numbers, a strict order relation (for instance $<$ in $\mathbb{N}$) may be used instead of the order relation (for instance $\leq$ in $\mathbb{N}$). A strict order relation may also order a set either partially or totally.

Order relation : Let a set M be given. A relation in M is called an order relation if it is reflexive, antisymmetric and transitive. Order relations are often represented by symbols like $\leq$ or $\sqsubseteq$. An order relation $\leq$ in M is thus a subset of the cartesian product $M \times M$ with the following properties for elements $a, b, c \in M$:

(1) $\leq$ is reflexive : $a \in M \quad \Rightarrow \quad a \leq a$

(2) $\leq$ is antisymmetric : $a \leq b \;\wedge\; b \leq a \quad \Rightarrow \quad a = b$

(3) $\leq$ is transitive : $a \leq b \;\wedge\; b \leq c \quad \Rightarrow \quad a \leq c$

Partially ordered set : Let a set M and an order relation $\leq$ be given. Then the domain $(M ; \leq)$ is called a partially ordered set. The set M is partially ordered by the relation $\leq$.

Example 1 : Every power set is partially ordered by inclusion.

Let a set M in the power set P(M) be given. The relation $\subseteq$ (is a subset of) is defined for elements $A, B \in P(M)$ in Section 2.1 :

$$A \subseteq B \quad \Rightarrow \quad (x \in A \;\Rightarrow\; x \in B)$$

This inclusion is an order relation, since for elements $A, B, C \in P(M)$ it possesses properties (1) to (3) :

(1) $\subseteq$ is reflexive : $A \in P(M) \quad \Rightarrow \quad A \subseteq A$

(2) $\subseteq$ is antisymmetric : $A \subseteq B \;\wedge\; B \subseteq A \quad \Rightarrow \quad A = B$

(3) $\subseteq$ is transitive : $A \subseteq B \;\wedge\; B \subseteq C \quad \Rightarrow \quad A \subseteq C$

Example 2 : Order relation ”is a divisor of”

The divisor relation $|$ (is a divisor of) in the set $\mathbb{N}'$ of natural numbers except zero is an order relation. The statement $a | b$ is true if a is a divisor of b. The statement is also true for the special cases $a = 1$ and $a = b$. For example, 2 is a divisor of 4. For elements $a, b, c \in \mathbb{N}'$:

(1) $|$ is reflexive : $a \in \mathbb{N}' \quad \Rightarrow \quad a | a$

(2) $|$ is antisymmetric : $a | b \;\wedge\; b | a \quad \Rightarrow \quad a = b$

(3) $|$ is transitive : $a | b \;\wedge\; b | c \quad \Rightarrow \quad a | c$

Comparable elements : Two elements a, b of a partially ordered set $(M\,;\,\leq)$ are said to be comparable if (a, b) or (b, a) is contained in the relation $\leq$, so that $a \leq b$ or $b \leq a$. In general, not every element of a partially ordered set can be compared with every other element of the set.

Total order relation : An order relation $\leq$ in a set M is said to be total (simple, linear, complete) if it allows any two elements of M to be compared. In addition to properties (1) to (3), a total order relation is also linear :

(4) $\leq$ is linear : $a, b \in M \quad \Rightarrow \quad a \leq b \;\vee\; b \leq a$

Totally ordered set : A partially ordered set $(M\,;\,\leq)$ is said to be totally ordered (simply ordered, linearly ordered, completely ordered, a chain) if the order relation $\leq$ is total. Any two elements of a totally ordered set are comparable.

Example 3 : The set $\mathbb{N}$ of natural numbers is totally ordered.

The set $\mathbb{N}$ of natural numbers is totally ordered by the relation $\leq$ (less than or equal to), since for elements $i, k, m \in \mathbb{N}$:

(1) $\leq$ is reflexive : $i \in \mathbb{N} \quad \Rightarrow \quad i \leq i$

(2) $\leq$ is antisymmetric : $i \leq k \;\wedge\; k \leq i \quad \Rightarrow \quad i = k$

(3) $\leq$ is transitive : $i \leq k \;\wedge\; k \leq m \quad \Rightarrow \quad i \leq m$

(4) $\leq$ is linear : $i, k \in \mathbb{N} \quad \Rightarrow \quad i \leq k \;\vee\; k \leq i$

Strict order relation : Let a set M be given. A relation in M is called a strict order relation if it is antireflexive, asymmetric and transitive. Strict order relations are often represented by symbols like $<$ or $\sqsubset$. A strict order relation $<$ in M is thus a subset of the cartesian product $M \times M$ with the following properties for elements $a, b, c \in M$:

(1) $<$ is antireflexive : $a \in M \quad \Rightarrow \quad \neg(a < a)$

(2) $<$ is asymmetric : $a < b \quad \Rightarrow \quad \neg(b < a)$

(3) $<$ is transitive : $a < b \;\wedge\; b < c \quad \Rightarrow \quad a < c$

Partially strictly ordered set : Let a set M and a strict order relation $<$ in M be given. Then the domain $(M\,;\,<)$ is called a partially strictly ordered set. The set M is partially strictly ordered by the relation $<$.

Example 4 : Every power set is partially strictly ordered by proper inclusion.

Let a set M with the power set P(M) be given. The relation $\subset$ (is a proper subset of) is defined for elements $A, B \in P(M)$ in Section 2.1 :

$$A \subset B \quad \Leftrightarrow \quad ((x \in A \;\Rightarrow\; x \in B) \;\wedge\; \bigvee_{y \in B} (\neg(y \in A)))$$

This proper inclusion is a strict order relation, since for elements $A, B, C \in P(M)$ it possesses properties (1) to (3) :

(1) $\subset$ is antireflexive : $A \in P(M) \quad \Rightarrow \quad \neg(A \subset A)$
(2) $\subset$ is asymmetric : $A \subset B \quad \Rightarrow \quad \neg(B \subset A)$
(3) $\subset$ is transitive : $A \subset B \ \wedge \ B \subset C \quad \Rightarrow \quad A \subset C$

Example 5 : Strict order of the divisors of natural numbers

The relation $\|$ (is a proper divisor of) in the set $\mathbb{N}'$ of natural numbers except zero is a strict order relation. The statement $a \| b$ is true if a is a divisor of b and $a \neq 1$ and $a \neq b$. For example, 2 is a proper divisor of 4. For elements $a, b, c \in \mathbb{N}'$:

(1) $\|$ is antireflexive : $a \in \mathbb{N}' \quad \Rightarrow \quad \neg(a \| a)$
(2) $\|$ is asymmetric : $a \| b \quad \Rightarrow \quad \neg(b \| a)$
(3) $\|$ is transitive : $a \| b \ \wedge \ b \| c \quad \Rightarrow \quad a \| c$

Comparable elements : Two different elements $a \neq b$ of a partially strictly ordered set $(M \,;\, <)$ are said to be comparable if either (a, b) or (b, a) is contained in the relation $<$, so that either $a < b$ or $b < a$. In general, not every element of a partially strictly ordered set can be compared with every other element of the set.

Total strict order relation : A strict order relation $<$ in a set M is said to be total (simple, connex, complete) if it allows any two elements of M to be compared. In addition to properties (1) to (3), a total strict order relation is also connex :

(4) $<$ is connex : $a, b \in M \quad \Rightarrow \quad (a \neq b \quad \Rightarrow \quad a < b \ \vee \ b < a)$

Totally strictly ordered set : A partially strictly ordered set $(M \,;\, <)$ is said to be totally (simply, completely) strictly ordered or a strict chain if the order relation $<$ is total. Any two elements of a totally strictly ordered set are comparable.

Example 6 : The set $\mathbb{N}$ of natural numbers is totally strictly ordered.

The set $\mathbb{N}$ of natural numbers is totally strictly ordered by the relation $<$ (less than), since for elements $i, k, m \in \mathbb{N}$:

(1) $<$ is antireflexive : $i \in \mathbb{N} \quad \Rightarrow \quad \neg(i < i)$
(2) $<$ is asymmetric : $i < k \quad \Rightarrow \quad \neg(k < i)$
(3) $<$ is transitive : $i < k \ \wedge \ k < m \quad \Rightarrow \quad i < m$
(4) $<$ is connex : $i \neq k \quad \Rightarrow \quad i < k \ \vee \ k < i$

Ordered set : A set M is said to be ordered if there is a domain $(M \,;\, \leq)$ with the order relation $\leq$ or a domain $(M \,;\, <)$ with the strict order relation $<$. The set may be partially or totally ordered. Ordered sets are graphically represented in order diagrams.

Intervals : In an ordered set M, the following intervals are defined with the elements $a, b \in M$. The properties of the intervals are indicated by different arrangements of the square brackets [ ].

closed : $[a, b] := \{x \in M \mid a \leq x \leq b\}$

open : $]a, b[:= \{x \in M \mid a < x < b\}$

closed on the left, open on the right : $[a, b[:= \{x \in M \mid a \leq x < b\}$

open on the left, closed on the right : $]a, b] := \{x \in M \mid a < x \leq b\}$

Successor and predecessor : An element b of an ordered set $(M ; \leq)$ or $(M ; <)$ is called a successor of the element $a \in M$ if $a \leq b$ or $a < b$, respectively, and the open interval $]a, b[$ is empty. The element a with $a \neq b$ is called a predecessor of b.

Order diagram : An order diagram represents an order structure of a set. Every element of the set is represented by a point in the plane. If a pair (a, b) of elements with $a \neq b$ is an element of the order relation, the point for element b is placed above the point for element a. The two points are joined by a line if element b is a successor of element a.

Associated order relations : An order relation $\leq$ and a strict order relation $<$ in the same set M are different subsets of the cartesian product $M \times M$. The order relations are said to be associated if they differ only by the diagonal $\{(x, x) \mid x \in M\}$ of the product $M \times M$.

$$\begin{aligned} \leq - < &= \{(x,x) \mid x \in M\} \\ \leq &= < \cup \{(x,x) \mid x \in M\} \\ < &= \leq - \{(x,x) \mid x \in M\} \end{aligned}$$

The following equivalences hold between the statements of associated order relations and the identity of elements :

$$\begin{aligned} x \leq y &\Leftrightarrow x < y \vee x = y \\ x < y &\Leftrightarrow x \leq y \wedge x \neq y \end{aligned}$$

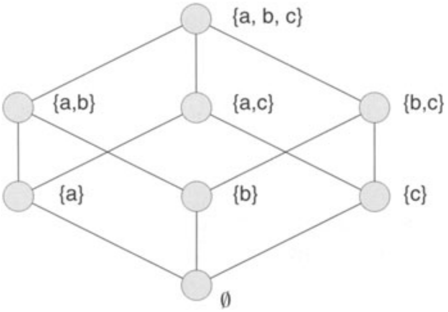
Associated order relations have identical order diagrams.

Example 7 : Order diagram of inclusion in a power set

Let the power set P(M) of a set M = {a, b, c} be given. According to Example 1, P(M) is partially ordered by the order relation $\subseteq$ (is a subset of). According to Example 3, P(M) is partially strictly ordered by the strict order relation $\subset$ (is a proper subset of).

The associated order relations $\subseteq$ and $\subset$ have the same order diagram. The order diagram shows that equipotent sets lie on the same level and are therefore not comparable elements of P(M). For example, {a, b} is not a subset or proper subset of {a, c}, and {a, c} is not a subset or proper subset of {a, b}.

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

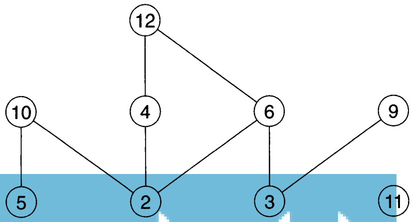


Example 8 : Order diagram of the divisor relation for natural numbers

Let the indicated set M of natural numbers be given. According to Example 2, the set M is partially ordered by the order relation | (is a divisor of). According to Example 4, the set M is partially strictly ordered by the strict order relation || (is a proper divisor of).

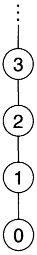
The associated order relations | and || have the same order diagram. The order diagram shows that natural numbers on the same level in the order diagram are not comparable. For example, 10 is not a divisor or proper divisor of 4, and 4 is not a divisor or proper divisor of 10.

$$M = \{2,3,4,5,6,7,9,10,11,12\} \subset \mathbb{N}$$



Example 9 : Order diagram of the comparison of natural numbers

According to Examples 3 and 6, the set $\mathbb{N}$ of natural numbers is totally ordered by the order relation $\le$ (less than or equal to) or by the strict order relation $<$ (less than). The associated order relations have the same order diagram in the form of a chain.



$$\mathbb{N} = \{0, 1, 2, 3, ...\}$$

$$0 \le 1 \le 2 \le 3 \le ...$$

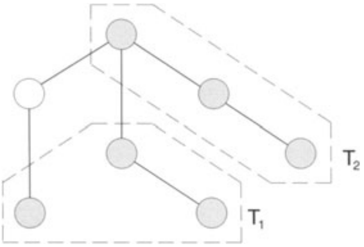
Subordering : The restriction of an order relation R in a set M to a subset S of M is called a subordering of M. While R is a subset of the cartesian product $M \times M$, the restriction of R to S is a subset of $S \times S$. The restriction contains exactly those elements of R which are contained in $S \times S$.

restriction of R to S : $\{(a, b) \in R \mid (a, b) \in S \times S\}$

If M is totally ordered by R, then S is totally ordered by the restriction of R to S. If M is only partially ordered by R, then S may be partially or totally ordered by the restriction of R to S.

Example 10 : Suborderings of a set

The following order diagram shows a partially ordered set with a partially ordered subset T_1 and a totally ordered subset T_2.



T_1	ordered set
T_2	totally ordered set

4.3 EXTREME ELEMENTS

Introduction : The total ordering of a set does not guarantee the existence of a least or greatest element in the set. For example, with respect to the relation $\leq$, the set of negative integers has the greatest element -1 but no least element, while the set of positive integers has the least element 1 but no greatest element, and the set of integers has neither a greatest nor a least element. Finite ordered sets which are not totally ordered also do not in general have a greatest or least element.

The extremality properties of a set A which is contained in a set B are described in two fundamentally different ways. Considering only extreme elements contained in A leads to the concepts of minimal / maximal element and least / greatest element of A. Considering extreme elements of A which are contained in the set M (including A) leads to the concepts of lower / upper bound and greatest lower / least upper bound of A in M.

Minimal element : Let a set M be partially ordered by the relation $\leq$. An element $a \in M$ is called a minimal element if every element $x \in M$ is not less than a, so that $x \leq a$ implies $x = a$. A minimal element of a set M is designated by minEl(M).

$$a \;=\; \text{minEl}(M) \;:\Leftrightarrow\; (x \in M \quad \wedge \quad x \leq a \quad \Rightarrow \quad x = a)$$

An ordered set M may contain zero, one or several minimal elements. If a set contains more than one minimal element, then these elements are not comparable. A minimal element does not have a predecessor in the order diagram.

Maximal element : Let a set M be partially ordered by the relation $\leq$. An element $a \in M$ is called a maximal element in M if a is not less than every element x in M, so that $a \leq x$ implies $a = x$. A maximal element of a set M is designated by maxEl(M).

$$a \;=\; \text{maxEl}(M) \;:\Leftrightarrow\; (x \in M \quad \wedge \quad a \leq x \quad \Rightarrow \quad a = x)$$

An ordered set M may contain zero, one or several maximal elements. If a set contains more than one maximal element, then these elements are not comparable. A maximal element does not have a successor in the order diagram.

Least element : Let a set M be partially ordered by the relation $\leq$. Then a minimal element $a \in M$ is called a least element in M if a is less than every other element in M. A least element of a set M is designated by leEl(M).

$$a \;=\; \text{leEl}(M) \quad :\Leftrightarrow\; (x \in M \quad \Rightarrow \quad a \leq x)$$

An ordered set need not contain a least element. If a least element exists, it is unique, for if a and b are least elements, then by definition $a \leq b$ and $b \leq a$, and hence $a = b$.

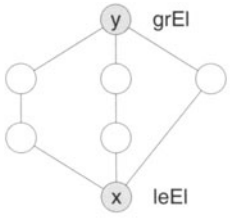
Greatest element : Let a set M be partially ordered by the relation $\leq$. Then a maximal element $a \in M$ is called a greatest element in M if every other element in M is less than a. A greatest element of a set M is designated by grEl(M).

$$a = \text{grEl}(M) \quad :\Leftrightarrow \quad (x \in M \quad \Rightarrow \quad x \leq a)$$

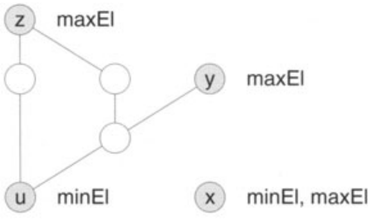
An ordered set need not contain a greatest element. If a greatest element exists, it is unique, for if a and b are greatest elements, then by definition $a \leq b$ and $b \leq a$, and hence $a = b$.

Example 1 : Extreme elements in finite sets

The following order diagrams for finite sets A and B show examples of extreme elements. The set A contains a minimal element x, which is also the least element in A, and a maximal element y, which is also the greatest element in A. The element x of the set B is also a minimal and maximal element in B. The set B contains the maximal elements x, y, z and the minimal elements u, x, but no least or greatest element.



ordered set A



ordered set B

Example 2 : Extreme elements in infinite sets

The infinite set $\mathbb{N}$ of natural numbers contains a minimal element 0, which is also the least element in $\mathbb{N}$. The set $\mathbb{Z}$ of integers contains no minimal, maximal, least or greatest element.

Upper bound : Let A be a subset of a partially ordered set $(M ; \leq)$. Then an element $a \in M$ is called an upper bound of A in M if every element $x \in A$ is less than or equal to a. An upper bound of a set A in a set M is designated by $\text{ub}_M(A)$.

$$a = \text{ub}_M(A) \quad :\Leftrightarrow \quad (x \in A \quad \Rightarrow \quad x \leq a)$$

A set A may possess zero, one or several upper bounds in M. If there is an upper bound a of A in M, then a is the greatest element in the set $A \cup \{a\}$. If the set A contains a greatest element a, then a is also an upper bound of A in M.

Lower bound : Let A be a subset of a partially ordered set $(M ; \leq)$. Then an element $a \in M$ is called a lower bound of A in M if a is less than or equal to every element $x \in A$. A lower bound of a set A in a set M is designated by $lb_M(A)$.

$$a = lb_M(A) \quad :\Leftrightarrow \quad (x \in A \quad \Rightarrow \quad x \leq a)$$

A set A may possess zero, one or several lower bounds in M. If there is a lower bound a of A in M, then a is the least element in the set $A \cup \{a\}$. If the set A contains a least element a, then a is also a lower bound of A in M.

Least upper bound : Let A be a subset of a partially ordered set $(M ; \leq)$. Then an element g of M is called a least upper bound (a supremum) of A in M if g is an upper bound of A in M and g is less than every other upper bound s of A in M. A least upper bound of a set A in a set M is designated by $lub_M(A)$.

$$g = lub_M(A) \quad :\Leftrightarrow \quad g = ub_M(A) \quad \wedge \quad (s = ub_M(A) \quad \Rightarrow \quad g \leq s)$$

A set A need not possess a least upper bound in M. If a least upper bound exists, it is unique, for if a and b are least upper bounds, then by definition $a \leq b$ and $b \leq a$, and hence $a = b$.

Greatest lower bound : Let A be a subset of a partially ordered set $(M ; \leq)$. Then an element g of M is called a greatest lower bound (an infimum) of A in M if g is a lower bound of A in M and every other lower bound s of A in M is less than g. A greatest lower bound of a set A in a set M is designated by $glb_M(A)$.

$$g = glb_M(A) \quad :\Leftrightarrow \quad g = lb_M(A) \quad \wedge \quad (s = lb_M(A) \quad \Rightarrow \quad s \leq g)$$

A set A need not possess a greatest lower bound in M. If a greatest lower bound exists, it is unique, for if a and b are greatest lower bounds, then by definition $a \leq b$ and $b \leq a$, and hence $a = b$.

4.4 ORDERED SETS WITH EXTREMALITY PROPERTIES

Introduction : Extreme elements are used to define special properties of ordered sets. Every subset of a noetherian / artinian ordered set contains a maximal / minimal element. Every subset of a well-ordered set contains a least element. Every well-ordered set is totally ordered (for instance the natural numbers ordered by the relation "less than"), but not every totally ordered set is well-ordered (for instance the integers ordered by the relation "less than").

Several types of ordered sets with extremality properties are defined in this section. The initial segment of an element a in a totally ordered set M contains all elements of M which are less than a. For any two elements a, b in a directed set there is an element which is greater or equal to a and b. A lattice is an ordered set whose subsets possess least upper and greatest lower bounds. For example, every power set ordered by inclusion is a complete lattice.

Noetherian ordered set : A partially ordered set M is said to be noetherian if every non-empty subset A of M contains a maximal element.

$$\bigwedge_{A \subseteq M} (A \neq \emptyset \quad \Rightarrow \quad \bigvee_{a \in A} (a = \text{maxEl}\,(A)))$$

Artinian ordered set : A partially ordered set M is said to be artinian if every non-empty subset A of M contains a minimal element.

$$\bigwedge_{A \subseteq M} (A \neq \emptyset \quad \Rightarrow \quad \bigvee_{a \in A} (a = \text{minEl}\,(A)))$$

Well-ordered set : A partially ordered set M is said to be well-ordered if every non-empty subset A of M contains a least element. The order relation of a well-ordered set is called a well-ordering.

$$\bigwedge_{A \subseteq M} (A \neq \emptyset \quad \Rightarrow \quad \bigvee_{a \in A} (a = \text{leEl}(A)))$$

Properties of well-ordered sets : A well-ordered set possesses the following properties:

(W1) Every subset of a well-ordered set is well-ordered.

(W2) Every well-ordered set is totally ordered.

Proof W1 : Every subset of a well-ordered set is well-ordered.

Let A be a subset of a well-ordered set M. Then by definition A contains a least element. Every subset of A is also a subset of M and hence contains a least element. Since every subset of A contains a least element, A is well-ordered.

Proof W2 : Every well-ordered set is totally ordered.

By definition, every subset of a well-ordered set M contains a least element. Hence for arbitrary elements $a, b \in M$ the subset $\{a, b\}$ of M contains a least element. It follows that a and b are comparable, and hence M is totally ordered.

Order diagram of a well-ordered set : The elements of a well-ordered set $\{a, b, c, ...\}$ are arranged consecutively on a line. The least element of the set is at the beginning of the line. Every element is less than all elements to its right.

$a \le b \le c$ ...

A well-ordered set may contain a greatest element. The greatest element is at the end of the line and has no successor. Every element except for the greatest element has a successor.

Example 1 : Order of the natural numbers and integers

The domain $(\mathbb{N} ; \le)$ with the natural numbers $\mathbb{N} = \{0, 1, 2, ...\}$ and the total order relation $\le$ is a well-ordered set, since every infinite subset of $\mathbb{N}$ possesses a least element. The domain $(\mathbb{Z} ; \le)$ with the integers $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$ and the total order relation $\le$ is not a well-ordered set, since the subset $\{-1, -2, ...\}$, for example, does not contain a least element.

Initial segment : Let $(M ; <)$ and $(M ; \le)$ be totally ordered sets. The subset A of M whose elements are less than $a \in M$ is called the strict initial segment of a in M. The subset B of M whose elements are less than or equal to $a \in M$ is called the initial segment of a in M.

$$\text{A is the strict initial segment of a in M} \quad :\Leftrightarrow \quad A = \{x \in M \mid x < a\}$$
$$\text{B is the initial segment of a in M} \quad :\Leftrightarrow \quad B = \{x \in M \mid x \le a\}$$

Successor : Let the domain $(M ; >)$ be totally ordered. If M has a greatest element g, then g does not have a successor. Assume that an element $x \in M$ is not the greatest element of M. Let A be the set of elements $z \in M$ which are greater than x. Then the element x has a unique successor if A has a greatest lower bound $y \ne x$.

$$A := \{z \in M \mid z > x\}$$
$$\text{y is the successor of x in M} \quad \Leftrightarrow \quad y = \text{glb}_M(A) \quad \wedge \quad y \ne x$$

In the general case, an element need not have a successor. However, for every element x of a well-ordered set M which is not the greatest element of M, the set A has a least element, and this is the successor of x.

Directed set : A partially ordered set $(M\,;\geq)$ is said to be directed if for any two elements $a,b\in M$ there is an element $c\in M$ which is greater than or equal to both a and b :

$$M \text{ is directed} \quad :\Leftrightarrow \quad \bigwedge_{a,b\in M}\ \bigvee_{c\in M} (c\geq a \ \wedge\ c\geq b)$$

Lattice : A partially ordered set $(M\,;\leq)$ is called a lattice if every subset of M which contains two elements possesses a least upper bound and a greatest lower bound in M. A lattice $(M\,;\leq)$ is said to be complete if every subset of M possesses a least upper bound and a greatest lower bound.

$$\begin{array}{llllll}
(M\,;\leq) \text{ is a lattice} & :\Leftrightarrow & a,b\in M & \Rightarrow & \bigvee\limits_{g\in M} (g = \text{lub}_M\{a,b\}) & \wedge \\
 & & & & \bigvee\limits_{h\in M} (h = \text{glb}_M\{a,b\}) & \\
(M\,;\leq) \text{ is complete} & :\Leftrightarrow & A\subseteq M & \Rightarrow & \bigvee\limits_{g\in M} (g = \text{lub}_M(A)) & \wedge \\
 & & & & \bigvee\limits_{h\in M} (h = \text{glb}_M(A)) &
\end{array}$$

Example 2 : A power set ordered by inclusion is a complete lattice.

In Example 1 of Section 4.2 it is shown that every power set P(M) is partially ordered by the inclusion $\subseteq$ (is a subset of). To prove that P(M) is a complete lattice, a subset A of P(M) is defined :

$$A := \{A_i\in P(M) \mid A_i\subseteq M \ \wedge\ i\in I\} \subseteq P(M)$$

(1) The set A possesses an upper bound $X = A_1\cup A_2\cup\ldots$ in P(M) :

$$A_i\in A \quad\Rightarrow\quad A_i\subseteq X$$

(2) Let the set Y be an upper bound of A in P(M). Then Y contains the set A_i for every $i\in I$. Hence $X = A_1\cup A_2\cup\ldots$ is a subset of Y, that is $X\subseteq Y$ for every upper bound Y of A in P(M). Hence X is the least upper bound of A in P(M).

(3) The set A possesses a lower bound $V = A_1\cap A_2\cap\ldots$ in P(M) :

$$A_i\in A \quad\Rightarrow\quad V\subseteq A_i$$

(4) Let the set W be a lower bound of A in P(M). Then W is contained in the set A_i for every $i\in I$. Hence W is a subset of $V = A_1\cap A_2\cap\ldots$, that is $W\subseteq V$ for every lower bound W of A in P(M). Hence V is a greatest lower bound of A in P(M).

Example 3 : Lattice structure of the power set P(M) with $M = \{a, b, c\}$

The elements of P(M) and the order diagram for the relation $\subseteq$ are shown in Example 7 of Section 4.2. The least upper and greatest lower bounds of some subsets of P(M) are determined using the formulas given in Example 2 :

$$A := \{\{a\}, \{a,c\}, \{a,b,c\}\}$$

$$\begin{aligned}
\text{lub}(A) &= \{a\} \cup \{a,c\} \cup \{a,b,c\} &&= \{a,b,c\} \\
\text{glb}\,(A) &= \{a\} \cap \{a,c\} \cap \{a,b,c\} &&= \{a\}
\end{aligned}$$

$$B := \{\{a\}, \{b\}, \{a,c\}\}$$

$$\begin{aligned}
\text{lub}\,(B) &= \{a\} \cup \{b\} \cup \{a,c\} &&= \{a,b,c\} \\
\text{glb}\,(B) &= \{a\} \cap \{b\} \cap \{a,c\} &&= \emptyset
\end{aligned}$$

$$C := \{\{a\}, \{a,c\}\}$$

$$\begin{aligned}
\text{lub}\,(C) &= \{a\} \cup \{a,c\} &&= \{a,c\} \\
\text{glb}\,(C) &= \{a\} \cap \{a,c\} &&= \{a\}
\end{aligned}$$

4.5 MAPPINGS OF ORDERED SETS

Introduction : In order to study the properties of ordered sets, the concept of "ordered sets with identical structure" is defined. Two ordered sets A and B are similar (isomorphic, similarly structured) if there is a bijective mapping $f : A \rightarrow B$ between them which is isotonic (homomorphic) in both directions. Ordered sets which are similar to the set of natural numbers are well-ordered.

The cardinalities of the sets of a given system of sets are defined in Section 2.7. Two sets A and B have the same cardinality if there is a bijective mapping $f : A \rightarrow B$. It follows from the definitions that similarly ordered sets have the same cardinality. It does not follow, however, that sets of the same cardinality are similarly ordered.

To characterize similarly ordered sets, the concept of order type is introduced as a class of similarly ordered sets. The relation $\le$ (less than or equal) is defined for comparing order types. In the special case of well-ordered sets, $\le$ is an order relation. The concept of ordinal number is introduced for the order types of well-ordered sets. Well-ordered sets can be counted through using the ordinal numbers of their subsets. Countable sets can only be counted through if they are well-ordered.

Isotonic mapping : Let the domains $(A ; \le_1)$ and $(B ; \le_2)$ be partially ordered sets. A mapping $f : A \rightarrow B$ is said to be isotonic (homomorphic) if $x \le_1 y$ for elements $x, y \in A$ implies $f(x) \le_2 f(y)$ for the images $f(x), f(y) \in B$.

$$f : A \rightarrow B \text{ is isotonic } :\Leftrightarrow \quad (x \le_1 y \quad \Rightarrow \quad f(x) \le_2 f(y))$$

Similarly ordered sets : Let the domains $(A ; \le_1)$ and $(B ; \le_2)$ be partially ordered sets. The ordered sets are said to be similar (isomorphic) if there is a bijective mapping $f : A \rightarrow B$ and both f and f^{-1} are isotonic. The similarity of sets is designated by $A \cong B$ (A is similar to B).

The similarity relation $\cong$ is an equivalence relation for a system $M = \{(A ; \le_1), (B ; \le_2), (C ; \le_3), ...\}$ of partially ordered sets :

(1) The relation is reflexive : By virtue of the identity mapping every ordered set is similar to itself.

(2) The relation is symmetric : $A \cong B$ implies $B \cong A$, since every isotonic mapping f has an isotonic inverse f^{-1}.

(3) The relation is transitive : $A \cong B$ and $B \cong C$ imply $A \cong C$, since for isotonic mappings $f : A \rightarrow B$ and $g : B \rightarrow C$:

$$x \le_1 y \quad \Rightarrow \quad f(x) \le_2 f(y) \quad \Rightarrow \quad g \circ f(x) \le_3 g \circ f(y)$$

Order type : Let a system $M = \{A, B,...\}$ of partially ordered sets be given. The set M is partitioned into disjoint classes of similarly ordered sets using the equivalence relation $\cong$ (similarly ordered). An element of the quotient set $M/\cong$ is called an order type of the system M of sets. The canonical mapping from M to $M/\cong$ is designated by otype :

$$\text{otype} : \quad M \to M/\cong \qquad \text{with} \qquad \text{otype}((A\,;\sqsubseteq)) = [(A\,;\sqsubseteq)]$$

$(A\,;\sqsubseteq)$ partially ordered set, element of M

$[(A\,;\sqsubseteq)]$ class with the representative $(A\,;\sqsubseteq)$

Comparable order types : The order type of a partially ordered set $(A\,;\sqsubseteq_1)$ is said to be less than or equal to (symbol $\le$) the order type of a partially ordered set $(B\,;\sqsubseteq_2)$ if A is similar to a subset S of B :

$$\text{otype}\,(A\,;\sqsubseteq_1) \le \text{otype}\,(B\,;\sqsubseteq_2) \quad :\Leftrightarrow \quad \bigvee_S (A \cong S \;\wedge\; S \subseteq B)$$

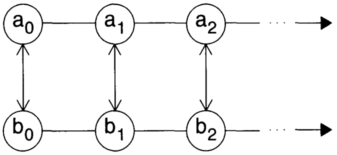
In the following, it is shown that the relation $\le$ for a system of well-ordered sets is an order relation. If the sets are not well-ordered, then $\le$ is generally not an order relation.

Order relation for well-ordered sets : For a system $M = \{A, B, C,...\}$ of well-ordered sets, the relation $\le$ (less than or equal to) has the properties of an order relation :

(1) The relation is reflexive, since every ordered set A is a subset of itself and similar to itself.

(2) The relation is transitive. In fact, otype $(A) \le$ otype (B) implies $A \cong S$ with $S \subseteq B$, and otype $(B) \le$ otype (C) implies $B \cong T$ with $T \subseteq C$. Hence there is a subset U of T such that $A \cong S \cong U$ and $U \subseteq T \subseteq C$. From $A \cong U$ and $U \subseteq C$ it follows that otype $(A) \le$ otype (C).

(3) The relation is antisymmetric. Each of the well-ordered sets of M is totally ordered and contains a least element. For the chains $A = < a_0, a_1, a_2,... >$ and $B = < b_0, b_1, b_2,... >$ the two conditions otype $(A) \le$ otype (B) and otype $(B) \le$ otype (A) can only both be satisfied if otype $(A) =$ otype (B).



$(A\,;\sqsubseteq_1)$

$f : A \to B$ and $f^{-1} : B \to A$

$(B\,;\sqsubseteq_2)$

Well-ordered order types : Let a system $M = \{A, B, ...\}$ of partially ordered sets with the quotient set $M/\cong$ for the equivalence relation $\cong$ (similarly ordered) be given. An equivalence relation may be defined in the quotient set $M/\cong$. Let two order types be equivalent if both contain only well-ordered sets or both contain only non-well-ordered sets :

(1) Let the sets $A, B \in M$ be similar, and therefore elements of the same class $[A]$ in $M/\cong$. Let the set A be well-ordered. Then the set B is also well-ordered. In fact, for every subset $\{x, y, z, ...\}$ of A with least element x the isotonic mapping $f : A \rightarrow B$ yields a subset $\{f(x), f(y), f(z), ...\}$ of B with least element $f(x)$. Every subset of B corresponds to exactly one subset of A.

(2) It follows from (1) that every class of $M/\cong$ contains either only well-ordered or only non-well-ordered sets. The equivalence relation therefore partitions the quotient set $M/\cong$ into the class $M_w/\cong$ of well-ordered order types and the class $M_n/\cong$ of non-well-ordered order types.

Ordinal numbers : Let the subset of well-ordered sets of a system $M = \{A, B, ...\}$ of sets be M_w. The order type of a well-ordered set $A \in M_w$ is called the ordinal number of A in the system M_w and is designated by ord(A). The mapping ord is the restriction of the mapping otype to the subset M_w of M :

$$\text{ord} : M_w \rightarrow M_w/\cong \qquad \text{with} \qquad \text{ord}\,(A\,;\sqsubseteq_1) = [(A\,;\sqsubseteq_1)]$$

$$\text{ord}(A\,;\sqsubseteq_1) \leq \text{ord}\,(B\,;\sqsubseteq_2) \quad :\Leftrightarrow \quad \bigvee_S \,(A \cong S \;\wedge\; S \subseteq B)$$

Well-orderings of a finite set : Every unstructured finite set $A = \{b, c, a, ...\}$ can be well-ordered. To this end, the subset $\{a\}$ is formed with an arbitrary element $a \in A$. In the difference $A - \{a\}$, an arbitrary element b is chosen, and the union $\{a\} \cup \{b\} = \{a, b\}$ is formed. In the difference $A - \{a, b\}$, an arbitrary element c is chosen, and the union $\{a, b\} \cup \{c\} = \{a, b, c\}$ is formed. By continuing this construction, a system of subsets is formed which is well-ordered with respect to the inclusion $\subset$:

$$\emptyset \subset \{a\} \subset \{a, b\} \subset \{a, b, c\} \subset ...$$

If the set A contains n elements, then this construction can be carried out in $n(n-1)(n-2)... = n!$ different ways. Hence there are $n!$ different well-orderings of a finite set of cardinality n. If subsets with the same number of elements are bijectively mapped onto each other, then the mappings between these well-orderings are isotonic. Hence the $n!$ different well-orderings are similar.

Finite ordinal numbers : The order types of well-ordered finite sets are ordinal numbers. The subsets $\emptyset, \{0\}, \{0, 1\}, \{0, 1, 2\}, ...$ of the natural numbers are chosen as representatives of the classes. The cardinal numbers $0, 1, 2, 3, ...$ of these subsets are used to designate the finite ordinal numbers.

Counting through a well-ordered set : To count through the elements of a well-ordered set $A = \{b, c, a, ...\}$, a system of subsets is formed. The least element $a \in A$ is used to form the first subset $\{a\}$. The least element b of the difference $A - \{a\}$ is determined, and the union $\{a\} \cup \{b\} = \{a, b\}$ is formed. The least element c of the difference $A - \{a, b\}$ is determined, and the union $\{a, b\} \cup \{c\} = \{a, b, c\}$ is formed. The subsets are well-ordered with respect to the inclusion $\subset$.

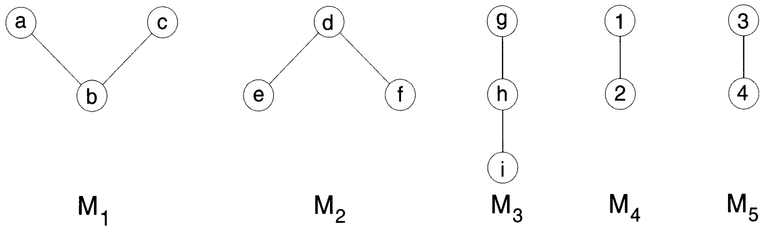
$$\emptyset \subset \{a\} \subset \{a, b\} \subset \{a, b, c\} \subset ...$$

Every subset is designated by its cardinal number. The difference of successive subsets contains exactly one element. The cardinal number of a given subset is mapped bijectively to the element that is contained in the difference between this subset and its predecessor. For example, the cardinal number 3 of the subset $\{a, b, c\}$ is mapped to the element c, since $\{a, b, c\} - \{a, b\} = \{c\}$. The elements of A are counted through using these cardinal numbers.

Similarity of finite well-ordered sets : Every well-ordered set is a chain with a least element. Finite well-ordered sets with the same cardinal number are therefore similar. If the cardinal numbers of two finite well-ordered sets are different, then the set with the lower cardinal number is similar to a subset of the other set. This subset is not unique.

Example 1 : Order types of a system of sets

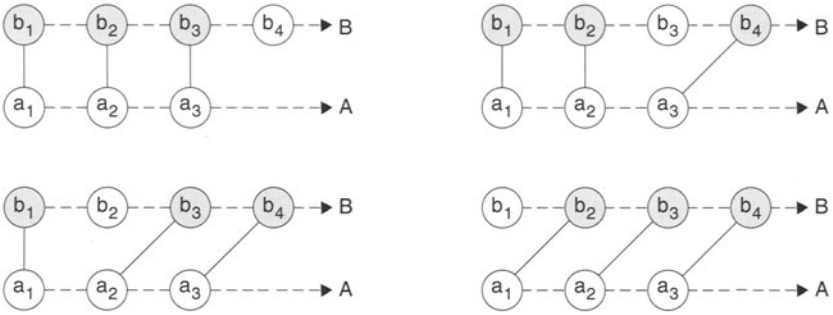
The following order diagrams show a system M of ordered sets $(M_i ; \sqsubseteq)$. None of the sets M_1, M_2, M_3 can be similar to one of the sets M_4, M_5, since there is no bijective mapping between finite sets with different numbers of elements. There are no isotonic mappings between the sets M_1, M_2, M_3. The sets M_4 and M_5 are similar. The domain $(M_1 ; \sqsubseteq)$ is not well-ordered, since the subset $\{a, c\}$ contains no least element. The domain $(M_2 ; \sqsubseteq)$ is not well-ordered, since M_2 contains no least element. The quotient set $M / \cong$ consists of four classes.



equivalence classes : $\{[\{a,b,c\}],[\{d,e,f\}], [\{g,h,i\}], [\{1,2\}]\}$
$[\{1,2\}] = \{\{1,2\}, \{3,4\}\}$

Example 2 : Similarity of finite well-ordered sets

Let the set $A = \{a_1, a_2, a_3\}$ be well-ordered with $a_1 < a_2 < a_3$. Let the set $B = \{b_1, b_2, b_3, b_4\}$ be well-ordered with $b_1 < b_2 < b_3 < b_4$. Then the set A is similar to every subset of B that consists of three elements.



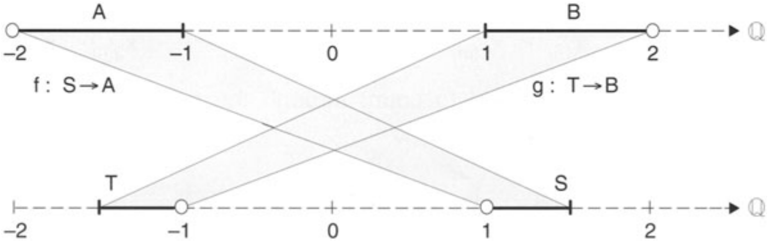
Example 3 : Order type of infinite non-well-ordered sets

Let A be the infinite set of rational numbers in the interval]–2, –1], and let B be the infinite set of rational numbers in the interval [1, 2[. The ordered sets (A ; ≤) and (B ; ≤) are not similar, since A contains a greatest element while B does not.

While the sets A and B are not similar, A is similar to the subset S =]1, 1.5] of B. Likewise, B is similar to the subset T = [–1.5, –1[of A. The corresponding isotonic mappings f and g are defined as follows :

$$f : \quad S \to A \quad \text{with} \quad f(x) = 2x - 4 \quad \Rightarrow \quad \text{otype}(A) \le \text{otype}(B)$$
$$g : \quad T \to B \quad \text{with} \quad g(x) = 2x + 4 \quad \Rightarrow \quad \text{otype}(B) \le \text{otype}(A)$$

For the order types of A and B, the relation ≤ is not an order relation. Since the sets A and B are not similar, otype (A) ≠ otype (B) although otype (A) ≤ otype (B) and otype (B) ≤ otype (A).



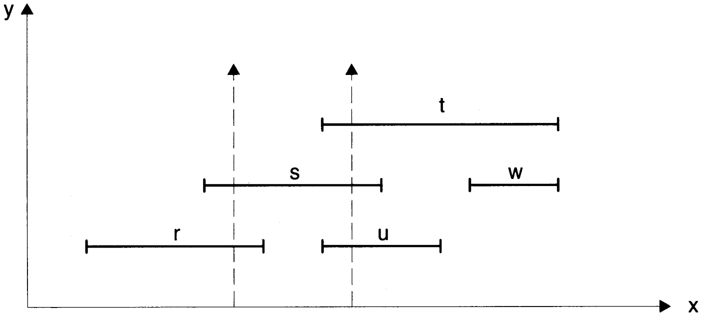
Example 4 : Counting through parallel line segments

Let M = {r, s, t, u, w} be a set of line segments parallel to the x-axis. Every segment is determined by the ordered pair (x_1, x_2) of the x-coordinates of its beginning and its end. The segments may be ordered according to their beginning x_1 or according to their end x_2. These orderings are not similar.

ordering according to beginning : $i \sqsubseteq_1 m \;:\Leftrightarrow\; x_{1(i)} \leq x_{1(m)}$

order according to end : $i \sqsubseteq_2 m \;:\Leftrightarrow\; x_{2(i)} \leq x_{2(m)}$

Let the statement "segment r covers segment s" be true if a straight line in the positive y-direction intersects first r and then s. Let the relation $\sqsubseteq_3$ be the set of pairs (r, s) for which the statement "r covers s" is true. The relation $\sqsubseteq_3$ is not an order relation, since it is not transitive : From "r covers s" and "s covers t" it does not generally follow that "r covers t". This is illustrated in the following example.



r covers s s covers t
r does not cover t

None of the relations $\sqsubseteq_1$ to $\sqsubseteq_3$ is a well-ordering. For $\sqsubseteq_1$, the subset {t, u} has no least element. For $\sqsubseteq_2$, the subset {t, w} has no least element. The relation $\sqsubseteq_3$ is not an order relation. Both coordinates of the pair (x_1, x_2) must be used in order to define a well-ordering $\sqsubseteq_4$. For example, segments with equal coordinates x_1 may be ordered according to the coordinate x_2 :

$$i \sqsubseteq_4 m \;:\Leftrightarrow\; x_{1(i)} < x_{1(m)} \;\vee\; (x_{1(i)} = x_{1(m)} \;\wedge\; x_{2(i)} \leq x_{2(m)})$$

The relation $\sqsubseteq_4$ may be used to count through the segments in the chain < r, s, u, t, w >.

4.6 PROPERTIES OF ORDERED SETS

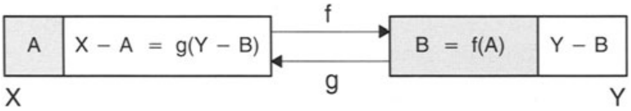
Introduction : The order type of a partially ordered set X determines whether a certain mapping $f: X \rightarrow X$ has a fixed point $f(x) = x$. Fixed points of mappings are used to prove equivalent properties of partially ordered sets. In particular, the following statements are proved to be equivalent :

- Zorn's Lemma : If every totally ordered subset of a partially ordered set X has a least upper bound, then X contains a maximal element.
- Zermelo's Theorem : Every set may be well-ordered.
- Axiom of Choice : For an indexed family of sets S_i with the index set I, there is a mapping $f : I \rightarrow \cup S_i$ with $f(i) \in S_i$.

Fixed points in mappings of ordered sets : A point $x \in A$ is called a fixed point of a mapping $f : A \rightarrow A$ if $f(x) = x$. If the set A is partially ordered, the existence of fixed points may be deduced from properties of the mapping and of the order relation. The following theorems serve this purpose :

(F1) Let the domain $(M ; \leq)$ be a complete lattice. Then every isotonic mapping $f : M \rightarrow M$ has a fixed point.

(F2) For arbitrary mappings $f : X \rightarrow Y$ and $g : Y \rightarrow X$ there are subsets $A \subseteq X$ and $B \subseteq Y$ such that $f(A) = B$ and $g(Y - B) = X - A$.



(F3) Let a set M be partially ordered by the relation $\geq$. Then every mapping $f : M \rightarrow M$ with $f(x) \geq x$ has a fixed point if every well-ordered subset of M has a least upper bound in M.

Proof F1 : Let the domain $(M ; \leq)$ be a complete lattice. Then every isotonic mapping $f : M \rightarrow M$ has a fixed point.

(1) Let A be the subset of M whose elements x are less than or equal to their image f(x). Since M is a complete lattice, A has a least upper bound a in M.

$$A := \{x \in M \mid x \leq f(x)\}$$

$$a := \text{lub}_M(A)$$

(2) Since the mapping f is isotonic, the image f(x) of every element x of A is also an element of A :

$$x \in A \quad \Rightarrow \quad x \leq f(x) \quad \Rightarrow \quad f(x) \leq f(f(x)) \quad \Rightarrow \quad f(x) \in A$$

(3) Since the mapping f is isotonic, the image of every element $x \in A$ is less than or equal to the image of the least upper bound a of A in M. Therefore f(a) is an upper bound of A in M :

$$x \in A \;\Rightarrow\; x \leq a \;\Rightarrow\; f(x) \leq f(a) \;\Rightarrow$$
$$x \leq f(x) \leq f(a) \;\Rightarrow\; f(a) = ub_M(A)$$

(4) Since a is a least upper bound and f(a) is an upper bound of A in M, it follows that $a \leq f(a)$. Hence the least upper bound a is an element of A, that is $a \in A$. In (2) it was shown that this implies that f(a) is also an element of A, that is $f(a) \in A$. Since a is the least upper bound of A in M, $a \in A$ and $f(a) \in A$ together imply $f(a) \leq a$. Thus $a \leq f(a) \leq a$, and hence $a = f(a)$. Therefore a is a fixed point of the mapping f.

Proof F2 : For arbitrary mappings $f : X \to Y$ and $g : Y \to X$ there are subsets $A \subseteq X$ and $B \subseteq Y$ such that $f(A) = B$ and $g(Y - B) = X - A$.

(1) The power set P(X) of the given set X is partially ordered by the inclusion $\subseteq$. Define a mapping $h : P(X) \to P(X)$ such that the image of a subset $S \subseteq X$ is the difference of X and the image of the difference of Y and f(S) under g :

$$h(S) := X - g(Y - f(S))$$

(2) The mapping h is isotonic. In fact, $S \subseteq T$ implies $f(S) \subseteq f(T)$ and hence $Y - f(T) \subseteq Y - f(S)$, so that $g(Y - f(T)) \subseteq (Y - f(S))$ and hence $h(S) \subseteq h(T)$.

(3) The power set P(X) is a complete lattice (see Example 2 in Section 4.4). By property (F1), the isotonic mapping h therefore has a fixed point $h(A) = A$. Let $B := f(A)$. Then the definition of h(S) under (1) yields :

$$g(Y - B) = g(Y - f(A)) = X - h(A) = X - A$$

Proof F3 : Let the domain $(M ; \geq)$ be a partially ordered set. Then every mapping $f : M \to M$ with $f(x) \geq x$ has a fixed point if every well-ordered subset of M has a least upper bound in M.

Note : The proof is complicated by the fact that the sets involved are allowed to be uncountable. To facilitate understanding, the following paragraphs describe the construction of countable sets which satisfy conditions (1a) and (1b) of the proof. However, these conditions are also suitable for characterizing uncountable sets.

For a freely chosen element $a \in M$, the image f(a) is determined. If $f(a) = a$, the desired fixed point has been found. Otherwise $f(a) > a$ by hypothesis. Choose f(a) as the successor of a. Analogously, f(a) is either a fixed point or may be used to determine a successor f(f(a)). The element a and its successors are combined in the chain $A := \{a, f(a), f(f(a)), \ldots\}$.

If the chain A is finite, its greatest element is by construction a fixed point of f. Otherwise, it has a least upper bound $b \in M$ by hypothesis. An analogous chain $B := \{b, f(b), f(f(b)), \dots\}$ is formed beginning with the element b. This process is continued. Since the least element of every chain is by construction greater than every element of the preceding chain, the union of the chains A, B,... is also a chain. The sets considered in the proof are initial segments of this union.

For the general case, the proof is carried out as follows :

(1) An element a is chosen in the set M. Let S be the set of the subsets of $Y \subseteq M$ with the following properties :

(a) Each of the subsets Y is well-ordered with least element a and successor function f_Y. Thus the successor of $y \in Y - \{lub_M(Y)\}$ is $f_Y(y)$. Here f_Y is the restriction of the given function f to the subset $Y - \{lub_M(Y)\}$.

(b) Let the strict initial segment of an arbitrary element $y \in Y$ with $y \neq a$ be $A_Y(y) = \{z \in Y \mid z < y\}$. Then the least upper bound of $A_Y(y)$ in M is an element of Y.

(2) The proof is carried out in the following steps :

(a) For different elements $Y, Z \in S$, either Z is the strict initial segment of an element in Y or Y is the strict initial segment of an element in Z.

(b) The union W of the sets $Y \in S$ is an element of S.

(c) The set W contains a greatest element, which is a fixed point of f.

(3) To prove (2a), consider the subset V of elements $x \in Y \cap Z$ whose initial segments in Y and Z coincide :

$V := \{ x \in Y \cap Z \mid B_Y(x) = B_Z(x) \}$

(a) First consider the case that V contains a greatest element v_0. Then $B_Y(v_0) = B_Z(v_0)$. If v_0 is not the greatest element of Y, then Y contains the successor $f(v_0)$. Likewise, if v_0 is not the greatest element of Z, then Z contains the successor $f(v_0)$. Now assume that v_0 is neither the greatest element of Y nor the greatest element of Z. Then both Y and Z contain the successor $f(v_0)$. It follows from $v_1 = f(v_0) \in Y \cap Z$ and $B_Y(v_0) = B_Z(v_0)$ that $B_Y(v_1) = B_Z(v_1)$, so that $f(v_0) \in V$. The successor $f(v_0) \in V$ is greater than $v_0 \in V$, contradicting the hypothesis that v is the greatest element of V. Hence, contrary to the assumption, v_0 is the greatest element of one of the sets Y and Z. This set is the strict initial segment of the other set with respect to the element $f(v_0)$.

(b) Now consider the case that V has no greatest element. Assume $Y \neq V$. Then the non-empty subset $Y - V$ of the well-ordered set Y contains a least element y_0, and V is the strict initial segment of y_0 in Y. Hence by property (1b) of $A_Y(y_0)$ the least upper bound v_2 of V in M is an element of Y :

$$\begin{aligned} y_0 = \text{leEl}(Y - V) \quad &\Rightarrow \quad V = A_Y(y_0) \\ &\Rightarrow \quad v_2 = \text{lub}_M(V) \in Y \end{aligned}$$

Analogously, the assumption $Z \neq V$ implies that v_2 is an element of Z. Thus $Y \neq V$ and $Z \neq V$ implies $v_2 \in Y \cap Z$. Since v_2 is the least upper bound of V in M, it follows that $v_2 \in V$. This contradicts the hypothesis that V has no greatest element. Hence $Y = V$ or $Z = V$, proving (2a).

(4) To prove (2b), consider an element y_0 of a set $Y \in S$.

(a) The strict initial segment of y_0 in Y is $A_Y(y_0) = \{y \in Y \mid y < y_0\}$. Let Z be an arbitrary other element of S. The initial segment of y_0 in $Y \cup Z$ is to be determined. By (3), either Y is a strict initial segment in Z or Z is a strict initial segment in Y. If Y is a strict initial segment in Z, then y_0 is an element of Z with the strict initial segment $A_Z(y_0) = A_Y(y_0)$. If $y_0 \in Z$ and Z is a strict initial segment in Y, then $A_Z(y_0) = A_Y(y_0)$ also holds. If Z is a strict initial segment in Y which does not contain y_0, then $Z \subseteq A_Y(y_0)$. Altogether, $A_{Y \cup Z}(y_0) = A_Y(y_0)$.

(b) The union $W = \bigcup\{y \mid y \in S\}$ is a chain, since arbitrary elements $y \in Y \in S$ and $z \in Z \in S$ are comparable. In fact, $Y \subseteq Z$ or $Z \subseteq Y$ by (2a). Without loss of generality, let $Z \subseteq Y$, so that $Y \cup Z = Y$. Then $y, z \in Y$. But Y is well-ordered by virtue of (1a), and hence y and z are comparable.

(c) The strict initial segment $A_W(y_0) = \{w \in W \mid w < y_0\}$ in the union W is formed. From $y_0 \in Y$ and (4a) it follows that $A_W(y_0) = A_Y(y_0)$. Hence $A_W(y_0)$, like $A_Y(y_0)$, is well-ordered with successor function f.

(d) The union W is well-ordered, since every subset W_s of the chain W has a least element. In fact, for a freely chosen element $w_0 \in W_s$ there is a $Y \in S$ with $w_0 \in Y$, and one obtains :

$$H := \{w \in W_s \mid w \leq w_0\} \subseteq B_W(w_0) = B_Y(w_0) \subseteq Y$$

Since Y is well-ordered, H contains a least element, which is also the least element of the chain W_s. Let the successor of an element $w \in W$ be $u \in Y \in S$. Then $w < u$ is contained in $B_W(u) = B_Y(u)$. Since f(w) is the successor of w in Y, f(w) is also the successor u of w in W. Hence W is well-ordered with successor function f.

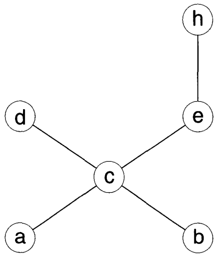
(e) By virtue of (1b), the least upper bound of $A_W(y_0) = A_Y(y_0)$ in M is an element of Y, and hence of W. Altogether, W possesses properties (1a) and (1b), so that $W \in S$.

(5) To prove (2c), consider the union $W = \bigcup \{Y \mid Y \in S\}$. It was shown in (4) that $W \in S$ is a well-ordered subset of M. By hypothesis, W therefore has a least upper bound $b = \text{lub}_M(W)$.

(a) Assume that the least upper bound b is not an element of W. Then $T := W \cup \{b\}$ is well-ordered with least element a and successor function f. The least upper bound of an arbitrary strict initial segment $A_T(y)$ is an element of T. Hence T possesses properties (1a) and (1b), so that $T = W \cup \{b\} \in S$. Thus, contrary to the assumption, $b \in W$. Hence b is the greatest element of W.

(b) Assume $f(b) > b$. It follows that f(b) is not an element of W, since b is the greatest element of W. Then $U := W \cup \{f(b)\}$ with $b \in W$ is well-ordered with least element a and successor function f. The least upper bound of an arbitrary strict initial segment $A_U(y)$ is an element of U. Hence U possesses properties (1a) and (1b), so that $U = W \cup \{f(b)\} \in S$. Hence, contrary to the assumption, $f(b) \in W$. Since b is the greatest element of W and $f(b) \geq b$ by hypothesis, it follows that $f(b) = b$. Thus the greatest element b of W is a fixed point of the mapping f.

Example 1 : Fixed point of a mapping with $f(x) \geq x$ in a partially ordered set $(M ; \geq)$ with property (F3)



The diagram shows a partial ordering of a set $M = \{a, b, c, d, e, h\}$. Let a mapping $f : M \rightarrow M$ be defined as follows :

$$f(a) = e \qquad f(c) = d \qquad f(e) = h$$
$$f(b) = c \qquad f(d) = d \qquad f(h) = h$$

For the element $a \in M$, the set S defined in Proof F3 contains the following well-ordered subsets of M :

$$S = \{\{a\}, \{a, e\}, \{a, e, h\}\}$$

The intersection of the sets $Y = \{a, e\}$ and $Z = \{a, e, h\}$ is $Y \cap Z = \{a, e\}$. The initial segments of a and e in Y and Z are equal :

$$B_Y(a) = B_Z(a) = \{a\}$$
$$B_Y(e) = B_Z(e) = \{a, e\}$$

The set defined in Proof F3 is $V = \{a, e\}$. It contains a greatest element e. The set Y is the strict initial segment of the element $h \in Z$. The union of the sets in S is $W = \{a, e, h\}$. The set W is an element of S and contains a greatest element h, which is a fixed point of the mapping.

The element $b \in M$ gives rise to the set $S = \{\{b\}, \{b, c\}, \{b, c, d\}\}$. This set of well-ordered subsets of M leads to the fixed point $f(d) = d$.

Equivalent properties of ordered sets

(E1) For every set X with the power set P(X) and $P_0(X) = P(X) - \emptyset$ there is a mapping $f : P_0(X) \rightarrow X$ with $f(A) \in A \subseteq X$.

(E2) A partially ordered set $(X\, ;\, \le)$ has a maximal element a if every well-ordered subset of X has a least upper bound in X.

(E3) Every partially ordered set $(X\, ;\, \le)$ contains a totally ordered subset which is not contained in any other totally ordered subset of X. This subset is called a maximal chain of X.

(E4) If every totally ordered subset (chain) of a partially ordered set $(X\, ;\, \le)$ has an upper bound, then X contains a maximal element (Maximality Principle, Zorn's Lemma).

(E5) Every set can be well-ordered (Zermelo's Theorem).

(E6) If a mapping $f : X \rightarrow Y$ is surjective, then there is an injection $g : Y \rightarrow X$ with $f \circ g(y) = y$.

(E7) Let $\{ S_i \mid i \in I \}$ be an indexed family of non-empty sets S_i. Then there is a choice function $f : I \rightarrow \bigcup \{ S_i \mid i \in I \}$ with $f(i) \in S_i$ (Axiom of Choice).

In the following, it is shown that each of the properties E2,...,E7 follows from the preceding property and that E1 follows from E7.

Proof : E1 $\Rightarrow$ E2

Assume that statement (E2) is false. Then $A := \{x \in X \mid x > a\} \neq \emptyset$ for every $a \in X$. By (E1) there is a mapping $f : P_0(X) \rightarrow X$ with $f(A) \in A$, so that $f(A) > a$. Hence the mapping $g : X \rightarrow X$ with $g(a) = f(A)$ has the property $g(a) > a$. By theorem (F3), however, g has a fixed point $g(x) = x$. Thus, contrary to the assumption, statement (E2) is true.

Proof : E2 $\Rightarrow$ E3

Let the set S of all chains in X be partially ordered by inclusion. Let the set C be a well-ordered subset of S, so that $K_1, K_2 \in C \Rightarrow K_1 \subseteq K_2 \ \vee \ K_2 \subseteq K_1$. The union $K = \bigcup \{ K_i \mid K_i \in C \}$ is a chain which includes every element of C. Since the chains are ordered by inclusion, K is the least upper bound of C. Since every well-ordered subset of S has a least upper bound in S, it follows by (E2) that S has a maximal element. This element is a maximal chain.

Proof : E3 $\Rightarrow$ E4
By statement (E3), every partially ordered set X contains a maximal chain. Choose a maximal chain K in X. By hypothesis, K possesses an upper bound x. This bound is a maximal element of X.

Proof : E4 $\Rightarrow$ E5
Let W be the set of well-ordered subsets $(M ; \sqsubset_M)$ of X. The set W is ordered using the relation $\le$, which is defined as follows for $(A, ; \sqsubset_A), (B ; \sqsubset_B) \in W$:

$$(A ; \sqsubset_A) = (B ; \sqsubset_B) \quad :\Leftrightarrow \quad A = B \quad \text{and} \quad \sqsubset_A = \sqsubset_B$$
$$(A ; \sqsubset_A) < (B ; \sqsubset_B) \quad :\Leftrightarrow \quad \text{A is a strict initial segment of B and } \sqsubset_A \text{ is the restriction of } \sqsubset_B \text{ to } A \times A$$

By part (4) of Proof F3, the union of the elements of a well-ordered subset T of W is an element of W, and hence an upper bound of T in W. It follows by (E4) that W contains a maximal element. Let this be $(C ; \sqsubset_C)$. To prove (E5), it is to be shown that $C = X$.

Assume $C \ne X$. The set $N = C \cup \{x\}$ is formed with an element $x \in X - C$. Since C is well-ordered, N is well-ordered if x is defined to be greater or equal to every element of C. Thus the set $(N ; \sqsubset_N)$ with $\sqsubset_N = \sqsubset_C \cup (C \times \{x\})$ is well-ordered. Contrary to the hypothesis, $(C ; \sqsubset_C)$ is not maximal, since $(C ; \sqsubset_C) < (N ; \sqsubset_N)$. The contradiction shows that, contrary to the assumption, $C = X$.

Proof : E5 $\Rightarrow$ E6
By statement (E5), the set X may be well-ordered. Since the mapping $f : X \to Y$ with $f(x) = y$ is surjective, the preimage $f^{-1}(y)$ contains at least one element of X. Since X is well-ordered, $f^{-1}(y)$ is well-ordered, . Define the mapping $g : Y \to X$ such that $g(y)$ is the least element of $f^{-1}(y)$. Then $f \circ g(y) = y$.

Proof : E6 $\Rightarrow$ E7
Let $S := \bigcup \{ S_i \mid i \in I\}$ and $X := \{(s, i) \in S \times I \mid s \in S_i\}$. Define the projections $p_S : X \to S$ with $p_S((s, i)) = s$ and $p_I : X \to I$ with $p_I((s, i)) = i$. Then p_I is surjective, since $S_i \ne \emptyset$. By (E6) there is an injection $g : I \to X$ with $g(i) = (s, i)$ for some $s \in S_i$. The mapping $f : I \to S$ with $f(i) = p_S \circ g(i)$ is a choice function since $f(i) = p_S((s, i)) = s$ for some $s \in S_i$.

Proof : E7 $\Rightarrow$ E1
The set $A \in P_0(X)$ is designated by S_A. Then $S = \{ S_A \mid A \in P_0(X)\}$ is an indexed family of non-empty sets $S_A \ne \emptyset$. Since for every element $x \in X$ there is a set $\{x\}$ in $P_0(X)$, the union of these sets is $\bigcup \{ S_A \mid A \in P_0(X)\} = X$. By (E7) there is a function $f : P_0(X) \to X$ with $f(A) \in A$ for every non-empty subset $A \in P_0(X)$.

4.7 ORDERED CARDINAL NUMBERS

Introduction : To compare the cardinal numbers introduced in Section 2.7, the order relation "less than or equal to" is defined. The cardinality of a set is less than the cardinality of its power set. The cardinality of the set $\mathbb{R}$ of real numbers is greater than the cardinality of the set $\mathbb{N}$ of natural numbers. The cardinality of the n-fold cartesian product $\mathbb{R}^n$ is equal to the cardinality of $\mathbb{R}$.

Comparison of cardinal numbers : Let $S = \{A, B, C, ...\}$ be a system of sets. The set S is partitioned into classes of equipotent sets using the equivalence relation $\sim$ (equipotent). The quotient set $S/\sim$ is the set of cardinal numbers for S. The cardinal number of A is said to be less than or equal to the cardinal number of B if A is equipotent with a subset $C \subseteq B$ in S.

$$\text{card } A \le \text{card } B \quad :\Leftrightarrow \quad \bigvee_{C \in S} (C \subseteq B \quad \wedge \quad A \sim C)$$

Order relation for cardinal numbers : Let $S = \{A, B, C, ...\}$ be a system of sets. The cardinal numbers of the sets are partially ordered by the relation $\le$ (less than or equal to), since $\le$ possesses the properties of an order relation :

(1) The relation $\le$ is reflexive, since for $A \sim A$:

$$\text{card } A \in S/\sim \quad \Rightarrow \quad \text{card } A \le \text{card } A$$

(2) The relation $\le$ is antisymmetric. By definition, for card A $\le$ card B there is an injection $f : A \to B$ and for card B $\le$ card A there is an injection $g : B \to A$. By the fixed point property (F2), there are sets $X \subseteq A$ and $Y \subseteq B$ such that $f(X) = Y$ and $g(B - Y) = A - X$. Since f and g are injections, the restricted mappings $f_X : X \to Y$ and $g_{B-Y} : (B - Y) \to (A - X)$ are bijective. Hence $h : A \to B$ with $h_X = f_X$ and $h_{A-X} = (g_{B-Y})^{-1}$ is also bijective, so that card A = card B.

$$\text{card } A \le \text{card } B \quad \wedge \quad \text{card } B \le \text{card } A \quad \Rightarrow \quad \text{card } A = \text{card } B$$

(3) The relation $\le$ is transitive. If card A $\le$ card B there is an injection $f : A \to B$. If card B $\le$ card C there is an injection $g : B \to C$. The composition of these injections is an injection $h : A \to C$ with $h = g \circ f$, and hence card A $\le$ card C.

$$\text{card } A \le \text{card } B \quad \wedge \quad \text{card } B \le \text{card } C \quad \Rightarrow \quad \text{card } A \le \text{card } C$$

Cardinality of a power set : The cardinality of the power set P(M) of a set M is greater than the cardinality of M. It is first proved indirectly that card P(M) $\le$ card M does not hold. Then card M $\le$ card P(M) is shown to hold. Altogether, it follows that card P(M) > card M :

$$\neg(\text{card } P(M) \le \text{card } M) \quad \wedge \quad (\text{card } M \le \text{card } P(M)) \quad \Rightarrow$$

$$\text{card } P(M) > \text{card } M$$

(1) Let a mapping $f : M \rightarrow P(M)$ be surjective. For every $x \in M$, $f(x)$ is the subset of M to which x is mapped. Consider the set T of those elements of M which are not contained in their image :

$$T := \{x \in M \mid x \notin f(x)\}$$

Since f is surjective and $T \in P(M)$, there is a preimage x_0 in M with $f(x_0) = T$. The definition of T implies $x_0 \in T \Leftrightarrow x_0 \notin f(x_0) = T$. The contradiction shows that f is not surjective. However, for card $P(M) \leq$ card M there is by definition a bijective mapping from a subset of M to P(M), and hence a surjective mapping from M to P(M). This yields $\neg$(card $P(M) \leq$ card M).

(2) The mapping $g : M \rightarrow P(M)$ with $g(x) = \{x\}$ is injective. Thus there is a bijective mapping from M to a subset of P(M), and hence card $M \leq$ card P(M).

Cardinality of the set of real numbers : In Section 2.7 it is shown that there exists no bijective mapping from the set $\mathbb{R}$ of real numbers to the set $\mathbb{N}$ of natural numbers. Hence card $\mathbb{R}$ = card $\mathbb{N}$ does not hold. Since $\mathbb{N}$ is a subset of $\mathbb{R}$, by definition card $\mathbb{N} \leq$ card $\mathbb{R}$. Altogether, it follows that card $\mathbb{N}$ < card $\mathbb{R}$:

$$\neg(\text{card}\,\mathbb{R} = \text{card}\,\mathbb{N}) \ \wedge\ (\text{card}\,\mathbb{N} \leq \text{card}\,\mathbb{R}) \ \Rightarrow$$
$$\text{card}\,\mathbb{N} < \text{card}\,\mathbb{R}$$

Cardinality of cartesian products of the set of real numbers : First, it is proved that card $\mathbb{R}^2$ = card $\mathbb{R}$. For this purpose, a bijective mapping from the open unit interval $I =]0,1[$ to $\mathbb{R}$ is introduced :

$$f : I \rightarrow \mathbb{R} \qquad \text{with} \qquad f(x) = \tan \pi \,(x - \tfrac{1}{2})$$

The existence of the bijection f implies card $\mathbb{R}$ = card I and card $\mathbb{R}^2$ = card I^2, and hence

$$\text{card}\ \mathbb{R}^2 = \text{card}\ \mathbb{R} \quad \Leftrightarrow \quad \text{card}\ I^2 = \text{card}\ I$$

The following bijection $g : I^2 \rightarrow I$ is constructed using the decimal representation of the real numbers :

$$x := 0.a_1 a_2 a_3 \ldots \qquad\qquad g(x,y) := 0.a_1 b_1 a_2 b_2 a_3 b_3 \ldots$$
$$y := 0.b_1 b_2 b_3 \ldots$$

This yields card I^2 = card I, and hence card $\mathbb{R}^2$ = card $\mathbb{R}$. It follows by induction that card $\mathbb{R}^n$ = card $\mathbb{R}$.

5 TOPOLOGICAL STRUCTURES

5.1 INTRODUCTION

Topology : A set M may be structured by distinguishing certain subsets of M from the remaining subsets. The set of these distinguished subsets is called a topology on M. The domain (M ; T) is called a topological space.

The power set of M contains every subset that can be formed in M. A topology is a subset of the power set which has the property that all finite intersections and all unions of elements of the topology are also elements of the topology. The empty set $\emptyset$ and the underlying set M are elements of any topology. An element of the topology is called an open set of the topological space (M ; T).

Euclidean space : Concrete examples of topological spaces are furnished by euclidean spaces. The points of a euclidean space form the underlying set M of the space. The set of points whose distance from a given point is less than a real value $\varepsilon > 0$ is called an ε-ball. Every ε-ball is an open set of the euclidean space. The ε-balls form a basis for the topology of the euclidean space. Every finite intersection of ε-balls and every union of ε-balls is also an element of the topology.

An arbitrary subset of a euclidean space is called a shape. Lines, surfaces and volumes are examples of such shapes. The topological properties of a shape result from the properties of the ε-neighborhoods of its points. For example, the interior of the shape is the subset of those points of the shape for which there is an ε-neighborhood which contains only points of the shape. If all boundary points belong to the shape, the shape is said to be closed.

Topological structure : The study of the structure of topological spaces is based on mappings of these spaces. A mapping between topological spaces is said to be continuous if the preimage of every open set of the target is an open set. Bijective mappings which are continuous in both directions are said to be topological. If there is a topological mapping between two spaces, they are said to be homeomorphic : They are indistinguishable with respect to their topological structure. A property of a topological space which remains invariant under topological mappings is called a topological invariant. Homeomorphic spaces possess the same topological invariants.

Types of topologies : The topology of a space may possess special properties in addition to the general properties of topologies. Examples of such topologies are furnished by the natural topology of metric spaces and by the discrete topology, which is defined for any set. Topologies with the same properties are subsumed under a topology type.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*
© Springer-Verlag Berlin Heidelberg 2001

Topologies for new spaces may be generated using a mapping of known topological spaces. Mappings with special properties generate topologies with special properties, and hence topology types. Quotient topologies, sum topologies, relative topologies and product topologies are treated as examples of such topologies. They play an important role in algebraic topology.

Connectedness : The connectedness of a set is a topological invariant. The concept of disjoint sets is not sufficient for defining connectedness. The concept of separated sets is therefore introduced. Two sets are separated if none of the sets contains points of the closure of the other set. A set is connected if it is not the union of non-empty separated sets.

Separation : Few of the properties of a topology are determined by the general definition of topologies. Topologies are therefore often further characterized using separation axioms. A separation axiom defines a relationship between points, closed sets and open sets of a topological space. In particular, these axioms lead to Hausdorff spaces as well as regular and normal spaces.

Convergence : An iterative mathematical procedure is said to be convergent if it identifies a point of a topological space. The convergence of sequences is often studied in metric spaces. A sequence is a mapping from the natural numbers to the underlying set of the space. The question of the existence and uniqueness of the limits of sequences and subsequences arises. A general description of convergence is based on the definition of nets. A net is a mapping from a directed set to the underlying set of the space. Nets are used in particular to study the compactness of topological spaces. Limits of sums of the terms of a real sequence are treated using series of numbers. The convergence of filters is studied in general topological spaces.

Compactness : The number of elements of a finite set is a topological invariant. It is suitable for characterizing the delimitation of the space. However, if a space contains infinitely many points, the number of points is not suitable for characterizing the topological delimitation. The concept of compactness is therefore defined. A set is said to be compact if every covering of the set with open sets contains a finite subcovering. The compactness of a space plays an important role in the study of convergence.

Continuity : In metric spaces, the continuity of mappings may be defined using the properties of accumulation points. This definition is compatible with the definition using the properties of open sets, but it is more convenient for some applications. The properties of limits and discontinuities of real functions are studied using this concept.

5.2 TOPOLOGICAL SPACES

Introduction : The concept of a topological space arose in connection with the study of continuous surfaces in euclidean space. The essential properties of such spaces, however, are best elucidated by a definition of topological spaces independent of geometry. For this purpose, a topological space (M ; T) is defined in this section as a domain (M ; T) with special properties. In contrast to algebraic and ordinal structures, topological structures are not specified in the form of a relation (such as $+$ or $\leq$) in a set M, but rather in the form of a set T of subsets of M. In the following, the central concepts of topology, open and closed set, neighborhood and neighborhood system are defined for topological spaces.

Topology : A subset T of the power set P(M) of a set M is called a topology on the underlying set M if :

(T1) The topology T contains the empty set $\emptyset$ and the underlying set M.

$$\emptyset \in T \quad \wedge \quad M \in T$$

(T2) The intersection of any two elements A and B of T is an element of T.

$$A \in T \quad \wedge \quad B \in T \quad \Rightarrow \quad A \cap B \in T$$

(T3) The union of an arbitrary number of elements A,B,... of T is an element of T.

$$A, B, \ldots \in T \qquad \Rightarrow \quad A \cup B \cup \ldots \in T$$

The last part of the definition is not limited to the union of a countable number of sets. Two topologies S and T on an underlying set M are said to be equal if they contain the same subsets from P(M).

$$S = T \quad :\Leftrightarrow \quad (A \in S \Leftrightarrow A \in T)$$

Topological space : A domain (M ; T) is called a topological space if T is a topology on the underlying set M. Every element of the underlying set M is called a point of the topological space. Every element of the topology T is called an open set of the topological space and is a subset of M.

The sets M and T may be finite or infinite. If the set M is infinite, elements of T may also be infinite sets. The infinite sets may be countable or uncountable.

To simplify notation, open sets are indexed in the following, as in T_i or T_k. This is not meant to imply that the set of sets under consideration is countable. The properties of a topology T on the underlying set M are then defined as follows :

(T1a) $\emptyset \in T \quad \wedge \quad M \in T$

(T2a) $T_i \in T \quad \wedge \quad T_k \in T \qquad \Rightarrow \qquad T_i \cap T_k \in T$

(T3a) $T_1, T_2, \ldots \in T \qquad \Rightarrow \qquad T_1 \cup T_2 \cup \ldots \in T$

Open and closed sets : A subset A of points of a topological space (M ; T) is called an open set if it is an element of the topology of the space. The subset A is called a closed set if its complement $\overline{A} = M - A$ in M is an element of the topology of the space.

$$\text{A is open in } (M\,;T) \quad :\Leftrightarrow \quad A \subseteq M \;\wedge\; A \in T$$
$$\text{A is closed in } (M\,;T) \quad :\Leftrightarrow \quad \overline{A} \subseteq M \;\wedge\; \overline{A} \in T$$

Note : Instead of the expression "A is open in (M ; T)", expressions like "A is open in M" or "A is open" are often used to improve legibility. The topological space (M ; T) in which the set A is open must be identifiable from the context in which the expression is used. Similar simplified expressions are used for closed sets.

Properties of open and closed sets :

(M1) The complement of an open set is a closed set. The complement of a closed set is an open set.

(M2) The empty set $\emptyset$ and the underlying set M of a topological space (M ; T) are open and closed sets.

(M3) The union of a finite or infinite number of open sets is an open set.

(M4) The intersection of a finite number of open sets is an open set. The intersection of an infinite number of open sets is not necessarily an open set.

(M5) The union of a finite number of closed sets is a closed set. The union of an infinite number of closed sets is not necessarily a closed set.

(M6) The intersection of a finite or infinite number of closed sets is a closed set.

Proof : Properties of open and closed sets

(M1) The complement of an open set $A \in T$ is $\overline{A} = M - A$. The complement $\overline{A} \subseteq M$ is a closed set since $M - \overline{A} = A \in T$. The complement of a closed set B is $\overline{B} = M - B$. The complement $\overline{B}$ is an open set since $\overline{B} \in T$ by definition.

(M2) By property (T1), the sets $\emptyset$ and M are elements of any topology; hence they are open sets. Since their complements $\overline{M} = \emptyset$ and $\overline{\emptyset} = M$ are open sets, the sets $\emptyset$ and M are also closed sets.

(M3) This property follows directly from item (T3) in the definition of a topology.

(M4) Property (T2) in the definition of a topology directly implies that the intersection of a finite number of open sets is an open set. The following example shows that the intersection of an infinite number of open sets is not necessarily open. The intersection E of open intervals $E_n =]-\frac{1}{n}, 1+\frac{1}{n}[$ with $n \in \mathbb{N}$ on the $\mathbb{R}$-axis is the interval [0,1], since the limits 0 and 1 are contained in each E_n; this interval is not open.

(M5) The union of a finite number of closed sets A_i is given by $\cup A_i = \cup(M - \bar{A}_i)$ $= M - \cap \bar{A}_i$. By (M1), each of the complements $\bar{A}_i$ is an open set. Hence by (M4) the intersection $\cap \bar{A}_i$ is an open set. By (M1), the complement $M - \cap \bar{A}_i$ is a closed set. The union $\cup A_i$ is therefore a closed set.

The following example shows that the union of infinitely many closed sets is not necessarily closed. The union E of the closed intervals $E_n = [\frac{1}{n+1}, \frac{1}{n}]$ with $n \in \mathbb{N}$ on the $\mathbb{R}$-axis is not closed, since the limit 0 is not contained in E.

(M6) The intersection of an arbitrary number of closed sets A_i is given by $\cap A_i = \cap(M - \bar{A}_i) = M - \cup \bar{A}_i$. By (M1), each complement $\bar{A}_i$ is an open set. Hence by (M3) the set $\cup \bar{A}_i$ is an open set. By (M1), the complement $M - \cup \bar{A}_i$ is a closed set. The intersection $\cap A_i$ is therefore a closed set.

Comparison of topologies : Two topologies T_1 and T_2 on the same underlying set M are comparable if one of the topologies is a subset of the other. If $T_1 \subset T_2$, then T_1 is said to be coarser than T_2, and T_2 is said to be finer than T_1. The coarsest topology on M is $\{\emptyset, M\}$. The finest topology on M is the power set P(M).

Neighborhoods : A subset $A \subseteq M$ of the underlying set of a topological space $(M; T)$ is called a neighborhood of a point $x \in M$ if there is a subset T_i of A which is open and contains x.

$$\text{A is a neighborhood of x} \quad :\Leftrightarrow \quad \bigvee_{T_i \in T} (x \in T_i \ \wedge \ T_i \subseteq A)$$

To simplify notation, different neighborhoods of a point x are indexed in the following, as in U_i and U_k. This does not imply that the set of neighborhoods of a point is countable. The set of neighborhoods of a point x is called the neighborhood system of x and is designated by U(x).

$$U(x) \ := \ \{U_i \subseteq M \mid \bigvee_{T_k \in T} (x \in T_k \ \wedge \ T_k \subseteq U_i)\}$$

A neighborhood A of a point $x \in M$ is said to be open if A is an open set. The neighborhood A is said to be closed if A is a closed set. In general, a neighborhood is neither open nor closed. The concept of an open neighborhood differs from the concept of an open set in that the open neighborhood is related to a point $x \in M$ and contains this point.

Properties of a neighborhood system : The definition of the neighborhood of a point leads to the following properties of its neighborhood system :

(1) The neighborhood system U(x) is not empty.

(2) The neighborhood system U(x) does not contain the empty set $\emptyset$.

(3) If the neighborhood system U(x) contains the neighborhoods U_i and U_k, then it also contains their intersection $U_m = U_i \cap U_k$.

(4) If the neighborhood system $U(x)$ contains the neighborhoods U_i and U_k, then it also contains their union $U_m = U_i \cup U_k$.

(5) Let U_m be a neighborhood of the point x. Then there is an open subset T_i of U_m such that U_m is also a neighborhood for every point y of the subset T_i.

(6) A set is a neighborhood for each of its points if and only if it is open.

Proof : Properties of a neighborhood system

(1) The neighborhood system $U(x)$ contains at least the underlying set M, since M is an element of the topology and x is an element of M.

(2) The neighborhood system $U(x)$ does not contain the empty set $\emptyset$ since every neighborhood U_m of x contains x as an element.

(3) If U_i and U_k are neighborhoods of x, then there are open sets T_i and T_k such that $x \in T_i \subseteq U_i$ and $x \in T_k \subseteq U_k$. The intersection $T_i \cap T_k$ is an element T_m of the topology. The intersection $U_i \cap U_k$ is an element U_m of the power set $P(M)$. From $x \in T_i \cap T_k$ and $T_i \cap T_k \subseteq U_i \cap U_k$ it follows that $x \in T_m \subseteq U_m$, so that U_m is a neighborhood of x.

(4) If U_i and U_k are neighborhoods of x, then there are open sets T_i and T_k such that $x \in T_i \subseteq U_i$ and $x \in T_k \subseteq U_k$. The union $T_i \cup T_k$ is an element T_m of the topology. The union $U_i \cup U_k$ is an element U_m of the power set $P(M)$. From $x \in T_i \cup T_k$ and $T_i \cup T_k \subseteq U_i \cup U_k$ it follows that $x \in T_m \subseteq U_m$, so that U_m is a neighborhood of x.

(5) For every neighborhood U_m of the point x there is an open set T_m such that $x \in T_m \subseteq U_m$. For every point y of T_m this implies $y \in T_m \subseteq U_m$, so that U_m is a neighborhood for every point y of T_m.

(6) If the neighborhood U_m is not one of the open sets of the topology T and T_i is the union of all open sets contained in U_m, then U_m is not a neighborhood for the points in the difference $U_m - T_i$. If, however, U_m is an open set, then the set U_m is a neighborhood for each of its points.

Neighborhood axioms : In the preceding section, the properties of a neighborhood system are derived from the definitions of topologies and neighborhoods. It is also possible to define the properties of a neighborhood system by the following neighborhood axioms and to derive the properties of topologies and neighborhoods from these axioms. These two definitions of a topological space are equivalent.

(U1) Every point x belongs to each of its neighborhoods.

(U2) The union of an arbitrary number of neighborhoods of a point x is a neighborhood of x. The underlying set M is a neighborhood of x.

(U3) The intersection of two neighborhoods of x is a neighborhood of x.

(U4) Every neighborhood U_m of a point x contains a neighborhood $U_i \subseteq U_m$ of x such that U_m is a neighborhood of every point of U_i.

5.3 BASES AND GENERATING SETS

Introduction : For a topological space $(M;T)$, the question arises whether the topology T may be constructed from a basis $B \subseteq T$. In the following, it is shown that while the topology T may be constructed by forming unions of sets of a basis, the basis B is generally not unique. The question also arises whether a basis may be constructed from a subset S of the power set $P(M)$. This is possible, since every subset of $P(M)$ may be used to construct a generating set.

Basis of a topology : Let $(M;T)$ be a topological space. A subset B of the topology T is called a basis of the topology if T contains exactly those sets which result from arbitrary unions of elements of B. To simplify notation, the sets of a basis B are indexed, as in B_i and B_k. This does not imply that every basis is countable.

Second countable topological space : A topological space $(M;T)$ is said to be second countable if there is a countable basis B of its topology T. For a countable basis $B = \{B_1, B_2, B_3, ...\}$ there is, for every open set $T_i \in T$, a countable index set N_i such that :

$$T_i \in T \quad :\Leftrightarrow \quad T_i = \bigcup_{n \in N_i} B_n$$

A second countable space is said to satisfy the second axiom of countability.

Generating set : A subset E of the power set $P(M)$ is called a generating set on the underlying set M if :

(E1) The underlying set M is a union of elements of E.

(E2) For every point x of the intersection A of two elements of E there is an element of E which contains x and is a subset of A.

To simplify notation, the sets of a generating set E are indexed, as in E_i and E_k. This does not imply that every generating set is countable. The properties of generating sets are then represented as follows :

(E1) $M = \bigcup E_i$

(E2) $\bigwedge_{E_i} \bigwedge_{E_k} ((x \in E_i \cap E_k) \Rightarrow \bigvee_{E_m} (x \in E_m \subseteq (E_i \cap E_k)))$

Construction of a topology : A generating set E is a basis for a topological space with the underlying set M. The set T which contains every union of elements of E is therefore a topology on M.

$$T = \{T_i \mid \bigvee_{E' \subseteq E} (T_i = \bigcup_{E_k \in E'} E_k)\}$$

Proof : A generating set is a basis of a topology.

(1) It follows from (E1) that T contains the underlying set M as an element. Since T contains the union of an empty set of elements of E, it follows from the definition of the generalized union that T contains the empty set $\emptyset$ as an element. Hence the set T possesses property (T1) of a topology.

(2) It follows from (E2) that every point of the intersection $E_i \cap E_k$ of two elements of E is contained in an element E_m of E which is a subset of $E_i \cap E_k$. Different points in $E_i \cap E_k$ are generally contained in different elements $E_m, E_n, ...$ The union of these elements is by definition an element of T. Hence the intersection $E_i \cap E_k$ is an element of T. For open sets T_r and T_s there are subsets E' and E'' of E such that T_r is the union of the sets in E' and T_s is the union of the sets in E''. Since all intersections $E_i \cap E_k$ are elements of T, $T_r \cap T_s$ is also an element of T. Hence the set T possesses property (T2) of a topology.

(3) Every element of T is by construction a union of sets in E. Every union of elements of T is therefore a union of elements of E. But every union of elements of E is by hypothesis an element of T. Hence the set T possesses property (T3) of a topology.

Since the set T possesses properties (T1) to (T3) of a topology, T is a topology on the underlying set M.

Subbasis of a topology : A subset S of the power set P(M) is called a subbasis on the underlying set M. To simplify notation, the sets of a subbasis S are indexed, as in S_i and S_k. This does not imply that every subbasis is countable.

Construction of a generating set : The set E of all intersections of a finite number of elements of a subbasis S is a generating set for a topology on the underlying set M of the subbasis. To prove this, conditions (E1) and (E2) are shown to be satisfied.

(1) Since E contains the intersection of an empty set of elements of E, it follows from the definition of the generalized intersection that E contains the underlying set M. Property (E1) is therefore satisfied.

(2) By construction, for two arbitrary sets E_i and E_k the set E contains their intersection $E_m = E_i \cap E_k$. Property (E2) is therefore satisfied.

Discrete topology : The power set P(M) is called the discrete topology on the underlying set M. A discrete topological space (M ; P(M)) has special properties:

(1) Every subset of M is open, since it is contained in P(M).

(2) Every subset of M is closed, since its complement is contained in P(M) and is therefore an open set.

(3) Mappings between discrete topological spaces are continuous, since the preimages of open sets are open sets. The mappings are, however, generally not bijective and hence not homeomorphic.

Subsets of a discrete topological space are thus both open and closed. Bijective mappings between discrete topological spaces are homeomorphic.

Equivalent bases : Let A and B be different generating sets on the underlying set M. Let S be the topology constructed by forming unions of sets in A. Let T be the topology constructed by forming unions of sets in B. Then the generating sets A and B are called equivalent bases if the topologies S and T are equal. Equivalent bases are generally not identical.

Establishing the equivalence of bases : Two generating sets A and B on an underlying set M are equivalent bases if :

(1) For every open set A_i of A and for every point x in A_i there is an open set B_k in B such that $x \in B_k \subseteq A_i$.

(2) For every open set B_m in B and for every point y in B_m there is an open set A_s in A such that $y \in A_s \subseteq B_m$.

Proof : Equivalence of bases

Let two generating sets A and B with properties (1) and (2) be given. Let the topology constructed from A be S with the open sets S_i. Then S_i is by definition the union of the sets of a subset A' of A :

$$S_i = \bigcup_{A_k \in A'} A_k$$

By (1), for every point $x \in S_i$ there are open sets $A_k \in A'$ and $B_m \in B$ such that $x \in B_m \subseteq A_k$. For every point $x \in S_i$, a set B_m is determined in this manner; these sets are collected in a subset B' of B. The union of the sets in B' is designated by T_n ; it is by definition an element of the topology T constructed from B :

$$B' := \{B_m \in B \mid \bigvee_{x \in S_i} (x \in B_m \subseteq A_k \subseteq S_i)\}$$

$$T_n := \bigcup_{B_m \in B'} B_m$$

Since every point $x \in S_i$ is contained in one of the open sets $B_m \in B'$, it follows that $S_i \subseteq T_n$. Since every set $B_m \in B'$ is a subset of a set $A_k \in A'$, it follows that $T_n \subseteq S_i$. From $S_i \subseteq T_n$ and $T_n \subseteq S_i$, it follows that $T_n = S_i$.

Analogously, (2) is used to show that for every open set $T_r \in T$ there is an identical open set $S_t \in S$. Hence the generating sets A and B lead to the same topologies S and T if conditions (1) and (2) are satisfied. Every generating set is a basis. Hence A and B are equivalent bases.

Neighborhood basis : Let (M ; T) be a topological space. A subset B(x) of the neighborhood system U(x) of a point $x \in M$ is called a neighborhood basis at x if every element of U(x) contains an element of B(x) as a subset. The definition of a neighborhood system guarantees that every element of the neighborhood basis B(x) contains at least one open set of the topology T as a subset.

$$\bigwedge_{U_i \in U} \bigvee_{B_k \in B} \bigvee_{T_m \in T} (T_m \subseteq B_k \subseteq U_i)$$

First countable topological space : A topological space (M ; T) is said to be first countable if every point $x \in M$ has a countable neighborhood basis. Not every first countable space is second countable. For example, it is shown in Section 5.4 that while every metric space is first countable, not every metric space is second countable. A first countable space is said to satisfy the first axiom of countability.

Example 1 : Construction of a topology

A topology T is constructed on the underlying set M = {a, b, c}. The subbasis S is chosen arbitrarily. The resulting neighborhood systems of the points of the topological space (M ; T) do not contain the subsets {a}, {c} and {a, c} of the power set P(M).

subbasis	:	S = {{a, b}, {b, c}}
generating set	:	E = {{a, b}, {b, c}, {b}}
topology	:	T = {∅ , {b}, {a, b}, {b, c}, {a, b, c}}
neighborhood systems	:	U(a) = {{a, b}, {a, b, c}}
		U(b) = {{b}, {a, b}, {b, c}, {a, b, c}}
		U(c) = {{b, c}, {a, b, c}}

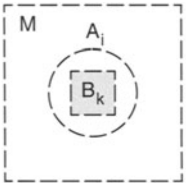
Example 2 : Comparison of topologies and neighborhoods

The set T $= \{\emptyset, \{b\}, \{b, c\}, \{a, b, c\}\}$ is a topology on the set M $= \{a,b,c\}$. The topology T is coarser than the one in Example 1, since it does not contain the element $\{a, b\}$. Hence $\{a, b\}$ is not an open set in Example 2. Nonetheless, $\{a, b\}$ is a neighborhood of the point b, since the condition $b \in \{b\} \subseteq \{a,b\}$ is satisfied : The point b lies in the open set $\{b\}$ which is contained in $\{a,b\}$. The neighborhood system of point a differs from U(a) in Example 1. The other neighborhood systems are the same.

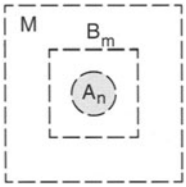
neighborhood systems : $U(a) = \{\{a, b, c\}\}$
$U(b) = \{\{b\}, \{a, b\}, \{b, c\}, \{a, b, c\}\}$
$U(c) = \{\{b, c\}, \{a, b, c\}\}$

Example 3 : Equivalence of bases

Let M be an open set in the euclidean plane $\mathbb{R}^2$. The set of all open disks around all points in M is a topological basis A on M. The set of all open squares around all points in M is a second topological basis on M. The bases A and B are equivalent. For each point in an open disk A_i there is an open square $B_k \subseteq A_i$ which contains this point. Likewise, for every point in an open square B_m there is an open disk $A_n \subseteq B_m$ which contains this point.



basis A : $x \in B_k \subseteq A_i$



basis B : $y \in A_n \subseteq B_m$

5.4 METRIC SPACES

Introduction : A structure is defined on a set by assigning a distance to every pair of points of the set. This mapping is called a metric. The metric structure of a space may be used to derive a topological structure, but the converse is not necessarily true. A metric topology is constructed by first defining ε-balls. These balls possess the properties of a generating set and therefore form a suitable basis for a topology. Different definitions of the metric generally lead to different topologies. Euclidean and discrete spaces are treated as examples of metric spaces.

Metric : A mapping $d : M \times M \to \mathbb{R}$ is called a metric on the set M if for all elements x, y, z of M :

(M1) $d(x, x) = 0$

(M2) $d(x, y) > 0 \quad \text{for} \quad x \neq y$

(M3) $d(x, y) = d(y, x)$

(M4) $d(x, z) \leq d(x, y) + d(y, z)$

The image $d(x, y)$ is called the distance of the points x and y. A mapping with the property $d(x, y) \geq 0$ instead of property (M2) is called a pseudometric.

Euclidean metric : The euclidean metric is defined by analogy with geometric distance. The underlying set of n-dimensional euclidean space contains the vectors $\mathbf{x} = (x_1, x_2, ..., x_n)$ of the real vector space $\mathbb{R}^n$. The real numbers x_i are called the coordinates of the vector **x**.

$$\mathbb{R}^n := \{\mathbf{x} = (x_1, x_2, ..., x_n) \mid x_i \in \mathbb{R}\}$$

The mapping $d : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is called the euclidean metric on $\mathbb{R}^n$ if the distance $d(\mathbf{x}, \mathbf{y})$ of the points $\mathbf{x} = (x_1, ..., x_n)$ and $\mathbf{y} = (y_1, ..., y_n)$ is determined by analogy with the distance in the space $\mathbb{R}^3$:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + ... + (x_n - y_n)^2}$$

Discrete metric : A mapping $d : M \times M \to \mathbb{R}$ for an arbitrary underlying set M is called a discrete metric if it yields the values $d(x, x) = 0$ and $d(x, y) = 1$ for arbitrary different points $x, y \in M$. Thus the discrete metric takes only two different values, which are chosen to be 0 and 1.

$$d(x, x) := 0$$

$$d(x, y) := 1 \quad \text{for} \quad x \neq y$$

ε-ball : Let a metric d be defined on a set M. The set of points of M whose distance from a fixed point $x \in M$ is less than a positive real number ε is called an ε-ball in M and is designated by D(x, ε).

$$D(x, \varepsilon) \ := \ \{y \in M \mid d(x, y) \ < \ \varepsilon\}$$

If the metric is euclidean, then different values of ε lead to different point sets. If the metric is discrete, then every ε-ball with $\varepsilon \le 1$ is the set {x}, and every ε-ball with $\varepsilon > 1$ is the entire set M.

Metric basis : For every point x of a set M on which a metric d is defined, there is at least one ε-ball $B_\varepsilon := D(x, \varepsilon)$. The set B of the ε-balls in M is a generating set, and hence a basis for a metric topology. The ε-balls are open sets of the metric topology.

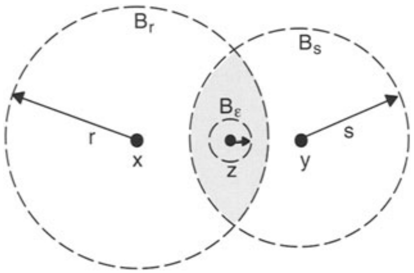
$$B \ := \ \{B_\varepsilon \ = \ D(x, \varepsilon) \mid \ x \in M\}$$

Proof : Construction of a metric basis

(1) By the definition of the set B, every point $x \in M$ is contained in at least one ε-ball. The union of the ε-balls therefore contains every point in M. Hence the set B possesses property (E1) of a generating set.

(2) Let $B_r = D(x, r)$ and $B_s = D(y, s)$ be different elements of the set B. If their intersection $B_r \cap B_s$ is empty (for instance for a discrete metric), then condition (E2) for generating sets is satisfied, since there is no point $x \in B_r \cap B_s$.

(3) If the intersection $B_r \cap B_s$ is not empty, then there is a point z in $B_r \cap B_s$ whose distance from x is less than r and whose distance from y is less than s :



$$\begin{aligned} z \in B_r \cap B_s \quad &\Rightarrow \quad z \in B_r \quad \wedge \quad z \in B_s \\ &\Rightarrow \quad d(x,z) < r \quad \wedge \quad d(y,z) < s \end{aligned}$$

There is an ε-ball $B_\varepsilon := D(x, \varepsilon)$ with $\varepsilon = \min\{r - d(x, z), s - d(y, z)\} / 2$ which is contained in $B_r \cap B_s$. Hence the set B also possesses property (E2) of a generating set.

$$\varepsilon \ = \ \min\{r - d(x, z), s - d(y, z)\} / 2$$

$$z \ \in \ B_\varepsilon \subseteq (B_r \cap B_s)$$

(4) Since the set B of ε-balls has the properties (E1) and (E2) of a generating set, B is a basis for a metric topology. The sets contained in a basis are by definition open sets in the topology constructed from this basis. Thus in a metric topology every ε-ball is an open set.

Metric topology : Let a set M and a metric $d: M \times M \to \mathbb{R}$ be given. Then the metric basis B associated with d generates a metric (natural) topology T on M. Every union of ε-balls B_ε of the metric basis B is an open set T_i of the natural topology T.

$$T_i \in T \quad :\Leftrightarrow \quad \bigvee_{B' \subseteq B} \left(T_i = \bigcup_{B_\varepsilon \in B'} B_\varepsilon\right)$$

Metric space : The domain (M ; d) with the underlying set M and the metric d is called a metric space. A metric space possesses the natural topology induced by its metric. In the following, a metric space is assumed to be equipped with this natural topology.

Euclidean space : The domain (M ; d) is called an n-dimensional euclidean space if the underlying set M is the real space $\mathbb{R}^n$ and d is the euclidean metric.

On the real line $\mathbb{R}^1$, the elements of the natural basis are open intervals. On the euclidean plane $\mathbb{R}^2$, the elements of the natural basis are open disks. In the euclidean space $\mathbb{R}^3$, the elements of the natural basis are open spheres.

Discrete metric space : Let a set M and a metric $d: M \times M \to \mathbb{R}$ be given. The metric basis B induced by d is said to be discrete if for every point $x \in M$ it contains an ε-ball which contains only the point x.

$$B := \{\{x\} \mid x \in M\}$$

A discrete metric basis generates a discrete topology.

(1) For different points $x, y \in M$, there are ε-balls whose intersection is empty.

(2) Every subset of M is an open set, since it is the union of the open sets of its points.

(3) Every subset of M is a closed set, since its complement in M is an open set.

(4) The underlying set M is an element of the topology T. Its intersection with the ε-balls D(x, 1) is not empty.

The discrete metric topology of a set M is therefore the power set P(M). The domain (M ; d) is called a discrete metric space if the metric d induces the discrete topology.

Neighborhood basis : The ε-balls $B_\varepsilon = d(x, \varepsilon)$ form a neighborhood basis of the point x. For every element B_ε there is an open set T_i (namely $T_i = B_\varepsilon$) such that the condition $T_i \subseteq B_\varepsilon$ is satisfied. For every neighborhood U_i of x there is an element $B_\varepsilon \subseteq U_i$ which contains at least the point x, and hence :

$$\bigwedge_{U_i} \bigvee_{B_\varepsilon} \bigvee_{T_m} (T_m = B_\varepsilon \subseteq U_i)$$

In the following study of topological properties, it suffices to consider the neighborhood basis B(x) instead of the complete neighborhood system U(x) of a point $x \in M$.

$$B(x) = \{B_\varepsilon \mid B_\varepsilon = D(x, \varepsilon)\}$$

Open initial segment : The set $\mathbb{Q}$ of rational numbers is totally ordered. For every number $q \in \mathbb{Q}$, there is therefore a unique subset S_q of $\mathbb{Q}$ which contains all elements $x \in \mathbb{Q}$ which are less than q. Such a subset is called an (open) initial segment of $\mathbb{Q}$.

$$S_q = \{x \in \mathbb{Q} \mid x < q\}$$

Open initial : The subset of $\mathbb{Q}$ for which $x < 0$ or $x^2 < 2$ holds is not an open initial segment in $\mathbb{Q}$. Since there is no rational solution of the equation $q^2 = 2$ (see Section 6.5), there is no number $q \in \mathbb{Q}$ which could be used to define an open initial segment S_q whose elements satisfy $x < 0$ or $x^2 < 2$. To characterize the subset of $\mathbb{Q}$ for which $x < 0$ or $x^2 < 2$ holds, the concept of an open initial is defined.

A subset A of a totally ordered set $(M ; \leq)$ is called an initial in M if for every $x \in M$ contained in A every $y \in M$ with $y \leq x$ is also contained in A. An initial without a greatest element is called an open initial.

$$\text{A is an open initial} \quad :\Leftrightarrow \quad \bigwedge_{x \in A} \bigwedge_{y \in M} (y \leq x \quad \Rightarrow \quad y \in A)$$

Example 1 : The real number $\sqrt{2}$

The real number $\sqrt{2}$ is an open initial B in the set $\mathbb{Q}$ of rational numbers :

$$B := \mathbb{Q}^- \cup \{x \in \mathbb{Q}_0^+ \mid x^2 < 2\}$$

$\mathbb{Q}^-$ negative rational numbers

$\mathbb{Q}_0^+$ positive rational numbers and zero

(1) From $x \in \mathbb{Q}^-$ and $y \leq x$ it follows that $y \in \mathbb{Q}^-$. Hence x satisfies the condition for an element of an initial.

(2) For $x \in \mathbb{Q}_0^+$ and $y \leq x$, either $y \in \mathbb{Q}^-$ or $y \in \mathbb{Q}_0^+$. If $y \in \mathbb{Q}^-$, then $y \in B$. If $y \in \mathbb{Q}_0^+$, then $y \leq x$ implies $(x + y)(x - y) \geq 0$, and thus $y^2 \leq x^2$. Then $x^2 < 2$ implies $y^2 < 2$, and hence $y \in B$. Altogether, $x \in \mathbb{Q}_0^+$ and $y \leq x$ with $x^2 < 2$ implies $y \in B$. Hence x satisfies the condition for an element of an initial.

(3) The set B contains no greatest element. For every rational number $x \in \mathbb{Q}_0^+$ with $x^2 < 2$ there is a number $y \in \mathbb{Q}_0^+$ with $y > x$ and $y^2 < 2$.

- Choose $y = \dfrac{4}{x + \frac{2}{x}}$
- The condition $y > x$ is satisfied if $4 > x^2 + 2$, that is $x^2 < 2$. By hypothesis $x^2 < 2$, and hence $y > x$.
- To prove $y^2 < 2$, the expression for y^2 is transformed. The inequality $\frac{1}{z} > 2 - z$ holds for every number $z \in \mathbb{Q}_0^+$ with $z < 1$. This property is used for $\frac{x^2}{2} < 1$:

$$y^2 \quad = \quad \frac{16}{x^2 + 4 + \frac{4}{x^2}} \quad < \quad \frac{16}{x^2 + 4 + 2(2 - \frac{x^2}{2})} \quad = \quad 2$$

From $x^2 < 2, y > x$ and $y^2 < 2$, it follows that B has no greatest element. Thus B is an open initial. The real number $B = \sqrt{2}$ is irrational, since there is no open initial segment for $\sqrt{2}$ in $\mathbb{Q}$.

First countability of metric spaces : A topological space (M ; T) is first countable if every point $x \in M$ has a countable neighborhood basis B(x). Every metric space (M ; d) is first countable. To prove this, consider the set of all ε-balls with center x and radius $q \in \mathbb{Q}^+$.

$$B(x) \ := \ \{ D(x, q) \mid q \in \mathbb{Q}^+ \}$$

It is required that every neighborhood of x contains an element of B(x). Since the neighborhood system of x contains ε-balls D(x, r) with an irrational radius r, it is to be proved that D(x, r) contains an element of B(x). The irrational number r is an open initial. The open initial contains a $q_0 \in \mathbb{Q}^+$ such that $D(x, q_0)$ is contained in D(x, r). Hence B(x) is a neighborhood basis of x. Since $\mathbb{Q}$ is countable, B(x) is also countable. Hence the metric space (M ; d) is first countable.

Second countability of metric spaces : A topological space (M ; T) is second countable if its topology T possesses a countable basis. Not every metric space is second countable. However, every euclidean space ($\mathbb{R}^n$; d) is second countable. A discrete metric space is second countable if its underlying set M is countable.

Proof : Second countability of euclidean spaces

As a basis B of the topology of the euclidean space ($\mathbb{R}^n$; d), choose the set of all ε-balls D(x, q) with rational radius $q \in \mathbb{Q}^+$ whose centers have rational coordinates $x_i \in \mathbb{Q}^+$. The set B is shown to have the properties of a generating set and hence of a basis of the topology. Since $\mathbb{Q} \times \mathbb{Q}^n$ is countable, B is also countable.

(1) Every rational point $x \in \mathbb{R}^n$ is contained in an ε-ball $D(x, q) \in B$. Every point $y \in \mathbb{R}^n$ some or all of whose coordinates $(y_1, ..., y_n)$ are irrational is contained in an ε-ball $D(0, q) \in B$ with $q > d(D, y)$. The union of the elements of B therefore contains every point in $\mathbb{R}^n$. Hence B possesses property (E1) of a generating set.

(2) In the construction of metric bases, it was shown that every non-empty intersection $B_i \cap B_k$ of two elements of B contains an ε-ball (w, r). If r is not rational, then the open initial for r contains a rational number q_0 such that the ε-ball$(w, q_0) \in B$ is contained in $B_i \cap B_k$. Hence B possesses property (E2) of a generating set.

Proof : Second countability of discrete metric spaces

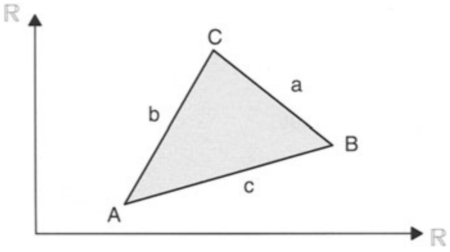
For every point $x \in M$, the topology of a discrete metric space (M ; T) contains the one-element set $D(x, 1) = \{x\}$. Therefore every basis of T must also contain every ε-ball $D(x, 1)$. Hence the cardinality of the basis is not less than the cardinality of M. The basis B which contains the ε-ball $D(x, 1)$ for every point $x \in M$ is chosen.

(1) The union of the elements of B contains every point in M. Hence B possesses property (E1) of a generating set.

(2) The intersection of two elements of B is empty. Hence B possesses property (E2) of a generating set.

B therefore has the properties of a generating set, and hence of a basis of the topology T. If the underlying set M is countable, then the basis B is also countable. If M is uncountable, then B is also uncountable.

Example 2 : Metric of the euclidean plane



$$a \le b + c$$

$$b \le c + a$$

$$c \le a + b$$

In the euclidean plane $\mathbb{R}^2$, the distance between points has the properties (M1) to (M4) of a metric :

(1) The distance from a point to itself is 0.

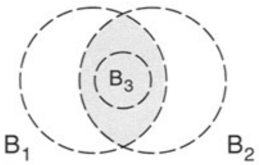
(2) The distance between different points is positive.

(3) The distance from A to B is equal to the distance from B to A.

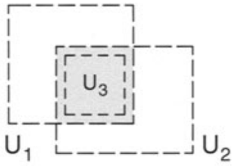
(4) The length of any side of a triangle is less than or equal to the sum of the lengths of the other two sides. If the corners are colinear (area 0), then the length of one side is equal to the sum of the lengths of the other two sides.

Example 3 : Bases in the euclidean plane

In the definition of the topology of a euclidean space, open disks were chosen as the basis elements of the metric topology. This choice is not unique. For example, open disks or open rectangles may be chosen as basis elements of the topology in the euclidean plane. However, the disks are usually preferred as a basis of the euclidean spaces, since they are readily represented using the euclidean metric.



disk basis
$B_3 \subseteq B_1 \cap B_2$



rectangle basis
$U_3 \subseteq U_1 \cap U_2$

Example 4 : Discrete metric topology

Let a metric space $(M\,;d)$ with the underlying set $M := \{a,b\}$ and the discrete metric d be given. The discrete basis of this space contains the ε-balls $D(a,1) = \{a\}$ and $D(b,1) = \{b\}$.

The set of all unions of elements of the basis $B = \{\{a\},\{b\}\}$ is the topology T of the discrete metric space. The topology is the power set P(M) of the underlying set M of the space.

$$T = \{\emptyset, \{a\}, \{b\}, \{a,b\}\} = P(M)$$

5.5 POINT SETS IN TOPOLOGICAL SPACES

Introduction : A point x of the underlying set M of a topological space (M;T) possesses properties with respect to a subset A of M which are determined by the relationships between the set A and the neighborhood system U(x) of the point x in the space (M ; T). These properties lead to the following definitions of point types and set types.

Point types : Points x of the underlying set M which possess special properties with respect to a subset A in the space (M ; T) form a class called a point type. Some point types are defined in the following.

Interior point : A point x is called an interior point of A if at least one neighborhood U of x is a subset of A.

$$\bigvee_{U \in U(x)} (x \in U \quad \wedge \quad U \subseteq A) \qquad \text{(P1)}$$

Exterior point : A point x is called an exterior point of A if at least one neighborhood U of x has no points in common with A.

$$\bigvee_{U \in U(x)} (x \in U \quad \wedge \quad U \cap A = \emptyset) \qquad \text{(P2)}$$

Contact point : A point x is called a contact point of A if every neighborhood U_i of x contains at least one point of A.

$$\bigwedge_{U_i \in U(x)} (U_i \cap A \neq \emptyset) \qquad \text{(P3)}$$

Boundary point : A contact point x is called a boundary point (frontier point) of A if every neighborhood U_i of x contains at least one point y not contained in A.

$$\bigwedge_{U_i \in U(x)} \; \bigvee_{y \in U_i} (U_i \cap A \neq \emptyset \quad \wedge \quad y \notin A) \qquad \text{(P4)}$$

Isolated point : A point x is called an isolated point of A if there is a neighborhood U of x whose intersection with A contains only x.

$$\bigvee_{U \in U(x)} (U \cap A = \{x\}) \qquad \text{(P5)}$$

Accumulation point : A contact point x of A which is not an isolated point is called an accumulation point (limit point) of A.

$$\bigwedge_{U_i \in U(x)} (U_i \cap A \neq \emptyset \quad \wedge \quad U_i \cap A \neq \{x\}) \qquad \text{(P6)}$$

Classification of points : The points of the set M of a topological space (M ; T) are assigned to the point types with respect to a set $A \subseteq M$ as follows :

(1) Every point in M is either an interior point or an exterior point or a boundary point. The point cannot be contained in more than one of these point types. For example, if x is an interior point, then x is neither an exterior point nor a boundary point.

(2) Every contact point of A is either an interior point or a boundary point.

(3) Interior points belong to A, and exterior points belong to the complement $\overline{A} = M - A$. Boundary points may belong to either A or $\overline{A}$.

(4) A point is a boundary point if and only if each of its neighborhoods contains both points in A and points in $\overline{A}$.

Example 1 : Point types of a finite set

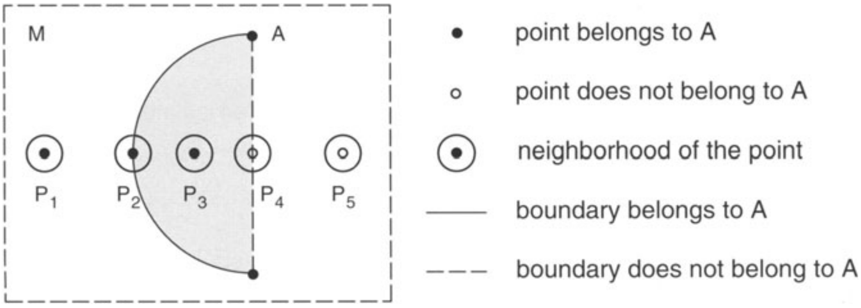
Let the underlying set of the topological space (M ; T) be $M = \{w, x, y, z\}$, and let the topology be $T = \{\emptyset, \{w\}, \{w,x\}, \{w,x,y\}, \{w,x,y,z\}\}$. Then the points of the space (M ; T) have the following neighborhood systems :

$$\begin{aligned} U(w) &= \{\{w\}, \{w,x\}, \{w,y\}, \{w,z\}, \{w,x,y\}, \{w,y,z\}, \{w,z,x\}, \{w,x,y,z\}\} \\ U(x) &= \{\{w,x\}, \{w,x,y\}, \{w,x,z\}, \{w,x,y,z\}\} \\ U(y) &= \{\{w,x,y\}, \{w,x,y,z\}\} \\ U(z) &= \{\{w,x,y,z\}\} \end{aligned}$$

With respect to the set $A = \{w,x\}$, the points have the following properties :

point w :	interior point :	$\{w\} \subset A$
	isolated point :	$w \in \{w\} \subset A$
point x :	interior point :	$\{w, x\} \subseteq A$
	contact point :	$\{w, x\} \cap A = \ldots = \{w,x,y,z\} \cap A = \{w, x\}$
	accumulation point :	$\{w, x\} \neq \{x\}$
point y :	contact point :	$\{w,x,y\} \cap A = \{w,x,y,z\} \cap A = \{w, x\}$
	boundary point :	$y \notin A$ and $y \in \{w,x,y\}, \{w,x,y,z\}$
	accumulation point :	$\{w,x\} \neq \{y\}$

Example 2 : Point types in the real euclidean plane


M
A
P_1
P_2
P_3
P_4
P_5
● point belongs to A
○ point does not belong to A
neighborhood of the point
—— boundary belongs to A
– – – boundary does not belong to A

interior point	: P_3	isolated point	: P_1
exterior point	: P_5	accumulation points	: P_2, P_3, P_4
boundary points	: P_1, P_2, P_4	contact points	: P_1, P_2, P_3, P_4

The underlying set M of the euclidean plane is $\mathbb{R}^2$. The subset A consists of a semidisk and the point P_1. The arc and its endpoints belong to A, the diameter does not. The properties of the points P_1 to P_5 of the underlying set M with respect to the subset A are specified.

Set types : Different subsets of the underlying sets of topological spaces may have identical properties. For example, there are subsets which consist of points of the same type. There are also subsets with the special property that they are formed from other subsets through the same operation. Subsets with identical special properties form a class, called a set type. In the following compilation of set types, A is a subset of the underlying set M of a topological space (M ; T).

Open set : A set A is open if every point x of A is an interior point.

$$\bigwedge_{x} (x \in A \quad \Rightarrow \quad x \text{ is an interior point}) \qquad \text{(A1)}$$

Closed set : A set A is closed if every point of the complement $\bar{A} = M - A$ is an exterior point.

$$\bigwedge_{x} (x \in \bar{A} \quad \Rightarrow \quad x \text{ is an exterior point}) \qquad \text{(A2)}$$

Bounded set : A subset A in a metric space is said to be bounded if every point x of A lies in a neighborhood D(y, r).

$$\bigvee_{y \in A} \; \bigvee_{r \in \mathbb{R}} (A \subseteq D(y,r)) \qquad \text{(A3)}$$

Interior of a set : The set of interior points of a set A is called the interior of A and is designated by I(A).

$$\mathrm{I}(A) \; := \; \{x \in A \mid x \text{ is an interior point}\} \qquad \text{(A4)}$$

Boundary of a set : The set of boundary points of a set A is called the boundary (frontier) of A and is designated by R(A).

$$R(A) := \; \{x \in A \mid x \text{ is a boundary point}\} \qquad \text{(A5)}$$

Closure of a set : The union of a set A with its boundary R(A) is called the closure of A and is designated by H(A).

$$H(A) := \; A \cup R(A) \qquad \text{(A6)}$$

Remarks about the set types :

(1) Every interior point x of A has a neighborhood which is a subset of A. By definition, this neighborhood contains an open set which contains x and is a subset of A. The interior points of A are contained in the union of these open sets. This union is by definition an element of the topology, and therefore an open set. Hence the interior of A is an open set.

(2) If some of the boundary points of a set A belong to the set A and some of the boundary points belong to the complement $\bar{A}$, then the set A is neither open nor closed.

(3) In the real euclidean space $\mathbb{R}^n$, a bounded set may be infinite.

(4) The closure of a set consists of the contact points of the set.

Relationships between set types : The following relationships hold between the interior $I(A)$, the boundary $R(A)$, the closure $H(A)$ and the complement $\bar{A}$ of a subset A of a topological space (M ; T) :

(1) The set A is closed if and only if the following equivalent conditions are satisfied :

$R(A) \subseteq A$: the set A contains its boundary R(A)

$H(A) = A$: the set A coincides with its closure H(A)

$I(\bar{A}) = \bar{A}$: the complement $\bar{A}$ is open

(2) The set A is open if and only if the following equivalent conditions are satisfied :

$R(A) \cap A = \emptyset$: the set A contains no points of its boundary R(A)

$I(A) = A$: the set A coincides with its interior I(A)

$R(\bar{A}) \subseteq \bar{A}$: the complement $\bar{A}$ is closed

(3) The boundary R(A) of a set A has the following properties :

$R(R(A)) \subseteq R(A)$: the boundary is closed

$R(A) = R(\bar{A})$: the set A and its complement $\bar{A}$ have the same boundary

$R(A) = H(A) \cap H(\bar{A})$: the boundary is the intersection of the closures of the set and its complement

(4) The closure of a set is a closed set.

The relationships (1) to (4) are proved by applying the definitions of the point types and set types and the properties of topological spaces. The proof of (4) is carried out as an example.

Proof : The closure of a set is a closed set.

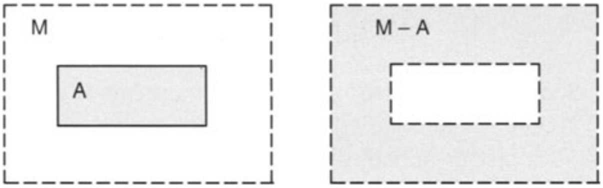
(1) The closure $H(A)$ is the set of interior points and boundary points of A, and hence its complement $\overline{H(A)}$ is the set of exterior points of A.

(2) For every exterior point $x \in M$ there is an open set B_x with $B_x \cap A = \emptyset$, and hence $B_x \cap H(A) = B_x \cap (A \cap R(A)) = B_x \cap R(A) \subseteq R(A)$.

(3) If the open set B_x contains a boundary point $y \in R(A)$, then B is a neighborhood of the boundary point y and therefore contains points in A. This contradicts the condition $B_x \cap A = \emptyset$ in (2). Thus B_x contains no boundary points. From $B_x \cap H(A) \subseteq R(A)$ and $B_x \cap R(A) = \emptyset$ it follows that $B_x \cap H(A) = \emptyset$.

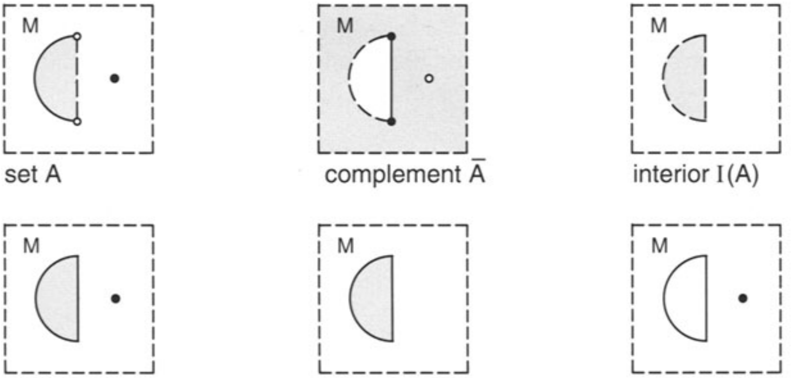
(4) Since $B_x \cap H(A) = \emptyset$, every exterior point of A is an exterior point of the closure H(A). Since by (1) the exterior points of A form the complement $\overline{H(A)}$, every point x of the complement $\overline{H(A)}$ is an exterior point of the closure H(A). Hence the closure H(A) is closed.

Example 3 : Open complement of a closed set in $\mathbb{R}^2$



Let the set A in the topological space $(M\,;T)$ be closed. Then its complement $\overline{A} = M - A$ is open.

Example 4 : Set types in the real euclidean plane



set A

complement $\overline{A}$

interior I(A)

closure H(A)

accumulation points

boundary R(A)

Example 5 : Boundaries in the one-dimensional euclidean space $\mathbb{R}$

Let M be the set $\mathbb{R}$ of real numbers, and let A be the set $\mathbb{Q}$ of rational numbers. Then $R(A) = \mathbb{R}$ and $R(\mathbb{R}) = \emptyset$, and thus $R(R(A)) = \emptyset \subseteq R(A)$.

5.6 TOPOLOGICAL MAPPINGS

Introduction : The study of the properties of topological spaces requires a definition of the concept of "topological spaces with identical structure". Two topological spaces are identically structured (isomorphic, homeomorphic) if there is a bijective mapping between them which preserves structure (is homomorphic) in both directions. Homomorphic mappings of topological spaces are called continuous mappings. Isomorphic mappings of topological spaces are called topological (homeomorphic) mappings. Homeomorphic spaces cannot be distinguished by topological means. Equivalent properties of homeomorphic spaces are topological invariants. These concepts are defined in the following.

Continuous mapping : Let (M ; T) and (N ; S) be spaces with the topologies T and S . A mapping $f : M \rightarrow N$ is said to be continuous on M if the preimage $f^{-1}(S_i)$ of every open set S_i of S is an open set of T.

$$\text{f is continuous} \quad :\Leftrightarrow \quad \bigwedge_{S_i \in S} (f^{-1}(S_i) \in T)$$

For a given topology S on N and a given continuous mapping $f : M \rightarrow N$, let $A := \{f^{-1}(S_i) \mid S_i \in S\}$ be the set of preimages of the open sets in S. Since f is continuous, $A \subseteq T$. If A is a proper subset of T, then the subset $T - A$ of open subsets in T is not determined by S and f. The set A is thus the coarsest topology on M which renders f continuous. Every topology which is finer than A also renders f continuous.

Locally continuous mapping : A mapping $f : M \rightarrow N$ from a topological space (M ; T) to a topological space (N ; S) is said to be (locally) continuous at a point $x \in M$ if for every element W_x of the neighborhood system of f(x) in N there is an element $U_x = f^{-1}(W_x)$ of the neighborhood system of x in M.

$$\text{f is continuous at } x \in M \quad :\Leftrightarrow \quad \bigwedge_{W_x} \bigvee_{U_x} (f^{-1}(W_x) = U_x)$$

A mapping $f : M \rightarrow N$ between topological spaces is continuous if and only if it is continuous at every point x of M. This theorem is often used to prove that a function is continuous.

Proof : $f : M \rightarrow N$ is continuous $\Leftrightarrow$ f is continuous at every point $x \in M$

(1) Let the mapping f be continuous. Let W_x be a neighborhood of f(x) in N. Then there is an open set $S_x \in S$ such that $f(x) \in S_x \subseteq W_x$. Since the mapping f is continuous, there is an open set $T_x = f^{-1}(S_x)$ such that $x \in T_x \subseteq f^{-1}(W_x)$. Hence for every neighborhood W_x of f(x) in N there is a neighborhood $U_x = f^{-1}(W_x)$ of x in M such that f is continuous at the point x.

(2) Let the mapping f be continuous at every point $x \in M$. Let S_i be an arbitrary open set in S. Then $f^{-1}(S_i)$ is a neighborhood of every point $x \in f^{-1}(S_i)$. Hence there is an open set T_x in T such that $x \in T_x \subseteq f^{-1}(S_i)$. The set $f^{-1}(S_i) = \bigcup_x T_x$ is thus a union of open sets and is therefore itself an open set. Since the preimage of every open set S_i in S is an open set $f^{-1}(S_i)$ in M, the mapping f is continuous.

Composition of continuous mappings : Let (A ; R), (B ; S) and (C ; T) be topological spaces. Let the mappings $f : A \to B$ and $g : B \to C$ be continuous. Then the composition $g \circ f : A \to C$ is also a continuous mapping.

$$(g \circ f)^{-1}(T_n) = f^{-1}(g^{-1}(T_n)) = f^{-1}(S_m) = R_i$$

Topological mapping : Let (M ; T) and (N ; S) be topological spaces. A mapping $f : M \to N$ is said to be topological (homeomorphic) if f is bijective and both f and the inverse mapping f^{-1} are continuous mappings.

$$\text{f is topological} \quad :\Leftrightarrow \quad \bigwedge_{T_k \in T} (f(T_k) \in S) \;\wedge\; \bigwedge_{S_m \in S} (f^{-1}(S_m) \in T)$$

Note : In the definitions of continuous mappings and their composition, $f^{-1}(S_i)$ designates the preimage of S_i. It is not assumed that the mapping f has an inverse. By contrast, in the definition of a topological mapping f^{-1} is the inverse of the mapping f.

Homeomorphic spaces : The spaces (M ; T) and (N ; S) are said to be homeomorphic if there is a topological mapping $f : M \to N$. The homeomorphism of the spaces (M ; T) and (N ; S) is designated by $M \sim N$. Homeomorphic spaces cannot be distinguished by topological means : They have identical topological structure.

Topological equivalence : The homeomorphism $\sim$ of topological spaces (A, T_A), (B, T_B), (C, T_C), ... is an equivalence relation. A set of topological spaces is partitioned into classes of homeomorphic spaces.

(1) The relation $\sim$ is reflexive since the identity mapping $i : A \to A$ is a topological mapping. The mapping i possesses the inverse i and is continuous. Every topological space is homeomorphic to itself.

(2) The relation $\sim$ is symmetric since for every topological mapping $f : A \to B$ there is a topological mapping $g : B \to A$. The mapping g is the inverse of f and is continuous. The inverse of g is the continuous mapping f.

(3) The relation $\sim$ is transitive : $A \sim B$ and $B \sim C$ imply $A \sim C$, since for topological mappings $f : A \to B$ and $g : B \to C$:

(a) $g \circ f$ is bijective since $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$ and g, f are bijective.

(b) $g \circ f$ is continuous since g and f are continuous.

(c) $(g \circ f)^{-1}$ is continuous since g^{-1} and f^{-1} are continuous.

Topological invariant : A property of a topological space (M ; T) is called a topological invariant if every space (N ; S) homeomorphic to (M ; T) has the same property. Several topological invariants are defined in the following sections.

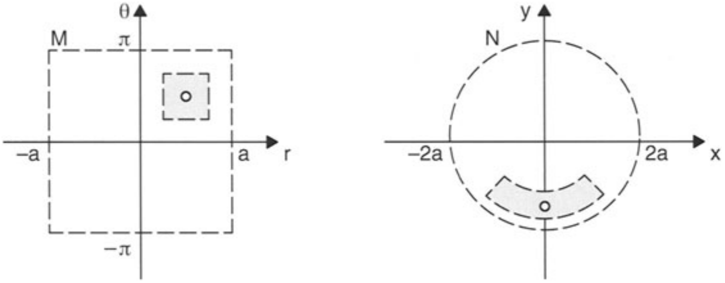
Open mapping : Let (M ; T) and (N ; S) be topological spaces. A mapping $f : M \rightarrow N$ is said to be open if the image of every open set T_i of T is an open set of S. If the mapping f is bijective, continuous and open, then the spaces M and N are homeomorphic.

$$f \text{ is open } :\Leftrightarrow \quad \bigwedge_{T_i} f(T_i) \in S$$

Closed mapping : Let (M ; T) and (N ; S) be topological spaces. A mapping $f : M \rightarrow N$ is said to be closed if the image of every closed set in M is a closed set in N.

Discrete mapping : A continuous mapping $f : M \rightarrow D$ from a topological space (M ; T) to a topological space (D ; S) is said to be discrete if the topology S on the target D is discrete.

Example 1 : Topological mapping



$$f : M \rightarrow N \qquad \text{with} \qquad f(r, \theta) = (x, y) \qquad \text{and} \qquad \begin{aligned} x &= (r + a)\cos(\theta + \pi) \\ y &= (r + a)\sin(\theta + \pi) \end{aligned}$$

The mapping $f : M \rightarrow N$ is topological : It is continuous and has a continuous inverse. Open sets of N are mapped to open sets of M. Open sets of M are mapped to open sets of N.

Example 2 : Non-topological mapping

Let $(M\,;T)$ and $(N\,;S)$ be topological spaces with the following sets :

$$M = \{a, b, c\} \qquad T = \{\emptyset, \{a\}, \{c\}, \{a,b\}, \{a,c\}, \{a,b,c\}\}$$

$$N = \{x, y, z\} \qquad S = \{\emptyset, \{x\}, \{x,y\}, \{x,y,z\}\}$$

Let the mapping $f : M \to N$ be bijective with $f(a) = x$, $f(b) = y$, $f(c) = z$. The mapping f is continuous since every open set S_i of S has an open preimage T_m in T :

$$f^{-1}(\emptyset) = \emptyset \qquad f^{-1}(\{x,y\}) = \{a,b\}$$

$$f^{-1}(\{x\}) = \{a\} \qquad f^{-1}(\{x,y,z\}) = \{a,b,c\}$$

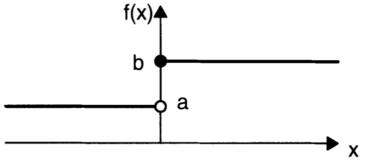
The inverse mapping $g : N \to M$ with $g(x) = a$, $g(y) = b$, $g(z) = c$ is not continuous since the open sets $\{c\}$ and $\{a, c\}$ of T do not have an open preimage S_i in S :

$g^{-1}(\{c\}) = \{z\}$ is not an element of S

$g^{-1}(\{a, c\}) = \{x, z\}$ is not an element of S

Although the mapping $f : M \to N$ is continuous and has an inverse, it is not topological since $f^{-1} : N \to M$ is not continuous.

Example 3 : Discontinuity of the Heaviside function



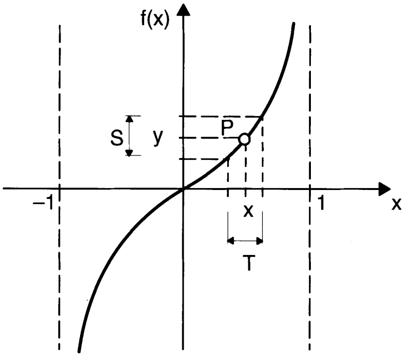
The Heaviside function is a mapping $f : \mathbb{R} \to \{a, b\}$ given by

$$x < 0 \quad \Rightarrow \quad f(x) = a$$

$$x \geq 0 \quad \Rightarrow \quad f(x) = b$$

Let the space $\mathbb{R}$ be equipped with its natural topology. Let the topology of $\{a,b\}$ be discrete, so that $\{a\}$ and $\{b\}$ are open sets. The preimage of the open set $\{a\}$ is $f^{-1}(\{a\}) = \,]-\infty, 0[$ and hence an open set. The preimage of the open set $\{b\}$ is $f^{-1}(\{b\}) = [0, \infty[$ and hence not an open set. On the real axis with its natural topology, the Heaviside function is therefore not a continuous function.

Example 4 : Topological invariance



The mapping $f : \,]-1, 1[\; \rightarrow \mathbb{R}$ with $f(x) = \tan \frac{\pi x}{2}$ is topological. For every point P of the graph the mapping $y = f(x)$ may be inverted to yield $x = f^{-1}(y)$. The open sets S and T are mapped to each other. The set $]-1, 1[$ is bounded, the set $\mathbb{R}$ is not bounded. Although the sets $]-1, 1[$ and $\mathbb{R}$ are homeomorphic, one of the sets is bounded and the other is not. Boundedness is not a topological invariant.

The example shows that boundedness is a metric property which cannot be described by topological concepts alone. This is hardly surprising, since the metric topology is derived from the metric, and thus the metric contains more information than the topology.

5.7 CONSTRUCTION OF TOPOLOGIES

5.7.1 FINAL AND INITIAL TOPOLOGIES

Introduction : Topologies for new spaces may be generated using a mapping of known topological spaces. In applications, this procedure is used particularly to construct continuous mappings. Continuous mappings are the basis for constructing homeomorphic spaces. Mappings with special properties generate topologies with special properties.

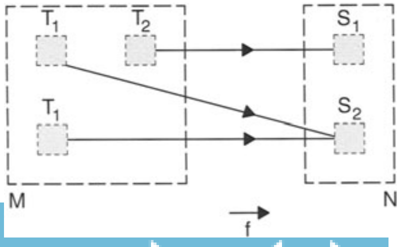
If a mapping $f : M \rightarrow N$ and a topology T on M are given, the question arises for which topologies on N the mapping f is continuous. To answer this question, the final topology S_{fin} on N is constructed. An example of a final topology is furnished by the quotient topology of a quotient set M/E of the underlying set M. The sum topology on the union of disjoint sets is closely related to final topologies.

If a mapping $f : M \rightarrow N$ and a topology S on N are given, the question arises for which topologies on M the mapping f is continuous. To answer this question, the initial topology T_{init} on M is constructed. An example of an initial topology is furnished by the relative topology of a subspace. The product topology of a cartesian product is closely related to initial topologies.

The essential difference between final and initial topologies is the following : In the construction of a final topology, the topology on the domain of the mapping is known and the topology on the target is to be determined, while in the construction of an initial topology the topology on the target is known and the topology on the domain is to be determined.

Final topology : Let $(M ; T)$ be a space with known topology T, and let $f : M \rightarrow N$ be a surjective mapping. Let the image of an open set T_m in T be $S_i = f(T_m)$. The set of the images of T is called the final topology induced (generated) on N by f and T and is designated by S_{fin}.

$$S_{fin} \quad := \quad \{ f(T_m) \mid T_m \in T \}$$



The mapping f : M → N between the spaces (M ; T) and (N ; S_{fin}) is continuous by virtue of the construction of S_{fin}. The mapping f is also continuous if a subset of S_{fin} is taken as the topology on N. For a given topology T on M, S_{fin} is the finest topology on N for which the mapping f is continuous.

If the mapping f is bijective, every open set S_i is the image of an open set T_i. Then the inverse mapping $f^{-1}: N \to M$ is also continuous with respect to the topologies T and S_{fin}. For a bijective mapping f, the spaces (M ; T) and (N ; S_{fin}) are therefore homeomorphic.

Quotient topology : Let (M ; T) be a topological space, and let E be an equivalence relation in M. The canonical mapping k : M → M/E is a surjective mapping from the underlying set M to the quotient set M/E. It induces a final topology T_E on the quotient set. The final topology T_E is called the quotient topology with respect to the equivalence relation E. The space (M/E ; T_E) is called a quotient space of the topological space (M ; T).

$$k : M \to M/E$$

$$T_E := \{ k(T_m) \mid T_m \in T \}$$

Sum topology : Let (M ; T) and (N ; S) be topological spaces whose underlying sets are disjoint, that is $M \cap N = \emptyset$. The union $M \cup N$ is taken as the underlying set of a new topological space. The finest topology V is chosen on $M \cup N$ which renders the injections $i_M : M \to M \cup N$ and $i_N : N \to M \cup N$ continuous. This is an extension of the concept of a final topology.

$$\begin{array}{llll} i_M : & M \to M \cup N & \text{with} & i_M(a) = a \\ i_N : & N \to M \cup N & \text{with} & i_N(b) = b \end{array}$$

Since V is the finest topology which renders i_M and i_N continuous, the basis B of V contains exactly the open sets $T_k \in T$ and $S_m \in S$. If the basis contains elements T_k and T_n of T, then it also contains their intersection $T_k \cap T_n$, since T is a topology.

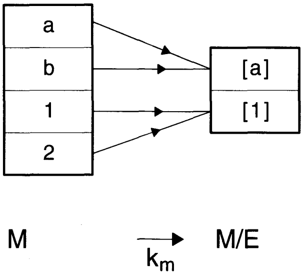
$$B := \{ B_i \mid B_i \in T \ \vee \ B_i \in S \}$$

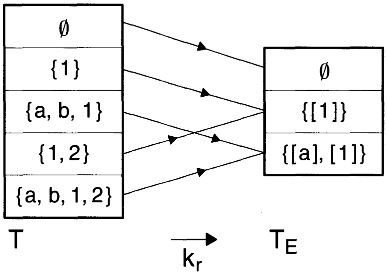
By property (T3) of topologies, V also contains all unions of elements of B. The space ($M \cup N$; V) is called the topological sum (union space) of the topological spaces (M ; T) and (N ; S). The topology V is called the sum topology (union topology).

$$V := \{ \bigcup_{B_i \in B'} B_i \mid B' \subseteq B \}$$

Example 1 : Quotient topology

Let (M ; T) be a topological space, and let E be an equivalence relation which partitions the underlying set M into the classes [a] and [1]. The underlying set M/E and the topology T_E of the quotient space follow from the definition of the canonical mappings $k_m : M \to M/E$ and $k_r : T \to T_E$.





Example 2 : Sum topology

Let the topological spaces (M ; T) and (N ; S) be given :

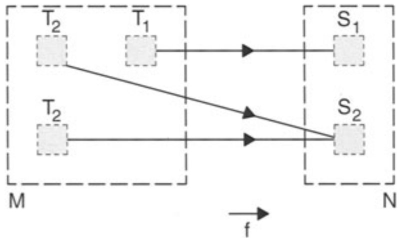
$$M = \{a,b\} \qquad T = \{\emptyset, \{a\}, \{b\}, \{a,b\}\}$$
$$N = \{1,2\} \qquad S = \{\emptyset, \{1\}, \{1,2\}\}$$

Then the topological sum (M ∪ N ; V) has the following underlying set M ∪ N, basis B and sum topology V :

$$\begin{aligned}
M \cup N &= \{a,b,1,2\} \\
B &= \{B_i \in T \ \vee \ B_i \in S\} \\
&= \{\{a\}, \{b\}, \{a,b\}, \{1\}, \{1,2\}\} \\
V &= \{\bigcup_{B_i \in B'} B_i \mid B' \subseteq B\} \\
&= \{\emptyset, \{a\}, \{b\}, \{a,b\}, \{1\}, \{1,2\}, \\
&\quad \{a,1\}, \{a,1,2\}, \{b,1\}, \{b,1,2\}, \{a, b, 1\}, \{a, b, 1, 2\}\}
\end{aligned}$$

Initial topology : Let (N ; S) be a space with known topology S, and let f : M → N be a mapping. Let the subset of M which f maps onto the open set $S_i \in S$ be $T_i = f^{-1}(S_i)$. Since the mapping f is generally not bijective, f^{-1} here does not designate the inverse of f : A point in S_i may be the image of more than one point in T_i. The set of preimages T_i of the elements of S is called the initial topology induced (generated) on M by f and S and is designated by T_{init}.

$$T_{init} := \{ f^{-1}(S_i) \mid S_i \in S \}$$



The mapping $f : M \to N$ between the spaces $(M ; T_{init})$ and $(N ; S)$ is continuous by virtue of the construction of T_{init}. The mapping is also continuous if a finer topology than T_{init} is chosen on M. The initial topology T_{init} is the coarsest topology on M which renders f continuous given the topology S.

If the mapping $f : M \to N$ is bijective, the spaces $(M ; T_{init})$ and (N ; S) are homeomorphic, since the mapping f and its inverse $f^{-1} : M \to N$ are continuous.

Relative topology : Let (N ; S) be a topological space, and let $M \subseteq N$ be a subset of N. The coarsest topology T on M is chosen which renders the injection $i : M \to N$ with i(a) = a continuous. This is the initial topology T induced on M by i. If M contains only a subset of the points of an open set $S_i \in S$, then the preimage of S_i is the intersection $M \cap S_i$. The space (M ; T) is called a subspace of the topological space (N ; S). The topology T is called the relative topology (subspace topology).

$$i : M \to N \quad \text{with} \quad i(a) = a$$

$$T := \{ M \cap S_i \mid S_i \in S \}$$

Product topology : Let (M ; T) and (N ; S) be topological spaces. The cartesian product $M \times N$ is taken as the underlying set of a new topological space. The coarsest topology P on $M \times N$ is chosen which renders both of the projections $p_M : M \times N \rightarrow M$ with $p_M((a, b)) = a$ and $p_N : M \times N \rightarrow N$ with $p_N((a, b)) = b$ continuous. This is an extension of the concept of an initial topology.

To determine a basis for the topology P, cartesian products $T_i \times S_k$ of open sets $T_i \in T$ and $S_k \in S$ are considered. Each of these products is a set of ordered pairs (t, s). The set of cartesian products $T_i \times S_k$ which can be formed with the elements of T and S is designated by $T \times S$:

$$T_i \times S_k = \{ (t, s) \mid t \in T_i \quad \wedge \quad s \in S_k \}$$

$$T \times S = \{ T_i \times S_k \mid T_i \in T \quad \wedge \quad S_k \in S \}$$

The set $T \times S$ is suitable as a basis of a topology if for any two elements $T_i \times S_k$ and $T_m \times S_n$ it also contains their intersection. The intersection of the elements is :

$$\begin{aligned} (T_i \times S_k) \cap (T_m \times S_n) &= \{ (t, s) \mid (t, s) \in T_i \times S_k \quad \wedge \quad (t, s) \in T_m \times S_n \} \\ &= \{ (t, s) \mid t \in T_i \cap T_m \quad \wedge \quad s \in S_k \cap S_n \} \\ &= (T_i \cap T_m) \times (S_k \cap S_n) \end{aligned}$$

Since by definition $T_u = T_i \cap T_m$ is an element of the topology T and $S_w = S_k \cap S_n$ is an element of the topology S, $T \times S$ contains the intersection $T_u \times S_w$ of $T_i \times S_k$ and $T_m \times S_n$. Hence $T \times S$ is suitable as a basis.

The set $B := T \times S$ is chosen as a basis of the topology P on the underlying set $M \times N$. Then the projections p_M and p_N are continuous, since every open set in T and every open set in S has a preimage in B and hence in P. The topology P contains all unions of elements of B.

$$\begin{aligned} p_M &: \quad M \times N \rightarrow M \qquad \text{with} \qquad p_M((a, b)) = a \\ p_N &: \quad M \times N \rightarrow N \qquad \text{with} \qquad p_N((a, b)) = b \\ B &:= \quad T \times S = \{ T_i \times S_k \mid T_i \in T \wedge S_k \in S \} \\ P &:= \quad \{ \bigcup_{B_i \in B'} B_i \mid B' \in B \} \end{aligned}$$

The space $(M \times N ; P)$ is called the product space of the topological spaces (M ; T) and (N ; S). The topology P is called the product topology. P is the coarsest topology on $M \times N$ which renders p_M and p_N continuous. For if the basis B does not contain all sets in $T \times S$, then either at least one intersection of elements of B is not contained in B, or at least one of the projections p_M and p_N is not continuous.

Example 3 : Relative topology of euclidean space

Let the euclidean space $\mathbb{R}^3$ be equipped with its natural topology T^3. The open sets of the basis of T^3 are open balls. The intersections of the open balls with the euclidean plane $\mathbb{R}^2$ are open disks, which are the basis elements of the natural topology on $\mathbb{R}^2$. The intersections of the open balls with the euclidean line $\mathbb{R}^1$ are open intervals, which are the basis elements of the natural topology on $\mathbb{R}^1$. The spaces $\mathbb{R}^1$ and $\mathbb{R}^2$ are subspaces of $\mathbb{R}^3$: They are equipped with the relative topology.

Example 4 : Product topology

Let the topological spaces (M ; T) and (N ; S) from Example 2 be given :

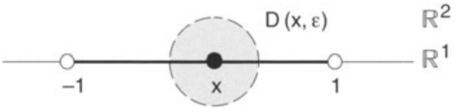
$$\begin{aligned} M &= \{a,b\} & T &= \{\emptyset, \{a\}, \{b\}, \{a,b\}\} \\ N &= \{1,2\} & S &= \{\emptyset, \{1\}, \{1,2\}\} \end{aligned}$$

Then the product space $(M \times N\ ; P)$ has the following underlying set M, basis B and product topology P :

$$\begin{aligned} M \times N &= \{(a,1), (a,2), (b,1), (b,2)\} \\ B &= \{T_i \times S_k \mid T_i \in T \ \wedge\ S_k \in S\} \\ &= \{\emptyset, \{a\} \times \{1\}, \{a\} \times \{1,2\}, \{b\} \times \{1\}, \{b\} \times \{1,2\}, \\ &\qquad \{a,b\} \times \{1\}, \{a,b\} \times \{1,2\}\} \\ &= \{\emptyset, \{(a,1)\}, \{(a,1), (a,2)\}, \{(b,1)\}, \{(b,1), (b,2)\}, \\ &\qquad \{(a,1), (b,1)\}, \{(a,1), (a,2), (b,1), (b,2)\}\} \\ P &= \{\bigcup_{B_i \in B'} B_i \mid B' \in B\} \\ &= \{\emptyset, \{(a,1)\}, \{(b,1)\}, \\ &\qquad \{(a,1), (a,2)\}, \{(b,1), (b,2)\}, \{(a,1), (b,1)\}, \\ &\qquad \{(a,1), (a,2), (b,1)\}, \{(a,1), (b,1), (b,2)\}. \\ &\qquad \{(a,1), (a,2), (b,1), (b,2)\}\} \end{aligned}$$

5.7.2 SUBSPACES

Introduction : The concept of the relative topology of a subset of a topological space introduced in Section 5.7.1 leads to a more precise treatment of the point types and set types defined in Section 5.5. The interval $]-1,1[$, which is open in $\mathbb{R}^1$, is considered as an example. In the euclidean space $\mathbb{R}^2$, this point set is not open, for at an arbitrary point $x \in]-1,1[$ the open set $D(x, \varepsilon)$ of the euclidean space $\mathbb{R}^2$ contains points of $\mathbb{R}^2$ which do not lie in the interval $]-1,1[$.



The example demonstrates that the definitions and rules of Section 5.5 hold only with respect to the subspace under consideration and its relative topology. Properties of subspaces are treated in this section.

Subspace : A topological space (M ; T) is called a subspace of the topological space (N ; S) if M is a subset of N and T is the relative topology induced on M by S.

$$T = \{S_i \cap M \mid S_i \in S\}$$

Interior point : Let A be a subset of a subspace (M ; T). Then a point $x \in M$ is called an interior point of A in M if there is an open set T_i of the relative topology T which contains x and is contained in A.

$$\text{x is an interior point of A in M} \quad :\Leftrightarrow \quad \bigvee_{T_i \in T} (x \in T_i \subseteq A)$$

Exterior point : Let A be a subset of a subspace (M ; T). Then a point $x \in M$ is called an exterior point of A in M if there is an open set T_i of the relative topology T which contains x and is contained in the complement $\bar{A} = M - A$.

$$\text{x is an exterior point of A in M} \quad :\Leftrightarrow \quad \bigvee_{T_i \in T} (x \in T_i \subseteq M - A)$$

Boundary point : Let A be a subset of a subspace (M ; T). Then a point $x \in M$ is called a boundary point of A in M if every open set T_x of the relative topology T which contains x contains at least one point of A and one point of $\bar{A}$.

$$\text{x is a boundary point of A in M} \quad :\Leftrightarrow \quad \bigwedge_{T_x \in T} (T_x \cap A \neq \emptyset \quad \wedge \quad T_x \cap \bar{A} \neq 0)$$

Interior : Let A be a subset of a subspace (M ; T). The set of all inner points of A in M is called the interior of A in M and is designated by I(A) or I_M(A). The interior of A is the union of the open sets $T_i \in T$ which are contained in A.

$$I(A) = \bigcup_{T_i \in T} (T_i \subseteq A)$$

Exterior : Let A be a subset of a subspace (M ; T). The set of exterior points of A in M is called the exterior of A in M and is designated by E(A) or E_M(A). The exterior of A is the union of the open sets $T_i \in T$ which are contained in $\bar{A} = M - A$.

$$E(A) = \bigcup_{T_i \in T} (T_i \subseteq M - A)$$

Boundary : Let A be a subset of a subspace (M ; T). The set of all boundary points of A in M is called the boundary of A in M and is designated by R(A) or δA or $\delta_M A$. The boundary of A is the underlying set M without the interior and exterior of A.

$$R(A) = \delta A = M - I(A) - E(A)$$

Closure : Let A be a subset of a subspace (M ; T). The set of all inner points and boundary points of A is called the closure of A in M and is designated by H(A) or H_M(A). The closure of A is the underlying set M without the exterior of A.

$$H(A) = M - E(A)$$

Properties of subspaces :

(U1) If the topology S of the space (N ; S) has a basis A, then the topology T of the subspace (M ; T) has the basis

$$B = \{B_i = A_i \cap M \mid A_i \in A\}$$

(U2) If $(M_1 ; T_1)$ and $(M_2 ; T_2)$ are disjoint subspaces of (N ; S), then their topological sum $(M_1 \cup M_2 ; V)$ with the sum topology V is also a subspace of (N ; S).

(U3) If (M ; T) is a subspace of (N ; S) and (N ; S) is a subspace of (U ; R), then (M ; T) is a subspace of (U ; R).

(U4) A subset A of a subspace (M ; T) of a topological space (N ; S) is open in M if and only if there is an open set P in N such that A is the intersection of M and P :

$$\text{A is open in M} \quad \Leftrightarrow \quad \bigvee_{P \in S} (A = M \cap P)$$

(U5) A subset A of a subspace (M ; T) of a topological space (N ; S) is closed in M if and only if there is a closed set C in N such that A is the intersection of M and C :

$$\text{A is closed in M} \quad \Leftrightarrow \quad \bigvee_{\bar{C} \in S} (A = M \cap C)$$

Proof : Properties of subspaces

(U1) Every element S_k of the topology S is a union $\cup A_i$ of elements of the basis A. For the element S_k of S there is a corresponding element $T_k = M \cap S_k$ of the relative topology T. Therefore :

$$T_k = M \cap S_k = M \cap \bigcup_i A_i = \bigcup_i (M \cap A_i) = \bigcup_i B_i$$

Every element $T_k \in T$ is a union of elements $B_i = A_i \cap M \in B$. Since for two elements A_i and A_m the basis A also contains their intersection $A_n = A_i \cap A_m$, it follows that for two elements $B_i = A_i \cap M$ and $B_m = A_m \cap M$ the set B also contains their intersection $B_i \cap B_m = (A_i \cap M) \cap (A_m \cap M) = (A_i \cap A_m) \cap M$. Hence B is a basis of the topology T.

(U2) The topologies of the subspaces are $T_1 = \{S_i \cap M_1\} \mid S_i \in S\}$ and $T_2 = \{S_m \cap M_2 \mid S_m \in S\}$, respectively. A general element of the sum topology V is $\{S_i \cap M_1\} \cup \{S_m \cap M_2\} = \{S_i \cup S_m\} \cap \{M_1 \cup M_2\}$. Since S_i and S_m are elements of S, $S_i \cup S_m$ is by definition an element of S, and therefore $\{S_i \cup S_m\} \cap \{M_1 \cup M_2\}$ is an element of the relative topology of S with respect to $M_1 \cup M_2$. Hence $(M_1 \cup M_2 ; V)$ is a subspace of (N ; S).

(U3) By definition, the underlying sets of the spaces satisfy $M \subseteq N \subseteq U$. The topologies of the subspaces are $T = \{S_i \cap M \mid S_i \in S\}$ and $S = \{R_i \cap N \mid R_i \in R\}$. Substituting $S_i = R_i \cap N$ yields $T_i = S_i \cap M = (R_i \cap N) \cap M = R_i \cap (N \cap M) = R_i \cap M$. Therefore $T = \{R_i \cap M \mid R_i \in R\}$. Hence T is the relative topology induced on M by R, and (M ; T) is a subspace of (U ; R).

(U4) Let the set A be open in a subspace (M ; T) of (N ; S). Then for every point $x \in A$ there is an open set $T_x = S_x \cap M \subseteq A$ which contains x, so that $x \in S_x$. Let P be the union of all open sets $S_x \in S$ which contain a point $x \in A$. Then P is an open set in N, since P is a union of open sets. The intersection of P with the subspace M is a union of open sets of the relative topology T :

$$P = \bigcup_{x \in A} S_x$$

$$P \cap M = \bigcup_{x \in A} (S_x \cap M) = \bigcup_{x \in A} T_x$$

$T_x \subseteq A$ implies $P \cap M \subseteq A$. The fact that every point $x \in A$ is contained in an open set T_x implies $A \subseteq P \cap M$. From $A \subseteq P \cap M \subseteq A$ it follows that $A = P \cap M$. Hence there is a set P open in N whose intersection with the subspace M is the given set A open in M.

Conversely, let P be an open set in the space (N ; S) and $A = M \cap P$. Then for every point $x \in A \subseteq P$ there is an open set $S_x \in S \subseteq P$ which contains x. By definition, the relative topology T of the subspace (M ; T) contains the open set $T_x = S_x \cap M$. Since $x \in A$ and $x \in S_x$, also $x \in M$ and $x \in M \cap S_x = T_x$. From $T_x \subseteq M$ and $T_x \subseteq S_x \subseteq P$ it follows that $T_x \subseteq M \cap P = A$. Thus for every point $x \in A$ there is an open set T_x which contains x and is contained in A. Hence A is open in M.

(U5) Let the set A be closed in a subspace (M ; T) of (N ; S). Then the set M – A is open in M. By (U4) there is an open set $\overline{C}$ in N such that $M - A = M \cap \overline{C}$. This yields the set $C = N - \overline{C}$, which is closed in N and satisfies $A = M - (M \cap \overline{C}) = M \cap (N - \overline{C}) = M \cap C$.

Conversely, let C be a closed set in the space (N ; S) with $A = M \cap C$. Then the set $\overline{C} = N - C$ is open in N. By (U4), the set $M \cap \overline{C} = M \cap (N - C) = M - (M \cap C) = M - A$ is open in M. Hence the set A is closed in M.

Dense subspace : A subspace (M ; T) of a topological space (N ; S) is said to be dense in N if N is the closure of M. A subspace M is said to be nowhere dense in N if the interior of the closure of M is empty.

M dense in N	:⇔	H(M)	=	N
M nowhere dense in N	:⇔	I(H(M))	=	∅

5.7.3 PRODUCT SPACES

Introduction : The topology of the product of two topological spaces is defined in Section 5.7.1. In this section this concept is extended to the product of a finite number of topological spaces. Rules for continuous mappings to such product spaces are derived.

Product space of a finite number of spaces : Let a finite number of topological spaces $(A\,;\,R),\ldots,(Z\,;\,S)$ be given. The elements of the underlying sets are designated by $a_i \in A,\ldots,z_n \in Z$, the open sets of the topologies are designated by $R_i \in R,\ldots,S_n \in S$. The cartesian product $X := A \times \ldots \times Z$ is taken as the underlying set of a new topological space. The elements of X are n-tuples $x = (a_i,\ldots,z_n)$. The product topology is chosen to be the coarsest topology P which renders all projections $p_A : X \rightarrow A$ with $p_A(x) = a_i$ to $p_Z : X \rightarrow Z$ with $p_Z(x) = z_n$ continuous.

The set of all cartesian products $R_i \times \ldots \times S_n$ whose factors are open sets of the topologies $R,\ldots,S$ is chosen as a basis B of the topology P. This set of cartesian products is suitable as a basis since, as in the special case of two spaces, the intersection of any two open sets $R_i \times \ldots \times S_n$ and $R_k \times \ldots \times S_m$ is also an open set $R_j \times \ldots \times S_u$.

$$B = \{R_i \times \ldots \times S_n \mid R_i \in R \wedge \ldots \wedge S_n \in S\}$$

$$R_i \times \ldots \times S_n = \{(a_i,\ldots,z_n) \mid a_i \in R_i \wedge \ldots \wedge z_n \in S_n\}$$

$$\begin{aligned}(R_i \times \ldots \times S_n) \cap (R_k \times \ldots \times S_m) &= (R_i \cap R_k) \times \ldots \times (S_n \cap S_m) \\ &= R_j \times \ldots \times S_u\end{aligned}$$

The product topology P contains all unions of elements of the basis B. The space $(A \times \ldots \times Z\,;\,P)$ is called the product space of the topological spaces $(A\,;\,R),\ldots,$ $(Z\,;\,S)$. The topology P is the coarsest topology on $X = A \times \ldots \times Z$ which renders the projections $p_A,\ldots,p_Z$ continuous.

$$P = \{\bigcup_{B_i \in B'} B_i \mid B' \in B\}$$

$$\begin{array}{llll} p_A : X \rightarrow A & \text{with} & p_A(x) = a_i & \text{is continuous} \\ p_Z : X \rightarrow Z & \text{with} & p_Z(x) = z_n & \text{is continuous} \end{array}$$

Continuous mappings to a product space : Let a topological space $(M\,;\,T)$ and the product space $(X\,;\,P)$ of a finite number of topological spaces $(A\,;\,R),\ldots,$ $(Z;\,S)$ be given. A mapping $f : M \rightarrow X$ to the underlying set $X = A \times \ldots \times Z$ of the product space is continuous if and only if the compositions $p_A \circ f : M \rightarrow A$ to $p_Z \circ f : M \rightarrow Z$ involving the projections $p_A : X \rightarrow A$ to $p_Z : X \rightarrow Z$ are continuous.

$$\begin{tikzcd}
M \arrow[r, "f"] \arrow[dr, "p_A \circ f"'] & X \arrow[d, "p_A"] \\
& A
\end{tikzcd}$$

Proof : Continuous mappings to a product space

(1) Let the mapping f be continuous. By the definition of the product topology, each of the projections $p_A,...,p_Z$ is continuous. Therefore the mappings $p_A \circ f,...,p_Z \circ f$ are compositions of continuous mappings, and hence continuous.

(2) Let each of the compositions $p_A \circ f,...,p_Z \circ f$ be continuous. It is to be proved that for the mapping $f : M \rightarrow X$ the preimage of every set open in the space (X ; P) is open in M. To prove this, it is sufficient to show that the preimage of every element of the basis $B = \{R_i \times ... \times S_n \mid R_i \in R \wedge ... \wedge S_n \in S\}$ is open in M. By the definition of a product topology, $p_A : X \rightarrow A$ is continuous, so that the preimage $R_i \times ... \times S_n$ of the set R_i open in A is a set open in X. Since $p_A \circ f$ is continuous by hypothesis, the preimage of R_i in M is an open set T_k. But by the definition of the composition $p_A \circ f$, the open set $T_k \subseteq M$ is also the preimage of the open set $R_i \times ... \times S_n \in B$ under f. Hence the mapping f is continuous.

Continuous mapping between product spaces : Let the mappings $f_1 : M_1 \rightarrow N_1$ and $f_2 : M_2 \rightarrow N_2$ be continuous. Let the cartesian products $M_1 \times M_2$ and $N_1 \times N_2$ be equipped with the product topologies. Then the mapping $g : M_1 \times M_2 \rightarrow N_1 \times N_2$ with $g(a,b) = (f_1(a), f_2(b))$ is continuous.

Proof : Continuous mapping between product spaces

Let $S_1 \times S_2$ be an element of the basis of $N_1 \times N_2$. Since $N_1 \times N_2$ is equipped with the product topology, the projections $S_1 = p_1(S_1 \times S_2)$ and $S_2 = p_2(S_1 \times S_2)$ are open sets. Their preimages $f_1^{-1}(S_1)$ and $f_2^{-1}(S_2)$ are open sets since f_1 and f_2 are continuous. Since $M_1 \times M_2$ is equipped with the product topology, the preimage $g^{-1}(S_1 \times S_2) = f_1^{-1}(S_1)) \times f_2^{-1}(S_2)$ of $S_1 \times S_2$ with respect to g is an open set of $M_1 \times M_2$. Hence the mapping g is continuous.

5.8 CONNECTEDNESS OF SETS

5.8.1 DISCONNECTIONS AND CONNECTEDNESS

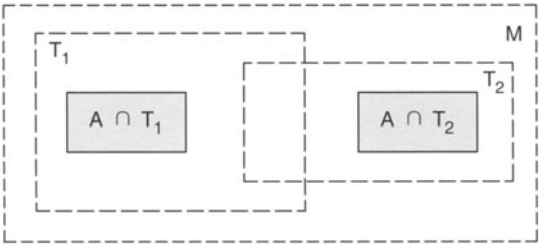
Introduction : A physical body is connected if forces can be conveyed between its parts. By contrast, mathematical connectedness is a topological property. The connectedness of two sets can only be studied if they are subsets of the underlying set of a topological space.

To define the connectedness of two sets, it is not sufficient to consider the intersection of these sets. For example, let A be a set with a boundary point x which does not belong to A. Let the same point x be a boundary point of the set B, and let it be contained in B. Let the intersection $A \cap B$ be empty. Then the union $A \cup B$ contains the point x although $A \cap B$ is empty. Hence the sets A and B are connected although their intersection is empty.

The mathematical definition of connectedness is based on the concepts of disconnections and separated sets. In a disconnection, the open sets of the topology of the space are used to assess connectedness. A set is connected if there is no disconnection for the set. For separated sets, the topological set types of boundary and closure are used to assess connectedness. In contrast to disjoint sets, separated sets contain no points of each other's closure. A set is connected if it is not the union of non-empty separated sets.

The concepts of disconnections, separated sets, connectedness of a set and connectedness of a space are treated in this section. Disconnections and separated sets are constructed from given sets. Different equivalent definitions of the connectedness of sets are presented.

Disconnection of a set : Let A be a subset of a topological space (M ; T), and let T_1, T_2 be open sets of the topology T. The sets T_1 and T_2 form a disconnection of the set A in the space (M ; T) if the intersections $A \cap T_1$ and $A \cap T_2$ are non-empty disjoint sets whose union is A. The disconnection is designated by $T_1 \& T_2$.



$$(A \cap T_1) \neq \emptyset \qquad (A \cap T_1) \cap (A \cap T_2) = \emptyset$$

$$(A \cap T_2) \neq \emptyset \qquad (A \cap T_1) \cup (A \cap T_2) = A$$

Properties of a disconnection : A set $A \subseteq M$ possesses a disconnection $T_1 \& T_2$ in the topological space $(M;T)$ if and only if it possesses a disconnection $S_1 \& S_2$ in the subspace $(A;S)$ with the relative topology $S = \{S_i \mid S_i = A \cap T_i \ \wedge \ T_i \in T\}$. The set A possesses a disconnection in the subspace $(A;S)$ if and only if there is a set $S_1 \in S$ with $\emptyset \subset S_1 \subset A$ which is both open and closed in $(A;S)$.

Proof : Properties of a disconnection

(1) Let $T_1 \& T_2$ be a disconnection of the set A in the space $(M;T)$. The open sets $S_i = A \cap T_i$ of the relative topology S are used in the defining properties of the disconnection $T_1 \& T_2$.

$$
\begin{array}{llll}
(A \cap T_1) \neq \emptyset & \wedge & (A \cap T_1) \cap (A \cap T_2) = \emptyset & \wedge \\
(A \cap T_2) \neq \emptyset & \wedge & (A \cap T_1) \cup (A \cap T_2) = A & \Leftrightarrow \\
\end{array}
$$

$$S_1 \neq \emptyset \ \wedge \ S_2 \neq \emptyset \ \wedge \ S_1 \cap S_2 = \emptyset \ \wedge \ S_1 \cup S_2 = A$$

By definition, the subsets S_1 and S_2 form a disconnection $S_1 \& S_2$ of A in the subspace $(A;S)$, since $S_i = S_i \cap A$. From $S_1 \cap S_2 = \emptyset$ and $S_1 \cup S_2 = A$ it follows that $\bar{S}_1 = A - S_1 = S_2$ and $\bar{S}_2 = A - S_2 = S_1$. Then $S_2 \neq \emptyset$ and $\bar{S}_1 = S_2$ implies $\bar{S}_1 \neq \emptyset$, and $S_1 \neq \emptyset$ and $\bar{S}_2 = S_1$ implies $\bar{S}_2 \neq \emptyset$.

$$S_1 \neq \emptyset \ \wedge \ S_2 \neq \emptyset \ \wedge \ \bar{S}_1 = S_2 \ \wedge \ S_2 = S_1 \qquad \Leftrightarrow$$

$$S_1 \neq \emptyset \ \wedge \ S_2 \neq \emptyset \ \wedge \ \bar{S}_1 \neq \emptyset \ \wedge \ S_2 \neq \emptyset$$

It follows from $\bar{S}_1 = S_2$ that $\bar{S}_1$ is an open set of the relative topology S. Since S_1 and $\bar{S}_1$ are open sets in $(A;S)$, the set S_1 is both open and closed. From $S_1 \neq \emptyset$ and $\bar{S}_1 = A - S_1 \neq \emptyset$ it follows that $\emptyset \subset S_1 \subset A$. Thus for the disconnection $T_1 \& T_2$ there is a set $\emptyset \subset S_1 \subset A$ which is both open and closed in $(A;S)$.

(2) Let the set S_1 with $\emptyset \subset S_1 \subset A$ be both open and closed in the subspace $(A;S)$. Then $S_1 \neq \emptyset$ and $\bar{S}_1 = A - S_1 \neq \emptyset$. It follows from $S_2 = \bar{S}_1$ that S_2 is an open set in S and $S_2 \neq \emptyset$. The equivalences in (1) show that $T_1 \& T_2$ is a disconnection of A in $(M;T)$.

Separated sets : Two subsets A, B of a topological space $(M;T)$ are said to be separated in the space $(M;T)$ if the following conditions are satisfied :

(1)	The sets A and B are disjoint	: $A \cap B = \emptyset$
(2)	A contains no boundary point of B	: $A \cap R(B) = \emptyset$
(3)	B contains no boundary point of A	: $B \cap R(A) = \emptyset$

These conditions are satisfied if and only if A does not contain any points of the closure of B and B does contain any points of the closure of A :

A, B are separated sets in $(M;T)$:⇔

$$A \cap B = \emptyset \ \wedge \ A \cap R(B) = \emptyset \ \wedge \ B \cap R(A) = \emptyset \quad \Leftrightarrow$$

$$A \cap H(B) = \emptyset \ \wedge \ B \cap H(A) = \emptyset$$

Properties of separated sets : Two subsets A, B of a topological space (M ; T) are separated in (M ; T) if and only if they are separated in the subspace $(A \cup B\,;\,S)$ with the relative topology $S = \{(A \cup B) \cap T_i \mid T_i \in T\}$.

Proof : Properties of separated sets

(1) Let the subsets A, B be separated in (M ; T). Let their closures in the space (M ; T) be H(A) and H(B), respectively. Then the defining properties of separated sets imply that $A \cap H(B) = \emptyset$ and $B \cap H(A) = \emptyset$. Let the closures of the sets A and B in the subspace $(A \cup B\,;\,S)$ be $H_u(A)$ and $H_u(B)$, respectively. The closures $H_u(A)$ and $H_u(B)$ contain the points of the closures H(A) and H(B) which are contained in the underlying set $A \cup B$:

$$H_u(A) = (A \cup B) \cap H(A) = (A \cap H(A)) \cup (B \cap H(A)) = A$$
$$H_u(B) = (A \cup B) \cap H(B) = (B \cap H(B)) \cup (B \cap H(B)) = B$$

The sets A and B are separated in $(A \cap B\,;\,S)$ since the closures $H_u(A)$ and $H_u(B)$ satisfy the conditions in the definition of separated sets :

$$A \cap H_u(B) = A \cap B = \emptyset$$
$$B \cap H_u(A) = B \cap A = \emptyset$$

(2) Let the sets A and B be separated in the subspace $(A \cup B\,;\,S)$. Then by definition $A \cap B = \emptyset$. Every point $x \in M$ which is not contained in $A \cup B$ belongs neither to A nor to B. Such points do not influence the value of $A \cap H(B)$ or of $B \cap H(A)$. Therefore $A \cap H_u(B) = \emptyset$ and $B \cap H_u(A) = \emptyset$ in $A \cup B$ implies $A \cap H(B) = \emptyset$ and $B \cap H(A) = \emptyset$ in M. Hence the sets A and B are separated in (M ; T).

Construction of separated sets : Let $T_1 \& T_2$ be a disconnection of a set A in a topological space (M ; T). Then the intersections $A \cap T_1$ and $A \cap T_2$ are separated sets (M ; T).

Proof : Construction of separated sets

By the definition of the disconnection $T_1 \& T_2$, the sets $S_1 = A \cap T_1$ and $S_2 = A \cap T_2$ are not empty. The disconnection $T_1 \& T_2$ in the space (M ; T) corresponds to the disconnection $S_1 \& S_2$ in the subspace (A ; S). Since the sets S_1 and S_2 are closed in the subspace, $H(S_1) = S_1$ and $H(S_2) = S_2$. The definition of a disconnection implies $S_1 \cap S_2 = \emptyset$. Hence the sets S_1 and S_2 satisfy the conditions for separated sets in the subspace (A ; S) :

$$S_1 \cap H(S_2) = S_1 \cap S_2 = \emptyset$$
$$S_2 \cap H(S_1) = S_2 \cap S_1 = \emptyset$$

Since the sets $S_1 = A \cap T_1$ and $S_2 = A \cap T_2$ are separated in the subspace (A ; S), they are, by virtue of the properties of separated sets, also separated in the space (M ; T).

Construction of a disconnection : Let the non-empty sets A and B be separated. Then the complements T_1 and T_2 of the closures of B and A form a disconnection $T_1 \& T_2$ of the set $A \cup B$.

$$T_1 = \overline{H(B)}$$
$$T_2 = \overline{H(A)}$$

Proof : Construction of a disconnection

(1) The closures H(A) and H(B) are closed sets in the space (M ; T). Hence their complements are open sets in the space (M ; T).

$$T_1 = \overline{H(B)} \in T$$
$$T_2 = \overline{H(A)} \in T$$

(2) Since the sets A and B are separated, $A \cap H(B) = \emptyset$ and $H(A) \cap B = \emptyset$. The complement T_1 of the closure of B contains the set A, but no point of B. The complement T_2 of the closure of A contains the set B, but no point of A. It follows that :

$$(A \cup B) \cap T_1 = A$$
$$(A \cup B) \cap T_2 = B$$

(3) The sets $(A \cup B) \cap T_1$ and $(A \cup B) \cap T_2$ are not empty because the sets A and B are not empty. The intersection of the sets $(A \cup B) \cap T_1$ and $(A \cup B) \cap T_2$ is empty, since the sets A and B are separated. Hence the sets T_1 and T_2 form a disconnection $T_1 \& T_2$ of the set $A \cup B$ in the space (M ; T).

Connected set : A set A in a topological space (M ; T) is said to be connected if it is not disconnected. The set A is said to be disconnected if one of the following equivalent conditions is satisfied :

(Z1) The set A can be represented as a union of two non-empty separated sets.

(Z2) There is a disconnection for the set A.

Proof : Equivalence of the definitions of the connectedness of a set

(1) Let the set A be disconnected according to definition (Z1). Then A can be represented as a union of two non-empty separated sets B and C. The separated sets B and C may be used to construct the disconnection $T_1 \& T_2$ with the open sets $T_1 = \overline{H(C)}$ and $T_2 = \overline{H(B)}$. Hence condition (Z2) is satisfied.

(2) Let the set A be disconnected according to definition (Z2). Then there is a disconnection $T_1 \& T_2$ for A. Using the open sets T_1 and T_2, the non-empty separated sets $A \cap T_1$ and $A \cap T_2$ may be constructed, whose union is A. Hence condition (Z1) is satisfied.

Connected space : A topological space (M ; T) is said to be connected if the set M is connected. The set M is connected if and only if the following condition is satisfied :

(Z3) There is no set T_i with $\emptyset \subset T_i \subset M$ which is both open and closed in the space (M ; T).

Proof : Connected space

By definition, the space (M ; T) is connected if and only if the set M is connected in (M ; T). This is the case if and only if there is no disconnection of M in (M ; T). But by the properties of disconnections, the set M possesses a disconnection in (M ; T) if and only if (M ; T) contains a set T_i with $\emptyset \subset T_i \subset M$ which is both open and closed.

Example 1 : Disconnection of a set

Let (M ; T) be a space with the underlying set $M = \{a, b, c, d, e\}$ and the topology $T = \{\emptyset, \{c\}, \{a, b, c\}, \{c, d, e\}, M\}$. Then the set $A = \{a, d, e\}$ is disconnected. The open sets $T_1 = \{a, b, c\}$ and $T_2 = \{c, d, e\}$ form a disconnection of the subset A.

$$
\begin{aligned}
A \cap T_1 &= \{a\} \\
A \cap T_2 &= \{d, e\} \\
(A \cap T_1) \cup (A \cap T_2) &= A \\
(A \cap T_1) \cap (A \cap T_2) &= \emptyset
\end{aligned}
$$

Example 2 : Separation of sets in the euclidean space $\mathbb{R}^1$

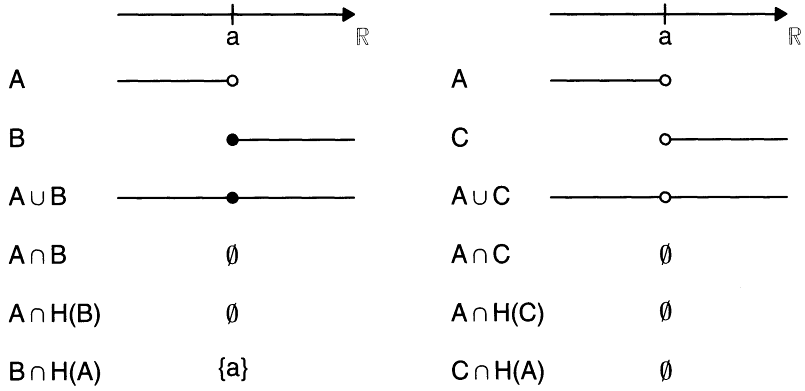
Let the following subsets of the real axis $\mathbb{R}^1$ with the euclidean topology be defined for a rational number $a \in \mathbb{Q}$:

$$
\begin{array}{lll}
A = \{x \in \mathbb{R} \mid x < a\} & R(A) = \{a\} & H(A) = \{x \in \mathbb{R} \mid x \le a\} \\
B = \{x \in \mathbb{R} \mid x \ge a\} & R(B) = \{a\} & H(B) = \{x \in \mathbb{R} \mid x \ge a\} \\
C = \{x \in \mathbb{R} \mid x > a\} & R(C) = \{a\} & H(C) = \{x \in \mathbb{R} \mid x \ge a\}
\end{array}
$$

The sets A and B are not separated, since the intersection of the closure of A with B is not empty. Although the intersection $A \cap B$ is empty, the union $A \cup B$ is the axis $\mathbb{R}^1$ including the point a.

$$H(A) \cap B = \{a\}$$

The sets A and C are separated, since the intersection $H(C) \cap A = \emptyset$ is empty. The union $A \cup C$ does not contain the point a of the real axis $\mathbb{R}^1$. The separation is illustrated in the following diagram.



Example 3 : Discrete disconnected space

The topology of a discrete space (M ; T) contains every subset of M as an open set. For elements a, b,... of M, the one-element sets {a},{b},... are by definition connected. Every set with more than one element is disconnected. For example, {a, b} possesses the disconnection {a}&{b}. This is the origin of the term ”discrete”.

Example 4 : Disconnected space

Let (M ; T) be a space with the underlying set M = {a, b, c, d, e} and the topology T = {∅, {a}, {c, d}, {a, c, d}, {b, c, d, e}, M}. Then M is the union of the disjoint non-empty open sets {a} and { b, c, d, e}, and is therefore disconnected. The set {a} and the set { b, c, d, e} are both open and closed, since their complements are also contained in T. In contrast to the set M, the subset A = { b, d, e} with the relative topology {∅, {d}, A} is connected, since ∅ and A are the only subsets of A which are both open and closed.

5.8.2 CONNECTEDNESS OF CONSTRUCTED SETS

Introduction : The connectedness of a given set in a topological space is studied in Section 5.8.1. The properties of sets which are constructed from given connected sets using operations on sets and mappings are treated in the following.

Properties of connected sets and spaces :

(E1) If a set A in the topological space (M ; T) possesses a disconnection T_1 & T_2, then every connected subset $B \subseteq A$ is either entirely contained in T_1 or entirely contained in T_2. One of the sets $B \cap T_1$ and $B \cap T_2$ is empty.

(E2) The union $A \cup B$ of two connected sets A and B of a topological space (M ; T) may be connected or disconnected.

(E3) The intersection $A \cap B$ of two connected sets A and B of a topological space (M ; T) may be connected or disconnected.

(E4) The union $A \cup B$ of two connected sets A and B of a topological space (M ; T) which are not separated is connected.

(E5) Let a topological space (M ; T) be connected, and let a surjective mapping $f : M \to N$ be continuous. Then the topological space (N ; S) is connected.

(E6) A topological space (M ; T) is connected if and only if every discrete mapping from M is constant.

(E7) In a topological space (M ; T), let $\{A_i\}$ be a finite or infinite family of subsets of M. Let every subset A_i be a connected set. Let none of the intersections $A_k \cap A_m$ of two sets from $\{A_i\}$ be empty. Then the union $\cup A_i$ of the sets from $\{A_i\}$ is a connected set.

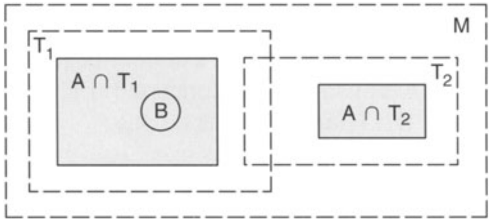
(E8) Let each of the topological spaces (M ; T) and (N ; S) be connected. Then the product space $(M \times N ; P)$ is connected.

(E9) The closure H(A) of a connected set A in a topological space (M ; T) and every intermediate set Z with $A \subseteq Z \subseteq H(A)$ are connected.

Proof E1 : Connected subset in a disconnection

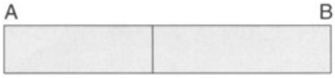
Let a connected set B in a topological space (M ; T) be a subset of a set $A \subseteq M$ with a disconnection $T_1 \,\&\, T_2$. This implies :

$$A \subseteq T_1 \cup T_2 \quad \wedge \quad B \subseteq A \quad \Rightarrow \quad B \subseteq T_1 \cup T_2$$
$$T_1 \cap T_2 \subseteq \overline{A} \quad \wedge \quad B \subseteq A \quad \Rightarrow \quad T_1 \cap T_2 \subseteq \overline{B}$$

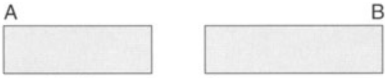


Let the intersections $B \cap T_1$ and $B \cap T_2$ be non-empty. Then T_1 and T_2 form a disconnection $T_1 \,\&\, T_2$ of B. The definition (Z2) of the connectedness of a set implies that B is disconnected. This contradicts the hypothesis that B is connected. Hence either $B \cap T_1$ or $B \cap T_2$ is empty. The set B is either entirely contained in T_1 or entirely contained in T_2.

Proof E2 : The union of two connected sets may be connected or disconnected.



$A \cup B$ connected



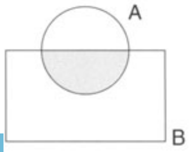
$A \cup B$ disconnected

The statement is proved by examples :

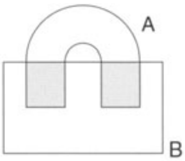
(1) On the real axis $\mathbb{R}$ each of the closed intervals [0,1] and [2,3] is connected in itself. By definition (Z1), the union of these intervals is disconnected, since the intervals are non-empty and separated.

(2) On the real axis $\mathbb{R}$ each of the closed intervals [1,2] and [2,3] is connected in itself. Their union is the connected interval [1,3].

Proof E3 : The intersection of two connected sets may be connected or disconnected.



$A \cap B$ connected



$A \cap B$ disconnected

The statement is proved by examples :

(1) On the real axis $\mathbb{R}$ each of the closed intervals [0,4] and [2,6] is connected in itself. The intersection of these intervals is the connected interval [2,4].

(2) In the real plane $\mathbb{R}^2$ the points ($x = r \cos \lambda$, $y = r \sin \lambda$) with $0.5 \le r \le 1.0$ and $0 \le \lambda \le \pi$ form a connected segment K of a circle. The points (x, y) with $-1.0 \le x \le 1.0$ and $0.0 \le y \le 0.1$ form a connected rectangle A. The intersection $A \cap K$ is disconnected by definition (Z1), since it consists of two point sets, one with $x < 0$ and one with $x > 0$, which are non-empty and separated.

Proof E4 : The union of connected sets which are not separated is connected.

Let each of the sets A and B in a topological space (M ; T) be connected in itself. Let their union $C := A \cup B$ be disconnected. Then by (Z2) C possesses a disconnection $T_1 \& T_2$. For the construction of separated sets it was proved that $C \cap T_1$ and $C \cap T_2$ are separated sets in (M ; T). But by (E1) each of the sets A and B is either entirely contained in T_1 or entirely contained in T_2. For example, let $A \cap T_2 = \emptyset$ and $B \cap T_1 = \emptyset$. This implies :

$$\begin{aligned} C \cap T_1 &= (A \cup B) \cap T_1 = (A \cap T_1) \cup (B \cap T_1) = A \cap T_1 \\ C \cap T_2 &= (A \cup B) \cap T_2 = (A \cap T_2) \cup (B \cap T_2) = B \cap T_2 \end{aligned}$$

Since $C \cap T_1$ and $C \cap T_2$ are separated, $A \cap T_1$ and $B \cap T_2$ are also separated. But by hypothesis in (E4) A and B are not separated. Contrary to the assumption, their union $A \cup B$ is therefore connected.

Proof E5 : The image of a connected space under a continuous mapping is connected.

Let the topological space (M ; T) be connected. Let a mapping $f : M \rightarrow N$ be continuous. Then the mapping f induces a topological space (N ; S) with the underlying set $N = f(M)$ and the topology $S = \{ S_i \mid S_i = f(T_i) \wedge T_i \in T \}$.

Let the space (N ; S) be disconnected. Then there is a set S_1 with $\emptyset \subset S_1 \subset N$ which is both open and closed in the space (N ; S). Since the mapping f is continuous, the space (M ; T) contains a set T_1 with $f(T_1) = S_1$ which is both open and closed in the space (M ; T). From $\emptyset \subset S_1 \subset N$ and $f(M) = N$ it follows that $\emptyset \subset T_1 \subset M$. Hence by (Z3) the space (M ; T) is disconnected. But (E5) contains the hypothesis that the space (M ; T) is connected. Contrary to the assumption, the space (N ; S) is therefore connected.

Proof E6 : A topological space is connected if and only if every discrete mapping from its underlying set is constant.

(1) Let a topological space $(M;T)$ be connected. For an arbitrary discrete mapping $d : M \rightarrow D$, let $y \in D$ be a point for which there is a preimage in M. The discrete space D contains the set $\{y\}$, which is both open and closed. Since the discrete mapping is by definition continuous, the preimage of the set $\{y\}$ in M is also both open and closed. By (Z3), the empty set $\emptyset$ and the underlying set M are the only sets of the connected space $(M;T)$ which are both open and closed. Since the preimage of $\{y\}$ is non-empty by hypothesis, the preimage of $\{y\}$ is M. Hence the mapping d is constant.

(2) Let every discrete mapping $d : M \rightarrow D$ be constant. Let the topological space $(M;T)$ be disconnected. Then by condition (Z3) there is a set A other than $\emptyset$ and M with $\emptyset \subset A \subset M$ which is both open and closed in $(M;T)$. Hence there is a discrete mapping $d : M \rightarrow \{0,1\}$ with $d(A) = \{0\}$ and $d(M-A) = \{1\}$. This contradicts the hypothesis that every discrete mapping d is constant. Contrary to the assumption, the space $(M;T)$ is therefore connected.

Proof E7 : In a family $\{A_i\}$ of subsets of the underlying set of a topological space, let each of the subsets A_i be connected. Let none of the intersections of two sets from $\{A_i\}$ be empty. Then the union of the sets from $\{A_i\}$ is connected.

Let each of the subsets $A_i \subseteq M$ in a topological space $(M;T)$ be connected. Let the union of a finite or infinite family $\{A_i\}$ of such sets be $\cup A_i$. Let the mapping $d : \cup A_i \rightarrow D$ be discrete. For arbitrary points $x,y \in \cup A_i$, let $x \in A_k$ and $y \in A_m$. By (E6), the discrete mappings from the connected sets A_k and A_m are constant :

$$d(A_k) = \{c_k\} \quad \wedge \quad d(A_m) = \{c_m\}$$

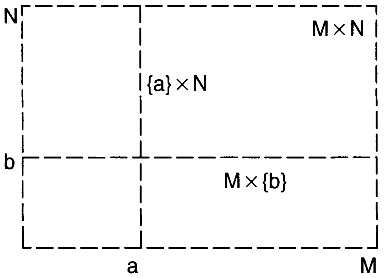
The intersection $A_k \cap A_m$ is non-empty by hypothesis. For a point $z \in A_k \cap A_m$ in the intersection of the sets, $d(z) = c_k \ \wedge \ d(z) = c_m$. Hence :

$$c_k = c_m = c \quad \wedge \quad d(x) = d(y) = d(z) = c$$

Since the points x and y are chosen arbitrarily, the mapping $d : \cup A_i \rightarrow D$ is constant with $d(\cup A_i) = \{c\}$. Hence the union $\cup A_i$ is connected by virtue of (E6).

Proof E8 : The product space of two connected spaces is connected.

Let each of the spaces $(M;T)$ and $(N;S)$ be connected. For arbitrary points $a \in M$ and $b \in N$, consider the subspaces $\{a\} \times N$ and $M \times \{b\}$ of the product space $M \times N$:



The open sets in $M \times \{b\}$ are of the form $T \times \{b\}$, where T is open in M. Since the space $(M;T)$ is connected, by (Z3) only the sets $\emptyset$ and M are both open and closed in M. Thus only the sets $\emptyset$ and $M \times \{b\}$ are both open and closed in $M \times \{b\}$. It follows by (Z3) that the space $M \times \{b\}$ is connected. The space $\{a\} \times N$ is likewise connected. The intersection of $\{a\} \times N$ and $M \times \{b\}$ is the point (a,b). It follows by (E7) that the union P_{ab} of the sets $\{a\} \times N$ and $M \times \{b\}$ is connected.

The product space $M \times N$ is the union of sets P_{xy} for different points (x,y). Each pair of sets P_{ab} and P_{cd} has a non-empty intersection $\{(a,d),(b,c)\}$. Since every set in the family $\{P_{xy}\}$ is connected, the pairwise intersection of the sets is non-empty and their union is the underlying set $M \times N$, it follows by (E7) that $M \times N$ is connected.

Proof E9 : The closure of a connected set A and every intermediate set Z with $A \subseteq Z \subseteq H(A)$ are connected.

Let the intermediate set Z be disconnected. Then by (Z2) there is a disconnection T_1 & T_2 of Z. By (E1), the connected subset A of Z is either entirely contained in T_1 or entirely contained in T_2. Let A be contained in T_1.

The sets $Z \cap T_1$ and $Z \cap T_2$ are separated. From $A \subseteq T_1$ and $A \subseteq Z$ it follows that $A \subseteq Z \cap T_1$. Hence the sets A and $Z \cap T_2$ are separated. Since $Z \subseteq H(A)$, the definition of separated sets yields :

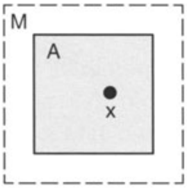
$$H(A) \cap (Z \cap T_2) = \emptyset \quad \Rightarrow \quad Z \cap T_2 = \emptyset$$

By the definition of the disconnection T_1 & T_2 of Z, however, the set $Z \cap T_2$ is non-empty. Contrary to the assumption, the intermediate set Z is therefore connected.

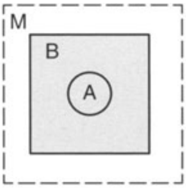
5.8.3 COMPONENTS AND PATHS

Introduction : Spaces are generally not connected. A disconnected space may however be partitioned into connected components. The concept of a path is useful in determining the components of a space. Every path-connected set is connected. The concepts of component, path and path-connectedness are defined in the following.

Component : A connected subset A of the underlying set M is called a component (connected component) of a space (M ; T) if there is no connected subset B of M which contains A as a proper subset.



component A



A is not a component

Properties of components :

(K1) Every component of a space (M ; T) is a closed set.

(K2) The union of all connected subsets of M which contain a common point x of M is a component of (M ; T).

(K3) The components of a space (M ; T) form a partition of M.

(K4) If a topological space (M ; T) has a finite number of components, then every component is both open and closed.

(K5) Every connected subset of M is entirely contained in a component of (M ; T).

(K6) Every non-empty connected subset of a space (M ; T) which is both open and closed is a component of the space.

Proof : Properties of components

(K1) The closure H(A) of a component A of (M ; T) contains the component A and is connected by virtue of property (E9). By the definition of a component, however, there is no connected subset of M which contains A as a proper subset. Hence A = H(A), and the component A is closed.

(K2) Let $\{A_i\}$ be the family of all connected subsets $A_i \subseteq M$ which contain a point x of the topological space $(M;T)$. Then the union $\cup A_i$ of these sets is connected by virtue of (E7). Every connected set B which contains $\cup A_i$ also contains the point x. But every connected set which contains the point x is by hypothesis a subset of $\cup A_i$. Hence $B \subseteq \cup A_i$ and $\cup A_i \subseteq B$, and therefore $B = \cup A_i$. Since there is thus no connected subset B of M which contains $\cup A_i$ as a proper subset, $\cup A_i$ is a component of $(M;T)$.

(K3) Let $\{C_i\}$ be the set of the components of a space $(M;T)$ which may be formed with all points of M by (K2). Then the underlying set $M = \cup C_i$ is the union of these components. Hence the components form a partition of M if they are disjoint. Assume that the components C_i and C_m are not disjoint. Then they have a common point a in $C_i \cap C_m$. Let the component formed with the point a be C_a. Each of the connected sets C_i and C_m contains the point a and is therefore a subset of C_a. However, the sets C_i and C_m cannot be proper subsets of C_a since they are components. It follows that $C_a = C_i = C_m$. Contrary to the assumption, the components are therefore disjoint.

(K4) Let the components of the space $(M;T)$ be $C_1, ..., C_m$. By (K1), each of the components is a closed set. By (K3), the underlying set M is the union $\cup C_i$ of the disjoint components. The complement of an arbitrary component C_k is therefore the union of all components except for C_k:

$$\bar{C}_k = \bigcup_{i \neq k} C_i$$

The union of a finite number of closed sets is a closed set. The complement $\bar{C}_k$ is therefore a closed set; hence the component C_k is open. By (K1), C_k is also closed. Since C_k is an arbitrary component, each of the components $C_1, ..., C_m$ is both open and closed.

(K5) Let a set $A \subseteq M$ in a topological space $(M;T)$ be connected. An arbitrary point $x \in A$ is contained in exactly one component C_i of the space, since by (K3) the components form a partition of the space. The sets A and C_i have the point x in common. By (K2), the set A is a subset of the component C_i.

(K6) Let a subset A of a topological space $(M;T)$ be non-empty, connected and both open and closed. By (K5), the connected set A is entirely contained in a component C of the space. The component C is connected. By hypothesis the set A is both open and closed. By (Z3) it follows that $A = C$; hence A is a component of the space.

Totally disconnected space : A topological space $(M;T)$ is said to be totally disconnected if for every two points $x,y \in M$ with $x \neq y$ there is a disconnection T_1 & T_2 of M such that $x \in T_1$ and $y \in T_2$. The components of a totally disconnected space are the one-point subsets of M.

Proof : Components of a totally disconnected space

Let two points $x \neq y$ be contained in the same component C of a topological space (M ; T). Since the space M is by hypothesis totally disconnected, there is a disconnection T_1 & T_2 of M such that $x \in T_1$ and $y \in T_2$. Hence the sets $C \cap T_1$ and $C \cap T_2$ are non-empty. The disconnection T_1 & T_2 of M is also a disconnection of C :

$$\begin{array}{lll} (M \cap T_1) \cup (M \cap T_2) = M & \Rightarrow & C \cap ((M \cap T_1) \cup (M \cap T_2)) = C \cap M \\ & \Rightarrow & (C \cap T_1) \cup (C \cap T_2) = C \\ (M \cap T_1) \cap (M \cap T_2) = \emptyset & \Rightarrow & C \cap ((M \cap T_1) \cap (M \cap T_2)) = C \cap \emptyset \\ & \Rightarrow & (C \cap T_1) \cap (C \cap T_2) = \emptyset \end{array}$$

By (Z2), however, there exists no disconnection for the connected component C. Contrary to the assumption, the component C therefore contains exactly one point.

Locally connected space : A topological space (M ; T) is said to be connected at a point x of the space if every neighborhood of x contains a connected open set. The topological space (M ; T) is said to be locally connected if M is connected at every point $x \in M$. Every component of a locally connected space is open.

Proof : Open components of a locally connected space

By the definition of local connectedness, an arbitrary point x in a component C of a locally connected space (M ; T) belongs to at least one connected open set T_x of the space. Since the connected sets C and T_x have the point x in common, it follows from (K5) that $x \in T_x \subseteq C$. For every point $x \in C$ there is a connected open neighborhood T_x. The union $\cup T_x$ of these sets yields the component C. Since C is a union of open sets, C is itself open.

$$C = \bigcup_x (T_x \mid x \in C)$$

Path : Let $I = [0, 1]$ be the closed unit interval in $\mathbb{R}$, and let a, b be points of a topological space (M ; T). A continuous mapping $f : I \to M$ with $f(0) = a$ and $f(1) = b$ is called a path from a to b in M. The points a and b are said to be connectable in M.

$$f : I \to M \quad \text{with} \quad f(0) = a \quad \wedge \quad f(1) = b$$

a origin of the path
b endpoint of the path

Connectability of points : Two points a and b in a subset A of a topological space (M ; T) are said to be connectable in A if there is a path f from a to b whose image is entirely contained in A.

$$f : I \to M \quad \text{with} \quad f(0) = a \quad \wedge \quad f(1) = b \quad \wedge \quad f(I) \subseteq A$$

The connectability relation is an equivalence relation :

(1) The relation is reflexive. Every point a is connectable to itself via the constant path $f : I \to M$ with $f(x) = a$.

(2) The relation is symmetric. If a is connectable to b via the path f, then b is connectable to a via the path $g : I \to M$ with $g(x) = f(1-x)$.

(3) The relation is transitive. If a is connectable to b via f and b is connectable to c via g, then a is connectable to c via the following path :

$$h(x) = \begin{cases} f(2x) & 0 \le x < 1/2 \\ g(2x-1) & 1/2 \le x \le 1 \end{cases}$$

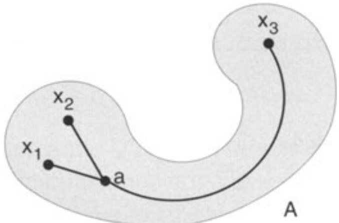
Path component : The equivalence classes with respect to the connectability relation are called the path components of the topological space (M ; T). According to Section 2.4, the path components form a partition of the underlying set M.

Path-connected set : A subset A of a topological space is said to be path-connected if every pair of points $a, b \in A$ is connectable in A.

$$\bigwedge_{a \in A} \bigwedge_{b \in A} \bigvee_{f : I \to M} (f(0) = a \quad \wedge \quad f(1) = b \quad \wedge \quad f(I) \subseteq A)$$

Every path-connected set is connected. Hence path-connected sets possess the properties of connected sets. There are, however, connected sets which are not path-connected.

Proof : Every path-connected set is connected



If the set A is empty, then it is connected. If A is non-empty and path-connected, then there is a point a in A which is connectable in A with every point x_m of A.

(1) The unit interval I is connected. Since the mapping $f_m : I \to A$ with $f_m(0) = a$ and $f_m(1) = x_m$ is continuous, it follows from (E5) that the image $f_m(I)$ is connected.

(2) Since the images $f_m(I)$ for different points x_m have the point a in common, their union is connected by virtue of (E7).

(3) Since every point x_m of A is the endpoint of a path f_m in A, the union of the images $f_m(I)$ is A.

Example 1 : Components of a set

Let $(M;T)$ be a space with the underlying set $M = \{a, b, c, d, e\}$ and the topology $T = \{\emptyset, \{b\}, \{c, d\}, \{b, c, d\}, \{a, c, d, e\}, M\}$.

(1) M is the union of the disjoint open sets $\{b\}$ and $\{a, c, d, e\}$. Hence M is disconnected.

(2) The open set $\{b, c, d\}$ is the union of the disjoint open sets $\{b\}$ and $\{c, d\}$. Hence $\{b, c, d\}$ is disconnected.

(3) Each of the sets $\emptyset$, $\{b\}$, $\{c, d\}$, and $\{a, c, d, e\}$ is connected, since they are the only sets which are both open and closed sets in their respective subspaces :

underlying set : $\emptyset$	topology : $\{\emptyset\}$
underlying set : $\{b\}$	topology : $\{\emptyset, \{b\}\}$
underlying set : $\{c, d\}$	topology : $\{\emptyset, \{c, d\}\}$
underlying set : $\{a, c, d, e\}$	topology : $\{\emptyset, \{c, d\}, \{a, c, d, e\}\}$

The set $\{c, d\}$, however, is not a component of the space, since it is contained in the connected set $\{a, c, d, e\}$.

(4) The sets $\{b\}$ and $\{a, c, d, e\}$ are the components of the space $(M;T)$.

(5) The components $\{b\}$ and $\{a, c, d, e\}$ form a partition of M.

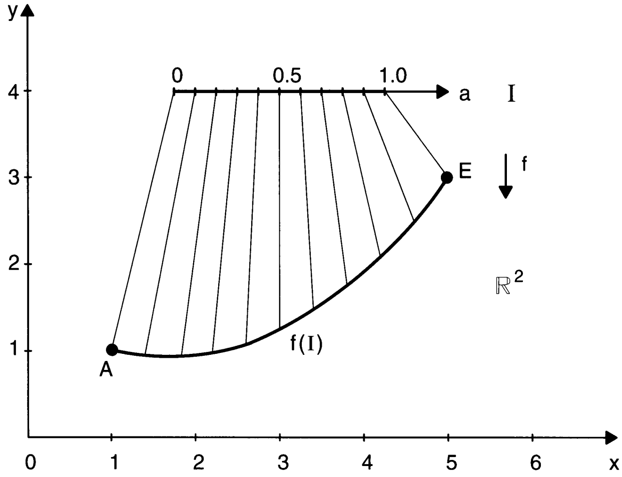
Example 2 : Totally disconnected space

The set $\mathbb{Q}$ of the rational numbers equipped with the relative topology of the natural topology on $\mathbb{R}^1$ is a totally disconnected space. For two arbitrary numbers $a, b \in \mathbb{Q}$ with $a < b$ there is an irrational number x with $a < x < b$. The open sets $T_1 := \{p \in \mathbb{Q} \mid p < x\}$ and $T_2 := \{p \in \mathbb{Q} \mid p > x\}$ with $a \in T_1$ and $b \in T_2$ form a disconnection $T_1 \,\&\, T_2$ of $\mathbb{Q}$. Hence the space $\mathbb{Q}$ is totally disconnected.

Example 3 : Locally connected space

Let $(M;T)$ be a discrete space. Then every one-element set $\{x\}$ is open and connected. Thus every neighborhood of the point x contains the connected open set $\{x\}$, and hence the discrete space $(M;T)$ is locally connected. For every two points $x \neq y$ in M there is a disconnection $\{x\} \,\&\, \{y\}$. Hence the space $(M;T)$ is totally disconnected. Thus the totally disconnected discrete space is locally connected !

Example 4 : Definition of a path $f : I \rightarrow \mathbb{R}^2$



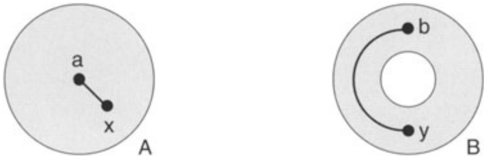
The illustrated path in the euclidean plane $\mathbb{R}^2$ is the following mapping of the closed unit interval I :

$$f : I \rightarrow \mathbb{R}^2 \quad \text{with} \quad f(a) = (x, y) = (4a + 1, 3a^2 - a + 1)$$

A origin of the path f

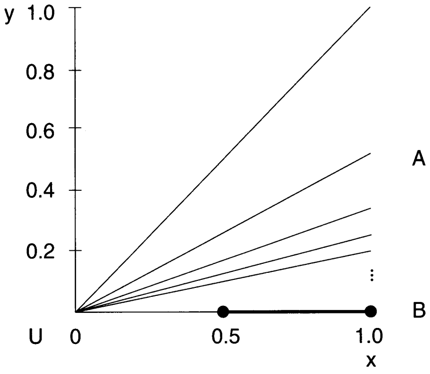
E endpoint of the path f

Example 5 : Path-connected sets



The two illustrated sets are path-connected in the euclidean plane $\mathbb{R}^2$. Every point x in A is connectable with the fixed point a in A. Every point y in B is connectable with the fixed point b in B.

Example 6 : A connected set which is not path-connected



$$A = \{(x, y) \in \mathbb{R}^2 \mid 0 \le x \le 1 \quad \wedge \quad y = \frac{x}{n} \quad \wedge \quad n \in \mathbb{N}'\}$$

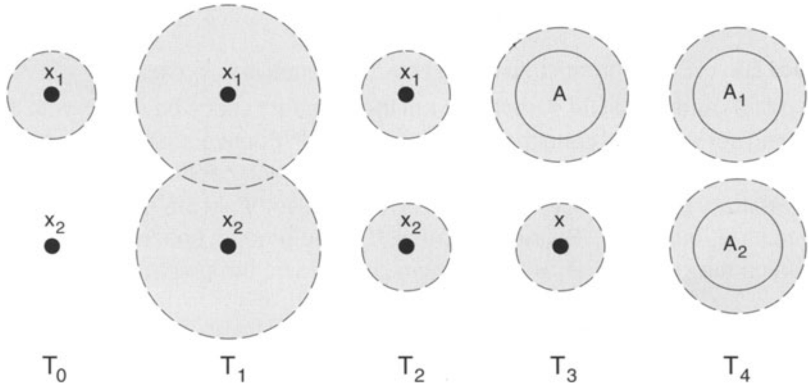
$$B = \{(x, 0) \in \mathbb{R}^2 \mid 0.5 \le x \le 1\}$$

Let A be the set of points on the segments joining the origin U of the plane $\mathbb{R}^2$ with the points $(1, \frac{1}{n})$. Let B be the closed interval [0.5, 1.0] on the x-axis. The set A is path-connected, since two arbitrary points $X, Y \in A$ can be connected by a path in A along the segments $\overline{XU}$ and $\overline{UY}$. The set B is path-connected. The set $A \cup B$ is connected, since every $(x, 0) \in B$ is an accumulation point of A. However, the set $A \cup B$ is not path-connected. There is no path with origin in A and endpoint in B whose image lies entirely in $A \cup B$.

5.9 SEPARATION PROPERTIES

Introduction : The definition of a topological space (M ; T) in Section 5.2 contains only a few general conditions for the open sets T_i of the topology T. Often the points and the open sets of a topological space satisfy additional conditions. These conditions are defined as separation axioms. The separation axiom determines the type of the topological space and some of its properties. In particular, the convergence of nets and sequences (see Section 5.10) in a topological space depends on the separation axiom which holds in the space. The separation axioms and the spaces associated with them are defined in the following.

Separation axioms : A relationship between the points x_i, the subsets A_i and the open sets T_i of a topological space (M ; T) is called a separation axiom. The following separation axioms are considered :



(T_0) A topological space is called a T_0-space if for any two points $x_1 \neq x_2$ of the space there is an open set which contains one of the two points but not the other.

(T_1) A topological space is called a T_1-space if for any two points $x_1 \neq x_2$ of the space there is an open set which contains x_1 but not x_2 and an open set which contains x_2 but not x_1.

(T_2) A topological space is called a T_2-space (Hausdorff space) if for any two points $x_1 \neq x_2$ of the space there are disjoint open sets S_1 and S_2 such that $x_1 \in S_1$ and $x_2 \in S_2$.

(T_3) A T_1-space is called a T_3-space (regular T_1-space) if for every point x and every closed set A with $x \notin A$ there are disjoint open sets S_1 and S_2 such that $x \in S_1$ and $A \subseteq S_2$.

(T_4) A T_1-space is called a T_4-space (normal T_1-space) if for any two disjoint closed sets A_1 and A_2 there are disjoint open sets S_1 and S_2 such that $A_1 \subseteq S_1$ and $A_2 \subseteq S_2$.

The separation axioms for the space types T_0 to T_4 are designed such that the separation axiom for the space type T_i implies the separation axiom for the preceding space type T_{i-1}. Hence a space of type T_i has all properties of spaces of type T_{i-1}. The essential properties of the space types are treated in the following.

T_0-space : A T_0-space has the following properties :

(B1) Every subspace of a T_0-space is a T_0-space.

(B2) Every product space of two T_0-spaces is a T_0-space.

Proof B1 : Every subspace of a T_0-space is a T_0-space.

Let $(M ; S)$ be a T_0-space, and let $(N ; V)$ be a subspace with $N \subseteq M$ and the relative topology V. For arbitrary points $x_1 \neq x_2$ of N there is an open set S_1 in the space $(M ; S)$ which contains x_1 but not x_2. Thus there is an open set $V_1 = N \cap S_1$ in the subspace $(N ; V)$ which contains x_1 but not x_2. Hence $(N ; V)$ is a T_0-space.

Proof B2 : Every product space of two T_0-spaces is a T_0-space.

Let $(M ; S)$ and $(N ; V)$ be T_0-spaces. Let their product space be $(W; P)$ with the underlying set $W = M \times N$ and the product topology P. For two arbitrary points $x_1 \neq x_2$ in M there is an open set S_1 in the space $(M ; S)$ which contains x_1 but not x_2. For two arbitrary points $y_1 \neq y_2$ in N there is an open set V_1 in the space $(N ; V)$ which contains y_1 but not y_2. By the construction of the product space, $(W; P)$ therefore contains the open set $P_1 = S_1 \times V_1$, which contains the point (x_1, y_1) but not the point (x_2, y_2). Hence $(W; P)$ is a T_0-space.

T_1-space : A T_1-space has the following properties :

(E1) Every T_1-space is a T_0-space.

(E2) A topological space is a T_1-space if and only if every one-element set in the space is closed.

(E3) Every subspace of a T_1-space is a T_1-space.

(E4) Every product space of two T_1-spaces is a T_1-space.

Proof E1 : Every T_1-space is a T_0-space.

Let $(M ; S)$ be a T_1-space. Then for two arbitrary points $x_1 \neq x_2$ there is an open set T_1 in the space $(M ; S)$ which contains x_1 but not x_2. Hence $(M ; S)$ is a T_0-space.

Proof E2 : One-element sets in T_1-spaces

(1) Let $(M;T)$ be a T_1-space. For a fixed point x and an arbitrary point $y \neq x$ of the underlying set M there is by definition an open set S_y which contains y but not x. The complement of the one-element set $\{x\}$ is therefore the union $\cup S_y$ of these open sets, and hence itself an open set. Since the complement $M-\{x\} = \cup S_y$ is open, every one-element set $\{x\}$ in a T_1-space is closed.

(2) Let a one-element set $\{x\}$ in a topological space $(M;T)$ be closed. Then the complement $M-\{x\}$ is open. The open set $M-\{x\}$ contains every point $y \neq x$ in M. This result holds for every choice of the point $x \in M$. Hence $(M;T)$ is a T_1-space.

Proof E3 : Every subspace of a T_1-space is a T_1-space.

Let $(M;S)$ be a T_1-space, and let $(N;V)$ be a subspace with $N \subseteq M$ and the relative topology V. For every point $x \in N$, $\{x\}$ is a closed set in M by (E2). By property (U5) of subspaces, $\{x\}$ is also a closed set in N. It follows from (E2) that $(N;V)$ is a T_1-space.

Proof E4 : Every product space of two T_1-spaces is a T_1-space.

Let $(M;S)$ and $(N;V)$ be T_1-spaces. Let their product space be $(W;P)$ with $W = M \times N$ and the product topology P. For points $x_1 \neq x_2$ in M there are open sets $S_1, S_2 \in S$ in the T_1-space $(M;S)$ with $x_1 \in S_1$ and $x_2 \in S_2$ but $x_1 \notin S_2$ and $x_2 \notin S_1$. Likewise, for points $y_1 \neq y_2$ in N there are open sets $V_1, V_2 \in V$ such that $y_1 \in V_1$ and $y_2 \in V_2$ but $y_1 \notin V_2$ and $y_2 \notin V_1$. For the points $(x_1, y_1) \neq (x_2, y_2)$ in W this yields the open sets $P_1 = S_1 \times V_1$ and $P_2 = S_2 \times V_2$ with $(x_1, y_1) \in P_1$ and $(x_2, y_2) \in P_2$ but $(x_1, y_1) \notin P_2$ and $(x_2, y_2) \notin P_1$. Hence $(W;P)$ is a T_1-space.

Hausdorff space : A Hausdorff space is defined such that every limit in the space is unique (see Section 5.10). Hausdorff spaces have the following properties :

(H1) Every Hausdorff space (T_2-space) is a T_1-space.

(H2) A topological space $(M;S)$ is a Hausdorff space if and only if the diagonal $D := \{(x,x) \mid x \in M\}$ is a closed set in the product space $M \times M$.

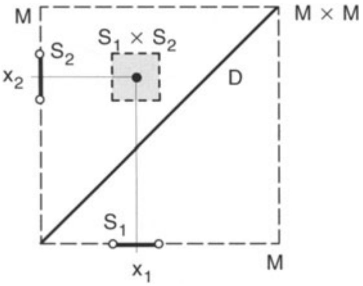
(H3) Every subspace of a T_2-space is a T_2-space.

(H4) Every product space of two T_2-spaces is a T_2-space.

Proof H1 : Every Hausdorff space is a T_1-space.

For two arbitrary points $x_1 \neq x_2$ of a Hausdorff space there are open sets S_1 and S_2 with $x_1 \in S_1 \wedge x_2 \in S_2 \wedge S_1 \cap S_2 = \emptyset$. Since the open set S_1 contains the point x_1 but not x_2 and the open set S_2 contains the point x_2 but not x_1, the Hausdorff space is a T_1-space.

Proof H2 : Closed diagonal in the product space of a Hausdorff space



(1) Let the diagonal $D := \{(x,x) \mid x \in M\}$ be closed. Then every point (x_1,x_2) in $M \times M$ with $x_1 \neq x_2$ lies in the open set $M \times M - D$. The products $S_k \times S_m$ of all open sets S_i of M form a basis of the product topology on $M \times M$; hence in particular there is such a basis element with $(x_1,x_2) \in S_1 \times S_2 \subseteq M \times M - D$. This implies $x_1 \in S_1$ and $x_2 \in S_2$. Since $S_1 \times S_2$ contains no point of D, S_1 contains no point of S_2, so that $S_1 \cap S_2 = \emptyset$. Hence the space is Hausdorff.

(2) Let the space (M ; T) be Hausdorff. By the definition of a Hausdorff space, for arbitrary points $x_1, x_2 \in M$ with $x_1 \neq x_2$ there are open sets S_1 and S_2 with $x_1 \in S_1$, $x_2 \in S_2$ and $S_1 \cap S_2 = \emptyset$. For an arbitrary point $(x_1,x_2) \in M \times M - D$ there is therefore an open set $S_1 \times S_2$ of the product topology with $(x_1,x_2) \in S_1 \times S_2$ and $S_1 \times S_2 \subseteq M \times M - D$. Thus every point of $M \times M - D$ is an inner point, so that the complement $M \times M - D$ of D is open. Hence the diagonal D is closed.

Proof H3 : Every subspace of a T_2-space is a T_2-space.

Let (M ; S) be a T_2-space, and let (N ; V) be a subspace with $N \subseteq M$ and the relative topology V. For points $x_1 \neq x_2$ in N there are disjoint open sets $S_1, S_2 \in S$ in the T_2-space (M ; S) with $x_1 \in S_1$ and $x_2 \in S_2$. The sets $V_1 = N \cap S_1$ and $V_2 = N \cap S_2$ are open sets of the subspace N. Hence :

$$\begin{array}{lll} x_1 \in S_1 \ \wedge \ x_1 \in N & \Rightarrow & x_1 \in N \cap S_1 = V_1 \\ x_2 \in S_2 \ \wedge \ x_2 \in N & \Rightarrow & x_2 \in N \cap S_2 = V_2 \\ S_1 \cap S_2 = \emptyset & \Rightarrow & V_1 \cap V_2 = N \cap (S_1 \cap S_2) = \emptyset \end{array}$$

From $x_1 \in V_1$, $x_2 \in V_2$ and $V_1 \cap V_2 = \emptyset$ it follows that (N ; V) is a T_2-space.

Proof H4 : Every product space of two T_2-spaces is a T_2-space.

Let (M ; S) and (N ; V) be T_2-spaces. Let their product space be (W ; P) with $W = M \times N$ and the product topology P. For points $x_1 \neq x_2$ in M there are disjoint open sets $S_1, S_2 \in S$ in the T_2-space (M ; S) with $x_1 \in S_1$ and $x_2 \in S_2$. Likewise, for points $y_1 \neq y_2$ in N there are disjoint open sets $V_1, V_2 \in V$ such that $y_1 \in V_1$ and $y_2 \in V_2$. For the points $(x_1,y_1) \neq (x_2,y_2)$ in W, this yields the disjoint open sets $P_1 = S_1 \times V_1$ and $P_2 = S_2 \times V_2$ with $(x_1,y_1) \in P_1$ and $(x_2,y_2) \in P_2$ but $(x_1,y_1) \notin P_2$ and $(x_2,y_2) \notin P_1$. Hence (W ; P) is a T_2-space.

Regular space : A topological space (M ; T) is said to be regular if for every point $x \in M$ and every closed set $A \subset M$ with $x \notin A$ there are disjoint open sets S_1 and S_2 such that $x \in S_1$ and $A \subseteq S_2$. A regular space is not necessarily a T_1-space. For example, the set $M = \{a,b,c\}$ with the topology $\{\emptyset, \{a\}, \{b,c\}, M\}$ is a regular space. However, this regular space is not a T_1-space since, for instance, the one-element set $\{c\}$ is not closed.

A regular T_1-space is called a T_3-space and has the following properties :

(R1) Every T_3-space is a Hausdorff space.

(R2) A T_1-space is regular if and only if the closed neighborhoods of every point form a neighborhood basis for that point.

(R3) Every subspace of a T_3-space is a T_3-space.

(R4) Every product space of two T_3-spaces is a T_3-space.

Proof R1 : Every T_3-space is a Hausdorff space.

Let (M ; S) be a T_3-space. Since the T_1-axiom holds in the T_3-space, for arbitrary different points $x_1, x_2 \in M$ there is a closed set $\{x_1\}$ which does not contain x_2. Since the T_3-space is regular, there are disjoint open sets S_1 and S_2 with $\{x_1\} \subseteq S_1$ and $x_2 \in S_2$. Thus the points $x_1 \neq x_2$ lie in disjoint open sets. Hence the T_3-space is a Hausdorff space.

Proof R2 : A T_1-space is a T_3-space if and only if the closed neighborhoods of every point form a neighborhood basis for that point.

(1) Let the space (M ; S) be a T_3-space. Since the T_1-axiom holds in the T_3-space, for every point $x \in M$ there is an open neighborhood U of x. The complement $A = M - U$ is a closed set. Since the T_3-space is regular, there are open sets S_1 and S_2 with $x \in S_1$, $A \subset S_2$ and $S_1 \cap S_2 = \emptyset$. The complement $M - S_2$ of the open set S_2 is closed. Every neighborhood of x contains an open set U. Since $M - S_2 \subset M - A = U$, every neighborhood of x also contains a closed neighborhood $M - S_2$ of x. Hence the closed neighborhoods form a neighborhood basis for x.

(2) Let a neighborhood basis of closed sets be given for every point x in a topological space (M ; T). Let an arbitrary point $x \in M$ not be contained in an arbitrary closed set $A \subset M$, that is $x \in M - A$. Since the set $M - A$ is open, by hypothesis the point x has a closed neighborhood $U \subset M - A$ which by definition contains an open set S_1 with $x \in S_1$ and $S_1 \subset U$. The open set $S_2 = M - U$ contains A and is disjoint from U, and thus also disjoint from S_1. From $x \in S_1$, $A \subseteq S_2$ and $S_1 \cap S_2 = \emptyset$, it follows that the space is a T_3-space.

Proof R3 : Every subspace of a T_3-space is a T_3-space.

Let (M ; S) be a T_3-space, and let (N ; V) be a subspace with $N \subseteq M$ and the relative topology V. Since M is a T_1-space, it follows from (E3) that N is also a T_1-space. By (R2), every point $x \in N$ has a closed neighborhood basis in M. By definition, a closed neighborhood U of x contains an open set S_1 in (M ; S) with $x \in S_1$. By property (U5) in Section 5.7.2, the intersection $N \cap U$ is closed. The intersection $V_1 = N \cap S_1$ is by definition an open set of the subspace (N; V). The closed neighborhood $N \cap U$ of x therefore contains the open set V_1 in N with $x \in V_1$. Hence the intersections of the elements of the neighborhood basis of x with N again form a closed neighborhood basis of x. It follows from (R2) that the space N is regular. Since N is a regular T_1-space, N is a T_3-space.

Proof R4 : Every product space of two T_3-spaces is a T_3-space.

Let (M ; S) and (N; V) be T_3-spaces. Let their product space be (W ; P) with $W = M \times N$ and the product topology P. For a point x and a closed set A there are disjoint open sets $S_1, S_2 \in S$ in the T_3-space M such that $x \in S_1$ and $A \subseteq S_2$. Likewise, for a point y and a closed set B there are disjoint open sets $V_1, V_2 \in V$ in the T_3-space N such that $y \in V_1$ and $B \subseteq V_2$. Hence for the point (x, y) and the closed set $A \times B$ the product space W possesses the disjoint open sets $P_1 = S_1 \times V_1$ and $P_2 = S_2 \times V_2$ with $(x, y) \in P_1$ and $A \times B \subseteq P_2$.

Let Q be an arbitrary closed set in the product space W which does not contain the point (x, y). Then the complement $\overline{Q}$ is an open set which contains (x, y) and is a union of basis elements B_i of the product topology P. At least one basis element $B_k := C \times D$ contains (x, y).

$$Q \subseteq M \times N \; - \; C \times D$$

The sets $\overline{C} \times N$ and $M \times \overline{D}$ are cartesian products of closed sets which do not contain (x, y). Thus by the above there are disjoint open sets T_C , S_C with $(x, y) \in T_C$ and $\overline{C} \times N \subseteq S_C$ as well as disjoint open sets T_D , S_D with $(x, y) \in T_D$ and $M \times \overline{D} \subseteq S_D$. The set $T_C \cap T_D$ is open and contains (x, y). The set $S_C \cap S_D$ is open and disjoint from $T_C \cap T_D$. It contains Q. Hence (W ; P) is a T_3-space.

$$Q \subseteq S_C \cup S_D \;=\; (M \times N \; - \; C \times N) \cup (M \times N \; - \; M \times D) \;=\; M \times N \; - \; C \times D$$

Normal space : A topological space is said to be normal if for every two disjoint closed sets A_1 and A_2 there are disjoint open sets S_1 and S_2 such that $A_1 \subseteq S_1$ and $A_2 \subseteq S_2$. A normal space is not necessarily a T_1-space. For example, the set $M = \{a, b, c\}$ with the topology $S = \{\emptyset, \{a\}, \{b\}, \{a,b\}, M\}$ is a normal space, but by (E2) it is not a T_1-space, since the one-element set {a} is not closed :

closed sets	:	$\emptyset$, {c}, {a,c}, {b,c}, {a,b,c}
disjoint closed sets	:	$A_1 = \emptyset$ and $A_2 \in \{\{c\}, \{a,c\}, \{b,c\}, \{a,b,c\}\}$
open sets	:	$\emptyset$, {a}, {b}, {a,b}, {a,b,c}
normality of (M;S)	:	$A_1 \subseteq \emptyset$ and $A_2 \subseteq \{a,b,c\}$

The normal space shown in the example is also not regular. The only open superset of the closed set {c} is {a, b, c} ; the point a, which is not contained in {c}, does not have an open neighborhood disjoint from {a, b, c}.

A normal T_1-space is called a T_4-space and has the following properties :

(N1) Every T_4-space is a T_3-space.

(N2) Every metric space is a T_4-space.

(N3) Subspaces of metric spaces are normal.

(N4) Product spaces of metric spaces are normal.

Proof N1 : Every T_4-space is a T_3-space.

In a T_4-space, let A be a closed set and let x be a point not contained in A. Since the T_1-axiom holds in the T_4-space, there is a closed set $\{x\}$ disjoint from A. Since the T_4-space is normal, there are disjoint open sets S_1 and S_2 with $\{x\} \subseteq S_1$, that is $x \in S_1$, and $A \subseteq S_2$. Hence the T_4-space is regular. Thus the T_4-space is a regular T_1-space, and hence a T_3-space.

Proof N2 : Every metric space is a T_4-space.

Let a metric space $(M ; d)$ be given. Let A and B be disjoint closed sets in M. First it is proved that for a fixed point $x \in A$ there is a distance $a_x > 0$ such that $d(x, y) > a_x$ for all $y \in B$. Assume this assertion to be false. Then every ε-ball around x contains a point of B. Hence x is a point in the closure of B, and therefore a point in the closed set B. The result $x \in B$ contradicts the hypothesis $A \cap B = \emptyset$. Hence the assertion is true.

The open set $D(x, 0.5\, a_x)$ does not intersect B. Thus the union $R = \bigcup_{x \in A} D(x, 0.5\, a_x)$ is an open set which contains A and does not intersect B. Analogously, an open set S is constructed which contains B and does not intersect A. Assume that the sets R and S are not disjoint. Then there are points $x \in A$ and $y \in B$ such that $D(x, 0.5\, a_x) \cap D(y, 0.5\, a_y) \neq \emptyset$, that is $0.5(a_x + a_y) > d(x, y)$. Without loss of generality, assume $a_x \geq a_y$, so that $a_x \geq 0.5(a_x + a_y) > d(x, y)$. The contradiction with $d(x, y) > a_x$ shows that contrary to the assumption the open sets R and S are disjoint. From $A \subset R$, $B \subset S$ and $R \cap S \neq \emptyset$, it follows that the metric space $(M ; d)$ is a T_4-space.

Proof N3 : Subspaces of metric spaces are normal.

Let $(M ; d)$ be a metric space, and let $(N ; d)$ be a subspace with $N \subseteq M$. Then N is also metric, and thus normal by (N2).

Proof N4 : Product spaces of metric spaces are normal.

Let $(M ; d)$ and $(N ; s)$ be metric spaces. Let their product space be $(W ; t)$ with $W = M \times N$. For points $x_i \in M$, $y_i \in N$ and $(x_i, y_i) \in W$, let

$$t^2\,((x_1, y_1), (x_2, y_2)) \;=\; d^2(x_1, x_2) + s^2(y_1, y_2)$$

Since d and s are metrics, the positive square root t also has the properties of a metric. The ε-balls of the metric t form the basis of a metric topology. Since the conditions for the equivalence of bases derived in Section 5.3 are satisfied, the topology induced by t coincides with the product topology of the space W. Hence the space $(W ; t)$ is metric, a product space and by (N2) normal.

Hierarchy of the separation properties : The properties of the space types T_0 to T_4 show that every T_i-space has the separation properties of the preceding space type T_{i-1}, and hence the properties of all preceding space types. In particular, metric spaces have all properties of the space types T_0 to T_4, since every metric space is a T_4-space and hence also a T_3-, T_2-, T_1- and T_0-space.

Subspaces and product spaces of a space of type T_0 to T_3 are of the same type. Subspaces and product spaces of a T_4-space may be normal. For example, the subspaces and product spaces of metric spaces are normal.

Example 1 : T_0-space

Let (M ; T) be a topological space with the underlying set M = {a,b} and the topology T = {∅, {a}, M}. The open set {a} satisfies the T_0-axiom, since it contains the point a but not b. The T_1-axiom is not satisfied, since there is no open neighborhood of b which does not contain a.

Example 2 : A finite topological space is a T_1-space if and only if it is discrete.

(1) In a finite T_1-space (M ; S), let a be an arbitrary fixed point. By the T_1-axiom, for every point x ≠ a in M there is an open neighborhood of a which does not contain x. The finite intersection of these open sets is the open set {a}, which by construction contains only the point a. Since the point a is arbitrary, the space (M ; S) is discrete.

(2) Let the space (M ; S) be discrete. Then arbitrary points a ≠ x in M have the disjoint open neighborhoods {a} and {x}. Hence every discrete space is a T_1-space.

Example 3 : Every discrete space is a T_4-space

In Section 5.3, it is shown that every subset of a discrete topological space (M ; T) is both open and closed. Let the subsets A and B be disjoint and closed. Then A and B are also disjoint open sets with $A \subseteq A$ and $B \subseteq B$. Hence M is a T_4-space.

Example 4 : T_1-space with finite-complement topology

Let (ℕ ; S) be a topological space. The underlying set ℕ contains the natural numbers {1,2,...}. The topology S contains ∅, ℕ and every open set S_i for which ℕ – S_i is finite. For example, S_i = {10, 11, 12,...} is an element of the topology. This topology is called the finite-complement topology.

If a ≠ b are arbitrary points of ℕ, then the open set S_1 = ℕ – {b} contains the point a but not b. Likewise, the open set S_2 = ℕ – {a} contains the point b but not a. Hence (ℕ ; S) is a T_1-space.

Any two non-empty open sets S_i and S_m in S contain an infinite number of common points. Therefore the points a and b cannot possess disjoint open neighborhoods. Hence (ℕ ; S) is not a T_2-space.

5.10 CONVERGENCE

5.10.1 SEQUENCES

Introduction : An iterative mathematical procedure is said to be convergent if it identifies a point of a topological space $(M;T)$. Convergence has inspired thought since antiquity : "Does an arrow ever reach its goal by repeatedly covering half of the remaining distance?" Today, a carefully defined concept of convergence is the basis for iterative approximation methods and search methods in structured sets.

There are iterative methods whose result depends on the order of the steps, and methods whose steps may be executed in an arbitrary order. An iterative method is said to be directed if the order of the steps is determined by a mapping $f : G \to M$ from a directed set G. An iterative method is said to be undirected if convergence is studied with a filter (a subset of the power set $P(M)$), whose elements may be considered in an arbitrary order.

Directed iterative methods are particularly suitable for metric spaces. The metric of the space serves to quantify convergence. If the well-ordered set $\mathbb{N}$ of the natural numbers is used as a directed set, then the mapping $f : \mathbb{N} \to M$ is called a sequence. If G is a general directed set, then the mapping $f : G \to M$ is called a net.

Undirected iterative methods are more general than directed methods, since no order structure is required in the filter $F \subset P(M)$. For convergence to a limit $x \in M$, it suffices that every neighborhood of x includes an element of the filter F. The order in which the neighborhoods of x and the elements of F are considered is irrelevant.

The concepts of sequence, net and filter and the related concepts of subsequence, series, subnet and filter basis are treated in this section. In particular, the convergence of iterative methods and the uniqueness of limits are studied. The uniqueness of limits is seen to depend critically on the separation properties of the space.

Sequence : A mapping f from the natural numbers $\mathbb{N}$ to the points of a topological space $(M;T)$ is called a sequence in M. The images x_n of the mapping are called the terms of the sequence.

$$f : \mathbb{N} \to M \quad \text{with} \quad f(n) = x_n$$

The image of a sequence is not the same as the sequence. For example, the sequence $< a, b, a, b, \ldots >$ has the image $\{a, b\}$. While a sequence is by definition infinite, in this example the image is finite.

A sequence is said to be constant if its terms are all equal and the image $\{c\}$ therefore consists of a single point. The sequence is said to be real if the underlying set M is real. The mapping f is called a sequence of functions if M is a set of functions.

Limit : A point x of a metric space (M ; d) is called the limit of a sequence f : $\mathbb{N} \to M$ if for every positive real number ε there is a natural number n_0 such that for $n \ge n_0$ the distance $d(x, x_n)$ is less than ε. The limit of a sequence is designated by lim. The limit is a point of the underlying set M, but it is not required to be a term of the sequence f.

$$\lim_{n\to\infty} x_n = x \quad :\Leftrightarrow \quad \bigwedge_{\varepsilon>0} \ \bigvee_{n_0\in\mathbb{N}} (n \ge n_0 \quad \Rightarrow \quad d(x, x_n) < \varepsilon)$$

This definition of the limit is alternatively formulated as follows :

(1) For all $\varepsilon > 0$, $d(x, x_n) < \varepsilon$ for almost all terms.

(2) For all $\varepsilon > 0$, $d(x, x_n) \ge \varepsilon$ for at most a finite number of terms.

Convergence in metric spaces : A sequence f : $\mathbb{N} \to M$ in a metric space (M ; d) is said to converge (be convergent) to a point x in M if x is the limit of the sequence. A sequence in a metric space has at most one limit. The sequence is said to be divergent if it does not have a limit.

Proof : A sequence in a metric space has at most one limit.

In a metric space (M ; d), let f : $\mathbb{N} \to M$ with $f(n) = x_n$ be a sequence with two different limits a and b. By property (M2) of a metric $d(a, b) = \delta > 0$. For every real number $\varepsilon > 0$ there are natural numbers n_a and n_b for which :

$$\begin{array}{lll} d(a, x_i) < \varepsilon & \text{for} & i \ge n_a \\ d(b, x_m) < \varepsilon & \text{for} & m \ge n_b \end{array}$$

With $n_0 = \max(n_a, n_b)$, it follows that for all $n \ge n_0$:

$$d(a, x_n) < \varepsilon \quad \text{and} \quad d(b, x_n) < \varepsilon$$

Property (M4) of a metric implies :

$$d(a, b) \le d(a, x_n) + d(b, x_n) < 2\varepsilon \qquad \text{for all} \quad n \ge n_0$$

The choice $\varepsilon = \frac{\delta}{2}$ leads to the contradiction $d(a, b) < 2\varepsilon = \delta$. Hence the limits a and b are equal. The sequence therefore has at most one limit.

Component sequence : Let a sequence f : $\mathbb{N} \to \mathbb{R}^m$ with $f(n) = \mathbf{x}_n$ be given. Let the k-th coordinate of the vector $\mathbf{x}_n$ be x_{nk}. Then the sequence $f_k : \mathbb{N} \to \mathbb{R}$ with $f_k(n) = x_{nk}$ is called the k-th component sequence of f. The following relationship holds between a sequence f and its component sequences f_k :

(K1) A sequence f : $\mathbb{N} \to \mathbb{R}^m$ in the euclidean space ($\mathbb{R}^m$; d) converges if and only if its component sequences in the space ($\mathbb{R}$; d) converge. The components of the limit of the sequence are the limits of the component sequences.

Proof K1 : Convergence of sequences and component sequences

(1) Let the sequence f converge to the point $\mathbf{a} \in \mathbb{R}^m$. Then for every $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ one has $|\mathbf{x}_n - \mathbf{a}| < \varepsilon$. For the euclidean metric this implies $|x_{nk} - a_k| < \varepsilon$. Hence f_k converges to a_k.

(2) Let the component sequence f_k converge to the limit $a_k \in \mathbb{R}$ for $k = 1,...,m$. Then for every $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$ one has $|x_{nk} - a_k| < \frac{\varepsilon}{\sqrt{m}}$. For the euclidean metric it follows that

$$|\mathbf{x}_n - \mathbf{a}| \leq \sqrt{\sum_{k=1}^{m} (x_{nk} - a_k)^2} < \sqrt{\frac{\varepsilon^2}{m} * m} = \varepsilon$$

Hence f converges to **a**.

Fundamental sequence : To prove the convergence of a sequence using the concepts defined up to now, the limit must be known beforehand. If the limit is not known, one can determine whether the sequence is a fundamental sequence (Cauchy sequence).

A sequence $f : \mathbb{N} \to M$ in a metric space (M ; d) is said to be fundamental if for every positive real number ε there is a natural number n_0 such that $d(x_i,x_m) < \varepsilon$ for all $i, m \geq n_0$.

$$\text{f is fundamental} \quad :\Leftrightarrow \quad \bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (i, m \geq n_0 \quad \Rightarrow \quad d(x_i,x_m) < \varepsilon)$$

Properties of fundamental sequences

(F1) In a metric space every convergent sequence is fundamental. Hence it is a necessary condition for the convergence of a sequence in a metric space that the sequence is fundamental.

(F2) In a metric space there are sequences which are not fundamental.

(F3) In the one-dimensional real space ($\mathbb{R}$; d) a sequence is convergent if and only if it is fundamental.

(F4) In the euclidean space ($\mathbb{R}^n$; d) a sequence is fundamental if and only if its component sequences are fundamental.

(F5) In the euclidean space ($\mathbb{R}^n$; d) a sequence is convergent if and only if it is fundamental.

Proof : Properties of fundamental sequences

(F1) If $f : \mathbb{N} \to M$ converges to the limit x in a metric space (M ; d), then for every positive real number ε there is a natural number n_0 such that $d(x, x_i) < 0.5\varepsilon$ for all $i \geq n_0$, and by property (M4) of metrics $d(x_i,x_m) \leq d(x_i, x) + d(x, x_m) < \varepsilon$ for all $i, m \geq n_0$. Hence the convergent sequence f is fundamental.

(F2) In the space ($\mathbb{R}$; d) with euclidean metric d, let the harmonic sequence f : $\mathbb{N} \to \mathbb{R}$ be defined as follows :

$$x_n = f(n) = \sum_{k=1}^{n} \frac{1}{k}$$

This sequence is not fundamental. With the choices $\varepsilon = 0.5$, $n = n_0$ and $m = 2n$ for an arbitrary natural number n_0, the distance between the terms x_n and x_m of the sequence is :

$$d(x_m, x_n) = \sum_{k=n+1}^{2n} \frac{1}{k} = \left(\frac{1}{n+1} + \ldots + \frac{1}{2n} \right) > n\left(\frac{1}{2n} \right) = \varepsilon$$

The sequence is not fundamental, since $d(x_m, x_n) > \varepsilon$.

(F3) It follows immediately from (F1) that every convergent sequence in ($\mathbb{R}$; d) is fundamental. Conversely, let a fundamental sequence f : $\mathbb{N} \to \mathbb{R}$ with $f(n) = x_n$ be given. A subset A of the rational numbers $\mathbb{Q}$ is constructed :

$$A := \{ q \in \mathbb{Q} \mid \bigvee_{\delta > 0} \bigvee_{n_0 \in \mathbb{N}} \bigwedge_{n \geq n_0} (x_n \geq q + \delta) \}$$

The subset A has the properties of an open initial :

(1) For $q, r \in \mathbb{Q}$, if $r < q$ then $q \in A \Rightarrow r \in A$.

(2) The set A has no greatest element, since $q \in A$ implies $(q + \frac{\delta}{2}) \in A$:

$$\bigwedge_{n \geq n_0} (x_n \geq q + \delta) \Rightarrow \bigwedge_{n \geq n_0} (x_n \geq (q + \frac{\delta}{2}) + \frac{\delta}{2})$$

An open initial in the rational numbers is a real number (see Chapter 6). To emphasize this aspect of A, the designation A is replaced by a. In the following the real number a is shown to be the limit of the sequence f : $\mathbb{N} \to \mathbb{R}$: For all $\varepsilon > 0$ one has $d(a, x_n) \geq \varepsilon$ for at most a finite number of terms.

The proof is carried out indirectly. Let there be an $\varepsilon > 0$ such that $d(a, x_n) \geq \varepsilon$ for an infinite number of terms. Then at least one of the following must be the case :

(1) $x_n \leq a - \varepsilon$ for an infinite number of terms

This case cannot occur. Since every rational number q in the range $a - \varepsilon < q < a$ is contained in A, there is a $\delta > 0$ and an $n_0 \in \mathbb{N}$ such that $x_n \geq q + \delta$ for $n \geq n_0$. Hence $x_n \leq a - \varepsilon$ can hold for at most a finite number of terms.

(2) $x_n \geq a + \varepsilon$ for an infinite number of terms

This case cannot occur either. Since f is fundamental :

$$\bigvee_{n_0 \in \mathbb{N}} \bigwedge_{i,m \geq n_0} (d(x_i, x_m) < \frac{\varepsilon}{2})$$

Since $x_n \ge a + \varepsilon$ for an infinite number of terms, there is an $n_1 \ge n_0$ with $x_{n_1} \ge a + \varepsilon$. From $n_1 \ge n_0$, it follows that all points x_k with $k \ge n_1$ lie in the interval $x_{n_1} - \frac{\varepsilon}{2} < x_k < x_{n_1} + \frac{\varepsilon}{2}$. Since $x_{n_1} - \frac{\varepsilon}{2} \ge (a + \varepsilon) - \frac{\varepsilon}{2} = a + \frac{\varepsilon}{2}$, this implies $x_k > a + \frac{\varepsilon}{2}$ for $k \ge n_1$.

For every rational number in the range $a < q < a + \frac{\varepsilon}{4}$ it follows that $x_k > a + \frac{\varepsilon}{2} > q + \frac{\varepsilon}{4}$ for all $k \ge n_1$. Hence $q \in A$, contradicting $q > a$.

It follows from (1) and (2) that there is no $\varepsilon > 0$ such that $d(a, x_n) \ge \varepsilon$ for an infinite number of terms. Thus for every $\varepsilon > 0$ one has $d(a, x_n) \ge \varepsilon$ for at most a finite number of terms : The real number a is the limit of the sequence f.

(F4) (1) Let the sequence $f : \mathbb{N} \to \mathbb{R}^n$ be fundamental. Then for every $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ such that $i, m \ge n_0$ implies $|\mathbf{x}_i - \mathbf{x}_m| < \varepsilon$. For the k-th component sequence $f_k : \mathbb{N} \to \mathbb{R}$ with $f_k(i) = x_{ik}$, the properties of the euclidean metric imply that $|x_{ik} - x_{mk}| < \varepsilon$. Hence f_k is fundamental.

(2) Let the component sequence f_k be fundamental for $k = 1,...,n$. Then for every $\varepsilon > 0$ there is an $n_{0k} \in \mathbb{N}$ such that $|x_{ik} - x_{mk}| < \varepsilon / \sqrt{n}$ for all $i, m \ge n_{0k}$. Then for all $i, m \ge n_0 = \max n_{0i}$, the euclidean metric yields :

$$|\mathbf{x}_i - \mathbf{x}_m| \le \sqrt{\sum_{k=1}^{n} (x_{ik} - a_{mk})^2} < \sqrt{\frac{\varepsilon^2}{m} * n} = \varepsilon$$

Hence $f : \mathbb{N} \to \mathbb{R}^n$ is fundamental.

(F5) This statement is proved using the equivalences (K1), (F3) and (F4) :

The sequence f is fundamental ⇔
Every component sequence f_k is fundamental ⇔
Every component sequence f_k is convergent ⇔
The sequence f is convergent

Complete space : A metric space (M ; d) is said to be complete if every fundamental sequence $f : \mathbb{N} \to M$ has a limit x in M. Thus in a complete metric space it is a sufficient condition for the convergence of a sequence that it is fundamental.

Improper convergence : A real sequence in the euclidean space ($\mathbb{R}$; d) is said to be improperly convergent if for every real number a there is a natural number n_0 such that $f(n) = x_n > a$ for all $n > n_0$ or $f(n) = x_n < a$ for all $n > n_0$.

$$\lim_{n \to \infty} x_n = \infty \quad :\Leftrightarrow \quad \bigwedge_{a \in \mathbb{R}} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad x_n > a)$$

$$\lim_{n \to \infty} x_n = -\infty \quad :\Leftrightarrow \quad \bigwedge_{a \in \mathbb{R}} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad x_n < a)$$

Monotonic and bounded sequences : Let a total order relation be defined on the underlying set M of a topological space. Then a sequence f : $\mathbb{N} \rightarrow M$ is said to be (strictly) monotonic if for every $n \in \mathbb{N}$ one of the following relationships holds :

increasing	:	$x_n \le x_{n+1}$	strictly increasing	:	$x_n < x_{n+1}$
decreasing	:	$x_n \ge x_{n+1}$	strictly decreasing	:	$x_n > x_{n+1}$

A sequence f : $\mathbb{N} \rightarrow M$ is said to be bounded from above if there is an upper bound $a \in M$ such that $x_n \le a$ for every $n \in \mathbb{N}$. The sequence f is said to be bounded from below if there is a lower bound $b \in M$ such that $x_n \ge b$ for every $n \in \mathbb{N}$. A sequence is bounded from above/below if and only if its image is bounded from above/below. A set bounded from above has a least upper bound. A set bounded from below has a greatest lower bound.

Proof : A set bounded from above has a least upper bound.

Let A be the set of upper bounds of a set $M \subset \mathbb{R}$ bounded from above. In the following A is shown to be closed. Therefore the set A is either empty (contrary to the hypothesis that M is bounded from above), or it contains its lower boundary point ; this point is the least upper bound of M.

A point y of the complement $\overline{A}$ is not an upper bound of M. Hence there is a point $x \in M$ with $x > y$ such that $]-\infty, x[$ is an open neighborhood of y which is entirely contained in $\overline{A}$. Thus $\overline{A}$ is open, and hence A is closed.

Monotonic real sequences : The one-dimensional euclidean space ($\mathbb{R}$; d) is totally ordered. Hence it is possible to determine whether a given real sequence f : $\mathbb{N} \rightarrow \mathbb{R}$ is monotonic. In a general metric space, a total order relation must be defined before the monotonicity of a sequence can be studied.

Convergence of monotonic real sequences : Every monotonic real sequence is either convergent or improperly convergent. If a monotonically increasing real sequence is bounded from above, then it converges to its least upper bound a. If a monotonically decreasing real sequence is bounded from below, then it converges to its greatest lower bound b.

increasing with $x_n \le a$: $\lim\limits_{n \to \infty} x_n = a$ with $\bigwedge\limits_{\varepsilon > 0} \bigvee\limits_{n_0 \in \mathbb{N}} (n > n_0 \Rightarrow d(a, x_n) < \varepsilon)$

decreasing with $x_n \ge b$: $\lim\limits_{n \to \infty} x_n = b$ with $\bigwedge\limits_{\varepsilon > 0} \bigvee\limits_{n_0 \in \mathbb{N}} (n > n_0 \Rightarrow d(b, x_n) < \varepsilon)$

Proof : Convergence of monotonic real sequences

(1) Let a real sequence $<x_1, x_2, ...>$ be monotonically increasing. If the sequence is bounded from above, it has a least upper bound a. Hence for every real number $\varepsilon > 0$ there is a natural number n_0 with $d(a, x_{n_0}) < \varepsilon$ and $x_{n_0} < a$. Since the sequence increases monotonically and a is the least upper bound, $x_{n_0} \leq x_n$ and $d(a, x_n) < \varepsilon$ for $n > n_0$. Hence the sequence converges to the limit a :

$$\bigwedge_{\varepsilon > 0} \ \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad d(a, x_n) < \varepsilon)$$

(2) Assume that the monotonically increasing real sequence $<x_1, x_2, ...>$ does not have a least upper bound. Then for every real number a there is a natural number n_0 such that $x_{n_0} > a$. Since the sequence increases monotonically, $n > n_0$ implies $x_{n_0} \leq x_n$. The sequence is improperly convergent :

$$\bigwedge_{a \in \mathbb{R}} \ \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad x_n > a)$$

(3) The proof for monotonically decreasing sequences is analogous.

Nested intervals : The closed intervals $I_n := [a_n, b_n]$ on the real axis $\mathbb{R}$ are said to be nested if the real sequences $f : \mathbb{N} \to \mathbb{R}$ with $f(n) = a_n$ and $g : \mathbb{N} \to \mathbb{R}$ with $g(n) = b_n$ have the following properties :

$$a_n < a_{n+1} < b_{n+1} < b_n \qquad \text{and} \qquad \lim_{n \to \infty} (b_n - a_n) = 0$$

Since the euclidean space $(\mathbb{R} ; d)$ is complete, the intersection of the nested intervals I_n contains exactly one point. This point p is the limit of the strictly monotonic sequences $<a_1, a_2, ...>$ and $<b_1, b_2, ...>$. The sequences do not contain the point p.

$$\cap I_n = \{p\} \quad \wedge \quad \lim_{n \to \infty} a_n = p \quad \wedge \quad \lim_{n \to \infty} b_n = p$$

Proof : Convergence of nested intervals

For $i > n$ it follows by induction that $a_i < b_i < b_n$. For $i < n$ it follows that $a_i < a_n < b_n$. Hence the monotonic sequence $<a_1, a_2, ...>$ is bounded from above by b_n. Let p be the least upper bound of $<a_1, a_2, ...>$. Then $a_n \leq p \leq b_n$, so that p is contained in the interval I_n, that is $p \in [a_n, b_n]$.

The point p is contained in the intersection $\cap I_n$ of all the nested intervals. If another point $x \neq p$ were contained in $\cap I_n$, then $\bigwedge_{n \in \mathbb{N}} (b_n - a_n \geq |p - x|)$ would hold, contradicting $\lim (b_n - a_n) = 0$. Hence p is the only point contained in $\cap I_p = \{p\}$. For every real number ε there is thus an $n_0 \in N$ such that $d(p, x_n) < \varepsilon$ for $n > n_0$. Hence p is the limit of the sequence $<a_1, a_2, ...>$. Then by item (1) of the following theorems for limits $\lim(b_n - a_n) = 0$ implies that p is also the limit of $<b_1, b_2, ...>$. Hence both sequences converge to the limit p.

Theorems for limits : Let the sequences $< a_1, a_2, \ldots >$ and $< b_1, b_2, \ldots >$ be convergent. Then derived sequences have the following limits :

$$\text{(1)} \quad \lim_{n \to \infty} (a_n \pm b_n) = \lim_{n \to \infty} (a_n) \pm \lim_{n \to \infty} (b_n)$$

$$\text{(2)} \quad \lim_{n \to \infty} (a_n \cdot b_n) = \lim_{n \to \infty} (a_n) \cdot \lim_{n \to \infty} (b_n)$$

$$\text{(3)} \quad \lim_{n \to \infty} \left(\frac{a_n}{b_n} \right) = \lim_{n \to \infty} (a_n) \;/\; \lim_{n \to \infty} (b_n) \quad \text{if} \quad b_n \neq 0, \; \lim_{n \to \infty} b_n \neq 0$$

$$\text{(4)} \quad \lim_{n \to \infty} (|a_n|) = \left| \lim_{n \to \infty} (a_n) \right|$$

Example 1 : Convergence of a sequence

Let a sequence be defined as follows in the euclidean space $(\mathbb{R}\,; d)$:

$$x_n = \frac{n+1}{2n} \qquad \text{with } n \in \mathbb{N}$$

In the following the limit of this sequence is shown to be $x = \frac{1}{2}$. The distance $d(x, x_n)$ is estimated :

$$d(x, x_n) = \frac{n+1}{2n} - \frac{1}{2} = \frac{1}{2n} < \frac{1}{n}$$

For arbitrary real $\varepsilon > 0$, choose $n_0 > \frac{1}{\varepsilon}$. The preceding inequality then implies the convergence of the sequence :

$$d(x, x_n) < \frac{1}{n} < \frac{1}{n_0} < \varepsilon \quad \text{for} \quad n \geq n_0$$

Example 2 : Fundamental sequence

For the sequence in Example 1, the distance $d(x_m, x_n)$ is estimated with $m, n > n_0$ for arbitrary $n_0 \in \mathbb{N}$:

$$d(x_m, x_n) = \left| \frac{m+1}{2m} - \frac{n+1}{2n} \right| = \left| \frac{n-m}{2mn} \right| < \frac{1}{2n_0}$$

The sequence is fundamental, since for arbitrary real $\varepsilon > 0$ a number $n_0 > \frac{1}{2\varepsilon}$ may be chosen. Then

$$d(x_m, x_n) < \varepsilon \qquad \text{for all} \quad m, n > n_0$$

Example 3 : Monotonic sequence

The sequence in Example 1 is strictly monotonically decreasing :

$$x_n - x_{n+1} = \frac{n+1}{2n} - \frac{n+2}{2n+2} = \frac{1}{2n(n+1)} > 0$$

The sequence is bounded from below by $x_n > \frac{1}{2}$. It converges to the greatest lower bound $b = \frac{1}{2}$.

$$x_n = \frac{n+1}{2n} \quad \Rightarrow \quad x_n > \frac{n}{2n} = \frac{1}{2}$$

Example 4 : Limits of sequences

$$\lim_{n\to\infty} \frac{n^k}{n!} = 0 \qquad \text{for} \qquad k \in \mathbb{N}$$

$$\lim_{n\to\infty} n^k x^n = 0 \qquad \text{for} \qquad k \in \mathbb{N} \qquad \text{and} \qquad |x| < 1$$

$$\lim_{n\to\infty} \frac{x^n}{n!} = 0 \qquad \text{for} \qquad x \in \mathbb{R}$$

$$\lim_{n\to\infty} \sqrt[n]{n} = 1$$

$$\lim_{n\to\infty} \left(1 + \frac{1}{n}\right)^n = e = 2.718281...$$

Example 5 : Geometric sequences

A sequence $f : \mathbb{N} \to \mathbb{R}$ in the euclidean space $(\mathbb{R}\,; d)$ is said to be geometric if the ratio $c = x_{n+1} \,/\, x_n$ of consecutive terms of the sequence is constant. If the ratio c is positive, the sequence is monotonic. If the ratio c is negative, the sequence is said to alternate.

$$\text{sequence } f : \mathbb{N} \to \mathbb{R} \text{ is geometric} \quad :\Leftrightarrow \quad f(n) = wc^n \quad \wedge \quad c,w \in \mathbb{R}$$

The convergence of geometric sequences is determined by the ratio c :

$c \le -1$	:	divergent
$-1 < c < 1$	:	convergent with limit 0
$c = 1$	:	convergent with limit w
$c > 1$	:	improperly convergent with limit $\pm\infty$

(a) For $c \le -1$ and $c = 1$ the statement is self-evident.

(b) To prove convergence for $-1 < c < 1$ in the euclidean space $(\mathbb{R}\,; d)$, let $|c| = \frac{1}{1+s}$ with $s > 0$ and $(1+s)^n > 1 + ns$, and hence $|c^n| < \frac{1}{1+ns}$. For every number $\varepsilon > 0$ there is a natural number $n_0 > (x_1 - \varepsilon)/\varepsilon s$ such that $d(0, x_1 c^n) < \varepsilon$ for all $n > n_0$. Hence the limit is 0.

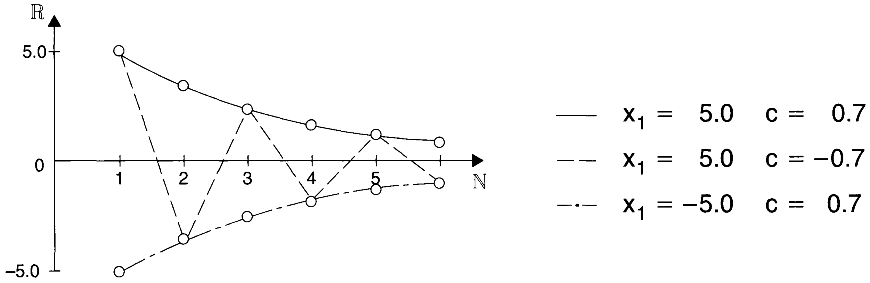
$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (n \ge n_0 \quad \Rightarrow \quad d(0, x_1 c^n) < \varepsilon)$$

(c) To prove improper convergence for $c > 1$, let $c = 1 + s$ with $s > 0$ and $(1 + s)^n > 1 + ns$. For $x_1 > 0$ and every real number $a \geq x_1$ there is a natural number $n_0 > (a - x_1) / sx_1$ such that $x_1 c^n > a$ for all $n \geq n_0$. Hence the sequence converges improperly to ∞. For $x_1 < 0$ and every real number $a \leq x_1$ there is a natural number n_0 such that $x_1 c^n < a$ for all $n \geq n_0$. Hence the sequence converges to $-\infty$.

$$x_1 > 0 \; : \; \bigwedge_{a \in \mathbb{R}} \; \bigvee_{n_0 \in \mathbb{N}} \; (n \geq n_0 \quad \Rightarrow \quad x_1 c^n > a)$$

$$x_1 < 0 \; : \; \bigwedge_{a \in \mathbb{R}} \; \bigvee_{n_0 \in \mathbb{N}} \; (n \geq n_0 \quad \Rightarrow \quad x_1 c^n < a)$$

The following diagram shows two monotonic geometric sequences with $c = 0.7$ but different terms $x_1 = -5.0$ and $x_1 = 5.0$, along with an alternating geometric sequence with $c = -0.7$ and $x_1 = 5.0$.



5.10.2 SUBSEQUENCES

Introduction : In a metric space a convergent sequence has exactly one limit. A divergent sequence does not have a limit. However, in special cases a divergent sequence may be divided into convergent subsequences. This leads to the concept of accumulation points of sequences and thus to the concepts of the superior limit and inferior limit of a sequence of numbers.

Subsequences : Let a mapping $f : \mathbb{N} \to M$ be an arbitrary sequence in a metric space $(M ; d)$. Let a mapping $k : \mathbb{N} \to \mathbb{N}$ with $k(n) = k_n$ be a strictly monotonically increasing sequence. Then the composition $f \circ k$ is called a subsequence of the sequence f.

$$f \circ k : \mathbb{N} \to M \qquad \text{with} \qquad f \circ k(n) = f(k(n)) = f(k_n) = x_{k_n}$$

Subsequences in metric spaces have the following properties :

(T1) A sequence converges if and only if each of its subsequences converges. In this case every subsequence converges to the same limit.

(T2) If a sequence has two subsequences which converge to different limits, the sequence is divergent.

Proof : Properties of subsequences

(T1) By definition, a sequence $f : \mathbb{N} \to M$ which converges to the limit x satisfies

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad d(x, x_n) < \varepsilon)$$

Since the sequence $k : \mathbb{N} \to \mathbb{N}$ is strictly monotonically increasing, there is a number i_0 such that $k(i_0) \geq n_0$. Hence the composition $f \circ k$ is a convergent subsequence of f with limit x :

$$\bigwedge_{\varepsilon > 0} \bigvee_{i_0 \in \mathbb{N}} (i > i_0 \quad \Rightarrow \quad d(x, x_{k_i}) < \varepsilon)$$

Conversely, let every subsequence $f \circ k$ be convergent. Then for every subsequence there is exactly one limit a, and for every real number $\varepsilon > 0$ there is a natural number $i_0 \in \mathbb{N}$ such that :

$$\begin{array}{lll} f \circ k : \mathbb{N} \to M & \text{with} & f \circ k(i) = f(k_i) = x_{k_i} \\ d(a, x_{k_i}) < \varepsilon & \text{for all} & i \geq i_0 \end{array}$$

The sequence f is a subsequence of itself and hence convergent by hypothesis. Its limit is the limit x of the sequence f.

$$\begin{array}{lll} f : \mathbb{N} \to M & \text{with} & f(m) = x_m \\ d(x, x_m) < \varepsilon & \text{for all} & m \geq m_0 \end{array}$$

With $n_0 = \max(k(i_0), m_0)$ it follows that

$$\begin{array}{lll} d(a,x_{k_i}) < \varepsilon & \text{for all} & k_i \geq n_0 \\ d(x,x_n) < \varepsilon & \text{for all} & n \geq n_0 \end{array}$$

Let the limits a and x be different. By property (M2) of metrics, their distance is $d(a,x) = \delta > 0$. Property (M4) of metrics yields :

$$d(a,x) \leq d(a,x_{k_i}) + d(x,x_{k_i}) < 2\varepsilon \qquad \text{for all } k_i \geq n_0$$

The result $d(a,x) < 2\varepsilon$ contradicts $d(a,x) = \delta > 0$, since for every $\delta > 0$ there is an $\varepsilon > 0$ such that $2\varepsilon < \delta$. Hence the limits a and x are equal. All subsequences $f \circ k$ and the sequence f have the same limit x.

(T2) Let the limits a and b of the subsequences $f \circ g$ and $f \circ k$ of a sequence f in a metric space (M ; d) be different. Let the sequence f converge to the limit x. Then by (T1) the limits of the sequence f and of the subsequences are equal, that is $a = b = x$. This contradicts the hypothesis $a \neq b$. Hence contrary to the assumption the sequence f is divergent.

Accumulation point of a sequence : A point x of a topological space (M ; T) is called an accumulation point of the sequence $f : \mathbb{N} \to M$ if every neighborhood of x contains an infinite number of terms of the sequence. These terms may be scattered throughout the sequence. A sequence may have more than one accumulation point. Accumulation points need not be terms of the sequence f, but they must be points of the topological space M.

A point x of a metric space (M ; d) is an accumulation point of the sequence $f : \mathbb{N} \to M$ if for every real number $\varepsilon > 0$ and every natural number n_0 there is a term x_n of the sequence with $n > n_0$ such that $d(x,x_n) < \varepsilon$. In a sequence f with an accumulation point x there is a subsequence $f \circ k$ with the limit x.

$$x \text{ is an accumulation point of } f \quad :\Leftrightarrow \quad \bigwedge_{\varepsilon > 0} \; \bigwedge_{n_0 \in \mathbb{N}} \; \bigvee_{n > n_0} (d(x, x_n) < \varepsilon)$$

Proof : In a sequence f with an accumulation point x there is a subsequence $f \circ k$ with the limit x.

A strictly monotonically increasing mapping $k : \mathbb{N} \to K$ with $k(n) = k_n$ is constructed for the sequence f. The terms of k are chosen such that $d(x, x_{k_n}) < \varepsilon_n = 2^{-n}$ for $x_{k_n} = f(k_n)$. Then $f \circ k$ is a subsequence. Its limit is x, since $d(x, x_{k_n}) < \varepsilon_{n_0}$ for $n \geq n_0$.

Accumulation point of a set : An accumulation point of a set is an inner point or boundary point of the set which is not isolated. This concept must not be confused with the concept of an accumulation point of a sequence in the set. Every accumulation point of a set $A := \{x_n \mid n \in \mathbb{N}\}$ in a metric space is also an accumulation point of a sequence $f : \mathbb{N} \to A$. The converse of this statement is false ! For example, x is an accumulation point of the sequence $f : \mathbb{N} \to \{x\}$ with $f(n) = x$, but not an accumulation point of the one-element set $\{x\}$.

Proof : In a metric space every accumulation point of a set is also an accumulation point of a sequence.

In a metric space $(M\,;d)$, let x be an accumulation point of a set $A \subseteq M$. Let the open ball with center x and radius $\frac{1}{n}$ for $n \in \mathbb{N}$ be $D_n := D(x, \frac{1}{n})$. A point x_n may be chosen in each of the intersections $A \cap D_n$. These points are the terms of a sequence f :

$$f : \mathbb{N} \to \mathbb{R}^n \qquad \text{with} \qquad f(n) = x_n \in A \cap D_n$$

For every real number $\varepsilon > 0$ and every natural number $n_0 > 0$ there is an $n > n_0$ such that $\frac{1}{n} < \varepsilon$. Hence the center x is an accumulation point of the sequence f.

$$\bigwedge_{\varepsilon>0} \bigwedge_{n_0 \in \mathbb{N}} \bigvee_{n>n_0} (d(x,x_n) < \varepsilon)$$

Improper accumulation point : A real sequence $f : \mathbb{N} \to \mathbb{R}$ with $f(n) = x_n$ has the improper accumulation point ∞ if for every real number a and every natural number n_0 there is a term x_n of the sequence with $n > n_0$ and $x_n > a$. The improper accumulation point $-\infty$ is defined analogously.

Superior limit : The greatest accumulation point a of a sequence $<x_1, x_2,...>$ in the set of real numbers is called the superior limit (limit superior, lim sup) of the sequence. For every real number $\varepsilon > 0$ and every natural number $n_0 > 0$ there is an $n > n_0$ such that $x_n > a - \varepsilon$. For every real number $\varepsilon > 0$ there is also a natural number n_1 such that $x_n < a + \varepsilon$ for all $n > n_0$. The superior limit may be an improper accumulation point.

$$a = \limsup <x_1, x_2,...> \quad :\Leftrightarrow \quad \bigwedge_{\varepsilon>0} \bigwedge_{n_0 \in \mathbb{N}} \bigvee_{n>n_0} (x_n > a - \varepsilon) \quad \wedge$$
$$\bigwedge_{\varepsilon>0} \bigvee_{n_1 \in \mathbb{N}} \bigwedge_{n>n_1} (x_n < a + \varepsilon)$$

Inferior limit : The least accumulation point b of a sequence $<x_1, x_2,...>$ in the set of real numbers is called the inferior limit (limit inferior, lim inf) of the sequence. For every real number $\varepsilon > 0$ and every natural number $n_0 > 0$ there is an $n > n_0$ such that $x_n < b + \varepsilon$. For every real number $\varepsilon > 0$ there is also a natural number n_1 such that $x_n > b - \varepsilon$ for all $n > n_0$. The inferior limit may be an improper accumulation point.

$$b = \liminf <x_1, x_2,...> \quad :\Leftrightarrow \quad \bigwedge_{\varepsilon>0} \bigwedge_{n_0 \in \mathbb{N}} \bigvee_{n>n_0} (x_n < b + \varepsilon) \quad \wedge$$
$$\bigwedge_{\varepsilon>0} \bigvee_{n_1 \in \mathbb{N}} \bigwedge_{n>n_1} (x_n > b - \varepsilon)$$

Example : Accumulation points of sequences

(1) The sequence $<1, -1, 1, -1, ...>$ has the accumulation points 1 and -1. The image $\{1, -1\}$ of the sequence consists of the two points -1 and 1.

(2) The sequence $<1, \frac{1}{2}, 1, \frac{1}{3}, 1, \frac{1}{4}, ...>$ has the accumulation points 1 and 0. The accumulation point 1 is a term of the sequence, but the accumulation point 0 is not.

(3) The sequence $<1, \frac{1}{2}, \frac{1}{3}, ...>$ has exactly one accumulation point 0, which is not a term of the sequence.

(4) The sequence $<1, -1, 2, -2, 3, -3 ...>$ does not have any proper accumulation points, but it has the improper accumulation points ∞ and $-\infty$.

5.10.3 SERIES

Introduction : In applications, one often needs the sum of the terms of a real sequence. This cannot be determined directly, since the sequence is infinite. One therefore defines a series whose terms are partial sums over a finite number of terms of the real sequence. If the series possesses a limit, then this is the desired sum. Several convergence tests for series are treated in the following.

Series : In the euclidean space $(\mathbb{R}\,;d)$, let $f : \mathbb{N} \to \mathbb{R}$ be a real sequence with $f(n) = x_n$. The sum of the first m terms of f is called the m-th partial sum of the sequence f and is designated by s_m. The sequence $< s_1, s_2, ... >$ of partial sums is called the series associated with the sequence f and is designated by Σx_n. The difference $s_m - s_n$ of the m-th and n-th partial sums is called a segment of the series. If the series $< s_1, s_2, ... >$ converges to a limit s, then s is called the sum of the sequence f.

$$s_m = \sum_{n=1}^{m} x_n$$

$$s = \lim_{m \to \infty} s_m$$

Absolutely convergent series : A series Σx_n is said to converge absolutely (be absolutely convergent) if the sequence $\Sigma\, |x_n|$ of the partial sums of the absolute values of its terms converges. A series which converges but does not converge absolutely is said to be conditionally convergent.

Proof of convergence : There is no general method for proving the convergence of a series and determining its limit. However, there are tests for the convergence of series. There are also comparison tests which allow the convergence of series to be inferred from the convergence of other series. Some of these tests are treated in the following.

Convergence tests : Of the following tests, (K1) and (K5) are necessary conditions for the convergence of a series, and (K1) to (K4) are sufficient conditions.

(K1) Cauchy test : A series Σx_n is convergent if and only if for every real number $\varepsilon > 0$ there is an index n_0 beyond which the absolute values $|s_m - s_n|$ of all segments of the series are less than ε.

$$\Sigma x_n \text{ converges} \quad \Leftrightarrow \quad \bigwedge_{\varepsilon > 0} \; \bigvee_{n_0 \in \mathbb{N}} \; (m \geq n \geq n_0 \quad \Rightarrow \quad |s_m - s_n| < \varepsilon)$$

$$\Leftrightarrow \quad \bigwedge_{\varepsilon > 0} \; \bigvee_{n_0 \in \mathbb{N}} \; (m \geq n \geq n_0 \quad \Rightarrow \quad \left| \sum_{k=n}^{m} x_k \right| < \varepsilon)$$

(K2) Monotonicity test : A series Σx_n with terms $x_n \geq 0$ converges if and only if the sequence $< s_1, s_2, ... >$ of partial sums is bounded.

(K3) Leibniz test : Let a sequence $< x_1, x_2, ... >$ be monotonically decreasing with limit 0. Then the sequence of partial sums of the alternating sequence $< x_1, -x_2, x_3, -x_4, ... >$ is convergent.

(K4) Ratio test : For a sequence $< x_1, x_2, ... >$, let every term $x_n \neq 0$. If there is a real number c such that beyond an index n_0 the absolute value of the ratio of consecutive terms is less than 1, that is $|x_{n+1} / x_n| \leq c$ with $c < 1$ for $n > n_0$, then the series Σx_n is absolutely convergent.

(K5) Trivial test : If the series Σx_n converges, then the sequence $< x_1, x_2, ... >$ converges to 0. The converse of this statement is false : If the terms of the sequence $< x_1, x_2, ... >$ converge to 0, it cannot be inferred that the series Σx_n converges.

Proof : Convergence tests

(K1) Cauchy test : By (F3), a sequence in the euclidean space $(\mathbb{R} ; d)$ is convergent if and only if it is fundamental. A series which satisfies the Cauchy test also satisfies the condition for a fundamental sequence, and conversely.

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (i, m \geq n_0 \quad \Rightarrow \quad d(s_i, s_m) < \varepsilon)$$

(K2) Monotonicity test : For the sequence $< s_1, s_2, ... >$ of the partial sums of a sequence $< x_1, x_2, ... >$ with $x_n \geq 0$ it follows that $s_m \geq s_i$ for $m > i$. Hence the sequence $\Sigma x_n = < s_1, s_2, ... >$ of partial sums is monotonically increasing.

(a) Let the series Σx_n be bounded. Then Section 5.10.1 shows that Σx_n converges.

(b) Let the series Σx_n be unbounded. Then Section 5.10.1 shows that Σx_n converges improperly, and hence diverges.

(K3) Leibniz test : The partial sums s_n of the alternating sequence $< x_1, -x_2, x_3, -x_4, ... >$ are

$$s_n = \sum_{k=1}^{n} (-1)^{k+1} x_k$$

The sequence $< s_1, s_2, ... >$ of partial sums converges if it passes the Cauchy test (K1), that is if

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (m \geq n \geq n_0 \quad \Rightarrow \quad \left| \sum_{k=n}^{m} (-1)^{k+1} x_k \right| < \varepsilon)$$

Since the sequence $<x_1, x_2,...>$ decreases monotonically and $x_n \rightarrow 0$, it follows that $x_n > 0$ and $(x_n - x_{n+1}) \geq 0$ for all $n \in \mathbb{N}$. For odd values of $m - n$:

$$\begin{aligned}\left|\sum_{k=n}^{m} (-1)^{k+1} x_k\right| &= \left|(-1)^{n+1} (x_n - x_{n+1} + x_{n+2} - ... + x_{m-1} - x_m)\right| \\ &= \left|(x_n - x_{n+1}) + (x_{n+2} - x_{n+3}) + ... + (x_{m-1} - x_m)\right| \\ &= (x_n - x_{n+1}) + (x_{n+2} - x_{n+3}) + ... + (x_{m-1} - x_m) \\ &= x_n - (x_{n+1} - x_{n+2}) - ... - (x_{m-2} - x_{m-1}) - x_m \\ &\leq x_n = |x_n|\end{aligned}$$

For even values of $m - n$:

$$\begin{aligned}\left|\sum_{k=n}^{m} (-1)^{k+1} x_k\right| &= \left|(-1)^{n+1} (x_n - x_{n+1} + ... + x_{m-2} - x_{m-1} + x_m)\right| \\ &= \left|(x_n - x_{n+1}) + ... + (x_{m-2} - x_{m-1}) + x_m\right| \\ &= (x_n - x_{n+1}) + ... + (x_{m-2} - x_{m-1}) + x_m \\ &= x_n - (x_{n+1} - x_{n+2}) - ... - (x_{m-1} - x_m) \\ &\leq x_n = |x_n|\end{aligned}$$

Since $x_n \rightarrow 0$, for every real number $\varepsilon > 0$ there is an index n_0 such that $|x_n| < \varepsilon$ for $n \geq n_0$. Hence the sequence $<s_0, s_1,...>$ passes the Cauchy test of convergence :

$$\left|\sum_{k=n}^{m} (-1)^k x_k\right| \leq |x_n| < \varepsilon \qquad \text{for all } m \geq n \geq n_0$$

(K4) Ratio test : The first n_0 terms of the sequence $<x_1, x_2,...>$ have no influence on the convergence of the series Σx_n and are therefore omitted without loss of generality. Thus, let $<x_1, x_2,...>$ be a sequence with $|x_{n+1}/x_n| \leq c$ for all $n \in \mathbb{N}$. For the sequence $<|x_1|, |x_2|,...>$, the difference of the partial sums s_m and s_n is formed and estimated for $0 < c < 1$ with $m \geq n \geq n_0$:

$$\begin{aligned}s_m - s_n &\leq c^n |x_1| + c^{n+1} |x_1| + ... + c^{m-1} |x_1| \\ &= (1 + c + ... + c^{m-n-1}) c^n |x_1| \\ &= \frac{1 - c^{m-n}}{1 - c} c^n |x_1|\end{aligned}$$

$$s_m - s_n \leq \frac{c^n}{1-c} |x_1| \qquad \text{and} \qquad s_m - s_n > 0$$

For every real number $\varepsilon > 0$ there is a natural number n_0 such that $c^{n_0} |x_1| < \varepsilon (1 - c)$. Hence the series $\Sigma |x_n|$ passes the Cauchy test of convergence :

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (m \geq n \geq n_0 \quad \Rightarrow \quad |s_m - s_n| < \varepsilon)$$

(K5) Trivial test : Let the series Σx_n be convergent. Then by (K1) for every real number $\varepsilon > 0$ there is an index n_0 such that the segment $|s_n - s_{n-1}| = |x_n|$ of the series satisfies $|x_n| < \varepsilon$ for $i = m = n > n_0$. Hence the terms x_n converge to 0.

Conversely, let $< x_1, x_2, ... >$ be a sequence whose terms converge to 0. Then the series Σx_n is not necessarily convergent. For example, in the proof of (F2) in Section 5.10.1 the sequence of partial sums of the sequence $<1, \frac{1}{2}, \frac{1}{3}, ...>$ is shown to diverge, although the terms $x_n = \frac{1}{n}$ converge to zero.

Comparison tests : The following tests allow the convergence of a series to be inferred from the convergence of another series.

(V1) Every absolutely convergent series is convergent. The absolute value $|\Sigma x_n|$ of the limit of an absolutely convergent series is less than or equal to the limit $\Sigma |x_n|$ of the series of the absolute values.

(V2) If the series Σx_n and Σy_n are convergent, then for arbitrary numbers $a, b \in \mathbb{R}$ the series $\Sigma (a x_n + b y_n)$ is also convergent.

(V3) If the series Σx_n and Σy_n are convergent and $x_n \le y_n$ for all n, then the limit of Σx_n is less than or equal to the limit of Σy_n.

(V4) If the series Σx_n is absolutely convergent and a sequence $< a_1, a_2, ... >$ is bounded, then the series $\Sigma a_n x_n$ is absolutely convergent.

(V5) If the series Σx_n is convergent and a monotonic sequence $< a_1, a_2, ... >$ is bounded, then the series $\Sigma a_n x_n$ converges.

(V6) If the series Σx_n is convergent and there is a natural number n_0 such that $0 \le |y_n| \le x_n$ for all $n > n_0$, then the series Σy_n is absolutely convergent.

Proof : Comparison tests

(V1) Let the series Σx_n be absolutely convergent. Then, according to the Cauchy test of convergence, for every real number $\varepsilon > 0$ there is a natural number n_0 such that for all $m \ge n \ge n_0$:

$$||x_n| + |x_{n+1}| + ... + |x_{m-1}|| < \varepsilon$$

From $|x_1 + x_2| \le |x_1| + |x_2|$, one obtains $|x_1 + ... + x_k| \le |x_1| + ... + |x_k|$ by induction. Hence $|\Sigma x_n| \le \Sigma |x_n|$. Also, the series Σx_n passes the Cauchy test :

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (m \ge n \ge n_0 : \quad |x_n + x_{n+1} + ... + x_{m-1}| < \varepsilon)$$

(V2) Let the limits of the convergent series Σx_n and Σy_n be x and y, respectively. Let the n-th terms of their sequences of partial sums be s_n and t_n, respectively. Then

$$\bigwedge_{\delta>0} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad |x - s_n| < \delta \quad \wedge \quad |y - t_n| < \delta)$$

For the sequence with the terms $(as_n + bt_n)$:

$$\begin{aligned} |(ax + by) - (as_n + bt_n)| &= |a(x - s_n) + b(y - t_n)| \\ &\leq |a|\,|x - s_n| + |b|\,|y - t_n| \\ &\leq (|a| + |b|)\,\delta \end{aligned}$$

The series $\Sigma(ax_n + by_n)$ converges to $(ax + by)$, since the condition for limits is satisfied by the choice $\delta = \varepsilon / (|a| + |b|)$:

$$\bigwedge_{\varepsilon>0} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad d(ax + by, as_n + bt_n) < \varepsilon)$$

(V3) Let the limits of the convergent series Σx_n and Σy_n be x and y, respectively. Then the proof of (V2) shows that the series $\Sigma(y_n - x_n)$ converges to a limit $d = y - x$. From $d_i = y_i - x_i \geq 0$ it follows that $d \geq 0$, and hence $x \leq y$.

(V4) For the bounded sequence $< a_1, a_2, \ldots >$ there is a real number $\gamma > 0$ such that $|a_n| < \gamma$ for all n. Let the i-th partial sum in the series $\Sigma |x_n|$ be s_i. Then for the i-th partial sum t_i in the series $\Sigma |a_n x_n|$ one obtains :

$$0 \leq t_i = \sum_{k=1}^{i} |a_k x_k| \leq \gamma \sum_{k=1}^{i} |x_k| = \gamma s_i$$

Since by hypothesis the series Σy_n converges absolutely, the partial sums s_i are bounded. Therefore the partial sums $t_i \leq \gamma s_i$ of the series $\Sigma |a_n x_n|$ are positive, monotonically increasing and bounded. Hence the series $\Sigma a_n x_n$ is absolutely convergent.

(V5) The partial sum t_i of the series $\Sigma a_n x_n$ is partially summed with the partial sums s_k of the series Σx_k and $s_0 := 0$.

$$\begin{aligned} t_i &= \sum_{k=1}^{i} a_k x_k = \sum_{k=1}^{i} a_k (s_k - s_{k-1}) = \sum_{k=1}^{i} a_k s_k - \sum_{k=1}^{i} a_k s_{k-1} \\ &= a_i s_i + \sum_{k=1}^{i-1} a_k s_k - \sum_{k=1}^{i-1} a_{k+1} s_k \\ t_i &= a_i s_i + \sum_{k=1}^{i-1} (a_k - a_{k+1}) s_k \end{aligned}$$

By the comparison test (V4), the series $<t_1, t_2,...>$ is absolutely convergent if the sequence $<s_1, s_2,...>$ is bounded and the series $\Sigma(a_k - a_{k+1})$ is absolutely convergent. Since the series Σx_n converges, the sequence $<s_1, s_2,...>$ is bounded. The partial sum w_k of the sequence $\Sigma|a_k - a_{k+1}|$ is determined using the monotonicity of the sequence $<a_1, a_2,...>$:

$$w_k = \sum_{i=1}^{k} |a_i - a_{i+1}| = \left|\sum_{i=1}^{k} (a_i - a_{k+i})\right| = |a_1 - a_{k+1}|$$

Since the sequence $<a_1, a_2,...>$ is monotonic and bounded, the sequence $<w_1, w_2,...>$ is also monotonic and bounded ; its terms are positive. By the monotonicity test (K2), the series $\Sigma(a_k - a_{k+1})$ is therefore absolutely convergent. Hence by (V4) the series $\Sigma\, a_n x_n$ is absolutely convergent.

(V6) The terms y_1 to y_{n_0} have no influence on the convergence of the series. Thus, consider a series $<x_1, x_2,...>$ with $0 \le |y_n| \le x_n$ for all $n \ge 1$, and hence $x_n \ge 0$. Since the convergent series Σx_n passes the Cauchy test (K1), for every real number $\varepsilon > 0$ there is a natural number k_0 such that

$$\begin{array}{lll} |x_n + x_{n+1} + ... + x_{m-1}| & < \varepsilon \text{ for all} & m \ge n \ge k_0 \\ x_n + x_{n+1} + ... + x_{m-1} & < \varepsilon & \\ |y_n| + |y_{n+1}| + ... + |y_{m-1}| & < \varepsilon \text{ for all} & m \ge n \ge k_0 \end{array}$$

Hence the series Σy_n is absolutely convergent.

Grouping in series : The sequence $<s_1, s_2,...>$ of the partial sums of a sequence $f : \mathbb{N} \to \mathbb{R}$ with $f(n) = x_n$ is formed according to the rule $s_0 := 0$ and $s_{n+1} = s_n + x_n$ for $n = 1,2,...$. In the partial sum $s_n = x_1 + x_2 + ... + x_n$, the additions are thus performed sequentally from left to right.

If terms of a sequence are grouped by parentheses, this defines a new sequence, and therefore also a new sequence of partial sums and a new series. For example, the grouping $(x_1 + x_2) + (x_3 + x_4) + ...$ leads to the sequence $<y_1, y_2,...>$ with $y_k = x_{2k-1} + x_{2k}$. The sequence $<y_1, y_2,...>$ is a mapping $g : \mathbb{N} \to \mathbb{R}$ which is not identical with the mapping $f : \mathbb{N} \to \mathbb{R}$ of the sequence $<x_1, x_2,...>$.

The convergence behavior of a series which is constructed from another series by grouping cannot in general be inferred from the convergence behavior of the original series. The following example shows that the series defined by the groupings $1 - 1 + 1 - 1 + ...$, $(1 - 1) + (1 - 1) + ...$ and $1 + (-1 + 1) + (-1 + 1) + ...$ have different convergence properties :

(1) The sequence $f : \mathbb{N} \to \mathbb{R}$ with the terms <1, –1, 1, –1,...> generates the divergent series <1, 0, 1, 0,...>.

(2) The sequence $g : \mathbb{N} \to \mathbb{R}$ with the terms <(1 – 1), (1 – 1),...> generates the convergent series <0, 0,...> with the limit 0.

(3) The sequence $h : \mathbb{N} \to \mathbb{R}$ with the terms <1, (– 1 + 1), (– 1 + 1),...> generates the convergent series <1, 1,...> with the limit 1.

If a convergent series Σx_n is used to construct a series Σy_n by grouping, then the latter is also convergent. The limits of the series coincide.

Proof : A series derived from a convergent series by grouping is convergent. Let a series Σx_n be convergent with limit a. For the sequence $< s_1, s_2, ... >$ of partial sums of this series :

$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} (n > n_0 \quad \Rightarrow \quad d(a, s_n) \; < \; \varepsilon)$$

Let a mapping $h : \mathbb{N} \to \mathbb{N}$ with $h(k) = n_k$ be strictly monotonically increasing with $n_1 = 0$. Then the numbers n_{k-1} and n_k define a segment of the sequence $< x_1, x_2, ... >$. Let the sum of the terms in this segment be a term y_k of a new series Σy_n. With $s_0 := 0$ and $k = 1,2,...$, it follows that

$$y_k := x_{n_{k-1}+1} + x_{n_{k-1}+2} + ... + x_{n_k} \quad = \quad s_{n_k} - s_{n_{k-1}}$$

Thus the sequence of partial sums of the sequence $< y_1, y_2, ... >$ is $< s_{n_1}, s_{n_2}, ... >$. By property (T1) in Section 5.10.2, this subsequence of the convergent sequence $< s_1, s_2, ... >$ converges to the same limit a.

Example 1 : Convergent series

$$\begin{array}{lllllllllllll}
e & = & \frac{1}{0!} & + & \frac{1}{1!} & + & \frac{1}{2!} & + & \frac{1}{3!} & + ... + & \frac{1}{n!} & + & ... \\
\frac{1}{e} & = & \frac{1}{0!} & - & \frac{1}{1!} & + & \frac{1}{2!} & - & \frac{1}{3!} & + ... \pm & \frac{1}{n!} & \mp & ... \\
2 & = & \frac{1}{1} & + & \frac{1}{2} & + & \frac{1}{4} & + & \frac{1}{8} & + ... + & \frac{1}{2^n} & + & ... \\
\frac{2}{3} & = & \frac{1}{1} & - & \frac{1}{2} & + & \frac{1}{4} & - & \frac{1}{8} & + ... \pm & \frac{1}{2^n} & \mp & ... \\
\ln 2 & = & \frac{1}{1} & - & \frac{1}{2} & + & \frac{1}{3} & - & \frac{1}{4} & + ... \pm & \frac{1}{n} & \mp & ... \\
\frac{\pi}{4} & = & \frac{1}{1} & - & \frac{1}{3} & + & \frac{1}{5} & - & \frac{1}{7} & + ... \pm & \frac{1}{2n-1} & \mp & ...
\end{array}$$

Example 2 : Geometric series

In Example 5 of Section 5.10.1, geometric sequences $f : \mathbb{N} \to \mathbb{R}$ with $f(n) = x_n$ are defined such that $x_{n+1} / x_n = c$. The partial sum s_n of a geometric sequence is given by

$$s_n = x_1(1 + c + c^2 + \ldots + c^{n-1}) = \frac{1-c^n}{1-c}\, x_1$$

The fixed ratio c determines the convergence of the series as follows :

$c \le -1$	:	The series diverges
$-1 < c < 1$	:	The series converges to $\lim\limits_{n\to\infty} \frac{1-c^n}{1-c} x_1 = \frac{x_1}{1-c}$
$c \ge 1$	:	The series converges improperly to ± ∞

Example 3 : Evaluation of alternating convergent series

Grouping is demonstrated using the example of the alternating harmonic sequence $< 1, -\frac{1}{2}, \frac{1}{3}, -\frac{1}{4}, \ldots >$, whose associated series passes the Leibniz test and hence converges. The exact limit of the series is $\ln 2 = 0.693147\ldots$. Parentheses are introduced in the partial sum as follows : $(1 - \frac{1}{2}) + (\frac{1}{3} - \frac{1}{4}) + \ldots$. Thus the partial sums of the following sequences are calculated :

$$f : \mathbb{N} \to \mathbb{R} \text{ with } f(n) = x_n = \frac{1}{n}(-1)^{n+1}$$

$$g : \mathbb{N} \to \mathbb{R} \text{ with } g(n) = y_n = \frac{1}{2n-1} - \frac{1}{2n} = \frac{1}{2n(2n-1)}$$

n	1	2	3	4	5
Σx_n	1.000000	0.500000	0.833333	0.583333	0.783333
Σy_n	0.500000	0.583333	0.616667	0.634524	0.645635

The series Σy_n converges monotonically, in contrast to the series Σx_n. There are other series for ln 2 which converge far more rapidly, for instance :

$$\ln x = 2\left[\frac{x-1}{x+1} + \frac{(x-1)^3}{3(x+1)^3} + \frac{(x-1)^5}{5(x+1)^5} + \ldots\right] \quad \text{with} \quad x > 0$$

n	1	2	3	4	5
Σz_n	0.666667	0.691358	0.693004	0.693135	0.693146

Convergence is further improved by using a known value e^z near 2. For example, let $2 = we^{0.1s}$ with $s \in \mathbb{N}$ and choose s such that $1 \le w < e^{0.1} = 1.105171$. This leads to $s = 6$ and $w = 1.097623$. With $\ln x = 0.6 + \ln w$, the series for ln w and ln x then yield in two steps :

n	1	2
ln w	0.093080	0.093147
ln x	0.693080	0.693147

5.10.4 NETS

Introduction : In the definition of the limit of a sequence $f : \mathbb{N} \rightarrow M$ in a metric space $(M ; d)$, the total ordering of the natural numbers $\mathbb{N}$ is used. The question arises whether a more general definition of convergence is possible involving a mapping $f : G \rightarrow M$ for which the set G is not necessarily totally ordered. As an example of such a definition, the convergence of nets is treated in this section. In a net, G is a directed set and $(M ; T)$ is a general topological space.

Directed set : A partially ordered set G is said to be directed if for any two elements $\alpha, \beta \in G$ there is a $\gamma \in G$ which is greater than or equal to α and β :

$$\text{G is directed} \quad :\Leftrightarrow \quad \bigwedge_{\alpha \in G} \ \bigwedge_{\beta \in G} \ \bigvee_{\gamma \in G} (\gamma \geq \alpha \quad \wedge \quad \gamma \geq \beta)$$

Net : A mapping f from a directed set G to the points of a topological space $(M ; T)$ is called a net in M. The images x_α of the mapping are called the terms of the net. The net is designated by $\{x_\alpha\}$. A sequence is a special case of a net, since the natural numbers $\mathbb{N}$ are a directed set.

$$f : G \rightarrow M \quad \text{with} \quad f(\alpha) = x_\alpha$$

Accumulation : A net $f : G \rightarrow M$ in a topological space $(M ; T)$ is said to accumulate in a set $A \subseteq M$ if for every element $\alpha \in G$ there is an element $\beta \geq \alpha$ in G such that $f(\beta) \in A$.

$$\text{f accumulates in A} \quad :\Leftrightarrow \quad \bigwedge_{\alpha \in G} \ \bigvee_{\beta \in G} (\beta \geq \alpha \quad \wedge \quad f(\beta) \in A)$$

Final segment : A net $f : G \rightarrow M$ in a topological space $(M ; T)$ is said to have a final segment in a set $A \subseteq M$ if there is $\alpha \in G$ such that $f(\beta) \in A$ for all $\beta \geq \alpha$.

$$\text{f has a final segment in A} \quad :\Leftrightarrow \quad \bigvee_{\alpha \in G} (\beta \geq \alpha \quad \Rightarrow \quad f(\beta) \in A)$$

Let H be the set of elements $\beta \in G$ for which $\beta \geq \alpha$ and $f(\beta) \in A$. Then the net $f_H : H \rightarrow M$ is called the final segment of f in A. A point $y \in M$ may occur more than once as a term of the final segment. However, by definition the point y is contained only once in the image E_α of the final segment.

$$E_\alpha := \{x_\beta = f(\beta) \mid \beta \geq \alpha\}$$

The final segments of a net have the following properties :

(E1) If a net hass final segments in two sets A and B, then the net has a final segment in the intersection $A \cap B$ of these sets.

(E2) If the intersection of two sets A and B is empty, then a net cannot have final segments in both A and B.

Proof : Properties of final segments

(E1) If a net $f : G \to M$ has a final segment in a set $A \subseteq M$, then there is an element $\alpha \in G$ such that $f(\delta) \in A$ for every $\delta \geq \alpha$. If the same net also has a final segment in the set $B \subseteq M$, then there is an element $\beta \in G$ such that $f(\varepsilon) \in B$ for every $\varepsilon \geq \beta$.

For $\alpha, \beta \in G$ the directed set G contains an element γ such that $\gamma \geq \alpha$ and $\gamma \geq \beta$. Since the directed set G is partially ordered, $\eta \geq \gamma$ and $\gamma \geq \alpha$ implies $\eta \geq \alpha$. Likewise, $\eta \geq \gamma$ and $\gamma \geq \beta$ implies $\eta \geq \beta$. Hence $f(\eta) \in A$ and $f(\eta) \in B$ for all $\eta \geq \gamma$. This implies $f(\eta) \in A \cap B$ for $\eta \geq \gamma$.

(E2) Let a net $f : G \to M$ haves final segments in the sets $A \subseteq M$ and $B \subseteq M$. Then by (E1) there is an element $\gamma \in G$ such that the term $f(\gamma)$ of the net lies both in A and in B. This contradicts the hypothesis that $A \cap B$ is empty. Hence, contrary to the assumption, the net does not have final segments in disjoint sets A and B.

Convergence : A net $f : G \to M$ is said to converge to a point x of the topological space (M ; T) if f has a final segment in every neighborhood $U_x \subseteq M$ of x. The point x is called a limit of the net f.

$$\text{f converges to x} \quad :\Leftrightarrow \quad \bigwedge_{U_x} \; \bigvee_{\alpha \in G} \; (E_\alpha \subseteq U_x)$$

Nets have the following convergence properties :

(K1) The closure of a set A is exactly the set of the limits of all convergent nets whose image lies in A.

(K2) A topological space is a Hausdorff space if and only if every convergent net has a unique limit.

Proof K1 : The closure of a set A is exactly the set of the limits of all convergent nets whose image lies in A.

(1) Let a net $f : G \to M$ with image $f(G) \subseteq A$ converge to a point $x \in M$. Then by the definition of convergence the net f has a final segment in every neighborhood U of x. Every neighborhood of x thus contains a point of A. Hence the limit x is a point of the closure H(A).

(2) Let the closure H(A) of a set $A \subseteq M$ be given. Let G be the set of open neighborhoods U, W,... of an arbitrary point $x \in H(A)$. The set G is ordered using the relation $\subseteq$, that is $U \geq W :\Leftrightarrow U \subseteq W$. From $U \cap W \subseteq U$ and $U \cap W \subseteq W$ it follows that $U \cap W \geq U$ and $U \cap W \geq W$; hence the set G is directed. Every intersection $U \cap W$ is an open neighborhood of x. By the definition of the closure, the intersection $U \cap A$ is not empty. Hence there is a net $f : G \to A$ with $f(U) = x_u$ and $x_u \in U \cap A$, and thus $x_u \in A$.

The net f has a final segment in every open neighborhood U of x, since $Z \geq U$ with $x \in Z \subseteq U$ implies $f(Z) = x_Z \in U$. Hence f converges to the point x. Since the choice of x in H(A) was arbitrary, every point $x \in H(A)$ is a limit of a net whose image lies in A.

Proof K2 : A topological space is a Hausdorff space if and only if every convergent net has a unique limit.

(1) Let the space (M ; T) be a Hausdorff space. Let a net $f : G \rightarrow M$ converge to two different points $x \neq y$ of M. Then by the definition of a Hausdorff space these points have disjoint neighborhoods $U_x \cap U_y = \emptyset$. By property (E2), the net f does not have final segments both in U_x and in U_y. Therefore, contrary to the assumption, the net does not converge to two different points.

(2) Assume that the space (M ; T) is not a Hausdorff space. Then there are points $x \neq y$ in M which cannot be separated by open sets. Let S_x and T_x be open neighborhoods of x, and let S_y and T_y be open neighborhoods of y. The cartesian product G of the neighborhood systems of x and y contains ordered pairs such as $\alpha := (S_x, S_y)$ and $\beta := (T_x, T_y)$. The set G is ordered by the relation $\geq$, that is $\beta \geq \alpha :\Leftrightarrow T_x \subseteq S_x \wedge T_y \subseteq S_y$. The intersection $T_x \cap T_y$ is not empty, since the points x and y cannot be separated by open sets. With the directed set G, there is thus a net $f : G \rightarrow M$ with $f(\beta) \in T_x \cap T_y$:

$$(\beta \geq \alpha \quad \Rightarrow \quad f(\beta) \in T_x \subseteq S_x \subseteq U_x) \quad \Rightarrow \quad \text{net f converges to x}$$

$$(\beta \geq \alpha \quad \Rightarrow \quad f(\beta) \in T_y \subseteq S_y \subseteq U_y) \quad \Rightarrow \quad \text{net f converges to y}$$

Hence the limit of a net is not unique in a space which is not a Hausdorff space.

Final mapping : Strictly monotonically increasing mappings are used to define subsequences in Section 5.10.2. Final mappings are now defined in order to transfer this concept to nets. A mapping $h : H \rightarrow G$ between directed sets G and H is said to be final if for every element $\alpha \in G$ there is an element $\beta \in H$ such that the images of all elements $\gamma \geq \beta$ of H are greater than or equal to α :

$$\text{h is a final mapping} \quad :\Leftrightarrow \quad \bigwedge_{\alpha \in G} \ \bigvee_{\beta \in H} \ (\gamma \geq \beta \quad \Rightarrow \quad h(\gamma) \geq \alpha)$$

Subnet : Let a mapping $f : G \rightarrow M$ be an arbitrary net in a topological space (M ; T). Let a mapping $h : H \rightarrow G$ be final. Then their composition $f \circ h$ is called a subnet of f.

$$f \circ h : H \rightarrow M \quad \text{with} \quad f \circ h(\gamma) = f(h(\gamma)) = f(\alpha_\gamma) \geq x_{\alpha_\gamma}$$

Convergent subnet : A net $f : G \rightarrow M$ possesses a subnet which converges to a limit $x \in M$ if and only if f accumulates in every neighborhood of x.

Proof : Convergent subnet

It is assumed that $f : G \rightarrow M$ is a net which accumulates in every neighborhood of a point $x \in M$, and it is shown that there is a final mapping $h : H \rightarrow G$, where the directed set H contains ordered pairs (α, A) with $\alpha \in G$, $x \in A$ and $f(\alpha) \in A$. Then the composition $f \circ h : H \rightarrow M$ is by definition a subnet. The subnet converges to the point x if every neighborhood of x contains a final segment of $f \circ h$.

(1) Let $S := \{A, B, ...\}$ be the set of neighborhoods of the point x. In the cartesian product $G \times S$, consider the subset H of ordered pairs (α, A) for which $f(\alpha) \in A$.

$$H := \{ (\alpha, A) \in G \times S \mid f(\alpha) \in A \}$$

(2) The set H is ordered by the relation $(\beta, B) \geq (\alpha, A) \;:\Leftrightarrow\; \beta \geq \alpha \;\wedge\; B \subseteq A$. For arbitrary elements (α, A) and (β, B) of H there is a $\delta \in G$ such that $f(\gamma) \in A \cap B$ for all $\gamma \geq \delta$, since by hypothesis the net accumulates in every neighborhood of x. In particular, the directed set G contains an element $\gamma \geq \alpha, \beta, \delta$, and $(\gamma, A \cap B)$ is an element of H with $(\gamma, A \cap B) \geq (\alpha, A), (\beta, B)$. Hence H is directed.

(3) The mapping $h : H \rightarrow G$ with $h(\alpha, A) = \alpha$ is final. For every $\alpha \in G$ the directed set H contains the element (α, M). By definition $(\beta, B) \geq (\alpha, M)$ implies $\beta \geq \alpha$, and hence $h(\beta, B) \geq \alpha$. Thus by definition h is final.

(4) Let A be a neighborhood of x. Since by hypothesis the net f accumulates in A, there is $\alpha \in G$ such that $f(\alpha) \in A$. Consider elements $(\beta, B) \geq (\alpha, A)$. According to the partial ordering of H, $B \subseteq A$. With $h(\beta, B) = \beta$, it follows that $f \circ h(\beta, B) = f(\beta)$. The definition of H implies $f(\beta) \in B$. Hence $f \circ h(\beta, B) \in A$ for all $(\beta, B) \geq (\alpha, A)$. Thus $f \circ h$ has a final segment in every neighborhood A of x : The subnet converges to the limit x.

$$f \circ h \text{ has a final segment in } A \quad \Leftrightarrow \quad \bigvee_{(\alpha, A)} ((\beta, B) \geq (\alpha, A) \;\Rightarrow\; f \circ h\,(\beta, B) \in A)$$

Conversely, let $x \in M$ be a limit of a subnet $f \circ h$. Then every neighborhood of x contains a final segment of $f \circ h$. Hence f accumulates in every neighborhood of x.

Universal net : A net $f : G \rightarrow M$ is said to be universal if for every subset $A \subseteq M$ it possesses a final segment either in A or in $M - A$. Universal nets have the following properties :

(U1) The composition $p \circ f$ of a mapping $p : M \rightarrow N$ with a universal net $f : G \rightarrow M$ is a universal net $p \circ f : G \rightarrow N$.

(U2) For every net there is a universal subnet.

Proof U1 : The composition of a mapping $p : M \to N$ with a universal net $f : G \to M$ is a universal net $p \circ f : G \to N$.

Let A be a subset of N with the preimage $p^{-1}(A)$ in M. Since the net f is universal, it possesses a final segment either in the preimage $p^{-1}(A)$ or in the complement $M - p^{-1}(A) \supseteq p^{-1}(N - A)$. The image of every point of a final segment in $p^{-1}(A)$ lies in A. The composition $p \circ f$ therefore has a final segment either in A or in $N - A$. Hence the net $p \circ f$ is universal.

Proof U2 : For every net there is a universal subnet.

(1) Let a mapping $f : G \to M$ be a net. Consider the set F of the subsets F_i of the power set P(M) which satisfy the following conditions :

(a) $A \in F_i \quad \Rightarrow \quad$ f accumulates in A

(b) $A, B \in F_i \quad \Rightarrow \quad A \cap B \in F_i$

An example of such a subset is furnished by $F_i = \{M\} \subseteq P(M)$. The set $F = \{F_1, F_2, ...\}$ is partially ordered by inclusion. Let $F' = \{F_{k_1}, F_{k_2}, ...\}$ be an arbitrary totally ordered subset of F, that is $F_{k_i} \subseteq F_{k_j} \vee F_{k_j} \subseteq F_{k_i}$. The union of the elements of F' satisfies the conditions (a) and (b) and is therefore an element of F, and hence an upper bound of F' in F.

By property (E4) of ordered sets in Section 4.6, F contains a maximal element F_0. Thus F_0 is not properly contained in any of the elements F_i. If a further element of P(M) is added to F_0, the result is a set which properly contains F_0 and is therefore not contained in F. Hence this set does not satisfy conditions (a) and (b).

(2) The set $Z_0 := \{(A, \alpha) \in F_0 \times G \mid f(\alpha) \in A\}$ is partially ordered by the relation $\geq$ as follows :

$$(B, \beta) \geq (A, \alpha) \quad :\Leftrightarrow \quad B \subseteq A \ \wedge \ \beta \geq \alpha$$

The set Z_0 is directed. In fact, if (A, α) and (B, β) are two arbitrary elements of Z_0, then (b) yields $C := A \cap B \in F_0$. For α and β the directed set G contains an element γ with $\gamma \geq \alpha$ and $\gamma \geq \beta$. From (a) and $C \in F_0$ it follows that f accumulates in C, so that there is a $\delta \geq \gamma$ with $f(\delta) \in C$. Thus the set Z_0 contains an element (C, δ) with $(C, \delta) \geq (A, \alpha)$ and $(C, \delta) \geq (B, \beta)$, and it is therefore directed.

The mapping $h : Z_0 \to G$ with $h(\alpha, A) = \alpha$ is final. In fact, if α is an arbitrary element of G, then for a freely chosen set $B \in F_0$ there is a $\beta \geq \alpha$ with $f(\beta) \in B$, since by (a) f accumulates in B. Hence (B, β) is an element of Z_0. For every $(C, \gamma) \in Z_0$ with $(C, \gamma) \geq (B, \beta)$ it follows that $h(C, \gamma) = \gamma \geq \beta \geq \alpha$. Hence h is final, and $f \circ h$ is a subnet. In the following this subnet is shown to be universal.

(3) Let S be a subset of M in which the subnet $f \circ h : Z_0 \to M$ accumulates. Then for every element $(\alpha, A) \in Z_0$ there is an element $(\beta, B) \geq (\alpha, A)$ in Z_0 such that $f \circ h(\beta, B) \in S$. Then $B \subseteq A$, $\beta \geq \alpha$ and $f(\beta) = f \circ h(\beta, B) \in B$ implies $f(\beta) \in B \cap S \subseteq A \cap S$, and hence the net $f : G \to M$ accumulates in $A \cap S$ for every $A \in F_0$. Thus an accumulation of the subnet $f \circ h$ in an arbitrary subset S of M corresponds to an accumulation of the net f in the intersection $A \cap S$ for every element $A \in F_0$. Since F_0 is maximal, S and every $A \cap S$ are elements of F_0.

(4) Assume that the subnet $f \circ h$ also accumulates in the complementary set $M - S$. Then $M - S$, like S, is an element of the set F_0. By hypothesis the intersection $S \cap (M - S) = \emptyset$ must also be an element of F_0. Since the net f does not accumulate in the empty set, it follows that, contrary to the assumption, the subnet $f \circ h$ does not accumulate in $M - S$. Hence the subnet $f \circ h$ has a final segment in S, but no final segment in $M - S$.

(5) If S is a subset of M in which the subnet $f \circ h : Z_0 \to M$ does not accumulate, then $f \circ h$ has a final segment in $M - S$. Altogether, it follows that $f \circ h$ has a final segment either in S or in $M - S$. Hence $f \circ h$ is a universal subnet of f.

5.10.5 FILTERS

Introduction : In this section the concept of convergence in a topological space (M ; T) is generalized further. Instead of sequences f : $\mathbb{N} \to M$ with totally ordered sets $\mathbb{N}$ and nets f : G → M with directed sets G, a filter F is now considered.

A filter F contains subsets of the underlying set M of the space. The filter converges to a point $x \in M$ if every neighborhood of x contains an element of the filter. A convergent filter is thus a contracting system of subsets of M which points to a point x of M. Very general search processes can be described using this concept.

In analogy with a topology T on M, a filter F on M has a basis B which is easier to deal with. In general, it is the filter basis which is used to study convergence. Compared with the concepts of a sequence and of a net, the concept of a filter offers the advantage that the filter basis B and the filter F need not be ordered sets.

Filter basis : A non-empty subset $B = \{B_1, B_2, ...\}$ of the power set of the underlying set M of a topological space (M ; T) is called a filter basis on M if the elements of B satisfy the following conditions :

(B1) The empty set ∅ is not an element of the filter basis B.

(B2) Every intersection of a finite number of elements of the filter basis B is again an element of B.

$$B_1, ..., B_m \in B \quad \Rightarrow \quad B_1 \cap \ ... \ \cap B_m \in B$$

The definition of a filter basis implies that a filter basis contains no disjoint sets. For by (B2) the intersection ∅ of the sets would be an element of the filter basis, contradicting (B1).

Convergence of a filter basis : Let $B = \{B_1, B_2, ...\}$ be a filter basis in the topological space (M ; T), and let x be a point of M with the neighborhood system $U(x) = \{U_1, U_2, ...\}$. The filter basis B is said to converge to the point x if every neighborhood U_k of x includes an element B_m of the filter basis as a subset.

$$\text{B converges to x} \quad :\Leftrightarrow \quad \bigwedge_{U_k \in U(x)} \ \bigvee_{B_m \in B} \ (B_m \subseteq U_k)$$

The point x is called a limit of the filter basis B. The limit x need not be contained in the filter sets B_i. A filter basis may converge to more than one point. In a Hausdorff space, a convergent filter basis has exactly one limit. Not all filter bases in a Hausdorff space are convergent. A filter basis which does not converge is said to diverge.

Proof : A topological space is a Hausdorff space if and only if every convergent filter basis has a unique limit.

(1) Let a topological space $(M\,;T)$ be a Hausdorff space, and let B be a convergent filter basis on M. Let x and y be two different limits of the filter basis B. Then by the definition of a Hausdorff space the points x and y have disjoint neighborhoods U_x and U_y. Since the filter basis B converges, there are filter elements $B_k \subseteq U_x$ and $B_m \subseteq U_y$. From $U_x \cap U_y = \emptyset$ it follows that $B_k \cap B_m = \emptyset$. This contradicts properties (B1) and (B2) of the filter basis. Hence the filter basis B converges to at most one point of M.

(2) Assume that a topological space $(M\,;T)$ is not a Hausdorff space. Then there are two different points $x, y \in M$ which cannot be separated by open sets. The set B_x of open sets T_x which contain x does not contain the empty set $\emptyset$, but it contains all finite intersections of its elements. Hence B_x is a filter basis which converges to x. Likewise, the open sets $T_y \in T$ with $y \in T_y$ form a filter basis B_y which converges to y. Since x and y cannot be separated by open sets, all intersections $T_{xy} = T_x \cap T_y$ are non-empty. They form a set B_{xy}. The union of the sets B_x, B_y and B_{xy} is a filter basis which converges to x and to y. Hence there is at least one filter basis $B_x \cup B_y \cup B_{xy}$ on M which converges to more than one limit.

Comparison of filter bases : The concept of the fineness of a filter basis corresponds to the concepts of subsequences and subnets. A filter basis B is said to be finer than a filter basis C if every element C_k of C includes an element B_m of B as a subset.

$$\text{B is finer than C} \quad :\Leftrightarrow \quad \bigwedge_{C_k \in C} \; \bigvee_{B_m \in B} \; (B_m \subseteq C_k)$$

If a filter basis C converges to a limit x, then every finer filter basis B also converges to the limit x.

Proof : Convergence of a finer filter basis

In a topological space $(M\,;T)$, let $x \in M$ be a limit of the filter basis C. Then by definition every neighborhood U_x of x contains an element C_m of the filter basis C. If the basis B is finer than C, then for every neighborhood U_x of x there is an element B_k of B with $B_k \subseteq C_m \subseteq U_x$. Thus the filter basis B also converges to the limit x.

Filter : Let $B = \{B_1, B_2, ...\}$ be a filter basis in the topological space $(M\,;T)$. Then the set of the elements B_k and all their non-empty unions is called a filter on M and is designated by $F = \{F_1, F_2, ...\}$. In contrast to a topology, a filter does not possess property (T1) of topologies : The empty set $\emptyset$ must not be an element of a filter F, and the underlying set M may or may not be an element of F. The topology T and a filter F on M thus contain different elements. Since every filter is also a filter basis, no further definition of the convergence of filters is required.

Example 1 : Uniquely convergent filter basis

Let $T = \{\emptyset, \{a\}, \{b, c\}, \{a, b, c\}\}$ be a topology and let $B = \{\{a\}, \{a, b\}, \{a, b, c\}\}$ be a filter basis on the set $M = \{a, b, c\}$. The neighborhoods of the points of M are determined using the condition $x \in T_k \subseteq U_m$:

$$\begin{aligned} U(a) &= \{\{a\}, \{a, b\}, \{a, c\}, \{a, b, c\}\} \\ U(b) &= \{\{b, c\}, \{a, b, c\}\} \\ U(c) &= \{\{b, c\}, \{a, b, c\}\} \end{aligned}$$

The filter basis B converges to the point a of M, since each of the neighborhoods in U(a) includes the filter element $\{a\}$. The filter basis B does not converge to $b \in M$ and $c \in M$, since the neighborhood $\{b, c\}$ of these points includes no element of the filter basis.

Example 2 : Convergent filter basis with several limits

Let $T = \{\emptyset, \{a\}, \{a, b\}, \{a, b, c\}\}$ be a topology and let $B = \{\{a\}\}$ be a filter basis on the set $M = \{a, b, c\}$. The neighborhoods of the points of M are determined using the condition $x \in T_k \subseteq U_m$:

$$\begin{aligned} U(a) &= \{\{a\}, \{a, b\}, \{a, c\}, \{a, b, c\}\} \\ U(b) &= \{\{a, b\}, \{a, b, c\}\} \\ U(c) &= \{\{a, b, c\}\} \end{aligned}$$

The filter basis converges to each of the points $a, b, c \in M$, since every neighborhood of these points includes the filter element $\{a\}$.

Example 3 : Divergent filter basis

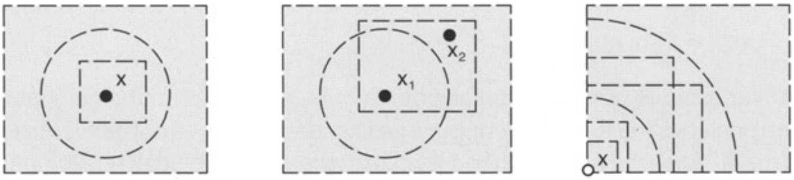
Let $T = \{\emptyset, \{c\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$ be a topology and let $B = \{\{a, b\}\}$ be a filter basis on the set $M = \{a, b, c\}$. The neighborhoods of the points of M are determined using the condition $x \in T_k \subseteq U_m$:

$$\begin{aligned} U(a) &= \{\{a, c\}, \{a, b, c\}\} \\ U(b) &= \{\{b, c\}, \{a, b, c\}\} \\ U(c) &= \{\{c\}, \{a, c\}, \{b, c\}, \{a, b, c\}\} \end{aligned}$$

The filter basis diverges, since the neighborhoods $\{a, c\}$ of a, $\{b, c\}$ of b and $\{c\}$ of c do not include an element of the filter basis as a subset.

Example 4 : Filter bases on the euclidean plane

The elements of the basis of the natural topology of the real euclidean plane $\mathbb{R}^2$ are the open disks $D(\mathbf{x}, r)$. Open rectangles with pairwise parallel edges are chosen as elements of the filter bases. The elements of the filter basis are said to be arbitrarily small if the distances of their corners determined by the metric of the euclidean plane are arbitrarily small.



a) convergent b) not convergent c) convergent

(a) If the intersection of all elements of the filter basis B contains only one point x for arbitrarily small elements, then the filter basis converges to x, since every open disk around x includes all open rectangles smaller than a certain size which contain x.

$\cap B_i = \{x\}$

(b) If the intersection of all elements of the filter basis B contains at least two points x_1 and x_2, then the filter basis does not converge : A neighborhood of x_1 which does not contain x_2 does not contain any element of the filter basis either.

$\cap B_i = \{x_1, x_2, ...\}$

(c) Although the point x in figure (c) is not contained in any of the arbitrarily small open elements of the filter basis B, the filter basis converges to $x \in \mathbb{R}^2$, since every neighborhood of x includes an element of the filter basis.

5.11 COMPACTNESS

5.11.1 COMPACT SPACES

Introduction : The number of elements of a finite set is a topological invariant. This invariant is a measure of the topological delimitation of the set. If a set contains an infinite number of points (for example in a real space), then the number of points is no longer suitable for characterizing the topological delimitation. The boundedness of a subset of a metric space cannot be used for such a characterization either, since boundedness is not a topological invariant (see Example 4 in Section 5.6).

Compactness of sets is defined in order to characterize the topological delimitation of infinite sets. For this purpose, coverings of a set with open sets of the topology of the space are considered. The set is compact if every open covering contains a finite subcovering. Compactness is a topological invariant. It may alternatively be defined in terms of closed sections of the set; this viewpoint is useful in proofs of convergence.

Compact spaces have important properties. In a compact space, every net possesses a convergent subnet. Every closed subset of a compact space is compact. Every compact subset of a Hausdorff space is closed. A finite product space is compact if and only if its factors are compact spaces.

Covering : Let A be a subset of the underlying set M of a topological space (M ; T). A family $C = \{C_i \mid C_i \subseteq M\}$ of sets is called a covering of the set A if the union of the sets C_i contains the set A.

$$C = \{C_i\} \text{ is a covering of } A \quad :\Leftrightarrow \quad A \subseteq C_1 \cup C_2 \cup ...$$

A covering is said to be finite if the family C of sets is finite. An infinite covering is designated by $\{C_1, C_2, ...\}$, a finite covering by $\{C_1, ..., C_n\}$. A covering is said to be open if each of the sets $C_i \in C$ is open, that is if $C_i \in T$. A subset of a covering which is also a covering of the set is called a subcovering.

Section : Let A be a subset of the underlying set M of a topological space (M ; T). A family $S = \{S_i \mid S_i \subseteq M\}$ of sets is called a section of A if A contains the intersection of the sets S_i.

$$S = \{S_i\} \text{ is a section of } A \quad :\Leftrightarrow \quad A \supseteq S_1 \cap S_2 \cap ...$$

A section is said to be finite if the family S of sets is finite. An infinite section is designated by $\{S_1, S_2, ...\}$, a finite section by $\{S_1, ..., S_n\}$. A section is said to be empty if the intersection $S_1 \cap S_2 \cap ...$ is empty. A section is said to be closed if the sets S_i are closed, that is if $M - S_i \in T$. A subset of a section which is also a section of A is called a subsection. The intersection of the sets of a subsection contains the intersection of the sets of the section.

Compact sets : A set A in a topological space (M ; T) is said to be compact if every open covering $C = \{C_i\}$ of A contains a finite subcovering $\hat{C} = \{C_{i_m}\}$.

$$\text{A is compact} \quad :\Leftrightarrow \quad \bigwedge_{C} \bigvee_{\hat{C}} (A \subseteq C_1 \cup C_2 \cup \dots \quad \Rightarrow \quad A \subseteq C_{i_1} \cup \dots \cup C_{i_n})$$

For a set to be compact, it is not sufficient that it possesses a finite open covering, for example the covering containing only the underlying set M. Rather, every open covering of the set must contain a finite open subcovering. To show that a set is not compact, it suffices to exhibit a single open covering which does not contain a finite open subcovering.

Compact space : A topological space (M ; T) is said to be compact if every open covering of M contains a finite open subcovering. If the set A is compact in the space M, then the space $(A ; R_A)$ with the relative topology $R_A = \{R_i = T_i \cap A \mid T_i \in T\}$ is compact. In contrast to some of the literature, a compact space is not assumed to be a Hausdorff space.

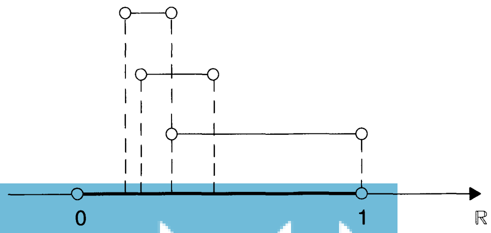
Continuum : A set A in a Hausdorff space (M ; T) is called a continuum if it is compact and connected. A part of physics deals with continua associated with matter, charge and energy.

Example 1 : Open covering of the open unit interval

Consider the open unit interval $\mathrm{I} = \,]0,1[$ on the real axis $\mathbb{R}$ with the natural topology. This unit interval has an open covering $\{C_1, C_2, \dots\}$ with the following sets :

$$C_k = \{x \in \mathbb{R} \mid \frac{1}{k+2} < x < \frac{1}{k} \quad \wedge \quad k \in \mathbb{N}'\}$$

This open covering contains no finite subcovering, since C_{k+1} is not contained in $C_1 \cup \dots \cup C_k$. Thus the open unit interval is not compact.



$C_3 : \frac{1}{5} < x < \frac{1}{3}$

$C_2 : \frac{1}{4} < x < \frac{1}{2}$

$C_1 : \frac{1}{3} < x < 1$

$\mathrm{I} : 0 < x < 1$

Example 2 : The closed unit interval is compact.

Let $C = \{C_1, C_2, ...\}$ be an open covering of the closed unit interval $I_1 = [0,1] \in \mathbb{R}$. Assume that C does not contain a finite subcovering. Then the subinterval $[0, \frac{1}{2}]$ or the subinterval $[\frac{1}{2}, 1]$ does not contain a finite open subcovering in C. Repeated bisection of the intervals leads to a sequence of nested intervals $I_1 \supset I_2 \supset ...$ for which there is no finite open subcovering in C.

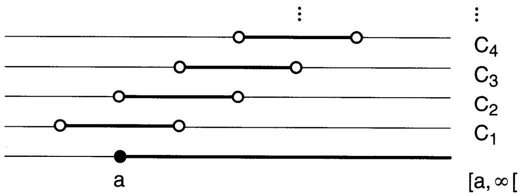
There is a point $w \in \mathbb{R}$ which is contained in each of the nested intervals. In the open covering C there is an open set $C_i = \{x \in \mathbb{R} \mid s_i < x < t_i\}$ which also contains the point w. Since the length of the nested intervals tends to zero due to the repeated bisection, there is a closed interval $I_n = [a_n, b_n]$ which is entirely contained in the open set C_i.

$$x \in I_n \subset C_i$$

Thus the closed interval I_n possesses the finite open covering C_i. This contradicts the construction of the nested intervals $I_1 \supset I_2 \supset ...$. Thus, contrary to the assumption, the open covering C of the closed unit interval I_1 contains a finite subcovering. Hence I_1 is compact.

Example 3 : Unbounded interval on $\mathbb{R}$

An unbounded interval $[a, \infty[$ on the real axis $\mathbb{R}$ with the natural topology is not compact. For example, the open covering $C = \{C_1, C_2, ...\}$ of $[a, \infty[$ with $C_i = \{x \in \mathbb{R} \mid a + i - 2 < x < a + i\}$ does not contain a finite subcovering of $[a, \infty[$.



Properties of compact sets and spaces

(K1) A topological space is compact if and only if every empty closed section $\{S_1, S_2, ...\}$ of the space contains an empty finite subsection $\{S_{i_1}, ..., S_{i_n}\}$.

(K2) A topological space is compact if and only if every closed section $\{S_1, S_2, ...\}$ which contains only non-empty finite subsections $\{S_{i_1}, ..., S_{i_n}\}$ is non-empty.

(K3) Every closed subset of a compact space is compact.

(K4) Every compact subset of a Hausdorff space is closed.

(K5) The image of a compact space under a continuous mapping is compact.

(K6) A topological space $(M;T)$ is compact if and only if every universal net in M converges.

(K7) The projection $\pi : M \times N \rightarrow N$ is closed if the space M is compact.

(K8) A net in a product space $M = M_1 \times ... \times M_n$ converges to a point x if and only if for every i its composition with the projection $p_i : M \rightarrow M_i$ converges to the i-th coordinate of x.

(K9) A product space is compact if its factors are compact spaces.

(K10) Every compact Hausdorff space is regular.

(K11) Every compact Hausdorff space is normal.

(K12) The union of a finite number of compact sets is compact.

(K13) The intersection of an infinite number of compact subsets of a compact Hausdorff space is a compact set.

Proof K1 : A topological space is compact if and only if every empty closed section $\{S_1, S_2, ...\}$ of the space contains an empty finite subsection $\{S_{i_1}, ..., S_{i_n}\}$.

(1) Let the topological space $(M;T)$ be compact. Let $\{S_1, S_2, ...\}$ be an empty closed section of M. The complement of the closed set S_i is the open set $C_i = M - S_i$. Since the section $\{S_1, S_2, ...\}$ is empty, $\{C_1, C_2, ...\}$ is an open covering of M. Since the space $(M;T)$ is compact, this open covering contains a finite subcovering $\{C_{i_1}, ..., C_{i_n}\}$. The complements $S_{i_1}, ..., S_{i_n}$ of these finitely many open sets form an empty finite closed subsection of $\{S_1, S_2, ...\}$.

$$\begin{aligned}
S_1 \cap S_2 \cap ... = \emptyset \quad & \Rightarrow \quad (M - C_1) \cap (M - C_2) \cap ... = \emptyset \\
& \Rightarrow \quad M - (C_1 \cup C_2 \cup ...) = \emptyset \\
& \Rightarrow \quad M = C_1 \cup C_2 \cup ... \\
& \Rightarrow \quad M = C_{i_1} \cup ... \cup C_{i_n} \\
& \Rightarrow \quad M = (M - S_{i_1}) \cup ... \cup (M - S_{i_n}) \\
& \Rightarrow \quad M = M - (S_{i_1} \cap ... \cap S_{i_n}) \\
& \Rightarrow \quad S_{i_1} \cap ... \cap S_{i_n} = \emptyset
\end{aligned}$$

Thus in a compact space every empty closed section $\{S_1, S_2, ...\}$ contains an empty finite subsection $\{S_{i_1}, ..., S_{i_n}\}$.

(2) Let every empty closed section $\{S_1, S_2, ...\}$ of a topological space $(M;T)$ contain an empty finite subsection $\{S_{i_1}, ..., S_{i_n}\}$. Let $C = \{C_1, C_2, ...\}$ be an arbitrary covering of M. The complement of the open set C_i is the closed set $S_i = M - C_i$. It follows that the section $\{S_1, S_2, ...\}$ of M is empty. By hypothesis, $\{S_1, S_2, ...\}$ therefore contains an empty finite subsection $\{S_{i_1}, ..., S_{i_n}\}$. The

complements $C_{i_1},...,C_{i_n}$ of these finitely many closed sets form a finite open subcovering of $\{C_1, C_2,...\}$.

$$\begin{aligned} C_1 \cup C_2 \cup ... = M \quad & \Rightarrow \quad (M - S_1) \cup (M - S_2) \cup ... = M \\ & \Rightarrow \quad M - (S_1 \cap S_2 \cap ...) = M \\ & \Rightarrow \quad S_1 \cap S_2 \cap ... = \emptyset \\ & \Rightarrow \quad S_{i_1} \cap ... \cap S_{i_n} = \emptyset \\ & \Rightarrow \quad (M - C_{i_1}) \cap ... \cap (M - C_{i_n}) = \emptyset \\ & \Rightarrow \quad M - (C_{i_1} \cup ... \cup C_{i_n}) = \emptyset \\ & \Rightarrow \quad C_{i_1} \cup ... \cup C_{i_n} = M \end{aligned}$$

Since every open covering of M contains a finite open subcovering, M is a compact space.

Proof K2 : A topological space $(M ; T)$ is compact if and only if every closed section $\{S_1, S_2,...\}$ which contains only non-empty finite subsections $\{S_{i_1},...,S_{i_n}\}$ is non-empty.

(1) Let a topological space $(M ; T)$ be compact. Let a closed section $\{S_1, S_2,...\}$ in M which contains only non-empty finite subsections be empty. This contradicts the statement (K1) that every empty closed section $\{S_1, S_2,...\}$ in a compact space contains an empty finite subsection. Hence, contrary to the assumption, $\{S_1, S_2,...\}$ is non-empty.

(2) Let every closed section which contains only non-empty finite subsections be non-empty. Assume that the topological space $(M ; T)$ is not compact. Then by (K1) there is an empty closed section $\{S_1, S_2,...\}$ which contains only non-empty finite subsections. This contradicts the hypothesis. Hence, contrary to the assumption, the space is compact.

Proof K3 : Every closed subset of a compact space is compact.

Let a topological space $(M ; T)$ be compact. Let an open covering $\{C_i\}$ of a closed subset $A \subseteq M$ with $A \subseteq C_1 \cup C_2 \cup ...$ and $C_i \in T$ be given. Since by hypothesis the set A is closed, the complement $M - A$ is open. The union $\{M - A\} \cup \{C_1, C_2,...\}$ with $M \subseteq (M - A) \cup C_1 \cup C_2 \cup ...$ is an open covering of M. Every such covering of the compact space M contains a finite open subcovering $\{M - A\} \cup \{C_{i_1},...,C_{i_n}\}$ with $M \subseteq (M - A) \cup C_{i_1} \cup ... \cup C_{i_n}$. Then every open covering $\{C_1, C_2,...\}$ of A also contains a finite open subcovering $\{C_{i_1},...,C_{i_n}\}$, and hence the closed set A is compact.

Proof K4 : Every compact subset of a Hausdorff space is closed.

For a fixed point $x \in M - A$ and every point $a \in A$ the Hausdorff space contains disjoint open sets T_a and C_a with $x \in T_a$ and $a \in C_a$. The open sets C_a form a covering of A, since every point $a \in A$ is contained at least in C_a. Since A is compact, this covering contains an open subcovering $\{C_{i_1},...,C_{i_n}\}$ with a corresponding set $\{T_{i_1},...,T_{i_n}\}$. The intersection of the finitely many open sets $T_{i_1},...,T_{i_n}$ is an open set which contains x and is disjoint from A. Therefore no $x \in M - A$ is an accumulation point of A, and hence A is a closed set.

Proof K5 : The image of a compact space under a continuous mapping is a compact set.

Let the space $(M;T)$ be compact. Let the mapping $f : M \to N$ to the space $(N;S)$ be continuous, and let $\{B_1,B_2,...\}$ with $f(M) \subseteq B_1 \cup B_2 \cup ...$ be an open covering of the image f(M). The preimages $C_i = f^{-1}(B_i)$ are open sets, since the mapping f is continuous. Hence they form an open covering $\{C_1,C_2,...\}$ of M. Since M is compact, $\{C_1,C_2,...\}$ contains a finite subcovering $\{C_{i_1},...,C_{i_n}\}$. Its image $\{B_{i_1},...,B_{i_n}\}$ is a finite subcovering of f(M) which is contained in the open covering $\{B_1,B_2,...\}$. Hence the image f(M) is compact. If the mapping f is surjective, the space $N = f(M)$ is compact.

Proof K6 : A topological space $(M;T)$ is compact if and only if every universal net in M converges.

(1) Let the space $(M;T)$ be compact. Assume that a universal net $f : G \to M$ does not converge. Then for every point $x_s \in M$ there is an open neighborhood $C_s \subseteq M$ such that the net f does not have a final segment in C_s. By definition this implies that the universal net has a final segment in $M - C_s$. Hence there is $\alpha_s \in G$ such that $f(\beta) \in M - C_s$ for all $\beta \ge \alpha_s$, and $f(\beta) \in M - C_s$ implies $f(\beta) \notin C_s$.

Traversing the points of M yields an open covering $\{C_1,C_2,...\}$ of M with $f(\beta) \notin C_s$ for $\beta \ge \alpha_s$. Since the space is compact, this covering contains a finite open covering $\{C_{i_1},...,C_{i_n}\}$ of M. By the definition of the directed set G there is $\gamma \in G$ with $\gamma \ge \alpha_{i_1},...,\alpha_{i_s}$. Then $f(\beta) \notin C_s$ for $\beta \ge \alpha_s$ and $\gamma \ge \alpha_s$ implies $f(\beta) \notin C_s$ for $\beta \ge \gamma$. With $M \subseteq C_{i_1} \cup ... \cup C_{i_n}$ this also implies $f(\gamma) \notin M$. By the definition of the net $f : G \to M$, however, $f(\gamma) \in M$. The contradiction shows that, contrary to the assumption, every universal net in M converges.

(2) Let every universal net in M be convergent. Property (K2) of compact sets is used to show that in this case the set M is compact.

Let $S = \{S_1, S_2,...\}$ be a closed section of M. Let every finite subsection of S be non-empty. Without loss of generality it is assumed that together with any two sets S_i, $S_m \in S$ their intersection $S_i \cap S_m$ is also contained in S. The set

S is ordered by inclusion, that is $S_i \geq S_k \;:\Leftrightarrow\; S_i \subseteq S_k$. The set S is directed, since for arbitrary elements S_i and S_k it contains the element $S_m = S_i \cap S_k$ with $S_m \geq S_i$ and $S_m \geq S_k$.

A mapping $f : S \to M$ with $f(S_i) = x_i \in S_i$ is a net. By property (U2) in Section 5.10.4 f possesses a universal subnet $f \circ h : H \to M$ with a final mapping $h : H \to S$ with $h(\alpha) = S_\alpha$ and $f \circ h(\alpha) = f(S_\alpha) = x_\alpha \in S_\alpha$. Since h is final, for every $S_\alpha \in S$ there is a $\beta \in H$ such that $h(\gamma) \geq S_\alpha$ for all $\gamma \geq \beta$, and hence $S_\gamma \subseteq S_\alpha$ and $x_\gamma \in S_\alpha$ for all $\gamma \geq \beta$.

By hypothesis the universal net $f \circ h$ converges to a limit x. Since by hypothesis the set S_α is closed, by property (K1) in Section 5.10.4 the limit x is a point of S_α. But $x \in S_\alpha$ for all α implies $x \in \bigcap_\alpha S_\alpha$. Hence the infinite intersection $S_1 \cap S_2 \cap \ldots$ is non-empty, and then property (K2) in the present section implies that the set M is compact.

Proof K7 : The projection $\pi : M \times N \to N$ is closed if the space M is compact.

Let $(M \times N\, ; P)$ be the product space of the topological spaces $(M\, ; T)$ and $(N\, ; S)$. By definition, the mapping $\pi : M \times N \to N$ is closed if the image $\pi(A)$ of every closed subset $A \subseteq M \times N$ is a closed set in $(N\, ; S)$. Thus every $y \in N - \pi(A)$ must have an open neighborhood $V \subseteq N - \pi(A)$.

Since the set A is closed, $M \times N - A$ is open. Fix a point $y \in N - \pi(A)$, and then consider the point $(x, y) \in M \times (N - \pi(A)) \subseteq M \times N - A$ for every point $x \in M$. This point possesses an open neighborhood $T_x \times S_x$ with $x \in T_x \in T$, $y \in S_x \in S$ and $(T_x \times S_x) \cap A = \emptyset$. The sets T_x for all $x \in M$ form an open covering of M, since every point of M is contained in at least one of these sets. Since the set M is compact, this covering contains a finite open subcovering $\{T_{x_1}, \ldots, T_{x_n}\}$. The intersection $V := S_{x_1} \cap \ldots \cap S_{x_n}$ of the corresponding open sets of N is open and contains the point y, that is $y \in V$. Using $(T_{x_i} \times S_{x_i}) \cap A = \emptyset$ one obtains :

$$\begin{aligned} (M \times V) \cap A &= ((T_{x_1} \cup \ldots \cup T_{x_n}) \times (S_{x_1} \cap \ldots \cap S_{x_n})) \;\; \cap A \\ &= ((T_{x_1} \times S_{x_1}) \cap A) \cap \ldots \cap ((T_{x_1} \times S_{x_n}) \cap A) \cup \ldots \cup \\ &\quad\;\; ((T_{x_n} \times S_{x_1}) \cap A) \cap \ldots \cap ((T_{x_n} \times S_{x_n}) \cap A) \\ &= \emptyset \end{aligned}$$

But $(M \times V) \cap A = \emptyset$ implies $\pi(M \times V \cup A) \subseteq \pi(M \times V) \cap \pi(A) = V \cap \pi(A) = \emptyset$, and hence $V \subseteq N - \pi(A)$. Thus every point $y \in N - \pi(A)$ possesses an open neighborhood V in $N - \pi(A)$. The set $N - \pi(A)$ is therefore open, and hence the set $\pi(A)$ is closed.

Proof K8 : A net in a product space $M := M_1 \times ... \times M_n$ converges to a point x if and only if for every i its composition with the projection $p_i : M \rightarrow M_i$ converges to the i-th coordinate of x.

(1) Let the net $f : G \rightarrow M$ with $f(\alpha) = x_\alpha = (x_{\alpha 1},...,x_{\alpha n})$ converge to the point $x = (x_1,...,x_n)$. The composition $p_i \circ f$ of the net with the projection $p_i : M \rightarrow M_i$ yields the net $f_i : G \rightarrow M_i$ with $f_i(\alpha) = p_i \circ f(\alpha) = p_i((x_{\alpha 1},...,x_{\alpha n})) = x_{\alpha i}$. Let U_{xi} be an open neighborhood of $x_{\alpha i}$. By virtue of the continuity of the projection p_i, the preimage U_x of U_{xi} is an open neighborhood of the point x_α, that is $U_{xi} = p_i(U_x)$. Since the net converges to x, f has a final segment in the neighborhood U_x of x : There is an $\alpha \in G$ such that $f(\beta) = x_\beta \in U_x$ for all $\beta \geq \alpha$ in G. The final segment of f in U_x corresponds to a final segment of f_i in U_{xi} with $f_i(\beta) \in U_{xi}$ for $\beta \geq \alpha$. Hence f_i converges to x_i.

(2) Let each of the nets $f_i : G \rightarrow M_i$ converge to a point $x_i \in M_i$. Then in every neighborhood U_{xi} of x_i the net f_i has a final segment with $f_i(\beta) \in U_{xi}$ for $\beta \geq \alpha_i$. The mapping $f : G \rightarrow M$ with $f(\alpha) = (x_{\alpha 1}, x_{\alpha 2},...)$ and $x_{\alpha i} = f_i(\alpha)$ is a net. Every neighborhood of a point $x = (x_1, x_2,...)$ in M contains a basis element $U_x = U_{x1} \times U_{x2} \times ...$ of the product space. For U_x the directed set G contains an $\alpha \geq \alpha_1, \alpha_2,...$ such that $f_i(\beta) \in U_{xi}$ for all $\beta \geq \alpha$, that is $f(\beta) \in U_x$ for $\beta \geq \alpha$. Hence the net f converges to the point x of M.

Proof K9 : A product space is compact if its factors are compact spaces.

Let a space M be the product $M_1 \times ... \times M_n$ of compact spaces M_i. Let the mapping $f : G \rightarrow M$ be a universal net, and let the mapping $p_i : M \rightarrow M_i$ be a projection. Then by property (U1) of universal nets in Section 5.10.4 the composition $p_i \circ f$ is also a universal net. By property (K6) the universal net $p_i \circ f$ converges, since the space M_i is by hypothesis compact. Let its limit be x_i. Then by property (K8) the net f converges to a point x whose i-th coordinate is x_i. Since the universal net $f : G \rightarrow M$ is arbitrary, every universal net in M is convergent. By property (K6) the space M is compact.

Proof K10 : Every compact Hausdorff space is regular.

Let C be a closed subset of a compact Hausdorff space (M ; T). For arbitrary points $x \in C$ and $y \in M - C$ the Hausdorff space M contains disjoint open neighborhoods T_x and S_x such that $x \in T_x$, $y \in S_x$ and $T_x \cap S_x = \emptyset$.

The closed subset C of the compact space M is compact by (K3). Hence there are points $x_1,...,x_n \in C$ such that $\{T_{x_1},...,T_{x_n}\}$ is a finite open covering of C. Let $T_y = T_{x_1} \cup ... \cup T_{x_n}$ and $S_y = S_{x_1} \cap ... \cap S_{x_n}$. Then $y \in S_y$, $C \subseteq T_y$ and $T_y \cap S_y = \emptyset$. Hence M has the properties of a regular space.

Proof K11 : Every compact Hausdorff space is normal.

Let C_1 be a closed subset of a compact Hausdorff space $(M ; T)$. Since the space M is regular by (K10), for an arbitrary point $x \in C_1$ and a closed subset $C_2 \subset M - C_1$ there are disjoint open sets T_x and S_x such that $x \in T_x$, $C_2 \subset S_x$ and $T_x \cap S_x = \emptyset$.

The closed subset C_1 of the compact space M is compact by (K3). Hence there are points $x_1,...,x_n \in C_1$ such that $\{T_{x_1},...,T_{x_n}\}$ is a finite open covering of C_1. Let $T_c = T_{x_1} \cup ... \cup T_{x_n}$ and $S_c = S_{x_1} \cap ... \cap S_{x_n}$. Then $C_1 \subseteq T_c$, $C_2 \subseteq S_c$ and $T_c \cap S_c = \emptyset$. Hence M has the properties of a normal space.

Proof K12 : The union of a finite number of compact sets is compact.

Let the sets $M_1,...,M_n$ be compact, and let their union be $M = M_1 \cup ... \cup M_n$. Let $C = \{C_1, C_2,...\}$ be an arbitrary open covering of M, and let the subset of C which forms an open covering of M_i be $A_i = \{A_{i1}, A_{i2},...\}$. Since M_i is compact, A_i contains a finite subcovering $E_i = \{A_{ik_1},..., A_{ik_s}\}$. The union $E = E_1 \cup ... \cup E_n$ is a finite subset of C which covers M. Since an arbitrary open covering C of M contains a finite subcovering E, the set M is compact.

Proof K13 : The intersection of an infinite number of compact subsets of a Hausdorff space is a compact set.

Let the subsets M_1, M_2,... of a Hausdorff space $(M ; T)$ be compact. Then by (K4) the sets M_1, M_2,... are closed. By (M6) in Section 5.2, the intersection of an infinite number of closed sets is a closed set $A = M_1 \cap M_2 \cap ...$. By (K3), the closed subset A of the compact set M_1 is a compact set.

Example 4 : The compact unit cube in $\mathbb{R}^n$

The point set $\mathrm{I}^n = \{(x_1, ..., x_n) \in \mathbb{R}^n \mid 0 \le x_i \le 1\}$ with the natural relative topology of $\mathbb{R}^n$ is called the unit cube. The unit cube is the n-fold product of the closed unit interval I^1, that is $\mathrm{I}^n = \mathrm{I}^1 \times ... \times \mathrm{I}^1$ (n-fold). In Example 2 the unit interval I^1 is shown to be compact. It follows by property (K9) of compact sets that the product space I^n is compact.

5.11.2 COMPACT METRIC SPACES

Introduction : In complete metric spaces, defined in Section 5.10.1, every fundamental sequence has a limit. In Section 5.11.1, every net in a compact space is shown to possess a convergent subnet. The question arises whether there is a connection between completeness and compactness of a metric space.

Every sequence in a metric space is a net, every subsequence is a subnet. By contrast, a net $n : G \rightarrow M$ is generally not a sequence. Thus the existence of convergent subnets in compact spaces does not necessarily imply the existence of convergent subsequences in compact metric spaces. A connection between completeness and compactness can only be drawn in totally bounded metric spaces.

Sequences and nets : A sequence $f : \mathbb{N} \rightarrow M$ is a mapping from the natural numbers $\mathbb{N}$, a net $n : G \rightarrow M$ is a mapping from a directed set G. The natural numbers $\mathbb{N}$ are a directed set. Thus every sequence is a net. A directed set G generally does not possess the order structure of the natural numbers $\mathbb{N}$. Thus not every net is a sequence. The properties of sequences can therefore generally not be transferred to nets.

Bounded set : A subset of a real space $\mathbb{R}^n$ is said to be bounded if it is contained in an open cuboid Q_n.

$$Q_n := \{(x_1,...,x_n) \in \mathbb{R}^n \mid x_i \in \ ]a,b[\quad \wedge \quad a,b \in \mathbb{R}\}$$

Totally bounded metric space : The completeness of a metric space does not imply that every open covering of the space contains a finite subcovering. Hence a complete metric space is not generally compact. To establish a connection between completeness and compactness, the concept of a totally bounded metric space is defined. A metric space (M ; d) is said to be totally bounded if for every real number $\varepsilon > 0$ the set M can be covered with a finite number of ε-balls.

Properties of compact metric spaces : For a metric space, the following statements are equivalent :

- The space is compact.
- The space is complete and totally bounded.
- Every sequence in the space contains a convergent subsequence.

This equivalence follows from a subset of the following statements :

(M1) In a compact metric space every sequence contains a convergent subsequence.

(M2) If every fundamental sequence in a metric space contains a convergent subsequence, then the space is complete.

(M3) If every sequence in a metric space contains a convergent subsequence, then the space is totally bounded.

(M4) Let a metric space be complete and totally bounded. Then every sequence contains a convergent subsequence.

(M5) Let a metric space be complete and totally bounded. Then the space is compact.

(M6) A subspace of a real space $(\mathbb{R}^n ; d)$ is compact if and only if it is closed and bounded.

Proof M1 : In a compact metric space every sequence contains a convergent subsequence.

Let a metric space $(M ; d)$ be compact, and let $f : \mathbb{N} \to M$ be a sequence which contains no convergent subsequence. Then every point $x \in M$ has an open neighborhood U_x which contains only a finite number of terms of the sequence f, since x is not an accumulation point of the sequence. The compact space M can be covered by a finite number of these neighborhoods U_x. Hence the sequence f is finite. Since this contradicts the definition of a sequence, it follows that, contrary to the assumption, the sequence contains a convergent subsequence.

Proof M2 : If every fundamental sequence in a metric space contains a convergent subsequence, then the space is complete.

Let the sequence $f : \mathbb{N} \to M$ be a fundamental sequence in a metric space $(M ; d)$. Then for every real number $\varepsilon > 0$ there is a natural number n_0 such that $d(x_i, x_m) < 0.5\,\varepsilon$ for $i,m \geq n_0$. Since f contains a subsequence with limit a, there is a natural number $s \geq n_0$ such that all terms x_n with $n \geq n_0$ as well as the limit a lie in the open ball $D(x_s, 0.5\,\varepsilon)$:

$$d(x_n, x_s) < 0.5\,\varepsilon \quad \wedge \quad d(a, x_s) < 0.5\,\varepsilon$$

For all $n \geq n_0$, property (M4) of a metric implies that the terms x_n lie in the open ball $D(a, \varepsilon)$. Thus f converges to the point a. Hence the space is complete.

$$d(a, x_n) \leq d(a, x_s) + d(x_n, x_s) < \varepsilon$$

Proof M3 : If every sequence in a metric space contains a convergent subsequence, then the space is totally bounded.

Let every subsequence in a metric space contain a convergent subsequence. Assume that the space is not totally bounded. Then by definition there is a real number $\varepsilon > 0$ such that M cannot be covered by a finite number of ε-balls. For n given points $x_1,...,x_n \in M$, a point x_{n+1} may therefore be chosen which does not lie in the union $D(x_1, \varepsilon) \cup ... \cup D(x_n, \varepsilon)$. In this way, a sequence $f : \mathbb{N} \to M$ with $f(k) = x_k$ is constructed such that $d(x_i, x_j) \ge \varepsilon$ for all i, j. In contradiction to the hypothesis, the sequence f does not contain a convergent subsequence. Thus contrary to the assumption the space M is totally bounded.

Proof M4 : Let a metric space be complete and totally bounded. Then every sequence contains a convergent subsequence.

Let $f : \mathbb{N} \to M$ be an arbitrary sequence in a metric space (M ; d). Since M is totally bounded, there is a finite open covering of M with 1-balls $D(x_i, 1)$. One of these 1-balls contains an infinite number of terms of the sequence f. Let this 1-ball be B_1.

If M is now covered with a finite number of $\frac{1}{2}$-balls, there is a $\frac{1}{2}$-ball B_2 such that $B_1 \cap B_2$ contains an infinite number of terms of the sequence f. By continuing this process for $n = 1, 2, 3,...$, one obtains an $\frac{1}{n}$-ball B_n such that $B_1 \cap ... \cap B_n$ contains an infinite number of terms of the sequence f.

For every radius $\frac{1}{n}$ with $n \in \mathbb{N}$, there is thus at least one set $B_1 \cap ... \cap B_n$ which contains an infinite number of terms of the sequence f. In each of these sets, a term of the sequence f is chosen and designated by x_n. Then $h := <x_1, x_2,...>$ is a subsequence of f. The subsequence h is fundamental. For every $\varepsilon > 0$ there is an $n_0 \in \mathbb{N}$ with $\frac{1}{n_0} < \frac{\varepsilon}{2}$. Let the center of B_{n_0} be a. Then

$$\bigwedge_{i,k \ge n_0} |x_i - x_k| \quad \le \quad |x_i - a| + |x_k - a| \quad < \quad \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \quad = \varepsilon$$

Since the space M is complete by hypothesis, the subsequence h converges. Hence every sequence in (M ; d) contains a convergent subsequence.

Proof M5 : Let a metric space be complete and totally bounded. Then the space is compact.

Let $C = \{C_1, C_2, ...\}$ be an arbitrary open covering of a metric space $(M ; d)$. Consider the unions $A_k := C_1 \cup ... \cup C_k$. Assume that $A_k \neq M$ for all $k \in \mathbb{N}'$. Then $x_k \notin A_k$ may be chosen for each $k \in \mathbb{N}'$. By (M4), the sequence $<x_1, x_2, ...>$ contains a convergent subsequence. Let x be the limit of this subsequence. Since C covers M, the point x lies in at least one open set C_m. Since C_m is open, an entire ε-ball B with center x lies in C_m. For $i \geq m$, the terms x_i do not lie in B since $B \subseteq C_m \subseteq A_i$. This contradicts the fact that a subsequence of $<x_1, x_2, ...>$ converges to x. Thus, contrary to the assumption, $A_k = M$ for some $k \in \mathbb{N}'$. Then $\{C_1, ..., C_k\}$ is a finite open subcovering of C. Hence M is compact.

Proof M6 : A subspace of a real space $(\mathbb{R}^n ; d)$ is compact if and only if it is closed and bounded.

(1) Let the subspace A of a real space $(\mathbb{R}^n ; d)$ be compact. By property (K4) of the Hausdorff space $\mathbb{R}^n$, the set A is closed. The ε-balls $D(0, r)$ with $r = 1, 2, ...$ form an open covering of A. Since A is compact, this covering contains a finite subcovering. This implies that A is contained in one of the ε-balls and is therefore bounded.

(2) Let the subspace A of a real space $(\mathbb{R}^n ; d)$ be closed and bounded. Then A is contained in an ε-ball $D(0, r)$. This ε-ball is a subset of the cube $W := [-r, r] \times ... \times [-r, r]$ (n-fold). There is a continuous mapping $f : \mathrm{I}^n \to W$ from the compact unit cube I^n (see Example 4 in Section 5.11.1) to W. By property (K5), the cube W is therefore compact. Hence A is a closed subset of a compact set, and therefore compact by (K3).

Example : Maximal values of real-valued mappings

Let the mapping $f : M \to \mathbb{R}$ from a compact space $(M ; T)$ be continuous. Then by property (K5) of compact spaces the image $f(M)$ is compact. The preceding proof shows that $f(M)$ is closed and bounded. Hence there is a finite least upper bound f_{max} of $f(M)$. Since the set $f(M)$ is closed, the value f_{max} is contained in $f(M)$.

5.11.3 LOCALLY COMPACT SPACES

Introduction : Euclidean spaces are not compact. However, every point of a euclidean space has a compact neighborhood. The question arises whether this local compactness has an essential influence on the properties of the non-compact space. In the following it is shown that such locally compact spaces may be compactified by forming the union of the underlying set with a one-point set disjoint from the underlying set.

Locally compact space : A topological space $(M ; T)$ is said to be locally compact if every point $x \in M$ has a compact neighborhood. Locally compact spaces have the following properties :

(L1) In a locally compact Hausdorff space $(M ; T)$ every neighborhood of a point $x \in M$ contains a compact neighborhood of x.

(L2) Every locally compact Hausdorff space is regular.

Proof L1 : In a locally compact Hausdorff space $(M ; T)$ every neighborhood of a point $x \in M$ contains a compact neighborhood of x.

Let U be an arbitrary neighborhood of x in M. Since M is locally compact, there is a compact neighborhood A of x in M. However, this is not necessarily contained in U. In the following, the intersection $A \cap U$ is shown to contain a compact neighborhood C of x.

By the definition of a neighborhood, there are open sets T_1, T_2 with $x \in T_1 \subseteq A$ and $x \in T_2 \subseteq U$, and hence there is an open set $T_3 = T_1 \cap T_2$ with $x \in T_3 \subseteq A \cap U$. By property (K4) in Section 5.11.1, the compact subset A of M is closed. The open set T_3 is contained in the closed set A. Hence the closure $H(T_3)$ is contained in A. The closed set $H(T_3)$ is compact by property (K3) in Section 5.11.1, since A is compact.

By property (H3) in Section 5.9, the subspace $H(T_3)$ of the Hausdorff space M is also Hausdorff. By (K10), the compact Hausdorff space $H(T_3)$ is regular. By property (R2) in Section 5.9, the regular Hausdorff space $H(T_3)$ possesses a closed neighborhood C of the point x which is contained in the open set T_3. Since $H(T_3)$ is compact, C is compact by property (K3). Since $C \subseteq T_3 \subseteq U$, it follows that every neighborhood U of x contains a compact neighborhood C of x.

Proof L2 : Every locally compact Hausdorff space is regular.

By (L1), the compact neighborhoods of a point x in a Hausdorff space $(M ; T)$ form a neighborhood basis at the point x. By property (K4) in Section 5.11.1, each of these compact neighborhoods is closed. Hence the closed neighborhoods of x form a neighborhood basis at the point x. By property (R2) in Section 5.9, the space M is therefore regular.

Compactification : A locally compact Hausdorff space $(M;T)$ may be extended to a compact space $(N;S)$. A point ∞ is defined which is not contained in M. The underlying set N of the compactified space is the union $M\cup\{\infty\}$ of the underlying set M with the one-point set $\{\infty\}$. The topology S contains all sets of the topology T as well as the complement $D_i := N - C_i$ of every compact subset C_i of M. The set of the complements $N - C_i$ contains the underlying set N. The compact Hausdorff space $(N;S)$ is called the one-point compactification of $(M;T)$.

$$N = M\cup\{\infty\} \quad \wedge \quad \infty \notin M$$
$$S = T\cup\{N - C_i \mid C_i \subseteq M \text{ is compact}\}$$

The following properties of this construction are proved in the following :

(C1) The set S is a topology.

(C2) The space $(N;S)$ is compact.

(C3) The space $(N;S)$ is a Hausdorff space.

Proof C1 : The set S is a topology.

(T1) The set S contains the underlying set N and the empty set $\emptyset\in T$.

(T2) The intersection of two elements of S is again an element of S. Three cases arise :

(a) The intersection of two elements of T is by definition again an element of T, and hence of S.

(b) The intersection of two elements D_i and D_k is given by $D_i\cap D_k = (N-C_i)\cap(N-C_k) = N-(C_i\cup C_k)$. Since by property (K12) the union of the compact sets C_i and C_k is compact, $N-(C_i\cup C_k)$ is by definition an element of S.

(c) The intersection of an element $T_i\in T$ and an element D_k is given by $T_i\cap D_k = T_i\cap(N-C_k) = T_i - C_k$. By (K4), the compact set C_k in the Hausdorff space M is closed. Therefore $T_i - C_k$ is an open set in M, and hence an element of S.

It follows by induction that the intersection of any finite number of elements of S is again an element of S.

(T3) The union of an arbitrary number of elements of S is again an element of S. Three cases arise :

(a) The union of an arbitrary number of elements of T is by definition again an element of T, and hence of S.

(b) The union of an arbitrary number of elements $D_i, D_k, \ldots$ is given by $D_i\cup D_k\cup\ldots = (N-C_i)\cup(N-C_k)\cup\ldots = N-(C_i\cap C_k\cap\ldots)$. By (K13) the intersection $C_i\cap C_k\cap\ldots$ is a compact set. Hence $N-(C_i\cap C_k\cap\ldots)$ is an element of S.

(c) For an arbitrary union of elements of S, results (a) and (b) are separately applied to the two types of elements to obtain a set T_i and a set $N - C_m$. The complement of the union of these sets is given by

$$N - (T_i \cup (N - C_m)) = (N - T_i) \cap C_m$$

The set $M - T_i$ is closed in M. By (K4), the compact set C_m is also closed in M. Thus $(M - T_i) \cap C_m$ is a closed subset of the compact set C_m, and is therefore compact by (K3). But $(M - T_i) \cap C_m = (N - T_i) \cap C_m$, since $\infty \notin C_m$. Thus the set $(N - T_i) \cap C_m$ is compact, and hence its complement $T_i \cup (N - C_m)$ is an element of S.

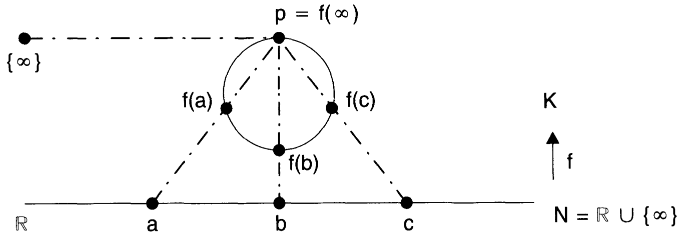
Proof C2 : The space (N ; S) is compact.

Let $\{U_1, U_2, ...\}$ be an open covering of the space (N ; S). Let the point ∞ be contained in the open set U_∞. Then U_∞ is of the form $N - C_k$. The set $N - U_\infty = C_k$ is compact, and hence there is a finite subcovering $\{U_{i_1}, ..., U_{i_n}\}$ of $N - U_\infty$. Thus the open covering $\{U_1, U_2, ...\}$ of N contains the finite subcovering $\{U_\infty, U_{i_1}, ..., U_{i_n}\}$. Hence N is compact.

Proof C3 : The space (N ; S) is a Hausdorff space.

In the locally compact set M, a point x has a compact neighborhood, which does not contain the point ∞, since ∞ is not a point of M. The complement of this neighborhood is open and contains ∞. Hence every pair of points x, ∞ satisfies the condition for a Hausdorff space. Since (M ; T) is already a Hausdorff space and $N = M \cup \{\infty\}$, it follows that (N ; S) is also a Hausdorff space.

Example : Compactification of the real axis ℝ

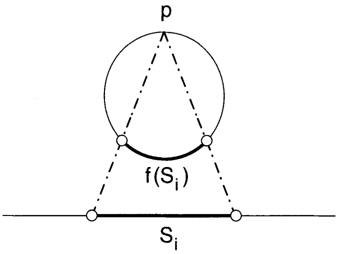


(1) The real axis ℝ with the natural topology T is unbounded and therefore not compact. Every point on ℝ has a compact neighborhood. Thus ℝ is locally compact. The real axis is compactified by forming the union $N = \mathbb{R} \cup \{\infty\}$. The point ∞ may be regarded as an infinitely remote point. The topology S of N is obtained by augmenting the natural topology T of ℝ by the complements of all compact subsets C_i of ℝ. The space (N ; S) is compact.

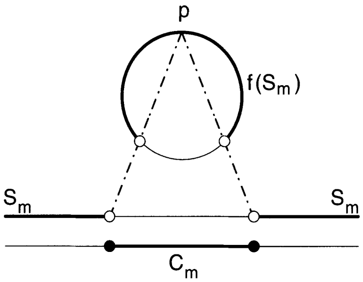
$$N = \mathbb{R} \cup \{\infty\}$$

$$S = T \cup \{N - C_i \mid C_i \text{ is a compact subset of } \mathbb{R}\}$$

(2) The boundary K of a circle with the relative topology Q with respect to the real plane $\mathbb{R}^2$ is a bounded, closed subset of $\mathbb{R}^2$, and hence a compact space (K ; Q). The mapping from the real axis $\mathbb{R}$ to the boundary K which is illustrated below is not topological, since it is not bijective : The pole p of K does not have a preimage in $\mathbb{R}$. By contrast, the mapping $f : N \rightarrow K$ from the compactification N of $\mathbb{R}$ to K is a topological mapping. The preimage of the pole p is the point ∞. The open sets of the topology Q on K are the images of the open sets of the topology S on N.



S_i is an open set of the natural topology T on $\mathbb{R}$



S_m is the complement of the compact subset C_m of $\mathbb{R}$

5.12 CONTINUITY OF REAL FUNCTIONS

Introduction : In Section 5.6 the continuity of a mapping $f : M \rightarrow N$ between general topological spaces $(M;T)$ and $(N;S)$ is defined using the properties of open sets. For metric spaces, there is a special definition of continuity based on the properties of accumulation points of sequences. This definition is particularly suitable for studying the continuity of real functions. The limits of real functions and the properties of continuous real functions are treated in the following.

Real functions : Let A and B be subsets of the real axis $\mathbb{R}$ equipped with the natural topology. A mapping $f : A \rightarrow B$ is called a real function.

$f : A \rightarrow B$	real function
$f(A) = \{f(x) \mid x \in A\}$	image of the function f
$G = \{(x, f(x)) \mid x \in A\}$	graph of the function f

Types of functions :

(1) Polynomial of n-th degree :

$$f \;:\; \mathbb{R} \rightarrow \mathbb{R} \qquad \text{with} \quad f(x) = \sum_{i=0}^{n} c_i x^i \qquad c_i \in \mathbb{R},\; i \in \mathbb{N},\; c_n \neq 0$$

(2) Identity function :

$$1_{\mathbb{R}} : \mathbb{R} \rightarrow \mathbb{R} \qquad \text{with} \quad f(x) = x$$

(3) Rational function :

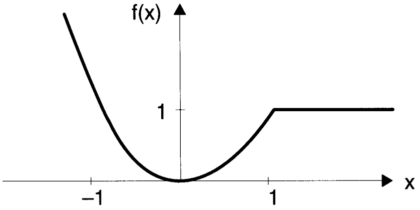
$$f \;:\; (\mathbb{R} - L) \rightarrow \mathbb{R} \qquad \text{with} \quad f(x) = \left\{ \sum_{i=0}^{n} c_i x^i \right\} / \left\{ \sum_{s=0}^{m} b_s x^s \right\} \qquad \begin{array}{r} c_i, b_s \in \mathbb{R} \\ i, s \in \mathbb{N} \\ c_n, b_m \neq 0 \end{array}$$

The set L contains the solutions of the equation $\Sigma\, b_s y^s = 0$:

$$L = \{ y \in \mathbb{R} \mid \sum_{s=0}^{m} b_s y^s = 0 \}$$

Bounded function : A function $f : A \rightarrow \mathbb{R}$ is said to be bounded from above if its image f(A) has an upper bound. The function is said to be bounded from below if f(A) has a lower bound. The function is said to be bounded if it is bounded from above and below.

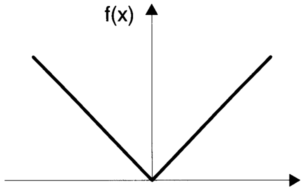
Special real functions :



function defined by cases

$f(x) = x^2$ for $x < 1$

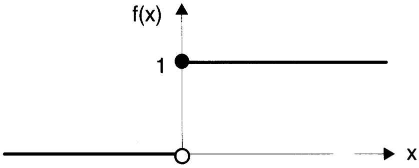
$f(x) = 1$ for $x \geq 1$



modulus function

$f(x) = -x$ for $x < 0$

$f(x) = x$ for $x \geq 0$



Heaviside function

$f(x) = 0$ for $x < 0$

$f(x) = 1$ for $x \geq 0$

Constructed functions : Real functions $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ may be used to construct new functions :

sum function	: $f + g$	: $A \cap B \to \mathbb{R}$	with	$(f + g)(x) = f(x) + g(x)$
difference function	: $f - g$	: $A \cap B \to \mathbb{R}$	with	$(f - g)(x) = f(x) - g(x)$
product function	: $f \cdot g$	: $A \cap B \to \mathbb{R}$	with	$(f \cdot g)(x) = f(x) \cdot g(x)$
quotient function	: f / g	: $C \to \mathbb{R}$ $C = A \cap B - \{y \mid g(y) = 0\}$	with	$(f / g)(x) = f(x) / g(x)$
composition	: $g \circ f$	: $A \to \mathbb{R}$	with if	$(g \circ f)(x) = g(f(x))$ $f(A) \subseteq B$
modulus function	: $\|f\|$	: $A \to \mathbb{R}$	with	$\|f\|(x) = \|f(x)\|$

Basic sequence : A sequence $g : \mathbb{N} \to M$ is called a basic sequence at the point $a \in M$ if the sequence g converges to a and a is not a term of the sequence. The point a is an accumulation point of $M - \{a\}$. There are no basic sequences at an isolated point a. There is more than one basic sequence at an accumulation point.

$$g \text{ is a basic sequence} \quad :\Leftrightarrow \quad g(n) \in M - \{a\} \quad \wedge \quad \lim_{n \to \infty} g(n) = a$$

Limit of a function : Let $f : A \to \mathbb{R}$ be a real function, and let a be an accumulation point of $A - \{a\}$. Then every basic sequence $< x_1, x_2, \ldots >$ at the point a has an image $< f(x_1), f(x_2), \ldots >$. The limits of the sequences $< f(x_1), f(x_2), \ldots >$ of function values will generally be different for different basic sequences. A point b is called the limit of the function f at a if the sequences of function values for all basic sequences have the same limit b. It is not required that the limit b be equal to the function value f(a).

$$\lim_{x\to a} f(x) = b \;:\Leftrightarrow\; ((\lim_{n\to\infty} x_n = a \;\wedge\; \bigwedge_{n\in N} (x_n \neq a)) \;\Rightarrow\; \lim_{n\to\infty} f(x_n) = b)$$

Left and right limits : The basic sequences at a point a on the real axis $\mathbb{R}$ may be restricted to points which lie only to the left or only to the right of the point a. In this case they are called left and right basic sequences at point a, respectively. A point b_1 is called the left limit of the function f at the point a if the sequences of function values of all left basic sequences at the point a have the limit b_1. A point b_2 is called the right limit of the function f at the point a if the sequences of function values of all right basic sequences at the point a have the limit b_2. The left limit b_1 and the right limit b_2 at the point a need not coincide.

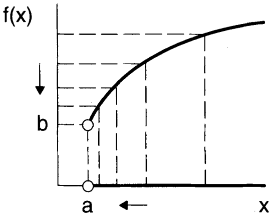
Establishing the limit of a function : Let $f : A \to \mathbb{R}$ be a real function, and let a be an accumulation point of $A - \{a\}$. The real number b is the limit of f at a if and only if for every real number $\varepsilon > 0$ there is a real number $\delta > 0$ such that for all $x \in A - \{a\}$ the inequality $|x - a| < \delta$ implies the inequality $|f(x) - b| < \varepsilon$ for the function values.

$$\bigwedge_{\varepsilon>0} \bigvee_{\delta>0} \bigwedge_{x\in A-\{a\}} (|x - a| < \delta \;\Rightarrow\; |f(x) - b| < \varepsilon)$$

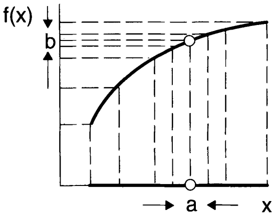
Theorems for limits : Let the limits of the functions $f : A \to \mathbb{R}$ and $g : B \to \mathbb{R}$ at the point a be $\lim\limits_{x\to a} f(x) = s$ and $\lim\limits_{x\to a} g(x) = t$, respectively. Then constructed functions have the following limits :

$$\begin{array}{llllll}
\lim\limits_{x\to a} (f(x) + g(x)) & := & \lim\limits_{x\to a} (f + g)(x) & = & s + t & \\
\lim\limits_{x\to a} (f(x) - g(x)) & := & \lim\limits_{x\to a} (f - g)(x) & = & s - t & \\
\lim\limits_{x\to a} (f(x) \cdot g(x)) & := & \lim\limits_{x\to a} (f \cdot g)(x) & = & s \cdot t & \\
\lim\limits_{x\to a} \dfrac{f(x)}{g(x)} & := & \lim\limits_{x\to a} (\tfrac{f}{g})(x) & = & \dfrac{s}{t} & \text{if} \quad t \neq 0 \\
\lim\limits_{x\to a} |f(x)| & := & \lim\limits_{x\to a} |f|(x) & = & |s| &
\end{array}$$

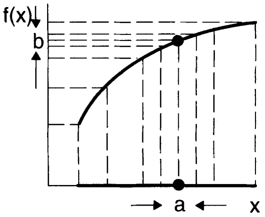
Example 1 : Limits of real functions



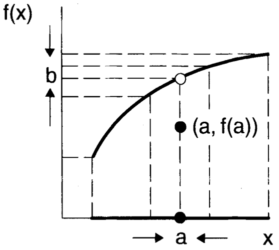
limit b
not defined on the left



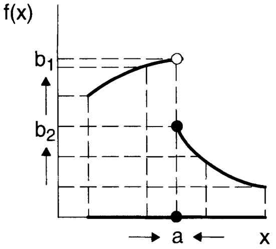
limit b
defined on the left and right



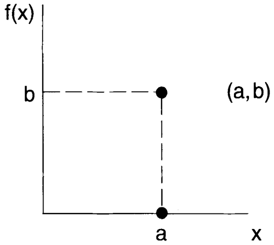
limit b
function value f(a) = b



limit b
function value f(a) ≠ b



left limit b_1
right limit b_2
function value f(a) = b_2



limit not defined
a is not an accumulation point

Continuity : Let $f: A \rightarrow \mathbb{R}$ be a real function with the limit $\lim\limits_{x \to a} f(x) = b$. The limit b and the function value f(a) at a point a need not coincide. Functions whose value f(a) at the point a may be approximated to arbitrary precision by sequences $<f(x_1), f(x_2), ...>$ of function values for suitable basic sequences $<x_1, x_2, ...>$ therefore possess a special property. In the proof of the following theorem, this property is shown to be the continuity of topological mappings already defined in Section 5.6.

Continuous mapping : Let X be a subset of the metric space (M ; T), and let Y be a subset of the metric space (N ; S). A mapping $f: X \rightarrow Y$ is continuous if and only if for every subset $A \subseteq X$ the image f(x) of every contact point x of A is a contact point of the set $f(A) \subseteq Y$.

A contact point of a set is either an inner point or a boundary point of that set. Hence the mapping $f: X \rightarrow Y$ is continuous if and only if for every subset $A \subseteq X$ the image f(x) of every point x of the closure H(A) is a point of the closure H(f(A)).

$$f: X \rightarrow Y \text{ is continuous} \quad \Leftrightarrow \quad \bigwedge_{A \subseteq X} \; \bigwedge_{x \in H(A)} [\, f(x) \in H(f(A)) \,]$$

$$\Leftrightarrow \quad \bigwedge_{A \subseteq X} [\, f(H(A)) \subseteq H(f(A)) \,]$$

Let a contact point $a \in A$ be an accumulation point of A. Let a real function $f : A \to \mathbb{R}$ be continuous. Then the function value $f(a)$ may be approximated to arbitrary precision by sequences $< f(x_1), f(x_2),... >$ of function values for suitable basic sequences.

Proof : A function $f : X \to Y$ is continuous if and only if for every subset $A \subseteq X$ the inclusion $f(H(A)) \subseteq H(f(A))$ holds.

(1) Let the mapping $f : X \to Y$ be continuous. From $f(A) \subseteq H(f(A))$ it follows that :

$$A \subseteq f^{-1} \circ f(A) \subset f^{-1}(H(f(A)))$$

Since the closure $H(f(A))$ is closed and the mapping is continuous, it follows that $f^{-1}(H(f(A)))$ is also closed. Hence :

$$A \subseteq H(A) \subseteq f^{-1}(H(f(A)))$$

$$H(f(A)) \subseteq f \circ f^{-1}(H(f(A))) = H(f(A))$$

(2) Let $f(H(A)) \subseteq H(f(A))$ for the preimage $A = f^{-1}(C)$ of every closed set $C \subseteq Y$. Then A is shown to be a closed set. Substituting $A = f^{-1}(C)$ yields :

$$f(H(A)) = f(H(f^{-1}(C))) \subseteq H(f \circ f^{-1}(C)) = H(C) = C$$

$$H(A) \subseteq f^{-1} \circ f(H(A)) \subseteq f^{-1}(C) = A$$

$A \subseteq H(A)$ and $H(A) \subseteq A$ implies $A = H(A)$. Thus the preimage of every set C closed in Y is a set A closed in X. It follows that the preimage of every set open in Y is open in X. Hence the mapping $f : X \to Y$ is continuous.

Establishing the continuity of a function : Let $f : A \to B$ be a real function, and let a be an arbitrary point in A. The function f is continuous in a if and only if for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every $x \in A$ and $|x-a| < \delta$ one has $|f(x) - f(a)| < \varepsilon$.

$$\lim_{x \to a} f(x) = f(a) \quad \Leftrightarrow$$

$$\bigwedge_{\varepsilon>0} \bigvee_{\delta>0} \bigwedge_{x \in A} (|x-a| < \delta \Rightarrow |f(x) - f(a)| < \varepsilon)$$

Proof : Establishing the continuity of a function

(1) Let the mapping $f : A \to B$ be continuous at the point $a \in A$. For any $\varepsilon > 0$ the ε-ball $D(f(a), \varepsilon)$ is an open set. Hence $C = B \cap D(f(a), \varepsilon)$ is an open set of the relative topology of B. Its preimage $f^{-1}(C)$ is open by hypothesis and contains the point a. Hence there is a neighborhood $D(a, \delta) \cap A \subseteq f^{-1}(C)$. For every point x of this neighborhood $x \in A$ and $|x-a| < \delta$. But $x \in A$ implies $f(x) \in C$, and hence $|f(a) - f(x)| < \varepsilon$.

(2) Assume that for every $\varepsilon > 0$ there is a $\delta > 0$ such that for every $x \in A$ and $|x - a| < \delta$ the condition $|f(x) - f(a)| < \varepsilon$ is satisfied. It is to be proved that the preimage of every set open in B is open in A. Let S be an open set of the relative topology of B. For a point $a \in f^{-1}(S)$ in its preimage, $f(a) \in S$. There is an $\varepsilon > 0$ such that $D(f(a), \varepsilon) \cap B \subseteq S$ is an element of the relative topology of B. For this ε there is by hypothesis a $\delta > 0$ such that for every point $x \in D(a, \delta) \cap A$ one has $f(x) \in D(f(a), \varepsilon) \cap B \subseteq S$. From $f(x) \in S$ for all $x \in D(a, \delta) \cap A$ it follows that $D(a, \delta) \cap A$ is contained in $f^{-1}(S)$. Thus every point $a \in f^{-1}(S)$ is contained in an element $D(a, \delta) \cap A$ of the relative topology. Hence the preimage $f^{-1}(S)$ of the open set S is open in A : The mapping f is continuous.

Uniform continuity : The value f(a) of a continuous real function $f : A \to \mathbb{R}$ at the point a may be approximated to an arbitrary precision ε. The approximation of f is regarded as equally good at all points of A if for all $\varepsilon > 0$ the value $\delta > 0$ may be chosen independently of $a \in A$. Then the following statement holds (note the order of the quantifiers, to be applied from left to right) :

$$\bigwedge_{\varepsilon>0} \bigvee_{\delta>0} \bigwedge_{a\in A} \bigwedge_{x\in A} (|x - a| < \delta \quad \Rightarrow \quad |f(x) - f(a)| < \varepsilon)$$

The function f is then said to be uniformly continuous on A. In the condition for general continuity on A, the order of the quantifiers for δ and a is reversed.

Discontinuity of a function : A real function $f : A \to \mathbb{R}$ is said to be discontinuous at the point a of A if f is not continuous at a. The function is discontinuous at a if and only if there is a basic sequence at the point a whose sequence of function values does not converge to f(a).

Jump discontinuity of a function : Let the function $f : A \to \mathbb{R}$ be real with a left limit b_1 and a right limit b_2 at a point a in A. If the limits b_1 and b_2 are different, then the function f is discontinuous at a. If one of the two limits coincides with the function value f(a), the point a is called a jump discontinuity of the function.

Sequence of functions : A sequence $g : \mathbb{N} \to F$ with $g(n) = f_n$ is called a sequence of real functions if its terms are real functions $f_n : \mathbb{R} \to \mathbb{R}$. At a point $a \in \mathbb{R}$, the sequence of functions leads to a real sequence $g_a : \mathbb{N} \to \mathbb{R}$ of function values with $g_a(n) = f_n(a)$.

Convergence of a sequence of functions : A sequence $< f_1, f_2, ... >$ of real functions $f_i : A_i \to \mathbb{R}$ is said to converge at a point a of the intersection $A = A_1 \cap A_2 \cap ...$ if the sequence $< f_1(a), f_2(a), ... >$ of function values at the point a converges.

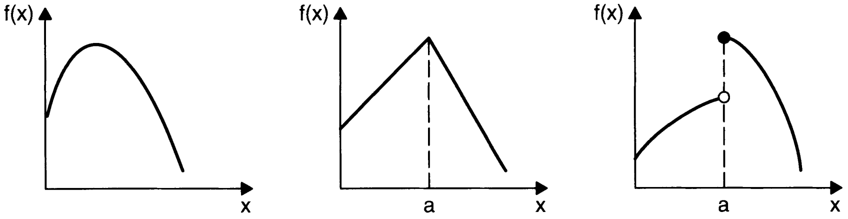
Limit function : A sequence $< f_1, f_2, ... >$ of functions $f_i : A_i \rightarrow \mathbb{R}$ is said to converge to a limit function $f : A \rightarrow \mathbb{R}$ with $A = A_1 \cap A_2 \cap ...$ if at every point a of A the sequence of function values converges to f(a). The limit function of a sequence of continuous functions is not necessarily continuous.

$$f : A \rightarrow \mathbb{R} \quad \text{with} \quad f(a) = \lim_{n \rightarrow \infty} f_n(a)$$

Uniform convergence of a sequence of functions : A sequence $g : \mathbb{N} \rightarrow F$ of functions with $g(n) = f_n$ is said to converge to f uniformly on A if for every real number $\varepsilon > 0$ there is a natural number n_0 independent of $a \in A$ such that for every $n > n_0$ the absolute value of the difference of the function values $f_n(a)$ and f(a) is less than ε. The limit function of a uniformly convergent sequence of continuous functions is continuous.

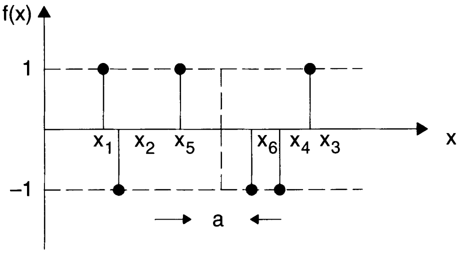
$$\bigwedge_{\varepsilon > 0} \bigvee_{n_0 \in \mathbb{N}} \bigwedge_{a \in A} \bigwedge_{n > n_0} (|f_n(a) - f(a)| < \varepsilon)$$

Example 2 : Continuous and discontinuous functions



f is continuous f is continuous f is discontinuous

Example 3 : Nowhere continuous function



$$f(x) = 1 \quad \text{for} \quad x \in \mathbb{Q}$$

$$f(x) = -1 \quad \text{for} \quad x \in (\mathbb{R} - \mathbb{Q})$$

The real function f(x) takes the value 1 if x is a rational number and the value -1 otherwise. For every point a of $\mathbb{R}$ there are basic sequences $< x_1, x_2, ... >$ with $x_{2i-1} \in \mathbb{Q}$ and $x_{2i} \in (\mathbb{R} - \mathbb{Q})$. The corresponding sequence of function values is $< f(x_1), f(x_2), ... > = < 1, -1, 1, -1, ... >$. This sequence diverges. Hence f is nowhere continuous in $\mathbb{R}$.

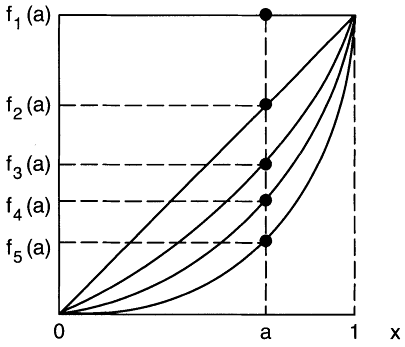
Example 4 : Continuous limit function

The function cos x is the limit function of a sequence $< f_1, f_2, ... >$ of functions. Both the functions f_i of the sequence and the limit function cos x are continuous. The following table shows the sequences of function values for different values of x.

function	π/6	π/4	π/2	3π/4	π
$f_1 = 1$	1.000000	1.000000	1.000000	1.000000	1.000000
$f_2 = f_1 - \frac{x^2}{2!}$	0.862922	0.691575	−0.233701	−1.775826	−3.934802
$f_3 = f_2 + \frac{x^4}{4!}$	0.866054	0.707429	0.019969	−0.491624	0.123910
$f_4 = f_3 - \frac{x^6}{6!}$	0.866025	0.707103	−0.000895	−0.729272	−1.211353
cos x	0.866025	0.707107	0	−0.707107	−1.000000

Example 5 : Discontinuous limit function

The sequence $g : \mathbb{N} \to F$ of functions with $g(n) = f_n$ and $f_n : \mathbb{R} \to \mathbb{R}$ with $f_n(x) = x^n$ is considered on the closed unit interval [0, 1]. Every function in the sequence $< f_1, f_2, ... >$ is continuous. Nevertheless, the limit function is discontinuous at the point x = 1.



$f_1 = 1$

$f_2 = x$

$f_3 = x^2$

$f_4 = x^3$

$f_5 = x^4$

$$\lim_{n \to \infty} f_n(x) = 0 \quad \text{for} \quad 0 \le x < 1$$

$$\lim_{n \to \infty} f_n(x) = 1 \quad \text{for} \quad x = 1$$

6 NUMBER SYSTEM

6.1 INTRODUCTION

The development of numerical algorithms requires knowledge of the algebraic structure, the order structure and the topological structure of the sets of numbers which form the number system. The representation of numerical values in computers and the errors related to numerical operations also depend on the structure of these sets of numbers.

The number system is described in the following sections. The axiomatically defined natural numbers form the basis of the number system. The integers are constructed such that the addition of integers is invertible, in contrast to the addition of natural numbers. The rational numbers are constructed such that the multiplication of rational numbers is invertible, in contrast to the multiplication of integers.

The roots of rational polynomials are not necessarily rational numbers. Some of the roots which are not rational are real numbers. These are defined as open sets of rational numbers. However, there are also real numbers such as e and π which are not roots of rational polynomials.

Not all roots of rational polynomials are real numbers. The complex numbers are therefore introduced as pairs of numbers with real and imaginary parts. In the set of complex numbers, the roots of every rational polynomial can be determined. The set of complex numbers is extended to the set of quaternions.

Different designations are introduced for the different sets of numbers. The sets of numbers are constructed by extension, so that :

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C} \subset \mathbb{H}$$

$\mathbb{N}$	set of natural numbers
$\mathbb{Z}$	set of integers
$\mathbb{Q}$	set of rational numbers
$\mathbb{R}$	set of real numbers
$\mathbb{C}$	set of complex numbers
$\mathbb{H}$	set of quaternions

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*
© Springer-Verlag Berlin Heidelberg 2001

6.2 NATURAL NUMBERS

Axioms : The set of natural numbers intuitively designated by $\mathbb{N} = \{0, 1, 2, ...\}$ is characterized by the following axioms :

(1) 0 is a natural number.

(2) Every natural number n has a successor n'.

(3) 0 is not the successor of a natural number.

(4) Natural numbers with equal successors are equal.

(5) A subset of $\mathbb{N}$ is identical with $\mathbb{N}$ if it contains the number 0 and if, for every natural number n it contains, the subset also contains the successor n'.

The set of natural numbers without zero is designated by $\mathbb{N}'$.

Algebraic structure : The inner operations + (addition) and • (multiplication) on the natural numbers are inductively defined as follows :

addition : $m + 0 := m$ and $m + n' = (m + n)'$

multiplication : $m \cdot 1 := m$ and $m \cdot n' = m \cdot n + n$

The natural number 0 is the identity element of addition. Addition is associative and commutative. Hence the domain ($\mathbb{N}$; +) is a commutative semigroup with identity element.

The natural number 1 is the identity element of multiplication. Multiplication is associative and commutative. Hence the domain ($\mathbb{N}$; •) is a commutative semigroup with identity element.

Multiplication is distributive with respect to addition. Hence the domain ($\mathbb{N}$; +, •) is a commutative semiring with 0 and 1 as identity elements. The cancellation law holds in this semiring.

Property	Addition +	Multiplication ·
associative	$(k + m) + n = k + (m + n)$	$(k \cdot m) \cdot n = k \cdot (m \cdot n)$
commutative	$m + n = n + m$	$m \cdot n = n \cdot m$
distributive	$k \cdot (m + n) = k \cdot m + k \cdot n$	$(m + n) \cdot k = m \cdot k + n \cdot k$
zero element	$m + 0 = m$	$m \cdot 0 = 0$
unit element	$m + 1 = m'$	$m \cdot 1 = m$
cancellation law	$m + k = n + k \Rightarrow m = n$	$k \neq 0 \Rightarrow (m \cdot k = n \cdot k \Rightarrow m = n)$

Ordinal structure : The relations ≤ (less than or equal to) and < (less than) in the set of natural numbers are defined as follows :

$$m \leq n \;:\Leftrightarrow\; \bigvee_{k} \,(m + k = n)$$

$$m < n \;:\Leftrightarrow\; (m \leq n) \;\wedge\; (m \neq n)$$

The relation ≤ is reflexive, antisymmetric, linear and transitive. The relation < is antireflexive, asymmetric, connex and transitive. Thus the natural numbers are totally ordered. They are also well-ordered, since every subset of ℕ has a least element. The ordinal and the algebraic structure of the natural numbers are compatible, since the monotonic laws for addition and multiplication hold.

Property	Relation ≤	Relation <
reflexive	$m \leq m$	
antireflexive		$\neg(m < m)$
antisymmetric	$m \leq n \;\wedge\; n \leq m \;\Rightarrow\; m = n$	
asymmetric		$m < n \;\Rightarrow\; \neg(n < m)$
linear	$m \leq n \;\vee\; n \leq m$	
connex		$m \neq n \;\Rightarrow\; m < n \;\vee\; n < m$
transitive	$k \leq m \;\wedge\; m \leq n \;\Rightarrow\; k \leq n$	$k < m \;\wedge\; m < n \;\Rightarrow\; k < n$
monotonic	$m \leq n \Leftrightarrow m + k \leq n + k$ $k > 0 \Rightarrow (m \leq n \Leftrightarrow k \cdot m \leq k \cdot n)$	$m < n \Leftrightarrow m + k < n + k$ $k > 0 \Rightarrow (m < n \Leftrightarrow k \cdot m < k \cdot n)$

Subtraction : In ℕ, the equation $m + x = n$ can only be solved for the variable x if $m \leq n$. The solution x is called the difference of n and m. It is designated by $n - m$.

Divisor : A natural number b is called a divisor of a natural number a if there is a natural number n such that $a = n \cdot b$. This is designated by $b \mid a$ (b divides a). A divisor b is said to be proper if $b \neq 1$ and $b \neq a$.

Prime : A natural number $m > 1$ is called a prime and is said to be irreducible (prime) if the only divisors of m are 1 and m. There are infinitely many primes. The set of primes is designated by P.

Proof : There are infinitely many primes (Euclid).
Assume that the set $P = \{p_1, ..., p_n\}$ of primes is finite. Then $m = p_1 \cdot ... \cdot p_n + 1$ is a natural number, and $m > 1$. Since m is not contained in the set P of primes, m has a proper divisor. However, none of the primes p_i is a divisor of m, since in that case p_i would also be a divisor of $m - p_1 \cdot ... \cdot p_n = 1$: This is impossible. Thus m has a proper divisor p which is not contained in P. The contradiction shows that there are infinitely many primes.

Prime factorization : Every natural number $m > 1$ is the product of primes. A given prime may occur more than once in this product. The product is unique up to the order of the factors.

$$m = p_1 \cdot p_2 \cdot ... \cdot p_n$$

Proof : Every natural number is the product of primes.

The assertion holds for $m \leq 3$. For $m > 3$, it is assumed that there is a unique prime factorization for every number less than m, and it is shown that in this case m also possesses a unique prime factorization.

(1) The number m is shown to be a product of primes. If m is a prime number, the product consists only of m itself. If m is reducible, then $m = a \cdot b$ with $1 < a, b < m$. Since the factors a and b are less than m, by the induction hypothesis they possess prime factorizations. Hence m possesses a prime factorization.

(2) It is proved that any two prime factorizations $m = p_1 \cdot ... \cdot p_n$ and $m = u_1 \cdot ... \cdot u_s$ of the same number m are identical. The designations and the order of the factors are chosen such that $p_i \leq p_{i+1}$, $u_k \leq u_{k+1}$ and $p_1 \leq u_1$. For $p_1 = u_1$, let $m' := p_2 \cdot ... \cdot p_n = u_2 \cdot ... \cdot u_s$ with $m' < m$. Since the factorization of m' is unique by the induction hypothesis, the factorization of $m = p_1 \cdot m'$ is also unique. For $p_1 < u_1$, there is a number $1 < k < m$ such that

$$k := m - p_1 u_2 ... u_s = (u_1 - p_1) u_2 ... u_s = p_1(p_2 ... p_n - u_2 ... u_s)$$

The factorization of k is unique by the induction hypothesis and contains the factor p_1. But $p_1 < u_1$ implies $p_1 \neq u_1$. Hence the term $(u_1 - p_1)$ in the product $k = (u_1 - p_1) u_2 ... u_s$ must have the divisor p_1. Hence $u_1 - p_1 = cp_1$, and thus $u_1 = (c + 1)p_1$. But the prime u_1 has no divisors. The contradiction shows that the prime factorization of m is unique.

6.3 INTEGERS

Invertibility of addition : The addition of natural numbers is not generally invertible, since in $\mathbb{N}$ the subtraction $x = n - m$ is only admissible for $n \geq m$. The set of integers intuitively designated by $\mathbb{Z} = \{..., -2, -1, 0, 1, 2, ...\}$ is constructed such that subtraction can be carried out without restrictions in $\mathbb{Z}$.

Construction of $\mathbb{Z}$: The pairs $(m, n) \in \mathbb{N} \times \mathbb{N}$ of natural numbers are partitioned into equivalence classes $[i, k]$. All pairs $(m, n) \in [i, k]$ have the same difference, that is $m - n = i - k$ for $i \geq k$ and $n - m = k - i$ for $k \geq i$, so that in both cases $m + k = n + i$. The set $\mathbb{Z}$ of integers is the set of all equivalence classes $[i, k]$ in $\mathbb{N} \times \mathbb{N}$.

Normal representation : The definition of the integers and the cancellation law for natural numbers imply $[n + i, i] = [n, 0]$ and $[i, n + i] = [0, n]$. Hence every integer may be represented in the normal form $[n, 0]$ or $[0, n]$ with $n \in \mathbb{N}$. Integers with the normal form $[n, 0]$ are said to be positive and are designated by n. Integers with the normal form $[0, n]$ are said to be negative and are designated by $-n$ (minus n). The integer with the normal form $[0, 0]$ is called zero and is designated by 0. The integer with the normal form $[1, 0]$ is called one and is designated by 1.

Algebraic structure : The inner operations $+$ (addition) and $\cdot$ (multiplication) on the integers are defined as follows with respect to the equivalence classes of pairs of natural numbers :

addition	:	$[i, k] + [m, n] := [i + m,\ k + n]$
multiplication	:	$[i, k] \cdot [m, n] := [i \cdot m + k \cdot n,\ i \cdot n + k \cdot m]$

The pair $[0, 0]$ of natural numbers is the identity element 0 of addition, since $[i, k] + [0, 0] = [i, k]$. The pairs $[i, k]$ and $[k, i]$ of natural numbers are additive inverses, since $[i, k] + [k, i] = [i + k,\ i + k]$ and the normal form of $[i + k,\ i + k]$ is $[0, 0]$. Addition is associative and commutative. Hence the domain $(\mathbb{Z}\ ;\ +)$ is a commutative group.

The pair $[1, 0]$ of natural numbers is the identity element 1 of multiplication, since $[i, k] \cdot [1, 0] = [i, k]$. Multiplication is associative and distributive. Hence the domain $(\mathbb{Z}\ ;\ \cdot)$ is a commutative semigroup with 1 as the identity element.

Multiplication is distributive with respect to addition. Hence the domain $(\mathbb{Z}\ ;\ +, \cdot)$ is a commutative ring with 1 as the unit element. The cancellation law holds as it does for natural numbers.

Extension of $\mathbb{N}$ **:** The injective mapping $i : \mathbb{N} \to \mathbb{Z}$ with $i(n) = [n, 0]$ preserves structure (is homomorphic), since $i(n + m) = i(n) + i(m)$ and $i(m \cdot n) = i(m) \cdot i(n)$. Hence the set $\mathbb{N}$ of natural numbers may be extended to the set $\mathbb{Z}$ of integers by adding the negative integers. The algebraic structure is thus extended from a commutative semiring to a commutative ring. Subsets of $\mathbb{Z}$ are designated as follows :

$\mathbb{Z}'$	integers without zero
$\mathbb{Z}^+$	positive integers
$\mathbb{Z}_0^+$	positive integers and zero
$\mathbb{Z}^-$	negative integers
$\mathbb{Z}_0^-$	negative integers and zero

Subtraction : In the set $\mathbb{Z}$ of integers, the equation $[i, k] + x = [m, n]$ can be solved without restrictions. The solution $x = [k + m, i + n]$ is called the difference of the integers and is designated by $[m, n] - [i, k]$. In the normal representation of the integers, correspondingly, $a + x = b$ is solved by $x = b - a$.

Ordinal structure : The relations $\leq$ (less than or equal to) and $<$ (less than) in the set of integers are defined as follows :

$$a \leq b \quad :\Leftrightarrow \quad b - a \in \mathbb{N}$$
$$a < b \quad :\Leftrightarrow \quad a \leq b \ \wedge \ a \neq b$$

Like the natural numbers, the integers are totally ordered. However, in contrast to the natural numbers, they are not well-ordered, since there is no least element in $\mathbb{Z}$. As in the case of the natural numbers, the ordinal and the algebraic structure of the integers are compatible, since the monotonic laws for addition and multiplication also hold in $\mathbb{Z}$.

Absolute value : The absolute value of an integer a is a positive integer, designated by $|a|$. The absolute value of an integer is determined as follows :

$$a \geq 0 \quad \Rightarrow \quad |a| = a$$
$$a < 0 \quad \Rightarrow \quad |a| = -a$$

Rules of calculation : The operations of addition, subtraction and multiplication can be carried out without restrictions in $\mathbb{Z}$. The rules of calculation for absolute values follow from the normal representation and the algebraic structure of $\mathbb{Z}$.

Sign	Absolute value	Operation
$a \in \mathbb{Z}^- \wedge b \in \mathbb{Z}^-$		$a + b = -(\lvert a\rvert + \lvert b\rvert)$
$a \in \mathbb{Z}^+ \wedge b \in \mathbb{Z}^-$	$\lvert a\rvert \geq \lvert b\rvert$	$a + b = \lvert a\rvert - \lvert b\rvert$
	$\lvert a\rvert \leq \lvert b\rvert$	$a + b = -(\lvert a\rvert - \lvert b\rvert)$
$a \in \mathbb{Z}^- \wedge b \in \mathbb{Z}^-$		$a \cdot b = \lvert a\rvert \cdot \lvert b\rvert$
$a \in \mathbb{Z}^+ \wedge b \in \mathbb{Z}^-$		$a \cdot b = -(\lvert a\rvert \cdot \lvert b\rvert)$

Prime factorization : The absolute value of an integer n with $|n| > 1$ is a natural number and therefore has a prime factorization $|n| = p_1 \dots p_r$. The factors p_i are brought into monotonically increasing order. Equal factors are combined into powers. The resulting representation of n is unique and is called the normal form of the prime factorization.

$$n = (-1)^{\varepsilon} q_1^{\alpha_1} \dots q_s^{\alpha_s} \qquad n \in \mathbb{Z} \wedge |n| > 1$$

$q_1 < \dots < q_s$ prime factors

$\alpha_1, \dots, \alpha_s > 0$ exponents

$(-1)^{\varepsilon}$ sign, $\varepsilon = 0$ for $n > 0$ and $\varepsilon = 1$ for $n < 0$

p-exponent : The occurrence of a prime p from the set P of prime numbers in the normal form of the prime factorization of an integer n is described by the p-exponent $v_p(n)$, which is defined as follows :

$$p = q_i \quad \Rightarrow \quad v_p(n) = \alpha_i$$

$$p \notin \{q_1, \dots, q_s\} \quad \Rightarrow \quad v_p(n) = 0$$

Greatest common divisor : Let $M = \{n_1, \dots, n_k\}$ be a finite non-empty set of integers with $n_i \neq 0$. For every number n_i a finite number of the p-exponents $v_p(n_i)$ with $p \in P$ differs from zero (almost all p-exponents are zero). For each prime $p \in P$, the least p-exponent $a_p = \min v_p(n_i)$ for $n_i \in M$ is determined. The product of the primes with the exponents a_p can be determined, since almost all exponents are 0. It is called the greatest common divisor of M and is designated by gcd. The gcd of the set M divides all numbers in M.

$$\gcd(n_1, \dots, n_k) = \prod_{p \in P} (p^{a_p}) \quad \wedge \quad a_p = \min_{n_i \in M} v_p(n_i)$$

Least common multiple : Let $M = \{n_1,...,n_k\}$ be a finite non-empty set of integers with $n_i \neq 0$. For every prime number $p \in P$ the greatest p-exponent $c_p = \max v_p(n_i)$ for $n_i \in M$ is determined. The product of the primes with the exponents c_p can be determined, since almost all exponents are 0. It is called the least common multiple of M and is designated by lcm. Every number in M divides the lcm.

$$\text{lcm}\,(n_1,...,n_k) = \prod_{p \in P}(p^{c_p}) \quad \wedge \quad c_p = \max_{n_i \in M} v_p(n_i)$$

Mutually prime numbers : The integers of a set $M = \{n_1,...,n_k\}$ are said to be mutually prime if there is no prime which divides all n_i. Since all natural numbers can be uniquely factorized into primes, there is no natural number which divides all the numbers of a mutually prime set. The greatest common divisor is therefore $\gcd(n_1,...,n_k) = 1$.

Division with remainder : For integers a and b with $b \neq 0$ there is a unique representation of a as a multiple of b with a non-negative remainder r which is less than the absolute value of b.

$$a = qb + r \quad \wedge \quad 0 \leq r < |b|$$

Proof : Existence and uniqueness of q and r in division with remainder

(1) It is shown that for given numbers a and b there are integers q and r which satisfy the division formula. Since $qb = (-q)(-b)$ and $a = qb + r \Leftrightarrow -a = (-q-1)\,b + (b-r)$, the proof for $a \geq 0$ and $b > 0$ suffices. The statement is true for $a = 0$. Let it be true for a, so that $a = qb + r$ and $0 \leq r < b$. Hence $a + 1 = qb + (r + 1)$. For $r + 1 < b$, this is the division formula. For $r + 1 = b$, the division formula is $a + 1 = (q + 1)b$. Hence the division formula holds for $a + 1$. It follows by induction that there are numbers q and r for all values of a.

(2) The numbers q and r are shown to be uniquely determined. If there are two representations $a = q_1b + r_1$ and $a = q_2b + r_2$ for the same number a, it follows that $(q_1 - q_2)b = r_2 - r_1$. But $0 \leq r_i < |b|$ implies $|r_2 - r_1| < |b|$, so that $q_1 - q_2 = 0$ and hence $r_2 - r_1 = 0$. Since $q_1 = q_2$ and $r_1 = r_2$, the division formula for a is unique.

6.4 RATIONAL NUMBERS

Invertibility of multiplication : In the set $\mathbb{Z}$ of integers, the equation $a\,x = b$ cannot generally be solved for the unknown x. If there is a solution, it is called the quotient of a and b and is designated by a/b. The equation $0 \cdot x = 0$ has no unique solution; the equation $0 \cdot x = b$ with $b \neq 0$ has no solution at all. A set $\mathbb{Q}$ of numbers is sought in which the equation $a\,x = b$ has a unique solution for any $a \neq 0$.

Construction of $\mathbb{Q}$: The pairs $(a, b) \in \mathbb{Z} \times \mathbb{Z}'$ of integers are partitioned into equivalence classes $[c, d]$. All pairs $(a, b) \in [c, d]$ have the same ratio $a : b = c : d$, so that $a \cdot d = b \cdot c$. The set $\mathbb{Q}$ of rational numbers is the set of all equivalence classes $[c, d]$ in $\mathbb{Z} \times \mathbb{Z}'$.

Normal representation : The definition of the rational numbers and the cancellation law for integers imply $[a, b] = [-a, -b]$ and $[ac, bc] = [a, b]$. Hence every rational number may be represented in the normal form $[a, b]$ with $b > 0$ such that a and b are mutually prime. A rational number is said to be positive if $a > 0$ in its normal representation, and negative if $a < 0$. A rational number with the normal form $[a, b]$ with $b > 0$ is also represented as a fraction $\frac{a}{b}$ and designated by p. The rational number with the normal form [0, 1] is called zero and designated by 0. The rational number with the normal form [1, 1] is called one and designated by 1.

Algebraic structure : The inner operations $+$ (addition) and $\cdot$ (multiplication) on the rational numbers are defined as follows with respect to the equivalence classes of pairs of integers :

addition : $[a, b] + [c, d] = [a \cdot d + b \cdot c,\ b \cdot d]$

multiplication : $[a, b] \cdot [c, d] = [a \cdot c,\ b \cdot d]$

The pair [0,1] is the identity element 0 of addition, since $[a, b] + [0,1] = [a, b]$. The pairs $[a, b]$ and $[-a, b]$ for $b \neq 0$ are additive inverses, since $[a, b] + [-a, b] = [0, b^2]$ and the normal form of $[0, b^2]$ is [0,1]. Addition is associative and commutative. Hence the domain $(\mathbb{Q}\ ; +)$ is a commutative group.

The pair [1,1] is the identity element 1 of multiplication, since $[a, b] \cdot [1,1] = [a, b]$. The pairs $[a, b]$ and $[b, a]$ for $a, b \neq 0$ are multiplicative inverses, since $[a, b] \cdot [b, a] = [a \cdot b,\ a \cdot b]$ and the normal form of $[a \cdot b,\ a \cdot b]$ is [1,1]. There is no pair of integers which is the multiplicative inverse of the pair $[0,1] =: 0$. Multiplication is associative and commutative. Hence the domain $(\mathbb{Q} - \{0\}\ ; \cdot)$ is a commutative group.

Multiplication is distributive with respect to addition. Hence the domain $(\mathbb{Q}\ ; +, \cdot)$ is a commutative field. The cancellation law holds as it does for integers.

Extension of $\mathbb{Z}$: The injective mapping $i : \mathbb{Z} \rightarrow \mathbb{Q}$ with $i(a) = [a, 1]$ preserves structure (is homomorphic), since $i(a + b) = i(a) + i(b)$ and $i(a \cdot b) = i(a) \cdot i(b)$. Hence the set $\mathbb{Z}$ of integers may be extended to the set $\mathbb{Q}$ of rational numbers by adding the fractions. The algebraic structure is thus extended from a commutative ring to a commutative field. In $\mathbb{Q}$, the equation $[a, 1] \cdot x = [b, 1]$ for $a \neq 0$ has the solution $x = [b, a]$, which corresponds to the fraction $\frac{b}{a}$. Subsets of $\mathbb{Q}$ are designated as follows :

$\mathbb{Q}'$	rational numbers without zero
$\mathbb{Q}^+$	positive rational numbers
$\mathbb{Q}_0^+$	positive rational numbers and zero
$\mathbb{Q}^-$	negative rational numbers
$\mathbb{Q}_0^-$	negative rational numbers and zero

Subtraction : In the set $\mathbb{Q}$ of rational numbers, the equation $[a, b] + x = [c, d]$ can be solved without restrictions. The solution $x = [bc - ad, bd]$ is called the difference of the rational numbers and is designated by $[c, d] - [a, b]$. In the normal representation, correspondingly, $p + x = s$ is solved by $x = s - p$.

Division : In the set $\mathbb{Q}$ of rational numbers, the equation $[a, b] \cdot x = [c, d]$ can be solved for any $a, d \neq 0$. The solution $x = [bc, ad]$ is called the quotient of the rational numbers and is designated by $x = [c, d] \ / \ [a, b]$. In the normal representation, correspondingly, $p \cdot x = s$ is solved by $x = s / p$.

Rules of calculation : The basic arithmetic operations of addition, subtraction, multiplication and division can be carried out in $\mathbb{Q}$ without restrictions except for division by zero. The rules of calculation follow from the normal representation and the algebraic structure of $\mathbb{Q}$.

addition	:	$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd}$	$b, d \neq 0$
subtraction	:	$\frac{a}{b} - \frac{c}{d} = \frac{ad - bc}{bd}$	$b, d \neq 0$
multiplication	:	$\frac{a}{b} \cdot \frac{c}{d} = \frac{ac}{bd}$	$b, d \neq 0$
division	:	$\frac{a}{b} \ / \ \frac{c}{d} = \frac{ad}{bc}$	$b, c, d \neq 0$

Ordinal structure : The relation $\le$ (less than or equal to) and the relation $<$ for rational numbers are defined as follows :

$$p \le s \;:\Leftrightarrow\; p - s \le 0$$
$$p < s \;:\Leftrightarrow\; p \le s \;\wedge\; p \ne s$$

Like the integers, the rational numbers are totally ordered. As for integers, the ordinal and the algebraic structure of the rational numbers are compatible, since the monotonic laws for addition and multiplication also hold in $\mathbb{Q}$. In contrast to the integers, there are always further rational numbers between two rational numbers p and s with $p \ne s$, such as $(p+s)/2$. Every open interval $]p, s[= \{q \in \mathbb{Q} \mid p < q < s\}$ is therefore an infinite set.

Absolute value : The absolute value (magnitude) of a rational number p is a positive rational number which is designated by $|p|$ and is determined as follows :

$$|p| := p \quad \text{for} \quad p \ge 0$$
$$|p| := -p \quad \text{for} \quad p < 0$$

Topological structure : The rational space $\mathbb{Q}$ is equipped with metric structure using the function $d : \mathbb{Q} \times \mathbb{Q} \to \mathbb{Q}_0^+$ with $d(p, s) = |p - s|$. The topology of the metric space $\mathbb{Q}$ is generated by the open intervals $]p, s[:= \{r \in \mathbb{Q} \mid p < r < s\}$.

6.5 REAL NUMBERS

Irrational numbers : In $\mathbb{Q}$, the linear equation $p \cdot x = s$ with $p \neq 0$ can be solved without restrictions. However, the non-linear equation $x \cdot x = 2$ does not have a rational solution x. For if there were a rational number x with the normal representation a / b and $a^2 = 2b^2$, then a^2 would be even, and hence a would be divisible by 2, so that $a = 2n$. Then substitution would yield $b^2 = 2n^2$, so that b would be even. That a and b have the common divisor 2 contradicts the assumption that a / b is the normal representation of a rational number.

Non-linear equations of the form $x^2 = 2$ arise frequently, for example in determining the length of the diagonal of a square of side length 1. Since they cannot be solved by rational numbers, a set $\mathbb{R}$ of numbers is sought which includes $\mathbb{Q}$ and contains solutions of equations of the form $x^p = s$ with $p, s \in \mathbb{Q}$. To define $\mathbb{R}$, the concepts of open initial segment and open initial are introduced.

Open initial segment : Let $(M ; \leq)$ and $(M ; <)$ be totally ordered sets. Then for every element $q \in M$ there is a unique subset which contains all elements of M which are less than q. Such a subset is called an (open) initial segment in M and is designated by S_q.

$$S_q := \{x \in M \mid x < q\} \subseteq M \quad \text{for} \quad q \in M$$

Open initial : Let $(M ; \leq)$ and $(M ; <)$ be totally ordered sets. Then a subset $A \subseteq M$ is called an initial in M if for every element $a \in A$ all elements $x \in M$ with $x \leq a$ also belong to A. An initial A without a greatest element is called an open initial in M (see Section 5.4).

Open initial segment and open initial in $\mathbb{Q}$: The set $\mathbb{Q}$ of rational numbers is totally ordered by the relations $\leq$ and $<$. Every open initial segment in $\mathbb{Q}$ is an open initial in $\mathbb{Q}$. But not every open initial in $\mathbb{Q}$ is an open initial segment in $\mathbb{Q}$.

Example 1 : Open initial segment and open initial in $\mathbb{Q}$

The set $S_2 = \{x \in \mathbb{Q} \mid x < 2\}$ contains all rational numbers which are less than 2. Hence it is an open initial segment in $\mathbb{Q}$. For every rational number $a \in S_2$, the set S_2 also contains all rational numbers $x \leq a$. Hence it is an initial in $\mathbb{Q}$. The set S_2 contains no greatest element, since for every positive rational number $\frac{c}{d}$ in S_2 the greater rational number $\frac{c}{d} + \frac{1}{2cd}$ also belongs to S_2 :

- A positive rational number $0 < x < 2$ is a quotient of natural numbers c / d with $c, d > 0$. From $0 < c/d < 2$ it follows that $0 < c < 2d$, and hence $c = 2d - n$ with the natural number $0 < n < 2d$.

- A positive rational number $y > x$ is formed according to the specified rule. The rule is transformed using $c = 2d - n$:

$$y = \frac{c}{d} + \frac{1}{2cd} = \frac{2d-n}{d} + \frac{1}{2cd} = 2 - \frac{1}{d}(n - \frac{1}{2c})$$

- The term $n - \frac{1}{2c}$ in the expression for y is positive since $n, c \geq 1$ and $\frac{1}{2c} < \frac{1}{2}$. For $n = 1$ and $c = 1$, it takes the least value $\frac{1}{2}$. This implies $y < 2$:

$$y = 2 - \frac{1}{d}(n - \frac{1}{2c}) \leq 2 - \frac{1}{2d} < 2$$

It follows from $x < 2$, $y > x$ and $y > 2$ that S_2 has no greatest element. Hence S_2 is an open initial.

Example 2 : Open initial in $\mathbb{Q}$

The set $A = \{x \in \mathbb{Q} \mid x < 0 \quad \vee \quad x^2 < 2\}$ contains all negative rational numbers and all non-negative rational numbers x with $x^2 < 2$. It is an initial in $\mathbb{Q}$. The set A contains no greatest element, since for every positive rational number $\frac{c}{d}$ in A the greater number $\frac{c}{d} + \frac{1}{3cd}$ is also contained in A :

- A positive rational number x with $x^2 < 2$ is a quotient of the natural numbers c/d with $c, d > 0$. It follows from $c^2/d^2 < 2$ that $c^2 < 2d^2$, and hence $c^2 = 2d^2 - n$ with the natural number $0 < n < 2d^2$.

- A positive rational number $y > x$ is formed according to the specified rule. The expression y^2 is transformed using $c^2 = 2d^2 - n$:

$$y = \frac{c}{d} + \frac{1}{3cd}$$

$$y^2 = \frac{c^2}{d^2} + \frac{2}{3d^2} + \frac{1}{9c^2d^2} = \frac{2d^2 - n}{d^2} + \frac{2}{3d^2} + \frac{1}{9c^2d^2}$$

$$y^2 = 2 - \frac{1}{d^2}(n - \frac{2}{3} - \frac{1}{9c^2})$$

- The term $n - \frac{2}{3} - \frac{1}{9c^2}$ is positive since $n, c \geq 1$ and $\frac{1}{9c^2} \leq \frac{1}{9}$. For $n = 1$ and $c = 1$, it takes the least value $\frac{2}{9}$. This implies $y^2 > 2$:

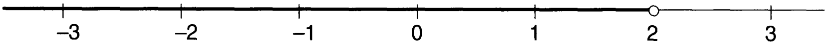
$$y^2 = 2 - \frac{1}{d^2}(n - \frac{2}{3} - \frac{1}{9c^2}) \leq 2 - \frac{1}{9d^2} < 2$$

It follows from $x^2 < 2$, $y < x$ and $y^2 > 2$ that A does not have a greatest element. Hence A is an open initial in $\mathbb{Q}$. A is not an open initial segment in $\mathbb{Q}$, since there is no rational number q with $q^2 = 2$.

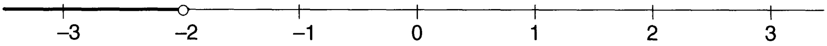
Construction of $\mathbb{R}$: A set $\mathbb{R}$ of real numbers is constructed such that every open initial in $\mathbb{R}$ is an open initial segment in $\mathbb{R}$. An open initial in the set $\mathbb{Q}$ of rational numbers is called a real number. The set $\mathbb{R}$ of real numbers is the set of open initials in $\mathbb{Q}$. Real numbers are designated by lowercase letters such as r and s.

Representation : A real number which is both an open initial A and an open initial segment S_q in $\mathbb{Q}$ is designated by the rational number q. Real numbers which are not open initial segments in $\mathbb{Q}$ are called irrational numbers. Irrational numbers are designated by their own symbols, such as $\sqrt{2}$ or π or e. The following diagram shows selected real numbers with their open initials in $\mathbb{Q}$ on the number line.

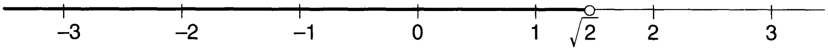
$$2 := \{x \in \mathbb{Q} \mid x < 2\} = S_2$$



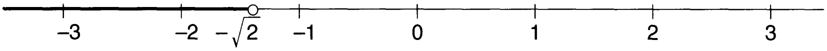
$$-2 := \{x \in \mathbb{Q} \mid x < -2\} = S_{-2}$$



$$\sqrt{2} := \{x \in \mathbb{Q} \mid x < 0 \lor x^2 < 2\}$$



$$-\sqrt{2} := \{x \in \mathbb{Q} \mid x < 0 \land x^2 < 2\} = \{x \in \mathbb{Q} \mid -x \in (\mathbb{Q} - \sqrt{2})\}$$



Ordinal structure : The relations $\leq$ (less than or equal to) and $<$ (less than) in the set of real numbers are defined as relations between open initials in $\mathbb{Q}$ as follows :

$$r \leq s \;:\Leftrightarrow\; r \subseteq s \qquad r, s \in \mathbb{R}$$
$$r < s \;:\Leftrightarrow\; r \subset s$$

Like the rational numbers, the real numbers are totally ordered. In contrast to the rational numbers, the least upper bound theorem holds for the real numbers : Every non-empty subset T of $\mathbb{R}$ bounded from above has a least upper bound (supremum). For example, $T = \{x \in \mathbb{R} \mid x^2 < 2\}$ has the least upper bound $\sqrt{2}$.

Algebraic structure : The inner operations + (addition) and • (multiplication) on the real numbers are defined as operations on open initials in $\mathbb{Q}$ as follows :

addition : $r + s := \{x + y \mid x \in r \land y \in s\}$

multiplication : $r \cdot s := \mathbb{Q}^- \cup \{x \cdot y \mid x \in (r - \mathbb{Q}^-) \land y \in (s - \mathbb{Q}^-)\}$

for $r, s \geq 0$

The open initial $S_0 = \{x \in \mathbb{Q} \mid x < 0\} = \mathbb{Q}^-$ is the identity element 0 of addition, since $r + 0 = \{x + y \mid x \in r \ \wedge \ y < 0\} = r$. The real numbers r and $-r$ are additive inverses, so that $r + (-r) = 0$. Their open initials satisfy :

$$\begin{array}{lll} r \text{ is rational} & \Rightarrow & -r = S_{-r} \\ r \text{ is irrational} & \Rightarrow & -r = \{x \mid -x \in (\mathbb{Q} - r)\} \end{array}$$

Addition is associative and commutative. Hence the domain $(\mathbb{R} \, ; +)$ is a commutative group.

The open initial $S_1 = \{x \in \mathbb{Q} \mid x < 1\}$ is the identity element 1 of multiplication, since for $r \geq 0$ by definition $r \cdot 1 = \mathbb{Q}^- \cup \{x \cdot y \mid x \in (r - \mathbb{Q}^-) \wedge 0 \leq y < 1\} = \mathbb{Q}^- \cup (r - \mathbb{Q}^-) = r$. The real numbers r and r^{-1} are multiplicative inverses, so that $r \cdot r^{-1} = 1$ for $r > 0$. Their open initials satisfy :

$$r > 0 \quad \Rightarrow \quad r^{-1} = \bigcup_{x \in \mathbb{Q} - r} S_{x^{-1}}$$

There is no real number which is the multiplicative inverse of $r = 0$. The multiplication of negative real numbers is reduced to the multiplication of non-negative real numbers and the formation of additive inverses using the relationships $r \cdot s = (-r) \cdot (-s) = -(r \cdot (-s)) = -((-r) \cdot s)$. The formation of multiplicative inverses of negative real numbers is reduced to the formation of multiplicative inverses of positive real numbers and the formation of additive inverses using the relationship $r^{-1} = -(-r)^{-1}$. Multiplication is associative and commutative. Hence the domain $(\mathbb{R} - \{0\} \, ; \ \cdot)$ is a commutative group.

Multiplication is distributive with respect to addition. Hence the domain $(\mathbb{R} \, ; +, \cdot)$ is a commutative field. The cancellation law holds as it does for rational numbers. The ordinal and the algebraic structure are compatible as in the case of rational numbers, since the monotonic laws for addition and multiplication also hold in $\mathbb{R}$.

Extension of $\mathbb{Q}$: The injective mapping $i : \mathbb{Q} \rightarrow \mathbb{R}$ with $i(q) = S_q$ preserves structure, since $i(r + s) = i(r) + i(s)$ and $i(r \cdot s) = i(r) \cdot i(s)$. Hence the set $\mathbb{Q}$ of rational numbers may be extended to the set $\mathbb{R}$ of real numbers by adding the irrational numbers. The order structure is completed while the algebraic structure is preserved. Subsets of $\mathbb{R}$ are designated as follows :

$\mathbb{R}'$	real numbers without zero
$\mathbb{R}^+$	positive real numbers
$\mathbb{R}_0^+$	positive real numbers and zero
$\mathbb{R}^-$	negative real numbers
$\mathbb{R}_0^-$	negative real numbers and zero

Powers and roots : The algebraic structure of $\mathbb{R}$ is extended by taking powers and roots of real numbers with integer exponents :

$$\begin{array}{llcll}
\text{powers :} & r^0 & := & 1 & \\
 & r^{n+1} & := & r \cdot r^n & n \in \mathbb{N}, \quad r \in \mathbb{R} \\
 & r^{-n} & := & 1 \,/\, r^n & n \in \mathbb{N}, \quad r \in \mathbb{R}' \\
\text{roots :} & \sqrt[n]{r} & := & \mathbb{Q}^- \cup \{x \in \mathbb{Q}_0^+ \mid x^n < r\} & n \in \mathbb{N}', \; r \in \mathbb{R}_0^+ \\
 & r^{m/n} & := & (\sqrt[n]{r})^m & m \in \mathbb{Z}^+, \; r \in \mathbb{R}_0^+ \\
 & & & & m \in \mathbb{Z}_0^+, \; r \in \mathbb{R}^+
\end{array}$$

The powers of real bases with real exponents are defined such that the result always lies in $\mathbb{R}_0^+$:

$$\begin{array}{lcllll}
r \geq 1 & : & r^s & := & \bigcup\limits_{p \in s} r^p & r \in \mathbb{R}_0^+, s \in \mathbb{R}, p \in \mathbb{Q} \\
0 < r < 1 & : & r^s & := & (r^{-1})^{-s} & \\
r = 0 & : & 0^0 & := & 1\,, \text{ otherwise } \; 0^s := 0 &
\end{array}$$

The following rules of calculation hold for powers :

$$\begin{array}{lclcl}
\text{same basis} & : & r^{s_1} \cdot r^{s_2} & = & r^{s_1 + s_2} \\
 & & r^{s_1} / r^{s_2} & = & r^{s_1 - s_2} \\
\text{same exponent} & : & r_1^s \cdot r_2^s & = & (r_1 \cdot r_2)^s \\
 & & r_1^s \,/\, r_2^s & = & (r_1 \,/\, r_2)^s \\
\text{multiple power} & : & (r^{s_1})^{s_2} & = & r^{s_1 \cdot s_2}
\end{array}$$

Logarithms : The solution x of the equation $r^x = s$ with $r, s \in \mathbb{R}^+$ and $r \neq 1$ is called the logarithm of s to base r and is designated by $x = \log_r s$. The abbreviation $\lg s$ is used for $\log_{10} s$, and the abbreviation $\ln s$ is used for $\log_e s$. The following rules of calculation hold for logarithms :

$$\begin{array}{lclcl}
\text{product} & : & \log_r (s_1 \cdot s_2) & = & \log_r s_1 + \log_r s_2 \\
\text{quotient} & : & \log_r (s_1 / s_2) & = & \log_r s_1 - \log_r s_2 \\
\text{power} & : & \log_r (s^t) & = & t \cdot \log_r s \\
\text{root} & : & \log_r \left(\sqrt[t]{s}\right) & = & \frac{1}{t} \cdot \log_r s
\end{array}$$

Topological structure of $\mathbb{R}$: In the set $\mathbb{Q}$ of rational numbers, not every fundamental sequence is convergent. For example, $\sqrt{2}$ cannot be represented by a convergent fundamental sequence in $\mathbb{Q}$. A set of numbers is desired in which every fundamental sequence of rational numbers converges. By property (F5) in Section 5.10.1, the real numbers defined as open initials have this property. Alternatively, the convergence of all fundamental sequences in $\mathbb{R}$ may be postulated; the real numbers are then derived as b-adic fractions.

b-adic fraction : A b-adic fraction is a series x_n with the basis b, the exponents i and the digits a_i. If the basis b is fixed, the fraction is uniquely described by its sequence of digits. The bases 2 (binary system) and 10 (decimal system) are often used.

$$x_n = \sum_{i=-k}^{n} a_i\, b^{-i} \qquad a_i,\ b,\ k,\ n \in \mathbb{N}$$

$$x_n = a_{-k} \ldots a_{-1}\, a_0 . a_1 \ldots a_n \qquad x = \lim_{n \to \infty} x_n \in \mathbb{R}$$

$$b \geq 2 \qquad \text{basis of the fraction}$$

$$0 \leq a_i < b \qquad \text{digits of the fraction}$$

Every b-adic fraction is a fundamental sequence, since for $n \geq m \geq -k$ and ε in $\mathbb{Q}^+$ and for sufficiently large m :

$$|x_n - x_m| = \sum_{i=m+1}^{n} a_i\, b^{-i} \leq \sum_{i=m+1}^{n} (b-1) b^{-i}$$

$$|x_n - x_m| \leq (b-1) b^{-m-1} \sum_{i=0}^{n-m-1} b^{-i} \leq (b-1) b^{-m-1} \frac{1}{1-b^{-1}} = b^{-m} < \varepsilon$$

Expansion of real numbers : Every non-negative real number x may be expanded into a b-adic fraction. The proof is carried out inductively.

Let k be the least natural number with $0 \leq x < b^{k+1}$. Consider the partitioning $0 = 0 \cdot b^k < 1 \cdot b^k < \ldots < b \cdot b^k = b^{k+1}$. There is exactly one natural number a_{-k} with $0 \leq a_{-k} < b - 1$ and $x_{-k} = a_{-k}\, b^k$ such that $x_{-k} \leq x < x_{-k} + b^k$.

Let the digits $a_{-k}, \ldots, a_n$ be known for $x_n \leq x < x_n + b^{-n}$. Consider the partitioning $x_n < x_n + b^{-n-1} < x_n + 2b^{-n-1} < \ldots < x_n + b \cdot b^{-n-1}$. There is exactly one natural number a_{n+1} with $0 \leq a_{n+1} < b - 1$ and $x_{n+1} = x_n + a_{n+1}\, b^{-n-1}$ such that $x_{n+1} \leq x < x_{n+1} + b^{-n-1}$. From $|x - x_n| < b^{-n}$ for $n \geq -k$ it follows that $\lim_{n \to \infty} x_n = x$.

Example 3 : Calculation of the square root of $a \in \mathbb{Q}^+$

The square root of a is the limit of the following convergent sequence :

$$x_{i+1} = \frac{1}{2}\left(x_i + \frac{a}{x_i}\right) \qquad i \in \mathbb{N},\ x_i \in \mathbb{R}^+$$

For $a = 5$ and $x_0 = 1$, the sequence has the following terms :

$$\begin{aligned}
x_0 &= 1.000\,000\,000 \\
x_1 &= 3.000\,000\,000 \\
x_2 &= 2.333\,333\,333 \\
x_3 &= 2.238\,095\,238 \\
x_4 &= 2.236\,068\,896 \\
x_5 &= 2.236\,067\,978
\end{aligned}$$

6.6 COMPLEX NUMBERS

Imaginary number : For non-negative real numbers $a \in \mathbb{R}_0^+$, the quadratic equation $x^2 = a$ has real solutions. For negative real numbers $a \in \mathbb{R}^-$, however, there are no real solutions. A set $\mathbb{C}$ of numbers is sought in which equations of the form $x^2 + 1 = 0$ can be solved. The solution i of the equation $x^2 = -1$ is called the imaginary unit. The numbers $b\,i$ with $b \in \mathbb{R}$ are called imaginary numbers.

Construction of $\mathbb{C}$: In analogy with the construction of the sets of integers and of rational numbers, pairs $(a, b) \in \mathbb{R} \times \mathbb{R}$ of real numbers are introduced. The set $\mathbb{C}$ of complex numbers is the set $\mathbb{R} \times \mathbb{R}$ of all pairs of real numbers.

Algebraic structure : The inner operations + (addition) and • (multiplication) on complex numbers are defined using pairs of real numbers as follows :

$$\begin{array}{lll} \text{addition} & : & (a,b) + (c,d) \ := \ (a+c,\ b+d) \\ \text{multiplication} & : & (a,b) \cdot (c,d) \ := \ (a \cdot c - b \cdot d,\ a \cdot d + b \cdot c) \end{array}$$

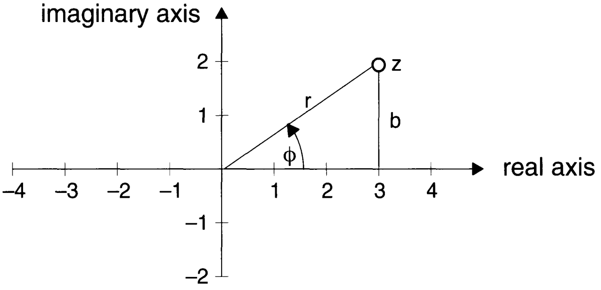
The pair (0, 0) is the identity element of addition, since $(a,b) + (0,0) = (a,b)$. The pairs (a,b) and (−a,−b) are additive inverses, since $(a,b) + (-a,-b) = (0,0)$. Addition is associative and commutative. Hence the domain $(\mathbb{R} \times \mathbb{R}\ ;\ +)$ is a commutative group.

The pair (1, 0) is the identity element of multiplication, since $(a,b) \cdot (1,0) = (a,b)$. The pairs (a,b) and $(\,a/(a^2 + b^2), -b/(a^2 + b^2))$ for $(a,b) \neq (0,0)$ are multiplicative inverses, since $(a,b) \cdot (\,a/(a^2 + b^2), -b/(a^2 + b^2)) = (1,0)$. Multiplication is associative and commutative. Hence the domain $(\mathbb{R} \times \mathbb{R}\ - \{(0,0)\}\ ;\cdot\,)$ is a commutative group.

Multiplication is distributive with respect to addition. Hence the domain $(\mathbb{R}\ \times \mathbb{R}\ ;\ +, \cdot\,)$ is a commutative field. The cancellation law holds as it does for real numbers. In Example 1 of Section 3.5, the set of complex numbers is shown to be a two-dimensional vector space over the field of real numbers.

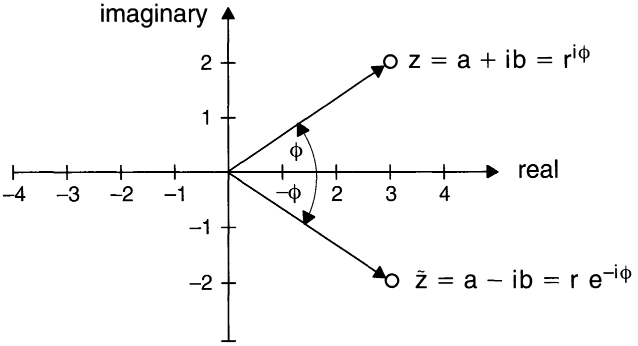
Extension of $\mathbb{R}$: The injective mapping $i : \mathbb{R} \to \mathbb{C}$ with $i(a) = (a, 0)$ preserves structure (is homomorphic), since $i(a+b) = i(a) + i(b)$ and $i(a \cdot b) = i(a) \cdot i(b)$. The set $\mathbb{R}$ of real numbers is extended to the set $\mathbb{C}$ of complex numbers by adding the pairs (a,b) of real numbers with $a, b \in \mathbb{R}$ and $b \neq 0$. If $(a,b) \in \mathbb{C}$ is a complex number, then $a \in \mathbb{R}$ is called its real part and $b \in \mathbb{R}$ is called its imaginary part. Instead of a pair $(a,b) \in \mathbb{C}$ of numbers, the notation $a + ib$, which derives from the solution of quadratic equations with real coefficients, is often used. For $a \in \mathbb{R}_0^+$, the quadratic equation $x^2 = a$ has the solutions $x_1 = (\sqrt{|a|}, 0) =: \sqrt{|a|}$ and $x_2 = (-\sqrt{|a|}, 0) = -\sqrt{|a|}$; for $a \in \mathbb{R}^-$ the solutions are $x_1 = (0, \sqrt{|a|}) =: i\sqrt{|a|}$ and $x_2 = (0, -\sqrt{|a|}) = -i\sqrt{|a|}$. The extension of $\mathbb{R}$ to $\mathbb{C}$ preserves the algebraic structure. The set $\mathbb{C}$ without the zero element $(0, 0) =: 0$ is designated by $\mathbb{C}'$.

Complex plane : Complex numbers $z \in \mathbb{C}$ are graphically represented in the complex plane. The number z is specified either by its cartesian coordinates a, b or by its polar coordinates r, ϕ. The normal representation can be used instead of the polar representation.



cartesian form	:	z	=	$a + ib$	
		a	=	$re(z)$	real part of z
		b	=	$im(z)$	imaginary part of z
polar form	:	z	=	$r\,(\cos\phi$	$+\ i \sin\phi)$
		r	=	$\lvert z \rvert$	absolute value of z : $\lvert z \rvert = \sqrt{a^2 + b^2}$
		ϕ	=	arg z	argument of z $(0 \le \phi < 2\pi)$
normal representation	:	z	=	$r\,e^{i\phi}$	exponential function

The reflection of the complex number z with respect to the real axis is called the corresponding conjugate complex number $\tilde{z}$. The numbers z and $\tilde{z}$ differ only in the sign of their imaginary part.



Rules of calculation : The basic arithmetic operations of addition, subtraction, multiplication and division can be carried out in $\mathbb{C}$ without restriction except for division by zero.

addition : $(a + ib) + (c + id) = (a + c) + i(b + d)$

subtraction : $(a + ib) - (c + id) = (a - c) + i(b - d)$

multiplication : $(a+ib) \cdot (c+id) = (ac-bd) + i(ad+bc)$

division : $(a+ib) \;/\; (c+id) = \dfrac{ac+bd}{c^2+d^2} + \dfrac{bc-ad}{c^2+d^2} \qquad c, d \neq 0$

If multiplication and division are carried out in the normal representation, the arguments of the results must be reduced mod 2π to values in the range $0 \leq \phi < 2\pi$.

multiplication : $z_1 \cdot z_2 = (r_1 e^{i\phi_1}) \cdot (r_2 e^{i\phi_2}) = r_1 r_2 e^{i(\phi_1 + \phi_2)}$

division : $z_1 / z_2 = (r_1 e^{i\phi_1}) / (r_2 e^{i\phi_2}) = \dfrac{r_1}{r_2} e^{i(\phi_1 - \phi_2)}$

Powers and roots : The power and the root of a complex number z with respect to a natural number n are reduced to arithmetic operations on real numbers :

normal representation : $z = re^{i\phi} = r(\cos\phi + i\sin\phi) \qquad z \in \mathbb{C}'$

power : $z^n = r^n e^{in\phi} \qquad n \in \mathbb{N}$

$|z^n| = |z|^n$

$\arg z^n = (n \cdot \arg z) \mod 2\pi \qquad 0 \leq \arg z^n < 2\pi$

root : $\sqrt[n]{z} = z^{1/n} \qquad n \in \mathbb{N}'$

For a complex exponent $w = a + ib$, the absolute value and argument of the products are separated :

power : $z^w = (re^{i\phi})^{a+ib} = r^a r^{ib} e^{i\phi a - \phi b} \qquad w \in \mathbb{C}$

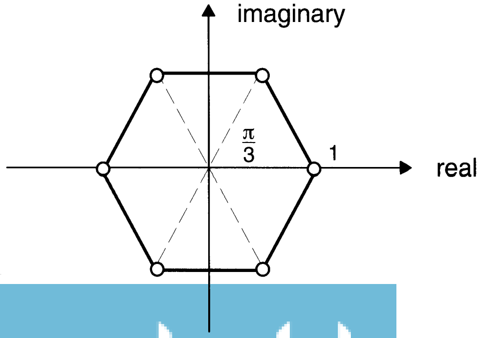
$z^w = (r^a e^{-\phi b}) e^{i(\phi a + b \ln r)}$

root : $\sqrt[w]{z} = z^{1/w} \qquad w \in \mathbb{C}'$

Roots of unity : The roots of $x^n = 1$ are called the n-th roots of unity. In the complex plane they lie on the unit circle and are the corners of a regular n-sided polygon. At least one of the roots of unity lies on the real axis.

Example : Roots of unity

$n = 6 : \quad z = e^{2\pi j/n} = e^{\frac{\pi}{3}j} \qquad j = 0, \ldots, n-1$



Logarithms : The solution $x = \ln r + i\phi$ of the equation $e^x = z = re^{i\Phi}$ is called the logarithm of z to base e and is designated by $\ln z$. Further solutions of the equation $e^x = z$ are given by $x = 2\pi k\, i + \ln z$ with $k \in \mathbb{Z}$. A solution $x = \log_w z$ of the equation $w^x = z$ with $w \neq 0, 1$ is given by $\log_w z = \ln z / \ln w$.

Algebraic and transcendental numbers : The zeros of the polynomials $\mathbb{Q}[x]$ in the unknown x with rational coefficients in $\mathbb{Q}$ are called algebraic numbers. Complex numbers which are not zeros of a polynomial in $\mathbb{Q}[x]$ are called transcendental numbers. Examples of transcendental numbers are π and e.

Ordinal structure : The relations $\sqsubset$ (less than) and $\sqsubseteq$ (less than or equal to) may be defined for complex numbers as follows :

$$y \sqsubset z \quad :\Leftrightarrow \quad |y| < |z| \quad \vee \quad (|y| = |z|) \quad \wedge \quad \arg y < \arg z)$$

$$y \sqsubseteq z \quad :\Leftrightarrow \quad y \sqsubseteq z \quad \vee \quad y = z$$

With these definitions, the complex numbers are totally ordered. However, the ordinal and the algebraic structure are not compatible, so that the monotonic laws for the addition and multiplication of complex numbers do not hold. The following example shows that the monotonic law $y \sqsubset z \Leftrightarrow z \cdot y \sqsubset z \cdot z$ for multiplication does not hold :

$$\begin{array}{lll lll lll}
y & = & e^{i\pi/4} & |y| & = & 1 & \arg y & = & \pi/4 \\
z & = & e^{i3\pi/2} & |z| & = & 1 & \arg z & = & 3\pi/2 \\
z \cdot y & = & e^{i7\pi/4} & |z \cdot y| & = & 1 & \arg z \cdot y & = & 7\pi/4 \\
z \cdot z & = & e^{i\pi} & |z \cdot z| & = & 1 & \arg z \cdot y & = & \pi
\end{array}$$

$$y \sqsubset z \quad \Leftrightarrow \quad (1 < 1) \quad \vee \quad ((1 = 1) \quad \wedge \quad (\pi/4 < 3\pi/2)) \quad \Leftrightarrow \quad \text{true}$$

$$z \cdot y \sqsubset z \cdot z \quad \Leftrightarrow \quad (1 < 1) \quad \vee \quad ((1 = 1) \quad \wedge \quad (7\pi/4 < \pi)) \quad \Leftrightarrow \quad \text{false}$$

Topological structure of $\mathbb{C}$: The complex space $\mathbb{C}$ is equipped with metric structure using the function $d : \mathbb{C} \times \mathbb{C} \to \mathbb{R}_0^+$ with $d(x, y) = |x - y|$. This induces a topological structure. Every fundamental sequence of complex numbers has a limit in $\mathbb{C}$.

6.7 QUATERNIONS

Introduction : After the construction of the set $\mathbb{C}$ of complex numbers as an extension of the set $\mathbb{R}$ of real numbers, the question arises whether $\mathbb{C}$ may be extended while preserving the algebraic structure. There is no extension of $\mathbb{C}$ which has the properties of a commutative field. There is, however, an extension of $\mathbb{C}$ which has the properties of a non-commutative field. This extension is called the set of quaternions and is designated by $\mathbb{H}$.

Construction of quaternions : Quaternions are represented by quadratic matrices of row and column dimension 2. In Section 3.6, the additive and multiplicative domain of quadratic matrices of row and column dimension m over a commutative field is shown to be a ring in which multiplicative inverses cannot always be formed. For special matrices with $m = 2$, however, a field in which every element has a multiplicative inverse can be constructed. A quaternion $A \in \mathbb{H}$ is defined as a quadratic matrix of row and column dimension 2 with the complex numbers $a, b \in \mathbb{C}$ and their conjugates $\tilde{a}, \tilde{b} \in \mathbb{C}$:

$$\text{quaternion } A := \begin{array}{|c|c|}\hline a & -\tilde{b} \\ \hline b & \tilde{a} \\ \hline\end{array} = \begin{array}{|c|c|}\hline a_1 + ia_2 & -b_1 + ib_2 \\ \hline b_1 + ib_2 & a_1 - ia_2 \\ \hline\end{array} \qquad \begin{array}{l} a, b \in \mathbb{C} \\ a_1, a_2, b_1, b_2 \in \mathbb{R} \end{array}$$

Algebraic structure : The inner operations + (addition) and $\cdot$ (multiplication) for quaternions are defined as the following matrix operations :

$$\text{addition} \quad : \quad A + B := \begin{array}{|c|c|}\hline a & -\tilde{b} \\ \hline b & \tilde{a} \\ \hline\end{array} + \begin{array}{|c|c|}\hline c & -\tilde{d} \\ \hline d & \tilde{c} \\ \hline\end{array} = \begin{array}{|c|c|}\hline e & -\tilde{f} \\ \hline f & \tilde{e} \\ \hline\end{array}$$

$$e = a + c \qquad f = b + d$$

$$\text{multiplication:} \quad A \cdot B := \begin{array}{|c|c|}\hline a & -\tilde{b} \\ \hline b & \tilde{a} \\ \hline\end{array} \cdot \begin{array}{|c|c|}\hline c & -\tilde{d} \\ \hline d & \tilde{c} \\ \hline\end{array} = \begin{array}{|c|c|}\hline e & -\tilde{f} \\ \hline f & \tilde{e} \\ \hline\end{array}$$

$$e = a \cdot c - \tilde{b} \cdot d \qquad f = b \cdot c + \tilde{a} \cdot d$$

$$\text{identity element of addition} \quad : \quad 0 = \begin{array}{|c|c|}\hline 0 & 0 \\ \hline 0 & 0 \\ \hline\end{array}$$

$$\text{identity element of multiplication} \quad : \quad 1 = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline\end{array}$$

$$\text{additive inverse} \quad : \quad -A = \begin{array}{|c|c|}\hline -a & \tilde{b} \\ \hline -b & -\tilde{a} \\ \hline\end{array}$$

multiplicative inverse :
$$A^{-1} = \begin{array}{|c|c|}\hline \tilde{a} & \tilde{b} \\ \hline -b & a \\ \hline \end{array} \cdot \frac{1}{a \cdot \tilde{a} + b \cdot \tilde{b}}$$

The sum of two quaternions is a quaternion. The quaternion 0 is the identity element of addition, since $A + 0 = A$. The quaternions A and $-A$ are additive inverses, since $A + (-A) = 0$. Addition is associative and commutative. Hence the domain $(\mathbb{H} ; +)$ is a commutative group.

The product of two quaternions is a quaternion. The quaternion 1 is the identity element of multiplication, since $A \cdot 1 = A$. The quaternions $A \neq 0$ and A^{-1} are multiplicative inverses, since $A \cdot A^{-1} = 1$. Multiplication is associative but not commutative. Hence the domain $(\mathbb{H} - \{0\} ; \cdot)$ is a non-commutative group.

Multiplication is distributive with respect to addition. Hence the domain $(\mathbb{H} ; +, \cdot)$ is a field. The cancellation law holds for quaternions.

A complex number $a_1 + ib_1 \in \mathbb{C}$ may be represented as a special quaternion with the real numbers $a_1, b_1 \in \mathbb{R}$ and $a_2 = b_2 = 0$. In this case multiplication is commutative. Hence the additive and multiplicative domain of these special quaternions is a commutative field, like $(\mathbb{C} ; +, \cdot)$. The commutative field $(\mathbb{C} ; +, \cdot)$ of complex numbers is thus a subfield of the field $(\mathbb{H} ; +, \cdot)$ of quaternions.

Vector space : The complex numbers form a two-dimensional vector space over the field of real numbers. The quaternions form a four-dimensional vector space over the field of real numbers. Every quaternion A may thus be represented as a linear combination of four basis quaternions with real coefficients. The definition of the quaternions implies :

$$A = a_1 \cdot 1 + a_2 \cdot I + b_1 \cdot J + b_2 \cdot K \qquad a_1, a_2, b_1, b_2 \in \mathbb{R}$$

$$1 = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 0 & 1 \\ \hline \end{array} \quad I = \begin{array}{|c|c|}\hline i & 0 \\ \hline 0 & -i \\ \hline \end{array} \quad J = \begin{array}{|c|c|}\hline 0 & -1 \\ \hline 1 & 0 \\ \hline \end{array} \quad K = \begin{array}{|c|c|}\hline 0 & i \\ \hline i & 0 \\ \hline \end{array}$$

If a complex number $a_1 + ib_1 \in \mathbb{C}$ is considered as a special quaternion with the real numbers $a_1, b_1 \in \mathbb{R}$ and $a_2 = b_2 = 0$, it may be represented as a linear combination of the basis quaternions 1 and I with real coefficients a_1 and a_2. The two-dimensional vector space of complex numbers is thus a subspace of the four-dimensional vector space of quaternions.

Extension of $\mathbb{C}$: The injective mapping $i := \mathbb{C} \rightarrow \mathbb{H}$ with $i(a) = A$ preserves structure (is homomorphic), since $i(a + b) = i(a) + i(b)$ and $i(a \cdot b) = i(a) \cdot i(b)$. The set $\mathbb{C}$ of complex numbers may be extended to the set $\mathbb{H}$ of quaternions. In this extension, the algebraic structure of a field is preserved, but the commutativity of the field is lost.

7 GROUPS

7.1 INTRODUCTION

7.1.1 GROUP THEORY

Introduction : The group is one of the fundamental concepts in mathematics. This is hardly surprising, since for example the integers form a group with respect to addition. Group theory provides a typical example of the mathematical method. Starting from a small number of definitions and axioms, a body of knowledge arises which has led to the solution of central problems in mathematics. Today, group theory belongs to the foundations of science and engineering. The present chapter is only an introduction to the amazingly rich properties of groups.

Characteristic property : A group is a relation in the threefold cartesian product $M \times M \times M$ of a set M. The elements of the relation are 3-tuples $(a, b, c) \in M \times M \times M$ with the following property : Two of the components a, b, c may be chosen freely in M, for example a and c. Then there is exactly one value of the third component (of b in this example) for which the tuple (a, b, c) belongs to the relation. The unique dependence of a third value on two given values is the characteristic property of a group.

Representation of groups : A group is usually not represented as a relation $R \subseteq M \times M \times M$, but as a mapping $f : M \times M \rightarrow M$ with $f(a,b) = c$. By definition, the mapping f assigns a unique element c of M to every pair (a,b) in $M \times M$. The image c is called the result of the operation for the pair (a,b), the mapping f is called the rule of the operation. Instead of the letter f, an operator symbol may be used, as in $a \circ b := f(a,b)$. In this case, the group is designated by $(G\,;\circ)$. However, this representation of a group captures the characteristic properties of a group only if additional conditions are satisfied : The group must contain an identity element 1, and for every element a it must contain an inverse element a^{-1}, since for $a \circ b = c$:

$$\begin{array}{lcl}
a, b \text{ known} & \Rightarrow & (a,b,c) = (a, \; b \;, a \circ b) \\
a, c \text{ known} & \Rightarrow & (a,b,c) = (a, a^{-1} \circ c \,, c \;) \\
a, c \text{ equal} & \Rightarrow & (a,b,c) = (a, \; 1 \;, c \;)
\end{array}$$

In expressions with more than two operands the operation $\circ$ is assumed to be associative.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*
© Springer-Verlag Berlin Heidelberg 2001

Structure : The aim of group theory is an understanding of the algebraic structure of groups. For example, the following questions are studied :

(1) Can the elements of a group $(G ; \circ)$ be generated by applying the group operation to the elements of a subset of G? This question is suggested by the concept of a basis for a vector space. It turns out that the concept of a basis is not sufficient for general groups.

(2) Can the set G be partitioned into classes of similar elements using the operation $\circ$? This question is suggested by the concept of type formation. It leads to the concepts of normal subgroups and quotient groups, which are of central importance in group theory.

(3) Are there mappings from a group $(G ; \circ)$ that preserve its structure? This question is suggested by the invariants of topological structures. It leads to the concept of isomorphic groups which cannot be distinguished by their group structure. It turns out, for example, that every finite group is isomorphic to a group of permutations.

(4) Are there finite groups of equal order which are not isomorphic? The answer is: yes. Does a group whose order is divisible by m always contain a subgroup of order m? For general groups, the answer is: no. For commutative groups, the answer is: yes.

The algebraic structure of groups is very diverse. The structure of groups with a commutative operation (abelian groups) is completely known and may be described by a small number of invariants (Betti number, torsion coefficients). The structure of the permutations of a finite set can also be analyzed satisfactorily using cycles and transpositions. However, the structure of general non-abelian groups is very complicated, due to the dependence of the value of an expression on the order of the operands.

Applications : Group theory is an essential tool in algebraic topology. For instance, the neighborhood relationships for cells and simplices are described using groups. The analysis of these groups provides insight into essential properties of these shapes and leads, for example, to the discovery of bubbles and of contradictions in the specification of solids. These results are required for the representation of shapes and the construction of new shapes on a computer.

Group theory is also an essential tool in the description of materials (molecular structure, crystal structure) and the formulation of material properties, in particular for the laws governing anisotropic materials. Symmetry groups allow a general description of the symmetry properties and a quantification of the degree of symmetry.

An early application of group theory regards the solubility of non-linear equations by radicals. Galois demonstrated the relationship between this question and the properties of the normal subgroups of groups. The question whether the circle can be squared using only compass and straightedge, which is also answered by group theory, is closely related.

There are further applications of group theory. For example, geometry (including its non-euclidean aspects) may be developed on the basis of group theory. However, these applications of group theory are not the subject of the present book.

7.1.2 OUTLINE

This introduction to group theory assumes knowledge of Chapters 1 to 3 (logic, set theory, algebraic structures) of this book. The material is divided into the following subject areas :

Subgroups : Subsets of a group may themselves exhibit a group structure. A subgroup has a generating set : Every element of the subgroup is a product of elements of the generating set and their inverses. The unique representation of a group as a product of subgroups is studied.

Examples of groups : Permutation groups, groups of covering operations of regular geometric shapes and groups generated by a subset are typical examples of groups. The simple structure of the cyclic groups generated by a single element and of their subgroups is studied. The cyclic groups are closely related to the additive group of integers and its subgroups.

Class structure : A group cannot be partitioned into disjoint subgroups, since every subgroup contains the identity element of the group. Thus the concept of a subgroup is not sufficient for classifying the elements of a group. Equivalence relations are therefore introduced to form classes. Different equivalence relations lead to different classifications of the group. The equivalence of pairs of elements with respect to a subgroup H, the equivalence of conjugate elements in the group G and the equivalence of H-conjugate sets in G are treated. This leads to the partitioning of a group into cosets of a subgroup and to the concept of a normal subgroup.

Group structure : The structure of two groups is compared by mapping one of the groups onto the other in a structurally compatible manner (homomorphically). Groups with identical structure are said to be isomorphic. Between two isomorphic groups there is a bijective mapping which preserves the group structure. For every normal subgroup N of a group G there is a natural homomorphism $f : G \rightarrow G/N$ which maps the group G to the quotient group G/N. This property leads to three isomorphism theorems, which form the basis for further study of the structure of groups. Isomorphic mappings of groups onto themselves (automorphisms) are important in this context. They form a permutation group.

Abelian groups : The result of an operation on elements of a commutative (abelian) group may be represented as a linear combination. In contrast to the case of real vector spaces, however, it is not possible to represent every element of an abelian group as a unique linear combination of basis elements. The concept of a direct sum of subgroups is therefore introduced. Every finitely generated abelian group can be represented as a direct sum of cyclic subgroups. New abelian groups can be constructed by direct addition of given abelian groups. Abelian groups can be analyzed into direct sums.

Permutations : A permutation is a bijective mapping of a set onto itself. Every permutation can be decomposed into a product of disjoint cycles. A cycle leaves a part of the permuted set unchanged and maps the remaining elements to their neighbors (assembly line). The decomposition of a permutation into cycles also yields a decomposition into transpositions. The number of transpositions determines the sign of the permutation, and hence the coordinates of the ε-tensors in Chapter 9. Even permutations form the alternating subgroup of the permutation group and play an important role in Galois theory.

General groups : Operations on elements of a non-commutative group can generally not be represented as linear combinations, since the value of an expression depends on the order of the operands. The structure of general groups is studied by partitioning them into the cosets of normalizers, centers and commutator groups. Centers and commutator groups are normal subgroups which give rise to chains of quotient groups, the central series and the derived series. The aim of the continued reduction of the quotient groups is to arrive at a simple quotient group without a proper normal subgroup. This group cannot be classified further, since it has only improper quotient groups with respect to itself or the trivial group $\{1\}$. A chain of normal subgroups which begins with the group itself and ends with the trivial group $\{1\}$ is called a composition series if its quotients have no proper normal subgroups. Galois was able to show that the non-linear equation $f(x) = 0$ can be solved by radicals if and only if its Galois group G contains a chain of subgroups $G = G_0 \supset G_1 \supset \ldots \supset G_n = \{1\}$ such that G_{i+1} is a normal subgroup in G_i and the quotient group G_i / G_{i+1} is cyclic.

Existence of subgroups : A group without proper subgroups is invariably a cyclic group of prime order. For every divisor m of its order, an abelian group contains a subgroup of order m. The same is not true for non-abelian groups : There are groups which contain no subgroup of order m although m is a divisor of the order of the group. By a theorem due to Sylow, however, for every divisor p^m of its order (with p prime) a finite group contains a subgroup of order p^m. The properties of groups of prime-power order are studied in detail.

Unique decomposition of abelian groups : Every finitely generated abelian group can be represented as the direct sum of cyclic groups whose orders form a divisor chain. However, the number of summands in this representation is not unique, so that the decomposition is not unique. Uniqueness of the decomposition is achieved by considering direct sums of irreducible subgroups. Every finitely generated abelian group is the direct sum of a unique finite number of irreducible subgroups of infinite order and a unique finite number of finite irreducible subgroups of unique prime-power order. These invariants determine the type of the abelian group. Abelian groups of equal type are isomorphic.

7.2 GROUPS AND SUBGROUPS

Introduction : Groups are domains with an inner operation which satisfies certain conditions. This operation may alternatively be represented by additive or multiplicative notation. The question arises whether a group contains subsets which likewise satisfy the conditions for a group with respect to the same operation. The question also arises whether a group can be represented in terms of its subgroups, and whether this representation is unique. The relevant properties of subgroups are treated in this section.

Inner operation : A mapping f from the cartesian product $M \times M$ to the set M is called an inner operation in M.

$$f: M \times M \rightarrow M \quad \text{with} \quad f(a,b) = c \quad \text{and} \quad a,b,c \in M$$

In additive notation, the inner operation is represented by the addition sign $+$. The pair (a,b) is mapped to the sum $a + b$. The elements a and b are called the summands of $a + b$. In multiplicative notation, the inner operation is represented by the multiplication sign $\circ$. The pair (a,b) is mapped to the product $a \circ b$. The elements a and b are called the factors of $a \circ b$. Generic operations will be represented as multiplications in the following. The additive notation is used primarily for abelian (commutative) groups.

additive : $f(a,b) := a + b$

multiplicative : $f(a,b) := a \circ b$

Identity element : In a set M with an inner operation f, an element $e \in M$ is called an identity element with respect to the operation f if for all $a \in M$:

$$f(a,e) = f(e,a) = a$$

There is at most one identity element for an operation in a set, for if e and x are identity elements, it follows that $e = x$:

x is an identity element : $f(e,x) = e$

e is an identity element : $f(e,x) = x$

together : $f(e,x) = e = x$

There are inner operations without an identity element. In additive notation, the identity element is designated by 0, in multiplicative notation by 1. This corresponds to addition and multiplication in the sets $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$ and $\mathbb{C}$ of numbers.

additive : $a + 0 = a = 0 + a$

multiplicative : $a \circ 1 = a = 1 \circ a$

Monoid : The domain $(M;\circ)$ is called a monoid if the inner operation $\circ$ is associative and the set M contains an identity element e with respect to the operation $\circ$.

$$(a\circ b)\circ c = a\circ(b\circ c) \qquad \text{for all } a,b,c\in M$$
$$a\circ e = a = e\circ a \qquad \text{for all } a\in M$$

Inverse element : Let $(M;\circ)$ be a monoid with identity element e. An element $a'\in M$ is said to be inverse to $a\in M$ if :

$$a\circ a' = e = a'\circ a$$

The element a is called invertible in M if M contains an element inverse to a. If an element a of the monoid $(M;\circ)$ is invertible, then there is only one element in M which is inverse to a, for if x and y are elements inverse to a, it follows that $x=y$:

$$x = e\circ x = (y\circ a)\circ x = y\circ(a\circ x) = y\circ e = y$$

In additive notation, the element inverse to a is designated by $-a$; in multiplicative notation it is designated by a^{-1}. The short form $a-b$ is used for the operation $a+(-b)$.

$$\text{additive} \quad : \quad a+(-a) = 0 = (-a)+a$$
$$\text{multiplicative} : \quad a\circ a^{-1} = 1 = a^{-1}\circ a$$

The following rules of calculation hold for inverse elements :

(1) $(a^{-1})^{-1} = a$

(2) $(a\circ b)^{-1} = b^{-1}\circ a^{-1}$

Proof : Rules of calculation for inverse elements

(1) $(a^{-1})^{-1}\circ a^{-1} = 1 \quad \Rightarrow \quad (a^{-1})^{-1}\circ a^{-1}\circ a = 1\circ a$
$\quad \Rightarrow \quad (a^{-1})^{-1} = a$

(2) $(a\circ b)^{-1}\circ(a\circ b) = 1 \quad \Rightarrow \quad (a\circ b)^{-1}\circ a\circ b\circ b^{-1}\circ a^{-1} = b^{-1}\circ a^{-1}$
$\quad \Rightarrow \quad (a\circ b)^{-1} = b^{-1}\circ a^{-1}$

Group : A domain $(M;\circ)$ is called a group if the following conditions are satisfied :

(1) The group operation $\circ$ is an inner operation in M.

(2) The associative law holds in the domain $(M;\circ)$.

(3) The set M contains an identity element.

(4) For every element a, the set M contains the inverse element a^{-1}.

Order of a group : The number of elements in the set M of a group $(M;\circ)$ is called the order of the group and is designated by ord M. If M is infinite, the order is said to be infinite.

Subgroups : A group $(H;\circ)$ is called a subgroup of the group $(G;\circ)$ if H is a non-empty subset of G, the operations $\circ$ for G and H are equal and $(H;\circ)$ has the properties of a group. The domain $(H;\circ)$ with $H \subseteq G$ is a subgroup of $(G;\circ)$ if and only if the following conditions are satisfied :

(U1) The identity element 1 of G is an element of H.

(U2) If H contains the elements a and b of G, then H also contains $a \circ b$:

$$a, b \in H \quad \Rightarrow \quad a \circ b \in H$$

(U3) If H contains an element a of G, then H also contains the element a^{-1} of G :

$$a \in H \quad \Rightarrow \quad a^{-1} \in H$$

Since the operation of the group G is restricted to the subset H in (U2), the subset H inherits the associative property of the operation $\circ$ from the group G. Every group G contains the subgroups $\{1\}$ and G. A subgroup H is said to be proper if $\{1\} \subset H \subset G$.

Note : The statement "$(H;\circ)$ is a subgroup of $(G;\circ)$" is often simplified to "H is a subgroup of $(G;\circ)$" or "H is a subgroup of G". In such cases, the subset H is always assumed to inherit the operation $\circ$ from the set G.

Properties of subgroups :

(E1) The identity elements of a group and its subgroups are identical.

(E2) A non-empty subset H of a group $(G;\circ)$ is a subgroup of G if and only if for any two elements a, b in H the product $a \circ b^{-1}$ also belongs to H.

(E3) The intersection $H \cap K$ of two subgroups H and K of a group $(G;\circ)$ is a subgroup of G.

Proof : Properties of subgroups

(E1) The identity element 1 of G satisfies $1 \circ 1 = 1$. By (U1), the subset H contains the element 1, and by (U2) it contains the element $1 \circ 1$. Thus $1 \circ 1 = 1$ also holds in H. Hence 1 is the identity element of H.

(E2) Let H be a subgroup of G. Then it follows from (U2) and (U3) that for arbitrary elements a and b of H, $a \circ b^{-1}$ is also an element of H.

Let H be a non-empty subset of G, and for all $a, b \in H$ let $a \circ b^{-1} \in H$. Then by hypothesis $a \circ a^{-1}$ belongs to H. Hence the identity element $1 = a \circ a^{-1}$ of G is an element of H, and (U1) is satisfied. If the elements a and 1 belong to H, then by hypothesis $1 \circ a^{-1} = a^{-1}$ also belongs to H, and (U3) is satisfied. If a and b belong to H, then by (U3) a and b^{-1} also belong to H. Then by hypothesis $a \circ (b^{-1})^{-1} = a \circ b$ belongs to H, and (U2) is satisfied.

(E3) By (U1), the identity element 1 of G is an element of every subgroup of G. Hence the intersection $H \cap K$ is non-empty. Arbitrary elements a,b of $H \cap K$ are also elements of the subgroup H and of the subgroup K. By (U2) and (U3), $a \circ b^{-1}$ is therefore an element of H and of K, and hence an element of $H \cap K$. Thus by (E2) the non-empty intersection $H \cap K$ is a subgroup of G.

Generating set of a subgroup : Let M be a subset of a group $(G ; \circ)$. The intersection H(M) of all subgroups of G which contain M is called the subgroup of G generated by M. The subset M is called a generating set of the subgroup H. If M is empty, then H is the trivial group {1}.

The subgroup H(M) generated by M is the least subgroup of G which contains M. Every other subgroup of G which contains M also contains H(M). If M is non-empty, then H(M) is the set of all products which can be formed using elements of M and their inverses.

Proof : Generating set of a subgroup

(1) Let H be a subgroup of G which contains M. Since H is one of the subgroups whose intersection is H(M), it follows that $H(M) \subseteq H$.

(2) Let T be the set of all products which can be formed using elements of M and their inverses. If M is non-empty, then $M \subseteq T$ implies that T is non-empty. The construction of T implies that if a and b are elements of T, then $a \circ b^{-1}$ is also an element of T. By property (E2), T is a subgroup.

(3) By part (1) of the proof, $M \subseteq T$ for the subgroup T implies that $H(M) \subseteq T$. If T contains the product $a \circ b$, then the definition of T implies that M contains at least the element a or a^{-1} and the element b or b^{-1}. Then by the definition of a group every subgroup of G which contains M contains the elements a, a^{-1}, b and b^{-1}, as well as their products. Thus the intersection of these subgroups contains the element $a \circ b$, and therefore $T \subseteq H(M)$. From $H(M) \subseteq T$ and $T \subseteq H(M)$ it follows that $T = H(M)$.

Plain product of subsets : Let A and B be non-empty subsets of a group $(G ; \circ)$. The product $A \circ B$ of these subsets contains all elements of G which are products $a \circ b$ of an element $a \in A$ and an element $b \in B$.

$$A \circ B := \{x \in G \mid x = a \circ b \quad \wedge \quad a \in A \quad \wedge \ b \in B\}$$

The product $A \circ B$ is said to be plain if the representation $a \circ b$ of every element of the product is unique, that is if $a_1 \circ b_1 = a_2 \circ b_2$ implies $a_1 = a_2$ and $b_1 = b_2$.

Inverse of a product : Let A and B be non-empty subsets of a group $(G ; \circ)$. The inverse $(A \circ B)^{-1}$ of the product $A \circ B$ contains all elements of G which are inverse elements $(a \circ b)^{-1}$ of the product of elements $a \in A$ and $b \in B$:

$$(A \circ B)^{-1} := \{x \in G \mid x = (a \circ b)^{-1} \quad \wedge \quad a \in A \quad \wedge \ b \in B\}$$

Products of subgroups : Let A and B be subgroups of a group $(G\,;\circ)$. Then their product $A \circ B$ has the following properties :

(P1) The product $A \circ B$ is plain if and only if $A \cap B = \{1\}$.

(P2) The product $A \circ B$ is a group if and only if $A \circ B = B \circ A$.

Proof : Properties of a product of subgroups

(P1) Let the product $A \circ B$ be plain. If x is an element of $A \cap B$, then x is an element of A and an element of B. Since A and B are groups, x^{-1} is also an element of A and of B. Thus x^{-1} is an element of $A \cap B$, and hence so is $1 = x \circ x^{-1}$. Since $1 = 1 \circ 1$ also holds, the uniqueness of the representation of the element 1 in the plain product $A \circ B$ implies that $x = 1$. Hence $A \cap B = \{1\}$.

Let the intersection of A and B be $A \cap B = \{1\}$. For elements $a_1, a_2 \in A$ and $b_1, b_2 \in B$, let $a_1 \circ b_1 = a_2 \circ b_2$. Then $a_2^{-1} \circ a_1 = b_2 \circ b_1^{-1}$. Since $a_2^{-1} \circ a_1$ is an element of A and $b_2 \circ b_1^{-1}$ is an element of B, it follows that $a_2^{-1} \circ a_1 = b_2 \circ b_1^{-1}$ is an element of $\{1\}$, that is $a_2^{-1} \circ a_1 = 1$ and $b_2 \circ b_1^{-1} = 1$. The product $A \circ B$ is plain, since $a_1 = a_2$ and $b_1 = b_2$.

(P2) Let $A \circ B$ be a subgroup. Then if $A \circ B$ contains the element $a \circ b$, it also contains the inverse element $(a \circ b)^{-1} = b^{-1} \circ a^{-1}$. Hence for the elements $a, a^{-1} \in A$ and $b, b^{-1} \in B$:

$$\begin{array}{llllll}
a \circ b & \in & A \circ B & \Rightarrow & b^{-1} \circ a^{-1} & \in & A \circ B \\
a \circ b^{-1} & \in & A \circ B & \Rightarrow & b \circ a^{-1} & \in & A \circ B \\
a^{-1} \circ b^{-1} & \in & A \circ B & \Rightarrow & b \circ a & \in & A \circ B \\
a^{-1} \circ b & \in & A \circ B & \Rightarrow & b^{-1} \circ a & \in & A \circ B
\end{array}$$

The product $A \circ B$ contains every element of the product $B \circ A$, that is $B \circ A \subseteq A \circ B$. Likewise, $A \circ B \subseteq B \circ A$, and hence $A \circ B = B \circ A$.

Conversely, assume $A \circ B = B \circ A$. Then $(A \circ B) \circ (A \circ B) = A \circ (B \circ A) \circ B = A \circ (A \circ B) \circ B = (A \circ A) \circ (B \circ B) = A \circ B$, since by definition a subgroup contains all products of its elements. Further $(A \circ B)^{-1} = B^{-1} \circ A^{-1} = B \circ A = A \circ B$, since by definition a subgroup contains the inverse of each of its elements. From $(A \circ B) \circ (A \circ B) = (A \circ B)$ and $(A \circ B)^{-1} = A \circ B$, it follows that $A \circ B$ is a group, since for each of its elements $A \circ B$ also contains the inverse element, and together with any two elements it contains the product of these elements.

7.3 TYPES OF GROUPS

Introduction : Different mathematical and physical problems lead to different types of groups. Some types of groups which occur frequently are presented in this section :

(1) **Permutation groups :** A group of permutations $p : M \to M$ of an underlying set M equipped with the composition $p_i \circ p_m$ of permutations p_i and p_m is called a permutation group. Every finite group is isomorphic to a group of permutations.

(2) **Symmetry groups :** A mapping from a shape in euclidean space to itself is called a covering operation. Regular solids, for instance equilateral triangular plates, may change their position in space when a covering operation is applied. The number of different covering operations for a circle is infinite; for an equilateral triangle it is finite. The set of covering operations $a : F \to F$ for a shape F, equipped with the composition $a_i \circ a_m$ of covering operations, is called a symmetry group.

(3) **Generated groups :** Every element of a group $(G ; \circ)$ can be generated by applying the inner operation to elements taken from a subset $X \subseteq G$. The generating set X is generally not unique. If a generating set X for a group G is known, it may be used to study the structure of the group G.

(4) **Cyclic groups :** A cyclic group is a special case of a generated group. It is generated by a single element, so that every element of the group may be obtained by repeatedly applying the group operation to the generating element. There are finite and infinite cyclic groups. In subsequent sections, freely generated abelian groups are decomposed into cyclic groups. This decomposition is used, for instance, in algebraic topology.

(5) **Groups of integers :** The set $\{...,-2,-1,0,1,2,...\}$ of integers with addition as the inner operation is a countably infinite cyclic group. Every integer $a \in \mathbb{Z}$ generates a cyclic subgroup of $\mathbb{Z}$. Subgroups of $\mathbb{Z}$ can alternatively be generated using more than one element. The formation of such subgroups is related to the divisibility of integers.

(6) **Cyclic subgroups :** Every subgroup of a cyclic group is cyclic. However, there are also cyclic subgroups of groups which are not themselves cyclic. Every element of a general group generates a cyclic subgroup. The properties of the inner operation of the group determine whether this subgroup is finite or infinite.

7.3.1 PERMUTATION GROUPS

Introduction : Let a set M be finite. Every element of M can be mapped to itself or to another element of M such that the mapping is bijective. This type of mapping occurs frequently, for instance in the redistribution of tasks in a group of people : A takes over B's task, B takes over C's, C takes over A's, and D's task stays the same. Such mappings are called permutations. Their composition leads to the permutation groups.

Permutation : A bijective mapping of a finite set M onto itself is called a permutation of M. The mapping rule for a given permutation is represented in a permutation scheme consisting of two rows. The top row contains the elements x_i of the set M in an arbitrary order. In the bottom row, the elements of M are arranged such that the image $p(x_i)$ of the element x_i is directly underneath x_i. The scheme is enclosed in parentheses.

permutation : $p : M \rightarrow M$ bijective

$p(x_i) = x_{m_i}$ $\quad i, m_i \in \{1,2,...,n\}$

permutation scheme : $p = \begin{pmatrix} x_1 & x_2 & \dots & x_i & \dots & x_n \\ x_{m_1} & x_{m_2} & \dots & x_{m_i} & \dots & x_{m_n} \end{pmatrix}$

Permutation group : For a set with n elements, there are $k = n!$ different permutations. These permutations form a set $S_n = \{p_0, p_1, ..., p_{k-1}\}$. The composition of permutations is defined as an inner operation $\circ : S_n \times S_n \rightarrow S_n$ on the set S_n with $p_r \circ p_s = p_t$. Thus, performing the permutation p_r after the permutation p_s leads to the same mapping of the set M as the permutation p_t.

product : $p_r \circ p_s = p_t$ with $p_t(x_i) = p_r(p_s(x_i))$

$$\begin{pmatrix} x_1 & \dots & x_i & \dots & x_n \\ x_{r_1} & \dots & x_{r_i} & \dots & x_{r_n} \end{pmatrix} \circ \begin{pmatrix} x_1 & \dots & x_i & \dots & x_n \\ x_{s_1} & \dots & x_{s_i} & \dots & x_{s_n} \end{pmatrix} = \begin{pmatrix} x_1 & \dots & x_i & \dots & x_n \\ x_{t_1} & \dots & x_{t_i} & \dots & x_{t_n} \end{pmatrix}$$

identity element : $e = \begin{pmatrix} x_1 & \dots & x_n \\ x_1 & \dots & x_n \end{pmatrix}$

inverse element : $p_s = \begin{pmatrix} x_1 & \dots & x_i & \dots & x_n \\ x_{s_1} & \dots & x_{s_i} & \dots & x_{s_n} \end{pmatrix}$ $\qquad p_s^{-1} = \begin{pmatrix} x_{s_1} & \dots & x_{s_i} & \dots & x_{s_n} \\ x_1 & \dots & x_i & \dots & x_n \end{pmatrix}$

The domain $(S_n ; \circ)$ has the properties of a group and is called the permutation group (symmetric group) on the underlying set M. The group S_n has the following properties :

(1) The identity element is the identity mapping with $p(x_i) = x_i$.

(2) The inverse element p_m^{-1} is obtained from the permutation p_m by interchanging the two rows of the permutation scheme.

(3) The composition $\circ$ of permutations is associative.

(4) The composition $\circ$ of permutations is not commutative.

Example 1 : Permutations of a set of 3 elements

The set $M = \{a, b, c\}$ gives rise to the following permutations $S_3 = \{p_0, ..., p_5\}$:

$$p_0 = \begin{bmatrix} a & b & c \\ a & b & c \end{bmatrix} \qquad p_1 = \begin{bmatrix} a & b & c \\ c & a & b \end{bmatrix} \qquad p_2 = \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix}$$

$$p_3 = \begin{bmatrix} a & b & c \\ a & c & b \end{bmatrix} \qquad p_4 = \begin{bmatrix} a & b & c \\ c & b & a \end{bmatrix} \qquad p_5 = \begin{bmatrix} a & b & c \\ b & a & c \end{bmatrix}$$

The products $p_2 \circ p_4$ and $p_4 \circ p_5$ lead to the elements p_3 and p_2 :

$$p_2 \circ p_4 = p_3 : \quad \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ c & b & a \end{bmatrix} = \begin{bmatrix} a & b & c \\ a & c & b \end{bmatrix}$$

$$p_4 \circ p_5 = p_2 : \quad \begin{bmatrix} a & b & c \\ c & b & a \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ b & a & c \end{bmatrix} = \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix}$$

The associative law holds, for instance for $(p_2 \circ p_4) \circ p_5 = p_2 \circ (p_4 \circ p_5)$:

$$\begin{bmatrix} a & b & c \\ a & c & b \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ b & a & c \end{bmatrix} = \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix} = \begin{bmatrix} a & b & c \\ c & a & b \end{bmatrix}$$

The commutative law does not hold, for instance for p_2 and p_4 :

$$p_2 \circ p_4 = p_3 : \quad \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ c & b & a \end{bmatrix} = \begin{bmatrix} a & b & c \\ a & c & b \end{bmatrix}$$

$$p_4 \circ p_2 = p_5 : \quad \begin{bmatrix} a & b & c \\ c & b & a \end{bmatrix} \circ \begin{bmatrix} a & b & c \\ b & c & a \end{bmatrix} = \begin{bmatrix} a & b & c \\ b & a & c \end{bmatrix}$$

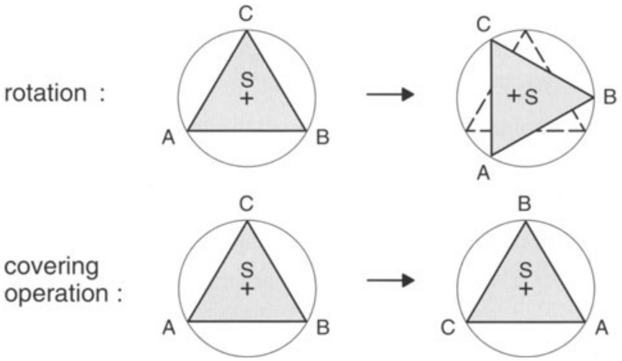
This permutation group has the following product table. The value of the first factor p_i of the product is associated with row i of the table, the value of the factor p_m is associated with column m. This convention also applies to all subsequent tables.

$\circ$	p_0	p_1	p_2	p_3	p_4	p_5
p_0	p_0	p_1	p_2	p_3	p_4	p_5
p_1	p_1	p_2	p_0	p_4	p_5	p_3
p_2	p_2	p_0	p_1	p_5	p_3	p_4
p_3	p_3	p_5	p_4	p_0	p_2	p_1
p_4	p_4	p_3	p_5	p_1	p_0	p_2
p_5	p_5	p_4	p_3	p_2	p_1	p_0

product table $p_i \circ p_m$

7.3.2 SYMMETRY GROUPS

Introduction : If an equilateral triangle is rotated in its plane about an axis through the center of gravity S, then the new position generally does not cover the old position of the triangle. In the special case of a rotation through a multiple of 120 degrees, however, the new position and the old position of the triangle cover each other. This special rotation of the triangle is called a covering operation. After the rotation, the points of the triangle are generally displaced. Nonetheless the triangle occupies the same part of the plane before and after the rotation. There are several covering operations which lead to the same position of the triangle. These covering operations differ by a rotation through a multiple of 360 degrees. These geometric observations are abstracted in the definition of symmetry groups.



Shapes and bodies : A compact connected subset of euclidean space is called a shape. The matter which occupies a shape at a given instant is called a body. A bijective mapping of the material points of a body to the points of a shape is called a position of the body.

Displacement : A bijective mapping of the material points of a body from one position to another position is called a displacement of the body. A general displacement is a non-linear mapping. Displacements with special common properties form a displacement type. Displacement types are defined by relationships between the initial shape and the final shape of the body :

(1) affine : straight fibers remain straight

(2) similar : angles between fibers are preserved

(3) congruent : distances between points are preserved

(4) covering : initial shape and final shape are identical

(5) trivial : initial position and final position are identical

Motion : A displacement of a body is called a motion if the initial shape and the final shape of the body are congruent. A body is called a rigid body if all shapes of the body are congruent. Thus all displacements of a rigid body are motions which leave the distance between the material points of the body unchanged.

Covering operation : A motion of a body is called a covering operation if the initial shape and the final shape of the body coincide. The mapping $a : F \to F$ of a covering operation maps the shape F to itself. The initial position of a body and its final position after a covering operation may, however, be different.

Symmetry group : Let $A = \{a_0, a_1, ..., a_n\}$ be the set of covering operations of a body. Let the result of the inner operation $\circ : A \times A \to A$ with $a_r \circ a_s = a_t$ be the covering operation a_t which leads to the same position of the body as is reached by performing the covering operation a_r after the covering operation a_s (composition of covering operations). The domain $(A ; \circ)$ has the properties of a group and is called the symmetry group of the body.

(1) The identity element of the symmetry group is the trivial mapping a_0.

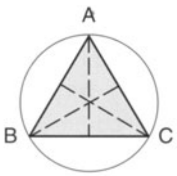
(2) For every covering operation a_m there is an inverse covering operation a_m^{-1}, which cancels the motion of the body produced by the covering operation a_m.

(3) The composition $\circ$ of covering operations is associative.

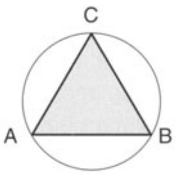
(4) The composition $\circ$ of covering operations is not commutative.

Example 1 : Symmetry group of an equilateral triangle

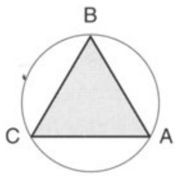
The symmetry group of equilateral triangles with respect to motions in a plane is treated in Example 4 of Section 3.3. If motions of equilateral triangles in space are considered, there are three further covering operations. Each of these covering operations is a rotation through an angle of 180 degrees about an axis which joins one corner of the triangle with the midpoint of the opposite side.



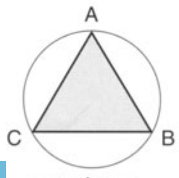
rotation a_0



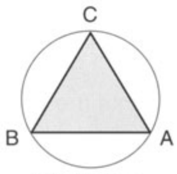
rotation a_1



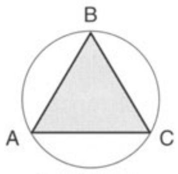
rotation a_2



rotation a_3



rotation a_4



rotation a_5

The rotations lead to permutations of the corners of the triangle. The product table for the symmetry group coincides with the product table for permutations in Section 7.3.1 if corresponding notation is used.

$\circ$	a_0	a_1	a_2	a_3	a_4	a_5
a_0	a_0	a_1	a_2	a_3	a_4	a_5
a_1	a_1	a_2	a_0	a_4	a_5	a_3
a_2	a_2	a_0	a_1	a_5	a_3	a_4
a_3	a_3	a_5	a_4	a_0	a_2	a_1
a_4	a_4	a_3	a_5	a_1	a_0	a_2
a_5	a_5	a_4	a_3	a_2	a_1	a_0

product table $a_i \circ a_m$

Example 2 : Symmetry group of a regular tetrahedron

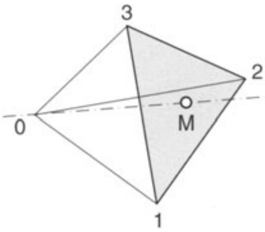
The rotations of a tetrahedron about its medians and its edge bisectors are considered. A median joins a corner of the tetrahedron with the center of the opposite face. An edge bisector joins the midpoints of two edges which have no endpoints in common. These rotations lead to 12 covering operations of the tetrahedron :

(1) The trivial mapping of the tetrahedron

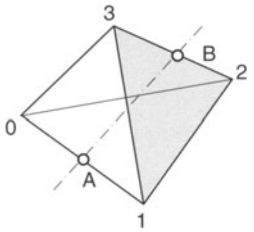
(2) Rotations through angles of 120 and 240 degrees about each of the medians of the tetrahedron

(3) Rotations through an angle of 180 degrees about each of the edge bisectors

The rotations lead to permutations of the corners of the tetrahedron; these permutations are compiled below.



median 0M



edge bisector AB

$$a_0 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 1 & 2 & 3 \end{bmatrix} \qquad a_1 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 2 & 3 & 1 \end{bmatrix} \qquad a_2 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 0 & 3 & 1 & 2 \end{bmatrix}$$

$$a_3 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 2 & 1 & 3 & 0 \end{bmatrix} \qquad a_4 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 1 & 0 & 2 \end{bmatrix} \qquad a_5 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 3 & 2 & 0 \end{bmatrix}$$

$$a_6 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 0 & 2 & 1 \end{bmatrix} \qquad a_7 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 2 & 0 & 3 \end{bmatrix} \qquad a_8 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 2 & 0 & 1 & 3 \end{bmatrix}$$

$$a_9 = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 1 & 0 & 3 & 2 \end{bmatrix} \qquad a_{10} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 2 & 3 & 0 & 1 \end{bmatrix} \qquad a_{11} = \begin{bmatrix} 0 & 1 & 2 & 3 \\ 3 & 2 & 1 & 0 \end{bmatrix}$$

If appropriate notation is used, the symmetry group of the tetrahedron coincides with a subgroup of the permutation group over sets with four elements. It contains proper subgroups for the following subsets of rotations :

(1) $H = \{ a_0, a_9, a_{10}, a_{11} \}$: edge bisectors

(2) $H_0 = \{ a_0, a_1, a_2 \}$: median through point 0

(3) $H_1 = \{ a_0, a_3, a_4 \}$: median through point 1

(4) $H_2 = \{ a_0, a_5, a_6 \}$: median through point 2

(5) $H_3 = \{ a_0, a_7, a_8 \}$: median through point 3

(6) $H_{01} = \{ a_0, a_9 \}$: edge bisector from 01 to 23

(7) $H_{02} = \{ a_0, a_{10} \}$: edge bisector from 02 to 13

(8) $H_{03} = \{ a_0, a_{11} \}$: edge bisector from 03 to 12

The symmetry group has the following product table :

$\circ$	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
a_0	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
a_1	a_1	a_2	a_0	a_{11}	a_7	a_3	a_9	a_{10}	a_6	a_8	a_4	a_5
a_2	a_2	a_0	a_1	a_5	a_{10}	a_{11}	a_8	a_4	a_9	a_6	a_7	a_3
a_3	a_3	a_{10}	a_8	a_4	a_0	a_9	a_1	a_5	a_{11}	a_7	a_6	a_2
a_4	a_4	a_6	a_{11}	a_0	a_3	a_7	a_{10}	a_9	a_2	a_5	a_1	a_8
a_5	a_5	a_7	a_9	a_{10}	a_2	a_6	a_0	a_{11}	a_3	a_4	a_8	a_1
a_6	a_6	a_{11}	a_4	a_8	a_9	a_0	a_5	a_1	a_{10}	a_2	a_3	a_7
a_7	a_7	a_9	a_5	a_1	a_{11}	a_{10}	a_4	a_8	a_0	a_3	a_2	a_6
a_8	a_8	a_3	a_{10}	a_9	a_6	a_2	a_{11}	a_0	a_7	a_1	a_5	a_4
a_9	a_9	a_5	a_7	a_6	a_8	a_1	a_3	a_2	a_4	a_0	a_{11}	a_{10}
a_{10}	a_{10}	a_8	a_3	a_2	a_5	a_4	a_7	a_6	a_1	a_{11}	a_0	a_9
a_{11}	a_{11}	a_4	a_6	a_7	a_1	a_8	a_2	a_3	a_5	a_{10}	a_9	a_0

product table $a_i \circ a_m$

7.3.3 GENERATED GROUPS

Introduction : The product $x_i \circ x_m$ of the elements x_i and x_m of a group $(G\,;\circ)$ is by definition a group element; the same is true for the product $x_i \circ x_m^{-1}$ of the element x_i with the inverse of x_m. The question arises whether the elements of G can be generated by operations using elements of a subset X of G, and which subsets of X are suitable for this purpose.

Generated groups : A group $(G\,;\circ)$ is said to be generated by the set X if every element of the group is the product of elements x_i of the set X and their inverses x_i^{-1}. The elements x_i and their inverses may occur an arbitrary number of times in arbitrary order in the product. The group generated by the set X is designated by gp(X).

$$\mathrm{gp}(X) \;:=\; \{x_{i_1}^{\alpha_1} \circ \ldots \circ x_{i_n}^{\alpha_n} \mid x_{i_s} \in X \;\wedge\; \alpha_s \in \{-1,1\}\}$$

Finitely generated groups : A group $(G\,;\circ)$ is said to be finitely generated if G is generated by a finite set X.

$$G \;=\; \mathrm{gp}(X) \;=\; \{x_{i_1}^{\alpha_1} \circ \ldots \circ x_{i_n}^{\alpha_n} \mid x_{i_s} \in \{x_1,\ldots,x_m\} \;\wedge\; \alpha_s \in \{-1,1\}\}$$

Generating set : The set X is called a generating set of the group $G = \mathrm{gp}(X)$. The generating set of a group G is not unique. The set of elements of the group G is a trivial generating set of this group.

Example 1 : Generating set of the symmetry group of a triangle

A generating set of the symmetry group of equilateral triangles in Example 1 of Section 7.3.2 contains at least one element from the subgroup $\{a_0, a_1, a_2\}$ of rotations in the plane and one element from the subgroup $\{a_3, a_4, a_5\}$ of rotations in space. If the generating set $X = \{a_1, a_3\}$ is chosen, for example, then the elements of the group are the following products of the generating elements :

$$\begin{array}{lll@{\qquad}lll}
a_0 & = & a_1 \circ a_1 \circ a_1 & a_3 & = & a_3 \\
a_1 & = & a_1 & a_4 & = & a_1 \circ a_3 \\
a_2 & = & a_1 \circ a_1 & a_5 & = & a_1 \circ a_1 \circ a_3
\end{array}$$

Example 2 : Generating set of the symmetry group of a tetrahedron

A generating set of the symmetry group of regular tetrahedra in Example 2 of Section 7.3.2 has at least two elements, for example $E = \{a_1, a_3\}$. The elements of the symmetry group are the following products of the generating elements :

$$\begin{array}{lll@{\qquad}lll@{\qquad}lll}
a_2 & = & a_1 \circ a_1 & a_0 & = & a_1 \circ a_1 \circ a_1 & a_7 & = & a_1 \circ a_3 \circ a_3 \\
a_4 & = & a_3 \circ a_3 & a_0 & = & a_3 \circ a_3 \circ a_3 & a_8 & = & a_3 \circ a_1 \circ a_1 \\
a_{10} & = & a_3 \circ a_1 & a_5 & = & a_1 \circ a_1 \circ a_3 & a_9 & = & a_1 \circ a_3 \circ a_3 \circ a_1 \\
a_{11} & = & a_1 \circ a_3 & a_6 & = & a_3 \circ a_3 \circ a_1 & a_9 & = & a_3 \circ a_1 \circ a_1 \circ a_3
\end{array}$$

Reduced X-product : A product of elements of the generating set X and their inverses may contain factors such as $x_i \circ x_i^{-1}$ and $x_i^{-1} \circ x_i$. Since such factors do not influence the value of the product, they may be deleted. A product is called a reduced X-product if no further deletions of such factors are possible.

$$g \quad = \quad x_1^{\alpha_1} \circ \ldots \circ x_n^{\alpha_n} \qquad\qquad x_i \in X,\ \alpha_i \in \{-1,1\}$$

$$g \quad \text{is a reduced X-product} \quad :\Leftrightarrow \quad \bigwedge_i \ (x_i = x_{i+1} \ \Rightarrow \ \alpha_i = \alpha_{i+1})$$

A finitely generated group with the identity element 1 is represented as follows using reduced X-products :

$$\mathrm{gp}(X) \quad = \quad \{g \mid g = 1 \ \vee \ g \ \text{ is a reduced X-product}\}$$

Further simplifications of a reduced X-product are possible if gp(X) is a commutative group. The simplifications which may be carried out in this case by changing the order of the operands are studied in the theory of abelian groups.

Equal X-products : The products $x_1^{\alpha_1} \circ \ldots \circ x_m^{\alpha_m}$ and $y_1^{\beta_1} \circ \ldots \circ y_n^{\beta_n}$ of elements x_i, y_i of a generating set X with $\alpha_i,\ \beta_i \in \{-1,1\}$ are said to be equal if $m = n$, $x_i = y_i$ and $\alpha_i = \beta_i$. Otherwise the products are said to be different.

Values of reduced X-products : The value of a reduced X-product is an element of gp(X). Equal X-products have the same value. Two different reduced X-products may also have the same value. Thus the representation of an element of gp(X) as a product of elements of the generating set and their inverses is generally not unique.

Freely generated group : The group gp(X) generated by a finite set X is said to be freely generated (free) if any two different reduced X-products are different elements of gp(X). In this case, the set X is called a free generating set, and the group gp(X) is called a free finitely generated group. A free finitely generated group $G = \mathrm{gp}(X)$ has the following properties :

(1) If the generating set X contains the element x, then it does not contain the inverse element $y = x^{-1}$, since otherwise x and y^{-1} would be two different reduced products for the same element of gp(X).

(2) The generating set X does not contain the identity element 1_G, since otherwise $1_G \circ x$ and x would be two different reduced X-products for the same element x of G.

(3) The reduction of an X-product containing a finite number of factors may be carried out in a finite number of steps. Hence it is possible to decide in a finite number of steps whether two X-products of a free group have the same value.

Example 3 : Reduction of X-products

Let the group G be generated by the set X = {x, y, z}. Then X-products are reduced as follows :

$$x \circ y \circ y^{-1} \circ z^{-1} \circ y \circ x = x \circ z^{-1} \circ y \circ x$$
$$x \circ z^{-1} \circ y \circ z \circ z^{-1} \circ x^{-1} = x \circ z^{-1} \circ y \circ x^{-1}$$
$$x \circ y \circ z \circ z^{-1} \circ y^{-1} = x \circ y \circ y^{-1} = x$$

Example 4 : X-products of a cyclic group

$\circ$	a_0	a_1	a_2
a_0	a_0	a_1	a_2
a_1	a_1	a_2	a_0
a_2	a_2	a_0	a_1

product table of a group $G = gp(a_1)$

The group G is generated by a single element, for example by a_1. Different X-products have the same value :

$$a_1 \circ a_1 = a_2 \qquad a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 = a_2$$
$$a_1 \circ a_1 \circ a_1 = a_0 \qquad a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 = a_0$$
$$a_1 \circ a_1 \circ a_1 \circ a_1 = a_1 \qquad a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 \circ a_1 = a_1$$

Example 5 : Decidability of the word problem

Let a character set $\{a_1,...,a_n, \bar{a}_1,...,\bar{a}_n\}$ with $n \in \mathbb{N}'$ for the construction of words be given. A sequence of a finite number of characters of the set is called a word and is designated by w. Repetitions of letters as in the word $w = a_1 a_2 \bar{a}_4 a_2$ are allowed. The word with no characters is called the empty word w_0. The concatenation of two words w_1 and w_2 is called the product of the words and is designated by $w_1 \circ w_2$.

Two words are said to be equivalent if one of the words is obtained from the other by deleting or inserting factors $a_i \bar{a}_i$ or $\bar{a}_i a_i$. For example, a_2 and $a_2 \bar{a}_3 a_1 \bar{a}_1 a_3$ are equivalent. The inverse w^{-1} of the word w contains the characters of w in inverse order, where a is replaced by $\bar{a}$ and $\bar{a}$ by a, so that $w \circ w^{-1}$ and $w^{-1} \circ w$ are equivalent to w_0.

Equivalent words form a class $[w_i]$. An inner operation $[w_i] \circ [w_m] = [w_i \circ w_m]$ is defined in the set G of all equivalence classes. Then $(G ; \circ)$ is a group with the identity element $[w_0]$ and the element $[w]^{-1} = [w^{-1}]$ inverse to [w]. In place of $\bar{a}_i$ one may therefore write a_i^{-1}.

The subset $A = \{[a_i],...,[a_n]\}$ is a generating set for the free finitely generated group $(G ; \circ)$. It is possible to decide in a finite number of steps whether two words are equivalent.

7.3.4 CYCLIC GROUPS

Introduction : The structure of a group which is generated by a single element is very simple. This is partly due to the fact that such groups are commutative. Every element of the group may therefore be represented as a power of the generating element. This leads to the properties of cyclic groups described in the following.

Powers of an element : The powers of an element a of a multiplicative group $(G\,;\circ)$ are defined as follows for integers $n \in \mathbb{Z}$:

$$\begin{array}{lllll} n > 0 : & a^n := & a \circ a \circ \ldots \circ a & : & n\text{-fold} \\ n = 0 : & a^0 := & 1 & & \\ n < 0 : & a^n := & a^{-1} \circ a^{-1} \circ \ldots \circ a^{-1} & : & |n|\text{-fold} \end{array}$$

Cyclic group : A group $(G\,;\circ)$ is said to be cyclic if it is generated by a single element. Every integer power a^n of the generating element a is an element of G. The identity element 1 is a^0. The cyclic group is designated by gp(a).

$$G = \text{gp}(a)$$

Properties of cyclic groups : The inner operation $\circ$ determines the properties of a cyclic group. In particular, the operation $\circ$ determines whether two different powers of the generating element are equal or not. This leads to the distinction of finite and infinite cyclic groups with the following properties :

(Z1) The general element a_i of a finite cyclic group $G = \text{gp}(a)$ of order m is a power a^i of the generating element a with $0 \le i < m$. Thus the group is given by $G = \{a_0, a_1, \ldots, a_{m-1}\}$.

(Z2) The elements a^r and a^s of a finite cyclic group of order m are equal if and only if the exponents r and s have the same remainder modulo m.

(Z3) In an infinite cyclic group gp(a), different powers a^r and a^s with $r \neq s$ are different elements of the group.

(Z4) Every cyclic group gp(a) is abelian (commutative).

Proof Z1 : The general element of a cyclic group gp(a) of finite order m is a power a^i of the generating element with $0 \leq i < m$.

(1) By definition, $a^0 = 1$. If the group is finite, then there are also other natural numbers n for which $a^n = 1$ holds. Let m be the least of these numbers, so that $a^m = 1$ and $a^i \neq 1$ for $0 < i < m$.

(2) The elements $a^0, a^1, \ldots, a^{m-1}$ are pairwise different. The proof is carried out indirectly. For $0 < r < s < m$, assume $a^r = a^s$ and therefore $a^{s-r} = 1$. Then there is a number $u = s - r$ with $0 < u < m$ such that $a^u = 1$. This is a contradiction, since m is the least positive number for which $a^m = 1$ holds. Hence the elements $\{a^0, a^1, \ldots, a^{m-1}\}$ are pairwise different.

(3) The group G contains only the elements $\{a^0, a^1, \ldots, a^{m-1}\}$, since $a^{km+i} = (a^m)^k\, a^i = a^i$ for an arbitrary element a^{km+i} with $0 \leq i < m$ and $k \in \mathbb{Z}$. Thus a finite cyclic group is completely described by the generating element a and the order m.

$$G(a, m) = \{ a^i \mid 0 \leq i < m \quad \wedge \quad i \in \mathbb{N} \}$$

Proof Z2 : The elements a^r and a^s of a finite cyclic group of order m are equal if and only if $r = s \bmod m$.

Using division with remainder, the exponents r and s may be represented in the forms $r = k_1 m + n_1$ and $s = k_2 m + n_2$, respectively, with $0 \leq k_1, k_2 < m$. The proof for (Z1) shows that $a^r = a^{n_1}$ and $a^s = a^{n_2}$. Hence the elements a^r and a^s are equal if and only if $n_1 = n_2$, that is if and only if $r = s \bmod m$.

Proof Z3 : In an infinite cyclic group gp(a) arbitrary powers a^r and a^s with $r \neq s$ are different.

If a cyclic group gp(a) is infinite, then there is no number $m \in \mathbb{Z}$ except for $m = 0$ for which $a^m = 1$ holds, for if there were such a number, then according to the preceding proof the group would be finite. Therefore $a^r = a^s$ implies $a^{r-s} = 1$ and $m = r - s = 0$, and hence $r = s$. Different powers of the generating element of an infinite cyclic group are therefore different elements of the group.

Proof Z4 : Every cyclic group is commutative.

This property follows directly from the representation of the elements of gp(a) as powers:

$$a_i \circ a_n = a^i \circ a^n = a^{i+n} = a^{n+i} = a^n \circ a^i = a_n \circ a_i$$

Example 1 : Finite cyclic group

The symmetry group $\{a_0, a_1, a_2\}$ of equilateral triangles in a plane is a finite cyclic group. Either the rotation a_1 through 120 degrees or the rotation a_2 through 240 degrees may be chosen as the generating element.

$\circ$	a_0	a_1	a_2
a_0	a_0	a_1	a_2
a_1	a_1	a_2	a_0
a_2	a_2	a_0	a_1

product table of the symmetry group

$$gp(a_1) = \{a_1^n \mid n \in \{0, 1, 2\}\} = \{a_0, a_1, a_2\}$$
$$gp(a_2) = \{a_2^n \mid n \in \{0, 1, 2\}\} = \{a_0, a_2, a_1\}$$

$$\begin{array}{lcllcl}
(a_1)^0 & = & a_0 & \qquad (a_2)^0 & = & a_0 \\
(a_1)^1 & = & a_1 & \qquad (a_2)^1 & = & a_2 \\
(a_1)^2 & = & a_2 & \qquad (a_2)^2 & = & a_1
\end{array}$$

Example 2 : Infinite cyclic groups

The additive group ($\mathbb{Z}$; +) of integers and each of its proper subgroups are infinite cyclic groups. For example, the subgroup $gp(3) = \{..., -6, -3, 0, 3, 6, ...\}$ is an infinite cyclic group.

Example 3 : Cyclic subgroups

A group which is not itself cyclic contains cyclic subgroups. For example, the permutation group shown in Example 1 of Section 7.3.1 contains the cyclic subgroup $gp(p_1) = \{p_0, p_1, p_2\}$.

7.3.5 GROUPS OF INTEGERS

Additive group of integers : The set $\mathbb{Z}$ of integers equipped with the inner operation + (addition) forms a group ($\mathbb{Z}$; +). If sum notation is used, the identity element of the group is designated by 0, and the inverse of $a \in \mathbb{Z}$ is designated by $-a$. Together with a_1 and a_2 the group also contains the sum $a_1 + a_2$ and the difference $a_1 - a_2 := a_1 + (-a_2)$. The group is commutative and countably infinite.

Multiples of an element : The multiples of an element a of an additive group (G ; +) are defined as follows for integers $n \in \mathbb{Z}$:

$$\begin{array}{llll} n > 0 : & na := & a + a + \ldots + a & : \quad n\text{-fold} \\ n = 0 : & na := & 0 & \\ n < 0 : & na := & (-a) + \ldots + (-a) & : \quad |n|\text{-fold} \end{array}$$

Subgroups of the group of integers : The subgroup of the integers generated by an integer a contains every multiple na of a with $n \in \mathbb{Z}$ and is designated by $\mathbb{Z}a$, that is $\mathbb{Z}a = \{\ldots, -2a, -a, 0, a, 2a, \ldots\}$. Thus the group ($\mathbb{Z}a$; +) contains the integers divisible by a. It is a cyclic group, since it is generated by an integer a. Every subgroup (H ; +) of the group ($\mathbb{Z}$; +) is a cyclic group ($\mathbb{Z}a$; +).

Proof : Every subgroup of the integers is cyclic.

(1) The subgroup $H = \{0\}$ has the form $H = \mathbb{Z} \cdot 0$. In the following, let $H \neq \{0\}$.

(2) Every subgroup H contains positive numbers, since with an element $x \neq 0$ it also contains the element $-x$ and one of the numbers $\{x, -x\}$ is positive. Let a be the least positive number in H. Since H is a group, $a \in H$ implies $\mathbb{Z}a \subseteq H$.

(3) An arbitrary element x of H is represented in the unique form $x = qa + r$ with $0 \le r < a$ by division with remainder. Since x and qa are elements of H, $r = x - qa$ is also an element of H. Since a is the least positive number in H and $0 \le r < a$, it follows that $r = 0$ and $x = qa$, so that $x \in \mathbb{Z}a$ and $H \subseteq \mathbb{Z}a$.

(4) $\mathbb{Z}a \subseteq H$ and $H \subseteq \mathbb{Z}a$ imply $H = \mathbb{Z}a$. Hence for every subgroup H of $\mathbb{Z}$ there is exactly one natural number a with $H = \mathbb{Z}a$.

Generating sets of integer groups : Let $H = gp(A)$ be the subgroup of the integers generated by a system $A = \{a_1, \ldots, a_s\}$ of natural numbers. The generating set of a subgroup is not unique. In fact, it was shown in the preceding proof that every subgroup H of the integers is generated by a single element a and is therefore cyclic. Every element na of H with $n \in \mathbb{Z}$ can therefore be generated either using the element a or using the system A.

$$\begin{array}{lllll} h & = & n_1 a_1 + \ldots + n_s a_s & = & na \qquad n_i \in \mathbb{Z},\ a_i \in \mathbb{N},\ h \in H \\ H & = & \mathbb{Z}a_1 + \ldots + \mathbb{Z}a_s & = & \mathbb{Z}a \qquad n \in Z,\ a \in N \end{array}$$

Properties of integers : The properties of the subgroups of the integers lead to important properties of the integers.

(G1) If a is the greatest common divisor of the natural numbers $a_1, ..., a_s$, then there are integers $n_1, ..., n_s$ such that

$$a \;=\; n_1 a_1 + ... + n_s a_s$$

(G2) The natural numbers $a_1, ..., a_s$ are mutually prime if and only if there are integers $n_1, ..., n_s$ such that

$$1 \;=\; n_1 a_1 + ... + n_s a_s$$

(G3) If a is the least common multiple of the natural numbers $a_1, ..., a_s$, then the cyclic group $\mathbb{Z}a$ is the intersection of the groups $\mathbb{Z}a_i$:

$$\mathbb{Z}a \;=\; \mathbb{Z}a_1 \cap ... \cap \mathbb{Z}a_s$$

Proof : Properties of integers

(G1) Let $H = gp(A)$ be the group generated by a set $A = \{a_1, ..., a_s\}$ of natural numbers. Since every subgroup of the integers is cyclic, there is an element $x \in H$ such that $H = gp(x)$. A general element $h \in H$ may alternatively be generated using A or using x :

$$h \;=\; k_1 a_1 + ... + k_s a_s \;=\; kx \qquad\qquad k, k_i \in \mathbb{Z}$$

Since every element $a_i \in A$ is an element of $H = gp(x)$, there is an integer $v_i \in \mathbb{Z}$ such that $a_i = v_i x$. Hence x is a common divisor of the numbers in A :

$$h \;=\; (k_1 v_1 + ... + k_s a_s)x \;=\; kx$$

If y is another common divisor of the numbers in A, then every number $a_i \in A$ may be represented in the form $a_i = m_i y$ with $m_i \in \mathbb{Z}$:

$$h \;=\; (k_1 m_1 + ... + k_s m_s)y \;=\; my$$

Hence every element $h \in H$ may be represented in the form $h = kx = my$. Since the group H is generated by x, it follows that $gp(x) \subseteq gp(y)$, and hence $y \le x$. Thus x is the greatest common divisor of the numbers in A; it is designated by a :

$$a \;:=\; x \;=\; \gcd(a_1, ..., a_s)$$

Since a is an element of $H = gp(a) = gp(A)$, there are integers $n_1, ..., n_s$ such that

$$a \;=\; n_1 a_1 + ... + n_s a_s \qquad\qquad n_i \in \mathbb{Z}$$

(G2) If the natural numbers $a_1, ..., a_s$ are mutually prime, then their greatest common divisor is $\gcd(a_1, ..., a_s) = 1$. By (G1), there are integers $n_1, ..., n_s$ such that $n_1 a_1 + ... + n_s a_s = 1$.

If $a_1, ..., a_s$ are natural numbers and there are integers $n_1, ..., n_s$ such that $n_1 a_1 + ... + n_s a_s = 1$, then the numbers $a_1, ..., a_s$ are mutually prime. In fact, $a_i = m b_i$ with the common divisor $m > 1$ would imply $m(n_1 b_1 + ... + n_s b_s) = 1$. This is impossible, since m and $(n_1 b_1 + ... + n_s b_s)$ are integers.

(G3) The intersection of the groups $\mathbb{Z}a_1, ..., \mathbb{Z}a_s$ is a subgroup of $\mathbb{Z}$, and hence a cyclic group $\mathbb{Z}a$ generated by a natural number a. The number a is a common multiple of $(a_1, ..., a_s)$, since $a \in \mathbb{Z}a_i$ for all $a_i \in A$. If v is another common multiple of the a_i, then v is an element of every group $\mathbb{Z}a_i$, and hence an element of the intersection $\mathbb{Z}a_1 \cap ... \cap \mathbb{Z}a_s$. Thus v is an element of $\mathbb{Z}a$, and therefore a multiple of a. Hence a is the least common multiple :

$$a = \operatorname{lcm}(a_1, ..., a_s)$$

Euclidean algorithm : Let the set $A = \{a_1, ..., a_s\}$ with the natural numbers a_i be a generating set of the group $H = gp(A)$. The generating element a of the group H is determined by the Euclidean algorithm in the following steps :

(1) First an algorithm for determining the greatest common divisor gcd(a, b) of two natural numbers a and b with $a \geq b > 0$ is developed. Let c be the remainder from dividing a by b :

$$a = qb + c \quad \text{and} \quad 0 \leq c < b \qquad q, c \in \mathbb{N}$$

For $c = 0$, b is the greatest common divisor gcd(a,b). For $c > 0$, the following proof shows that $\gcd(a,b) = \gcd(b,c)$. The procedure is therefore continued with the natural numbers b and c. It terminates after a finite number of steps since $c < b$.

(2) The determination of the greatest common divisor of s natural numbers $\{a_1, ..., a_s\}$ is reduced to the determination of the greatest common divisor of $(s-1)$ natural numbers $\{b_2, a_3, ..., a_s\}$ by taking $b_2 = \gcd(a_1, a_2)$ (see the following proof) :

$$\gcd(a_1, a_2, ..., a_s) = \gcd(b_2, a_3, ..., a_s) \quad \text{with} \quad b_2 = \gcd(a_1, a_2)$$
$$\gcd(b_2, a_3, ..., a_s) = \gcd(b_3, a_4, ..., a_s) \quad \text{with} \quad b_3 = \gcd(b_2, a_3)$$

The greatest common divisor $\gcd(a_1, ..., a_s)$ is determined after at most $s-1$ reductions. For $b_i = 1$, the algorithm terminates earlier.

(3) The generating element of the group H is $a = \gcd(a_1, ..., a_s)$.

Proof : Euclidean algorithm

(1) Since $a = q\,b + c$, every common divisor of b and c is also a common divisor of a and b. Since $c = a - qb$, every common divisor of a and b is also a common divisor of b and c.

$$a = qb + c \quad \Rightarrow \quad \gcd(a,b) = \gcd(b,c)$$

(2) If the pair (b,c) possesses a representation $\gcd(b,c) = u_1 b + u_2 c$, then the pair (a,b) has a representation $\gcd(a,b) = u_2 a + u_3 b$:

$$\begin{aligned} \gcd(a,b) &= \gcd(b,c) = u_1 b + u_2 c = u_1 b + u_2\,(a - q\,b) \\ \gcd(a,b) &= u_2 a + (u_1 - q\,u_2) b \end{aligned}$$

(3) If $c = 0$ in $a = qb + c$, then $\gcd(a,b) = b$. If $c > 0$, the procedure is continued with the pair (b, c). If $\gcd(b,c) = c$, then $\gcd(a,b) = c$. Otherwise the procedure is continued. It ends after a finite number of steps since $c < b$, so that the natural numbers decrease in every step.

(4) Let the greatest common divisor of the numbers $\{a_1, ..., a_s\}$ be g. Then $b_2 = \gcd(a_1, a_2)$ contains the number g as a factor, that is $b_2 = mg$. Hence $\{a_1, ..., a_s\}$ and $\{b_2, a_3, ..., a_s\}$ have the same greatest common divisor g.

Example 1 : Euclidean algorithm for pairs of numbers

The numbers 108 and 84 have the gcd(108, 84) = 12. In the first step the remainder 24 is determined; in the second step the remainder 12 is determined. In the third step the remainder is 0, and the greatest common divisor 12 is determined :

$$\begin{aligned} (108, 84) :&\quad 108 = 1 * 84 + 24 \\ (84, 24) :&\quad 84 = 3 * 24 + 12 \\ (24, 12) :&\quad 24 = 2 * 12 + 0 \end{aligned}$$

The two pairs of numbers have the same greatest common divisor :

$$\gcd(108, 84) = \gcd(84, 24) = \gcd(24, 12) = 12$$

The numbers 37 and 13 are mutually prime :

$$\begin{aligned} (37, 13) :&\quad 37 = 2 * 13 + 11 \\ (13, 11) :&\quad 13 = 1 * 11 + 2 \\ (11, 2) :&\quad 11 = 5 * 2 + 1 \\ (2, 1) :&\quad 2 = 2 * 1 + 0 \end{aligned}$$

$$\gcd(37, 13) = \gcd(13, 11) = \gcd(11, 2) = \gcd(2,1) = 1$$

Example 2 : Euclidean algorithm for four numbers

The greatest common divisor gcd$(60, 105, 21, 84)$ is to be determined. In the first step, gcd $(60, 105) = 15$ is determined :

$$\begin{aligned} 105 &= 1 * 60 + 45 \\ 60 &= 1 * 45 + 15 \\ 45 &= 3 * 15 \end{aligned}$$

Now gcd$(15, 21, 84)$ is to be determined. For this purpose gcd$(15, 21) = 3$ is determined in the second step :

$$\begin{aligned} 21 &= 1 * 15 + 6 \\ 15 &= 2 * \ \ 6 + 3 \\ 6 &= 2 * 3 \end{aligned}$$

Now gcd$(3, 84) = 3$ is determined : $84 = 28 * 3$. Combining these results yields gcd$(60, 105, 21, 84) = 3$. The greatest common divisor 3 may be represented as a combination of the numbers $60, 105, 21, 84$:

$$-60 + 105 - 6 * 21 + 84 = 3$$

7.3.6 CYCLIC SUBGROUPS

Introduction : Cyclic groups have a particularly simple structure, namely the structure of the additive group $(\mathbb{Z}\,;+)$ of integers or of one of its groups of residue classes. It follows that cyclic groups are completely characterized by their order. These properties become apparent after the definition of residue classes in Section 7.4.3 and of the isomorphism of groups in Sections 7.5.3 and 7.5.4.

In the following section, cyclic subgroups are treated without making use of the isomorphism mentioned above, since the required concepts have not yet been introduced. Such cyclic subgroups exist in cyclic groups, but also as subgroups of general groups with more than one generating element. In view of the close relationship with the group $(\mathbb{Z}\,;+)$ of integers, additive notation is used.

Properties of subgroups of a cyclic group

(U1) Every subgroup H of a cyclic group G is cyclic.

(U2) Let $(G\,;+)$ be a finite cyclic group of order m with a generating element a. Then for every divisor n of m there is exactly one subgroup H of order n in G. The generating element of H is $\frac{m}{n}$ a. There are no other subgroups of G.

$$H = gp(sa) \quad \text{with} \quad m = sn$$

(U3) Let $(G\,;+)$ be a finite cyclic group of order m with a generating element a. The group G is also generated by a multiple na of the element a if and only if m and n are mutually prime.

$$\text{ord}\, G = m \quad \wedge \quad \gcd(m,n) = 1 \quad \Leftrightarrow \quad gp(a) = gp(na)$$

(U4) If the group G is infinite and a is a generating element of G, then for every subgroup H of G there is exactly one natural number s with $H = gp(sa)$.

Proof : Properties of subgroups of a cyclic group

(U1) Let H be a subgroup, and let a be a generating element of the cyclic group G. Then every element of H has the form $h = ta$ with an element t of the additive group $(\mathbb{Z}\,;+)$ of integers. Hence $T := \{t \in \mathbb{Z} \mid ta \in H\}$ is a subgroup of $\mathbb{Z}$: If $t_1, t_2 \in T$ and $h_1 = t_1 a$, $h_2 = t_2 a$, then $t_3 = t_1 + t_2 \in T$ since $h_3 = h_1 + h_2 \in H$. Since all subgroups of $\mathbb{Z}$ are of the form $\mathbb{Z}u$, there is a natural number s with $T = \mathbb{Z}s$ and $t = ks$. It follows that every subgroup H of G is cyclic :

$$\begin{aligned} H &= \{ta \mid t \in T\} = \{ksa \mid k \in \mathbb{Z}\} = \{k(sa) \mid k \in \mathbb{Z}\} \\ H &= gp(sa) \end{aligned}$$

(U2) Let H be a subgroup of order n in a cyclic group $G = gp(a)$ of order m. By (U1), there is a least positive integer s such that $H = gp(sa)$. The identity element 0 of G is also the identity element of H. Hence $ma = nsa = 0$. By property (Z2) in Section 7.3.4, $ns = 0 \bmod m$, and hence $ns = km$ with $k \in \mathbb{N}'$. If k has a prime factor p, then p is a divisor of n or of s. If p is a divisor of n :

$$\frac{n}{p} s = \frac{k}{p} m \quad \Rightarrow \quad \frac{n}{p} sa = \frac{k}{p} ma = 0 \quad \text{since} \quad ma = 0$$

This is a contradiction, since sa generates the group H of order n. If p is a divisor of s :

$$n \frac{s}{p} = \frac{k}{p} m \quad \Rightarrow \quad n \frac{s}{p} a = \frac{k}{p} ma = 0$$

Hence the element $\frac{s}{p}$ a is at most of order n. But $p \frac{s}{p} a = sa$, and therefore $gp(sa) \subseteq gp(\frac{s}{p} a)$. Since ord $(gp(sa)) = n$, it follows that $gp(\frac{s}{p} a) = gp(sa) = H$. This contradicts the fact that s is the least positive integer with $gp(sa) = H$. The contradictions imply $k = 1$, and hence $m = ns$, so that the order n of H is a divisor of the order m of G. The expression $s = \frac{m}{n}$ shows that s, and hence $H = gp(sa)$, is unique for a given value of n.

Conversely, let n be a divisor of the order m of a cyclic group $G = gp(a)$. Then $m = ns$ with a natural number s. But $ma = 0$, and thus $nsa = n(sa) = 0$. Since m is the least positive integer for which $ma = 0$ holds, n is the least positive integer for which $n(sa) = 0$ holds. Hence there is a cyclic subgroup of G with the generating element $sa = \frac{m}{n}$ a and the order n.

(U3) Let the order m of a cyclic group gp(a) and a natural number n be mutually prime. Then by property (G2) in Section 7.3.5 there are integers c_1 and c_2 such that $c_1 m + c_2 n = 1$, and therefore $c_2(na) = a - c_1 ma$. Since the order m of gp(a) is finite, $ma = 0$ and hence $c_2(na) = a$. Every multiple of a may thus be represented as a multiple of the element na, so that $gp(a) = gp(na)$.

Assume that a finite cyclic group G is generated by the element a and also by the element na, that is $gp(a) = gp(na)$. Then every multiple of a may be represented as a multiple of na, in particular $a = tna$ with $t \in \mathbb{Z}$. By (Z2) in Section 7.3.4, this implies $1 = tn \bmod m$, and thus $sm + tn = 1$ with $s, t \in \mathbb{Z}$. Hence by property (G2) the natural numbers m and n are mutually prime.

(U4) By property (U1), every subgroup H of a cyclic group $G = gp(a)$ is a cyclic group $H = gp(sa)$. It is to be shown that different values of s lead to different subgroups H if the group G is infinite.

All elements na of the group G are different and therefore unique. For if $n_1 a = n_2 a$ for $n_1 \neq n_2$, the group with $(n_2 - n_1)a = 0$ would be finite, contrary to the hypothesis.

For every element na of G, the division $n = qs + r$ with $0 \le r < s$ leads to a unique remainder r. Hence every element na of G may be associated with one of the s residue classes $[r] := \mathbb{Z}s + r$. The elements of H have the form ksa, that is $n = ks$, and hence $r = 0$. These elements of G form the class [0].

Different values of s lead to different partitions of G into residue classes [r]. Therefore the class [0] is also different for different values of s. Hence the number s for a given subgroup H is unique.

Order of an element : Let the identity element of a general group (G ; +) be 0. If for an arbitrary element $a \in G$ there is at least one positive integer m with $ma = 0$, then the least positive integer k with $ka = 0$ is called the order of the element a in the group G. If there is no positive integer m with $ma = 0$, then the order of the element a in the group G is infinite. The order of the element a is designated by ord a.

$$(\text{ord } a) \cdot a = 0 \quad \vee \quad \text{ord}\, a = \infty$$

Properties of cyclic subgroups of an arbitrary group :

(E1) Every element a of an arbitrary group (G ; +) generates a cyclic subgroup gp(a) of G whose order coincides with the order of the element a in G.

$$\text{ord gp}(a) = \text{ord } a$$

(E2) Let a be an element of order m in an arbitrary group (G ; +). Then the element $na \in G$ with $n \in \mathbb{Z}$ is of order $m / \gcd(n,m)$.

$$\text{ord } a = m \quad \Rightarrow \quad \text{ord}(na) = \frac{m}{\gcd(m, n)}$$

Proof E1 : Order of a cyclic subgroup

Let the order k of the element a be finite, that is $ka = 0$. Then the element a generates exactly the elements $\{0, a, 2a, ..., (k-1)a\}$, since any number m may be represented as $m = qk + r$ with $0 \le r < k$, so that $ma = (qk + r)a = q(ka) + ra = ra$. Hence the order of the cyclic group $\text{gp}(a) = \{0, a, ..., (k-1)a\}$ is equal to the order of a in G. If the order of the element a is infinite, then the order of gp(a) is also infinite.

Proof E2 : Order of the multiples of an element

The number $c = \gcd(m, n)$ is a divisor of the order m of gp(a). By property (U2), gp(ca) is a cyclic group of order $\frac{m}{c}$. The numbers $\frac{m}{c}$ and $\frac{n}{c}$ are mutually prime. Hence, by property (U3), $\text{gp}(ca) = \text{gp}(\frac{n}{c}ca) = \text{gp}(na)$ with $\text{ord}(na) = \text{ord}(ca) = \frac{m}{c}$.

Example 1 : Subgroups of a finite cyclic group

$+$	a_0	a_1	a_2	a_3	a_4	a_5
a_0	a_0	a_1	a_2	a_3	a_4	a_5
a_1	a_1	a_2	a_3	a_4	a_5	a_0
a_2	a_2	a_3	a_4	a_5	a_0	a_1
a_3	a_3	a_4	a_5	a_0	a_1	a_2
a_4	a_4	a_5	a_0	a_1	a_2	a_3
a_5	a_5	a_0	a_1	a_2	a_3	a_4

$G = gp(a_1) = gp(a_5)$

$+$	a_0	a_2	a_4
a_0	a_0	a_2	a_4
a_2	a_2	a_4	a_0
a_4	a_4	a_0	a_2

$H_1 = gp(a_2) = gp(a_4)$

$+$	a_0	a_3
a_0	a_0	a_3
a_3	a_3	a_0

$H_2 = gp(a_3)$

The cyclic group G of order 6 has the property $6a_1 = a_0$. By property (U2), since the order of G has the divisors 2 and 3, G contains exactly two subgroups : the cyclic group H_1 of order 3 generated by the element $\frac{6}{3}a_1 = 2a_1 = a_2$ and the cyclic group H_2 of order 2 generated by the element $\frac{6}{2}a_1 = 3a_1 = a_3$.

Example 2 : Alternative generating elements

By property (E1) of cyclic subgroups, the group G in Example 1 is also generated by the elements $5a_1 = a_5$, $7a_1 = a_1$, $11a_1 = a_5$,..., since the numbers $\{5, 7, 11, ...\}$ are mutually prime to the order 6 of the group G. By contrast, elements such as $2a_1 = a_2$, $3a_1 = a_3$, $4a_1 = a_4$, $8a_1 = a_2$, $9a_1 = a_3$,... only generate subgroups of G, since the numbers $\{2, 3, 4, 8, 9,...\}$ have divisors in common with 6.

7.4 CLASS STRUCTURE

7.4.1 CLASSES

Introduction : The structure of groups is studied using structurally compatible (homomorphic) mappings. Such mappings are constructed by partitioning a group into disjoint classes of equivalent elements. Such a partition is called a classification of the group. Different equivalence relations lead to different classifications of the group. The class structure of general groups is studied in Section 7.4. Homomorphic mappings are treated in Section 7.5.

Partitioning : Only in very simple cases is it possible to partition a group into disjoint subsets of elements with equivalent properties by inspecting its operation table. The simple example of the symmetry group of the regular tetrahedron already illustrates the difficulties which arise. For example, a group cannot be partitioned into disjoint subgroups, since every subgroup contains the identity element of the group (property U2 in Section 7.2).

The equivalence relations treated in Section 2.4 may be used to partition a set into disjoint classes of equivalent elements. Every element of the quotient set of the equivalence relation is a class of the group. Using an equivalence relation, a systematic classification of the group may be derived from its operation table. Two equivalence relations which partition groups into cosets and into classes of conjugate elements, respectively, are mentioned in the following. These classifications are studied in detail in subsequent sections.

Cosets : One of the equivalence relations used for classifying groups is based on the relationship $a \circ b = c$ for elements a, b, c of a group $(G ; \circ)$. Let a subgroup H of G be given as one of the classes.

The elements a, c are equivalent with respect to H if there is an element $h \in H$ which satisfies $a \circ h = c$. Equivalent pairs (a, c) form a left coset with respect to H. It turns out that the number of elements in every left coset is equal to the number of elements in H. Analogously, the relationship $h \circ a = c$ leads to right cosets.

The number of left and right cosets is equal and is called the index of the subgroup H in the group G. If the left and right cosets coincide, H is called a normal subgroup. If an element n of a normal subgroup is transformed using an arbitrary element $g \in G$, then the transformed element $g^{-1} \circ n \circ g$ is also an element of the normal subgroup. Normal subgroups are of fundamental importance for the structure of a group.

Classes of conjugate elements : A second equivalence relation used for classifying groups is based on the relationship $a \circ b = b \circ c$ for elements a, b, c of a group $(G\,;\circ)$. The elements a, c are said to be conjugate if there is an element $b \in G$ which satisfies $a \circ b = b \circ c$. Conjugate elements of G form an equivalence class. The number of elements in different classes of conjugate elements of a group G is generally different. The only subgroup among the classes of conjugate elements is the trivial subgroup $\{1\}$.

Classes of conjugate subsets : The concept of conjugation may also be used to classify a set of subsets A, B, ... of a group G. A subset A is called the g-transform of a subset B if A contains the transformed element $a = g^{-1} \circ b \circ g$ for every element $b \in B$. Let a subgroup H of G for classifying the subsets A, B, ... be given. Then the subsets A and B are said to be H-conjugate if there is an element $h \in H$ such that A is the h-transform of B. H-conjugate subsets of G form an equivalence class.

The fundamentals of class structure are treated in the following.

7.4.2 COSETS AND NORMAL SUBGROUPS

Left cosets : The elements of a group $(G\,;\circ)$ are classified with respect to a given subgroup H of G by defining the relation "left-equivalent with respect to H". A pair (a, c) in the cartesian product $G \times G$ is said to be left-equivalent with respect to H if there is an element h of H for which $a \circ h = c$. The relation has the properties of an equivalence relation :

(1) Reflexive : Every element a of the group G is left-equivalent to itself, since the subgroup H contains the identity element 1_G.

$$a \circ 1_G = a$$

(2) Symmetric : If an element a is left-equivalent to an element c, then c is also left-equivalent to a, since for every element h the subgroup H also contains the inverse element h^{-1}.

$$a \circ h = c \quad \Rightarrow \quad a \circ h \circ h^{-1} = c \circ h^{-1} \quad \Rightarrow \quad c \circ h^{-1} = a$$

(3) Transitive : If a is left-equivalent to c and c is left-equivalent to e, then a is left-equivalent to e, since for every pair of elements h_1, h_2 the subgroup H also contains the product $h_1 \circ h_2 = h_3$.

$$a \circ h_1 = c \quad \wedge \quad c \circ h_2 = e \quad \Rightarrow \quad a \circ h_1 \circ h_2 = c \circ h_2 \quad \Rightarrow \quad a \circ h_3 = e$$

Thus elements of G which are left-equivalent with respect to H form an equivalence class. This is identified by a representative a and designated by $a \circ H$ (left coset of a with respect to H). The quotient set of the left cosets of H in G is designated by G / H.

$$a \circ H := \{c \in G \mid a \circ h = c \quad \wedge \quad h \in H\}$$

Properties of left cosets : The left cosets with respect to a subgroup H in a group $(G\,;\circ)$ have the following properties :

(1) If a is an element of the subgroup H, then the left coset $a \circ H$ coincides with H, since the product $a \circ b$ of elements a, b of the group H is also an element of H.

$$a \in H \quad \Rightarrow \quad a \circ H = H$$

(2) The mapping $f : H \to a \circ H$ from the subgroup H to the left coset $a \circ H$ is bijective. The mapping is surjective since by definition for every element $c \in a \circ H$ there is an $h \in H$ with $c = a \circ h$. The mapping is also injective since any two different elements $h_1, h_2 \in H$ are mapped to two different elements $a \circ h_1$, $a \circ h_2 \in a \circ H$.

$$a \circ h_1 \neq a \circ h_2 \quad \Leftrightarrow \quad a^{-1} \circ a \circ h_1 \neq a^{-1} \circ a \circ h_2 \quad \Leftrightarrow \quad h_1 \neq h_2$$

The mapping f is bijective since it is both surjective and injective.

(3) The left cosets are equivalence classes and therefore form a partition of the set G. One of the disjoint subsets is the subgroup H itself. Each of the other subsets can be mapped to H bijectively.

(4) Since the cosets form a partition of G, only one of the cosets $a \circ H$ contains the identity element 1_G. However, every subgroup of G by definition contains the identity element 1_G. Hence the subgroup H contains the element 1_G and is therefore the only coset in G/H which is a subgroup of G.

Right cosets : In order to classify the elements of a group $(G ; \circ)$ with respect to a given subgroup H, the relation ”right-equivalent with respect to H” is defined. The pair (a, c) in the cartesian product $G \times G$ is called right-equivalent with respect to H if there is an element h of H for which $h \circ a = c$. Thus right-equivalence differs from left-equivalence in the order of the factors a and h. Elements of G which are right-equivalent with respect to H form an equivalence class. This is identified by a representative a and designated by $H \circ a$ (right coset of a with respect to H). The quotient set of the right cosets of the subgroup H in G is designated by $G \backslash H$.

$$H \circ a \ := \ \{ c \in G \mid h \circ a = c \ \wedge \ h \in H \}$$

Properties of right cosets : In analogy with the left cosets, the right cosets with respect to a subgroup H in a group $(G ; \circ)$ have the following properties :

(1) If a is an element of the subgroup H, then the right coset $H \circ a$ is H itself.

(2) There is a bijective mapping $f : H \rightarrow H \circ a$ from the subgroup H to the coset $H \circ a$.

(3) The right cosets form a partition of the set G.

(4) The subgroup H is the only coset in $G \backslash H$ which is a subgroup of G.

Index of a subgroup : Let H be a subgroup of the finite group $(G ; \circ)$. The left and right cosets of H in G are determined. Each of the cosets may be bijectively mapped to H. Hence the order of every coset is equal to the order of H. Since the left and right cosets of H form partitions of G, the number of elements in G is a multiple of the number of elements in H. The number of left and right cosets of H in G is equal. The number of cosets of H in G is called the index of H in G and is designated by [G : H].

$$[G : H] := \text{ord } G/H = \text{ord } G \backslash H$$

Lagrange's Theorem : The order of every subgroup H of a finite group $(G ; \circ)$ is a divisor of the order of the group G.

$$\text{ord } G = [G : H] \cdot \text{ord } H$$

Lagrange's Theorem does not assert that if the order of a group has a divisor m the group necessarily contains a subgroup of order m. This is not the case. For example, the order 12 of the alternating group A_4 in Section 7.7.7 has the divisor 6, but the group A_4 does not contain a subgroup of order 6. The existence of subgroups is treated in the theorems of Sylow in Section 7.8.3.

Corollaries to Lagrange's Theorem :

(F1) Let the group $(G ; +)$ be finite. Each element a of G generates a cyclic subgroup gp(a). By Lagrange's Theorem, the order of the subgroup gp(a) is a divisor of the order of the group G. Hence the order of the element a is a divisor of the order of G.

$$a \in G \quad \Rightarrow \quad \text{ord } a \mid \text{ord } G$$

(F2) Let $(G ; +)$ be a finite group of order n with the identity element 0. Then the n-fold multiple of every element a of G is equal to the identity element 0 (Fermat's lesser theorem).

Let the order of the element a be k. Then by (F1) k divides the order n of the group G, that is $n = qk$. Since $ka = 0$, the n-fold multiple of a is obtained as :

$$na = (qk)a = q(ka) = q0 = 0$$

(F3) Every group of prime order is cyclic, and hence abelian. To prove this, consider the cyclic subgroup gp(a) generated by an element $a \neq 0$ of a group $(G ; +)$. By (F1), the order of gp(a) divides the order of G. But the order of G is prime, and hence $\text{ord} G = \text{ord} a$. Thus the group G is cyclic. By property (Z4) in Section 7.3.4, cyclic groups are abelian.

(F4) Let the subgroups H and K of a group $(G ; +)$ be nested, that is $K \subseteq H$. Then the index of K in G is the product of the index of H in G with the index of K in H.

$$K \subseteq H \subseteq G \quad \Rightarrow \quad [G : K] = [G : H] \cdot [H : K]$$

The proof follows directly from Lagrange's Theorem :

$$[G : H] \cdot [H : K] = \frac{\text{ord } G}{\text{ord } H} \cdot \frac{\text{ord } H}{\text{ord } K} = \frac{\text{ord } G}{\text{ord } K} = [G : K]$$

Order of a product of subgroups : Let H_1 and H_2 be finite subgroups of a group $(G\,;\circ)$. The number of elements in the product $H_1 \circ H_2$ is :

$$\text{ord}\,(H_1 \circ H_2) \;=\; \frac{\text{ord}\;H_1 \cdot \text{ord}\;H_2}{\text{ord}\;(H_1 \cap H_2)}$$

Proof : Order of a product of subgroups

The left cosets $a \circ H_2$ and $b \circ H_2$ are formed with different elements $a,b \in H_1$. They are equal if and only if $a^{-1} \circ b$ is an element of H_2 :

$$\begin{aligned} a \circ H_2 = b \circ H_2 \quad &\Leftrightarrow \quad H_2 = (a^{-1} \circ b) \circ H_2 \\ &\Leftrightarrow \quad a^{-1} \circ b \in H_2 \end{aligned}$$

From $a,b \in H_1$ it follows that $a^{-1} \circ b \in H_1$. Hence $a \circ H_2 = b \circ H_2$ holds if and only if $a^{-1} \circ b$ is an element of $H_1 \cap H_2$. This implies :

$$\begin{aligned} a \circ H_2 = b \circ H_2 \quad &\Leftrightarrow \quad a^{-1} \circ b \in H_1 \cap H_2 \\ &\Leftrightarrow \quad H_1 \cap H_2 = (a^{-1} \circ b) \circ (H_1 \cap H_2) \\ &\Leftrightarrow \quad a \circ (H_1 \cap H_2) = b \circ (H_1 \cap H_2) \end{aligned}$$

The equivalence shows that the number of different left cosets of H_2 formed with elements of H_1 is equal to the number of different left cosets of $H_1 \cap H_2$ formed with elements of H_1. Since $H_1 \cap H_2$ is a subgroup of H_1, this number is the index $n = [H_1 : H_1 \cap H_2]$. Lagrange's Theorem yields :

$$\text{ord}\;H_1 \;=\; n\;\text{ord}\,(H_1 \cap H_2)$$

The number of elements in every left coset of H_2 is ord H_2. The order of $H_1 \circ H_2$ is this number times the number n of different cosets of H_2 :

$$\text{ord}\,(H_1 \circ H_2) \;=\; n\;\text{ord}\;H_2 \;=\; \frac{\text{ord}\;H_1}{\text{ord}\;(H_1 \cap H_2)}\,\text{ord}\;H_2$$

Normal subgroup : The operation $\circ$ of a group $(G\,;\circ)$ is generally not symmetric. The left and right cosets of a subgroup H of G are therefore generally different. A subgroup $(N\,;\circ)$ is called a normal (invariant) subgroup of G if the left and right cosets of N in G coincide. This relationship is designated by $N \lhd G$ (N is a normal subgroup of G).

$$N \lhd G \;:=\; \bigwedge_{a \in G}\;\bigwedge_{n_1 \in N}\;\bigvee_{n_2 \in N}\,(a \circ n_1 \;=\; n_2 \circ a)$$

The cosets of a normal subgroup N are designated by $N \circ a = a \circ N = [a]$. The canonical mapping k from the group G to the quotient set G/N maps the group to the set of cosets.

$$k : \; G \rightarrow G/N \quad \text{with} \quad k(a) \;=\; [a]$$

Properties of normal subgroups :

(T1) A subgroup $(N\,;\circ)$ of a group $(G\,;\circ)$ is a normal subgroup in G if and only if $g \circ n \circ g^{-1} \in N$ for all $g \in G$ and all $n \in N$.

(T2) Let N and T be normal subgroups of a group $(G\,;\circ)$. Then their intersection $N \cap T$ is also a normal subgroup of G.

(T3) Let H be a subgroup and N a normal subgroup of a group $(G\,;\circ)$. Then their intersection $H \cap N$ is a normal subgroup of H.

Further properties of normal subgroups are treated in Section 7.5.5 (automorphisms).

Proof : Properties of normal subgroups

(T1) Let N be a normal subgroup in G. Then for arbitrary elements $g \in G$ and $n \in N$ there is an element $\hat{n} \in N$ with $g \circ n = \hat{n} \circ g$. Multiplying by g^{-1} from the right yields $g \circ n \circ g^{-1} = \hat{n} \in N$.

Conversely, assume that for arbitrary elements $g \in G$ and $n \in N$ the element $\hat{n} := g \circ n \circ g^{-1}$ is contained in N. Then multiplying by g from the right yields $\hat{n} \circ g = g \circ n$. Hence N is a normal subgroup in G.

$$\bigwedge_{g\in G}\ \bigwedge_{n\in N}\ \bigvee_{\hat{n}\in N} (g \circ n = \hat{n} \circ g) \quad \Leftrightarrow$$

$$\bigwedge_{g\in G}\ \bigwedge_{n\in N}\ \bigvee_{\hat{n}\in N} (g \circ n \circ g^{-1} = \hat{n}) \quad \Leftrightarrow$$

$$\bigwedge_{g\in G}\ \bigwedge_{n\in N} (g \circ n \circ g^{-1} \in N)$$

(T2) For arbitrary elements $g \in G$ and $n \in N$, (T1) yields $g \circ n \circ g^{-1} \in N$. For arbitrary elements $g \in G$ and $t \in T$, (T1) yields $g \circ t \circ g^{-1} \in T$. For arbitrary elements $g \in G$ and $s \in N \cap T$, this implies $g \circ s \circ g^{-1} \in N \cap T$.

$$\bigwedge_{g\in G}\ \bigwedge_{n\in N} (g \circ n \circ g^{-1} \in N) \ \wedge\ \bigwedge_{g\in G}\ \bigwedge_{t\in T} (g \circ t \circ g^{-1} \in T) \quad \Rightarrow$$

$$\bigwedge_{g\in G}\ \bigwedge_{s\in N\cap T} (g \circ s \circ g^{-1} \in N \cap T)$$

(T3) For arbitrary elements $h \in H$ and $a \in H \cap N$, the product $h \circ a \circ h^{-1}$ is an element of the group H. The product $h \circ a \circ h^{-1}$ is also an element of N, since $a \in N$, $h \in G$ and N is a normal subgroup in G. Thus $h \circ a \circ h^{-1}$ is an element of $H \cap N$, and hence $H \cap N$ is a normal subgroup in H.

Simple groups : A group $(G\,;\circ)$ with the identity element 1 is said to be simple if $G \neq \{1\}$ and the improper subgroups $\{1\}$ and G are the only normal subgroups in G.

Example 1 : Normal subgroups of the symmetry group of equilateral triangles

The symmetry group $G = \{a_0, ..., a_5\}$ of equilateral triangles is shown in Example 1 of Section 7.3.2. The subgroup $H = \{a_0, a_1, a_2\}$ contains the rotations in the plane. The rotations $\{a_3, a_4, a_5\}$ in space form a coset of H in G. The subgroup H is a normal subgroup in G, since $H \circ a_3 = a_3 \circ H = [a_3]$. The cosets $\{a_0, a_1, a_2\}$ and $\{a_3, a_4, a_5\}$ form a partition of G.

$$\begin{array}{lllllllllllll} H \circ a_0 & = & H \circ a_1 & = & H \circ a_2 & = & \{a_0, a_1, a_2\} & = & a_0 \circ H & = & a_1 \circ H & = & a_2 \circ H \\ H \circ a_3 & = & H \circ a_4 & = & H \circ a_5 & = & \{a_3, a_4, a_5\} & = & a_3 \circ H & = & a_4 \circ H & = & a_5 \circ H \end{array}$$

Example 2 : Normal subgroups of the symmetry group of regular tetrahedra

The symmetry group $G = \{a_0, ..., a_{11}\}$ of regular tetrahedra is shown in Example 2 of Section 7.3.2. The subgroup $H = \{a_0, a_9, a_{10}, a_{11}\}$ contains the rotations through 180 degrees about the edge bisectors. The rotations through 120 degrees about the medians and the rotations through 240 degrees about the medians form the cosets $\{a_1, a_4, a_5, a_8\}$ and $\{a_2, a_3, a_6, a_7\}$ of H in G. Since $H \circ a_1 = a_1 \circ H = [a_1]$ and $H \circ a_2 = a_2 \circ H = [a_2]$, the subgroup H is a normal subgroup in G. The cosets $\{a_0, a_9, a_{10}, a_{11}\}$, $\{a_1, a_4, a_5, a_8\}$ and $\{a_2, a_3, a_6, a_7\}$ form a partition of G.

$$\begin{array}{lllllll} H \circ a_0 & = & \{a_0, a_9, a_{10}, a_{11}\} & = & a_0 \circ H & = & [a_0] \\ H \circ a_1 & = & \{a_1, a_4, a_5, a_8\} & = & a_1 \circ H & = & [a_1] \\ H \circ a_2 & = & \{a_2, a_3, a_6, a_7\} & = & a_2 \circ H & = & [a_2] \end{array}$$

7.4.3 GROUPS OF RESIDUE CLASSES

Congruent integers : For pairs (a,n) and (b,n) of integers there is a unique division with remainder.

$$\begin{array}{lllllll} a & = & q_1 n + r_1 & \text{with} & 0 \le r_1 < n & \text{and} & q_1, r_1 \in \mathbb{Z} \\ b & = & q_2 n + r_2 & \text{with} & 0 \le r_2 < n & \text{and} & q_2, r_2 \in \mathbb{Z} \end{array}$$

The numbers a and b are said to be congruent modulo n if the remainders r_1 and r_2 with respect to division by n are equal. This relation is designated by $a \equiv b \bmod n$ (a is congruent with b modulo n). The remainder in the division of b by n is designated by b mod n.

$$a \equiv b \bmod n \quad \Rightarrow \quad a = q_1 n + r_1 \ \wedge \ b = q_2 n + r_2 \ \wedge \ r_1 = r_2$$

Set of residue classes : The subgroup $\mathbb{Z}n$ of the group $(\mathbb{Z}\,;+)$ of integers, defined in Section 7.3.5, contains the integers divisible by n. Since addition of integers is commutative, the left and right cosets of $\mathbb{Z}n$ in $\mathbb{Z}$ coincide, so that $\mathbb{Z}n$ is a normal subgroup in $\mathbb{Z}$. The coset [a] of the normal subgroup $\mathbb{Z}n$ contains all integers which yield the same remainder as the number a when divided by n; it is called a residue class. All numbers in [a] are therefore congruent with a modulo n. The set of residue classes for the number n is $\{[0], [1],...,[n-1]\}$; it is designated by $\mathbb{Z}/\mathbb{Z}n$ or $\mathbb{Z}_n$.

$$\begin{array}{lll} \mathbb{Z}n & = & \{nz \mid z \in \mathbb{Z}\} \\ {[a]} & = & \{c \in \mathbb{Z} \mid c \equiv a \bmod n\} \\ \mathbb{Z}/\mathbb{Z}n & := & \mathbb{Z}_n := \{[0],...,[n-1]\} \end{array}$$

Group of residue classes : The relation "congruent modulo n" is an equivalence relation. Thus the set of residue classes $\mathbb{Z}_n$ is a quotient set.

$$\begin{array}{lll} \text{reflexive} & : & a \equiv a \bmod n \\ \text{symmetric} & : & a \equiv b \bmod n \ \Leftrightarrow \ b \equiv a \bmod n \\ \text{transitive} & : & a \equiv b \bmod n \ \wedge \ b \equiv c \bmod n \ \Rightarrow \ a \equiv c \bmod n \end{array}$$

The inner operation + with $[k]+[m] = [k+m]$ is defined in the quotient set $\{[0],...,[n-1]\}$ with $n \ge 1$. This definition is valid since the compatibility condition $a \in [k], b \in [m] \Rightarrow a+b \in [k+m]$ is satisfied :

$$\begin{array}{lll} a \in [k] & \Rightarrow & k = q_1 n + r_1 \\ & & a = q_2 n + r_1 \\ & & \\ b \in [m] & \Rightarrow & m = q_3 n + r_2 \\ & & b = q_4 n + r_2 \end{array}$$

$$\begin{array}{llcll}
r_1 + r_2 < n: & k+m & = & (q_1+q_3)\,n & + \; (r_1+r_2) \\
 & a+b & = & (q_2+q_4)\,n & + \; (r_1+r_2)
\end{array}$$

$$\begin{array}{llcll}
r_1 + r_2 \geq n: & k+m & = & (q_1+q_3+1)\,n & + \; (r_1+r_2-n) \\
 & a+b & = & (q_2+q_4+1)\,n & + \; (r_1+r_2-n)
\end{array}$$

Hence $a + b \in [k+m]$. The domain $(\mathbb{Z}_n \; ; +)$ is a finite cyclic group of order n; it is called the group of residue classes modulo n. Its identity element is [0]. It is generated by the element [1].

Example 1 : Groups of residue classes modulo n

The normal subgroup $\mathbb{Z}n$ has the following n residue classes :

$$\begin{array}{lcl}
[0] & = & \{\ldots, -2n \;\;\;\;, -n \;\;\;\;, 0 \;, n \;\;\;\;, 2n \;\;\;\;, \ldots\} \\
[1] & = & \{\ldots, -2n+1, -n+1, 1, n+1, 2n+1, \ldots\} \\
[2] & = & \{\ldots, -2n+2, -n+2, 2, n+2, 2n+2, \ldots\} \\
\vdots & & \\
[n-1] & = & \{\ldots, -n-1 \;\;, -1, n-1 \;\;, 2n-1, 3n-1, \ldots\}
\end{array}$$

Thus the group of residue classes modulo 3 consists of the following sets :

$$\begin{array}{lcl}
[0] & = & \{\ldots, -6, -3, 0, 3, 6, \ldots\} \\
[1] & = & \{\ldots, -5, -2, 1, 4, 7, \ldots\} \\
[2] & = & \{\ldots, -4, -1, 2, 5, 8, \ldots\} \\
\mathbb{Z}_3 & = & \{[0], [1], [2]\}
\end{array}$$

For the elements $3 \in [0]$, $4 \in [1]$ and $7 \in [1]$, the compatibility conditions is satisfied as follows :

$$3+4 = 7 \quad \wedge \quad [0]+[1] = [0+1] = [1] \quad \wedge \quad 7 \in [1]$$

7.4.4 CONJUGATE ELEMENTS AND SETS

Transform of an element : An element a of a group $(G ; \circ)$ is called the transform of the element b with respect to the element g (g-transform of b) if the following equation holds :

$$a = g^{-1} \circ b \circ g \qquad\qquad a, b, g \in G$$

Conjugate elements : The transformation of the elements of a group $(G ; \circ)$ leads to a classification of the group. For this purpose, the relation "conjugate" is defined. A pair (a, c) in the cartesian product $G \times G$ is said to be conjugate if there is an element g in the group G which transforms a into c.

$$(a, c) \text{ is conjugate } :\Leftrightarrow \bigvee_{g \in G} (a \circ g = g \circ c) \qquad\qquad a, c, g \in G$$

The relation "conjugate" has the properties of an equivalence relation :

(1) Reflexive : Every element a of the group G is conjugate to itself.
$a \circ a^{-1} = a^{-1} \circ a$

(2) Symmetric : If the element a is conjugate to c with respect to the element g, then c is conjugate to a with respect to the inverse element g^{-1}.
$a \circ g = g \circ c \quad \Rightarrow \quad c \circ g^{-1} = g^{-1} \circ a$

(3) Transitive : If a is conjugate to c with respect to g and c is conjugate to e with respect to h, then a is conjugate to e, since together with the elements g and h the group G also contains their product $g \circ h$.
$a \circ g = g \circ c \quad \wedge \quad c \circ h = h \circ e \quad \Rightarrow \quad a \circ (g \circ h) = g \circ c \circ h = (g \circ h) \circ e$

Thus conjugate elements of G form an equivalence class. Such a class is called a conjugacy class. It is identified by a representative a and designated by [a]. The class $[1_G]$ consists only of the identity element, since $g^{-1} \circ 1_G \circ g = 1_G$ for every element g of G. The conjugacy classes form a partition of the set G.

$$[a] := \{ c \in G \mid c = g^{-1} \circ a \circ g \quad \wedge \quad g \in G \}$$

Note : The partition of a group $(G ; \circ)$ with respect to conjugacy must not be confused with its partition with respect to a normal subgroup. In the case of the partition with respect to a normal subgroup N, the condition $g \circ N = N \circ g$ holds for every element $g \in G$. Each of the resulting classes [a] contains the same number of elements (Lagrange's Theorem). The normal subgroup itself is the only class which is a subgroup of G. In the case of the partition with respect to conjugacy, the elements a and c already belong to the class [a] if there is but a single element $g \in G$ for which $a \circ g = g \circ c$. The number of elements in the conjugacy classes may be different. The only subgroup among the conjugacy classes is the trivial subgroup {1}.

Conjugacy classes and normal subgroups : A subgroup H of a group $(G ; \circ)$ is a normal subgroup in G if and only if H consists of entire conjugacy classes. Thus if H contains an element of a conjugacy class then H contains all elements of that conjugacy class. This property may be used to determine all normal subgroups of a group.

Proof : Conjugacy classes and normal subgroups

(1) Let H be a normal subgroup in G. Let the representative a of the conjugacy class $\{c \in G \mid c = g^{-1} \circ a \circ g \wedge g \in G\}$ be an element of H. Then by the definition of a normal subgroup for every $g \in G$ there is an element $c \in H$ such that $g \circ a = c \circ g$, that is $c = g^{-1} \circ a \circ g$. Hence the conjugacy class [a] is contained in H.

(2) Let every element of the conjugacy class [a] be contained in the subgroup H. For every element $g \in G$ there is an element $x \in [a]$ such that $x = g^{-1} \circ a \circ g$, and hence $g \circ x = a \circ g$ with $a, x \in [a] \subseteq H$. If further conjugacy classes are completely contained in H, analogous statements hold for the elements of these classes. If H is the union of entire conjugacy classes, then $g \circ H = H \circ g$ for every $g \in H$. Hence H is a normal subgroup in G.

Transform of a subset : A subset A of a group $(G ; \circ)$ is called the g-transform of the subset B of G if A contains exactly the transforms of the elements of B with respect to $g \in G$.

$$A = g^{-1} \circ B \circ g \quad :\Leftrightarrow \quad A = \{a \in G \mid a = g^{-1} \circ b \circ g \quad \wedge \quad b \in B\}$$

The transforms of a subset have the following properties :

(1) The transform $g^{-1} \circ H \circ g$ of a subgroup H of G is also a group.

(2) The mapping $f : B \rightarrow g^{-1} \circ B \circ g$ with $f(b) = g^{-1} \circ b \circ g$ is bijective.

(3) The transform of a finite subgroup H with respect to an element g of that subgroup is the subgroup H itself.

Proof : Properties of the transform of a subgroup H

(1) The transform $A = g^{-1} \circ H \circ g$ contains the identity element 1_G, the inverse a^{-1} for any element a and the product $a_1 \circ a_2$ for any two elements a_1, a_2 :

$$\bigwedge_{g \in G} (g^{-1} \circ 1_G \circ g = 1_G) \quad \wedge \quad 1_G \in H \quad \Rightarrow \quad 1_G \in g^{-1} \circ H \circ g$$

$$\bigwedge_{a,b \in H} (a = g^{-1} \circ b \circ g \quad \Rightarrow \quad a^{-1} = g^{-1} \circ b^{-1} \circ g)$$

$$\bigwedge_{a_i, b_i \in H} (a_1 = g^{-1} \circ b_1 \circ g \wedge a_2 = g^{-1} \circ b_2 \circ g \quad \Rightarrow \quad a_1 \circ a_2 = g^{-1} \circ b_1 \circ b_2 \circ g)$$

(2) The mapping $f : B \to g^{-1} \circ B \circ g$ is surjective, since by the definition of the transform for every element $a \in g^{-1} \circ B \circ g$ there is an element $b \in B$ with $a = g^{-1} \circ b \circ g$. The mapping is injective, since two different elements $b_1, b_2 \in B$ are mapped to two different elements $g^{-1} \circ b_1 \circ g$, $g^{-1} \circ b_2 \circ g$ of $g^{-1} \circ B \circ g$:

$$g^{-1} \circ b_1 \circ g \neq g^{-1} \circ b_2 \circ g \Leftrightarrow$$

$$g \circ g^{-1} \circ b_1 \circ g \circ g^{-1} \neq g \circ g^{-1} \circ b_2 \circ g \circ g^{-1} \Leftrightarrow b_1 \neq b_2$$

The mapping f is bijective, since it is surjective and injective.

(3) If b and g are elements of the subgroup H, then the inverse element g^{-1} is also an element of H. Hence the product $g^{-1} \circ b \circ g$ is an element of the group H. Since H is finite, it follows that $H = g^{-1} \circ H \circ g$.

Conjugate subsets : Let $\{M_1, ..., M_s\}$ be a set of subsets of a group $(G ; \circ)$ with a subgroup H. The relation "H-conjugate" is defined in order to classify the subsets M_i with respect to the subgroup H. A pair (M_i, M_k) in the cartesian product $\{M_1, ..., M_s\} \times \{M_1, ..., M_s\}$ is said to be H-conjugate if M_i is the transform of M_k with respect to an element h of H. M_i is called an H-conjugate of M_k. The relation "H-conjugate" is an equivalence relation.

$$M_i \text{ and } M_k \text{ are H-conjugate } :\Leftrightarrow \bigvee_{h \in H} (M_i = h^{-1} \circ M_k \circ h)$$

(1) Reflexive : Every set M_i is its own H-conjugate.

$$M_i = 1_G^{-1} \circ M_i \circ 1_G \quad \wedge \quad 1_G \in H$$

(2) Symmetric : If M_i is H-conjugate to M_k, then M_k is H-conjugate to M_i, since together with h the group H also contains the inverse h^{-1}.

$$M_i = h^{-1} \circ M_k \circ h \quad \Rightarrow \quad M_k = h \circ M_i \circ h^{-1}$$

(3) Transitive : If M_i is H-conjugate to M_k and M_k is H-conjugate to M_n, then M_i is H-conjugate to M_n, since together with the elements h_1 and h_2 the group H also contains their product $h_2 \circ h_1$.

$$M_i = h_1^{-1} \circ M_k \circ h_1 \quad \wedge \quad M_k = h_2^{-1} \circ M_n \circ h_2 \quad \Rightarrow$$

$$M_i = (h_2 \circ h_1)^{-1} \circ M_n \circ (h_2 \circ h_1)$$

Thus elements of $\{M_1, ..., M_s\}$ which are H-conjugate form an equivalence class. It is identified by a representative M_i and designated by $[M_i]$. The classes $[M_i]$ of H-conjugate subsets form a partition of the set $\{M_1, ..., M_s\}$.

$$[M_i] = \{M_n \in \{M_1, ..., M_s\} \mid \bigvee_{h \in H} (M_n = h^{-1} \circ M_i \circ h)\}$$

Example 1 : Conjugate elements of the tetrahedral symmetry group

The symmetry group $G = \{a_0, ..., a_{11}\}$ for regular tetrahedra is shown in Example 2 of Section 7.3.2. The transforms of the elements a_k with respect to the elements a_m are compiled in the following transformation table. The element in row k and column m of the table is the transform of a_m with respect to a_k.

T	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
a_0	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
a_1	a_0	a_1	a_2	a_7	a_8	a_4	a_3	a_6	a_5	a_{11}	a_9	a_{10}
a_2	a_0	a_1	a_2	a_6	a_5	a_8	a_7	a_3	a_4	a_{10}	a_{11}	a_9
a_3	a_0	a_8	a_7	a_3	a_4	a_1	a_2	a_6	a_5	a_{10}	a_{11}	a_9
a_4	a_0	a_5	a_6	a_3	a_4	a_8	a_7	a_2	a_1	a_{11}	a_9	a_{10}
a_5	a_0	a_8	a_7	a_2	a_1	a_5	a_6	a_3	a_4	a_{11}	a_9	a_{10}
a_6	a_0	a_4	a_3	a_7	a_8	a_5	a_6	a_2	a_1	a_{10}	a_{11}	a_9
a_7	a_0	a_5	a_6	a_2	a_1	a_4	a_3	a_7	a_8	a_{10}	a_{11}	a_9
a_8	a_0	a_4	a_3	a_6	a_5	a_1	a_2	a_7	a_8	a_{11}	a_9	a_{10}
a_9	a_0	a_4	a_3	a_2	a_1	a_8	a_7	a_6	a_5	a_9	a_{10}	a_{11}
a_{10}	a_0	a_5	a_6	a_7	a_8	a_1	a_2	a_3	a_4	a_9	a_{10}	a_{11}
a_{11}	a_0	a_8	a_7	a_6	a_5	a_4	a_3	a_2	a_1	a_9	a_{10}	a_{11}

transformation table $a_i = a_k^{-1} \circ a_m \circ a_k$ (row k, column m)

example : $a_2^{-1} \circ a_5 \circ a_2 \; = \; a_1 \circ a_5 \circ a_2 \; = \; a_8$

The conjugacy classes of the group may be read off in the columns of the table marked with a_0, a_1, a_2 and a_9 :

$[a_0] = \{a_0\}$: trivial mapping

$[a_1] = \{a_1, a_4, a_5, a_8\}$: rotations about medians, 120 degrees

$[a_2] = \{a_2, a_3, a_6, a_7\}$: rotations about medians, 240 degrees

$[a_9] = \{a_9, a_{10}, a_{11}\}$: rotations about edge bisectors, 180 degrees

Example 2 : Conjugate subsets of the tetrahedral symmetry group

Consider the following subsets of the tetrahedral symmetry group G of Example 1 :

$$\begin{array}{lll} M_1 = \{a_1, a_2\} & M_3 = \{a_5, a_6\} & M_5 = \{a_9, a_{10}\} \\ M_2 = \{a_3, a_4\} & M_4 = \{a_7, a_8\} & \end{array}$$

The classes of conjugate subsets for the subgroup $H = \{a_0, a_9, a_{10}, a_{11}\}$ are determined according to $[M_i] = \{M_n \in \{M_1, \ldots, M_5\} \mid \bigvee_{h \in H} (M_n = h^{-1} \circ M_i \circ h)\}$.

$$\begin{array}{lcllcl} a_0^{-1} \circ M_1 \circ a_0 & = & M_1 & a_0^{-1} \circ M_5 \circ a_0 & = & M_5 \\ a_9^{-1} \circ M_1 \circ a_9 & = & M_2 & a_9^{-1} \circ M_5 \circ a_9 & = & M_5 \\ a_{10}^{-1} \circ M_1 \circ a_{10} & = & M_3 & a_{10}^{-1} \circ M_5 \circ a_{10} & = & M_5 \\ a_{11}^{-1} \circ M_1 \circ a_{11} & = & M_4 & a_{11}^{-1} \circ M_5 \circ a_{11} & = & M_5 \end{array}$$

Hence the equivalence classes of H-conjugate sets are

$$[M_1] = \{M_1, M_2, M_3, M_4\} \quad \text{and} \quad [M_5] = \{M_5\}.$$

7.5 GROUP STRUCTURE

7.5.1 INTRODUCTION

Subgroups : A group $(G ; \circ)$ cannot be partitioned into disjoint subgroups, since by definition every subgroup contains the identity element 1. The question arises whether a group can be decomposed into subgroups such that any two of these subgroups have only the element 1 in common. The question also arises whether every element of the group G can be constructed by applying the group operation to elements of certain subgroups of G. If this is the case, then the properties of a group may be studied by comparing its generating subgroups with known groups. The set of different constructions of a group from its subgroups is called the group structure. In this section, the structure of given groups is compared using mappings. The construction of groups from subgroups is treated in Sections 7.6 to 7.8.

Homomorphism : Mappings are used to compare groups. A mapping is structurally compatible (homomorphic) if the order in which the mapping and the group operations are carried out may be changed. Thus, in the case of a homomorphic mapping, if two elements of the domain are mapped and the group operation is then performed in the target, the result obtained is the same as if the group operation is performed in the domain and the result is then mapped to the target.

Isomorphism : Studying groups by comparing their subgroups requires a definition of the concept of "groups with identical structure". Two groups are identically structured (isomorphic) if there is a bijective homomorphic mapping between them. It turns out that every infinite cyclic group is isomorphic to the additive group of the integers. Every finite cyclic group is isomorphic to a group of residue classes. Every finite group is isomorphic to a group of permutations.

Natural homomorphism : The canonical mapping $k : G \rightarrow G/N$ from a group $(G ; \circ)$ to its quotient group G/N with respect to a normal subgroup N of G is a homomorphism. This natural homomorphism relates the class structure of a group to its group structure. All elements of the normal subgroup N are mapped to the identity element of G/N. All elements of a coset $a \circ N$ are mapped to the element $[a]$ of G/N. If there is another homomorphism $f : G \rightarrow H$ which maps exactly the elements of N to the identity element 1_H (N is the kernel of f), then the groups G/N and H are isomorphic. This homomorphism theorem is the central theorem in the study of group structure.

Automorphisms : In studying the structure of a group $(G\,;\circ)$, its isomorphic mappings onto itself (automorphisms) are particularly important. The composition of the automorphisms of G leads to the automorphism group of G, a subgroup of the permutation group of G. The mappings which map the elements of G to their g-transforms with composition as the inner operation form a subgroup of the automorphism group (inner automorphisms). Subgroups which are mapped to themselves under all automorphisms are said to be characteristic. Every characteristic subgroup is a normal subgroup of G.

7.5.2 HOMOMORPHISM

Compatible structures : The structure of groups is studied by mapping the groups. A mapping is said to be structurally compatible (homomorphic) if the order in which mappings and operations on elements are carried out may be changed. If two elements $a, b \in G$ are mapped by a homomorphic mapping $f : G \rightarrow H$ and the operation $f(a) \circ f(b)$ is then performed in the target, the result is the same as if the operation $a \circ b$ is performed in the domain and the mapping is then applied to obtain the image $f(a \circ b)$ in the target. Thus, for a structurally compatible mapping, $f(a) \circ f(b) = f(a \circ b)$.

Regarding a homomorphic mapping from a group to another group, the question arises whether the images and the preimages of subgroups are also subgroups. In particular, the question arises whether the images and the preimages of normal subgroups are also normal subgroups. In the following, homomorphic mappings are shown to preserve the group properties and normality. For a group G with a normal subgroup N, this property yields a natural homomorphism which maps G to the quotient set G/N.

Homomorphic mappings : The mapping $f : G \rightarrow H$ from a group $(G\,; \circ_1)$ to a group $(H\,; \circ_2)$ is said to be homomorphic (structurally compatible) if the image $f(a \circ_1 b)$ of the product $a \circ_1 b$ of any two elements a and b of the set G is equal to the product $f(a) \circ_2 f(b)$ of their images $f(a)$ and $f(b)$ in the set H. Thus, in the case of a homomorphic mapping the order of operation and mapping may be changed. The indices which distinguish the operations $\circ_1$ and $\circ_2$ are usually omitted, since the distinction may be inferred from the context. A homomorphic mapping f is called a homomorphism.

$$f : G \rightarrow H \text{ is homomorphic} \quad :\Leftrightarrow \quad \bigwedge_{a,b \in G} (f(a \circ b) = f(a) \circ f(b))$$

Properties of homomorphic mappings : A homomorphic mapping $f : G \rightarrow H$ of groups has the following properties :

(1) The image $f(G)$ of the group G is a group.

(2) The identity element 1_H of the group H is the image of the identity element 1_G of the group G :

$$1_H = f(1_G)$$

(3) The inverse $f(a)^{-1}$ of the image $f(a)$ of an element a of the group G is equal to the image $f(a^{-1})$ of the inverse a^{-1} of the element a :

$$f(a)^{-1} := (f(a))^{-1} = f(a^{-1})$$

Proof : Properties of homomorphic mappings

Let G be a group. The image f(G) is shown to possess the group properties.

(a) For given elements b_1 and b_2 of the image f(G), there are elements a_1 and a_2 of the group G such that $b_1 = f(a_1)$ and $b_2 = f(a_2)$. Together with a_1 and a_2, the group G contains the product $a_1 \circ a_2$. Therefore the image f(G) contains the element $f(a_1 \circ a_2) = f(a_1) \circ f(a_2)$, and hence f(G) contains the product $b_1 \circ b_2$.

(b) The group G contains the identity element 1_G. For every element a of G, $a \circ 1_G = a$ and hence $f(a \circ 1_G) = f(a)$. Since the mapping f is homomorphic, $f(a \circ 1_G) = f(a) \circ f(1_G)$. Combining these two results yields $f(a) \circ f(1_G) = f(a)$, and hence $f(1_G)$ is the identity element of H. This proves property (2).

(c) The identity element 1_H is the image of the identity element 1_G, and hence the image of the product $a \circ a^{-1}$, that is $1_H = f(1_G) = f(a \circ a^{-1})$. Since the mapping f is homomorphic, $f(a \circ a^{-1}) = f(a) \circ f(a^{-1})$. Combining these results yields $1_H = f(a) \circ f(a^{-1})$, and hence $f(a^{-1})^{-1} = f(a)$.

(d) The image f(G) is a group, since it contains the identity element 1_H, the inverse b^{-1} for any element b and the product $b_1 \circ b_2$ for any two elements b_1, b_2. This proves property (1).

Composition of homomorphic mappings : Let G_1, G_2 and G_3 be groups, and let $f_1 : G_1 \rightarrow G_2$ and $f_2 : G_2 \rightarrow G_3$ be homomorphic mappings. Then the composition $f_2 \circ f_1 : G_1 \rightarrow G_3$ is a homomorphic mapping.

Proof : Composition of homomorphic mappings

The homomorphic mapping of arbitrary elements a,b of G_1 yields $f_1(a \circ b) = f_1(a) \circ f_1(b)$ with $f_1(a), f_1(b) \in G_2$. Then the homomorphic mapping of $f_1(a)$ and $f_1(b)$ yields :

$$\begin{aligned}(f_2 \circ f_1)(a \circ b) &= f_2(f_1(a \circ b)) = f_2(f_1(a) \circ f_1(b)) \\ &= f_2(f_1(a)) \circ f_2(f_1(b)) = (f_2 \circ f_1)(a) \circ (f_2 \circ f_1)(b)\end{aligned}$$

Kernel of a homomorphic mapping : Let a mapping $f : G \rightarrow H$ from the group G to the group $f(G) = H$ be homomorphic. The subset of elements of G which are mapped to the identity element 1_H of the group H is called the kernel of the homomorphic mapping f and is designated by ker f.

$$\ker f := \{a \in G \mid f(a) = 1_H\}$$

Injective homomorphism : A homomorphic mapping f from a group G to a group H is injective if and only if the kernel of f is trivial (contains only the identity element 1_G). Thus the homomorphic mapping f is injective if and only if $f(a) = 1_H$ for $a \in G$ implies $a = 1_G$.

$$f : G \rightarrow H \text{ is injective} \quad \Leftrightarrow \quad \ker f = \{1_G\}$$

Proof : Trivial kernel of an injective homomorphism

(1) Let the homomorphic mapping f be injective. For every mapping of a group, $f(1_G) = 1_H$. Since f is injective, 1_G is the only preimage of 1_H, and hence $\ker f = \{1_G\}$.

(2) Let the kernel of the homomorphic mapping f be $\{1_G\}$. Let the elements f(a) and f(b) of H be equal. Then $f(a \circ b^{-1}) = f(a) \circ f(b)^{-1} = f(a) \circ f(a)^{-1} = 1_H$. Since the kernel of f contains only the element 1_G, it follows that $a \circ b^{-1} = 1_G$ and therefore $a = b$. Hence the mapping is injective.

Induced homomorphism : For the groups G, H and Z, let $\phi : G \to H$ be an arbitrary homomorphism, and let $\pi : G \to Z$ be a surjective homomorphism, that is $\pi(G) = Z$. If $\ker \pi \subseteq \ker \phi$, then there is exactly one homomorphism $\sigma : Z \to H$ with $\sigma \circ \pi = \phi$. The mapping σ is called the homomorphism induced by ϕ. It has the following properties :

(1) If ϕ is surjective, then σ is also surjective.

(2) If $\ker \pi = \ker \phi$, then σ is injective.

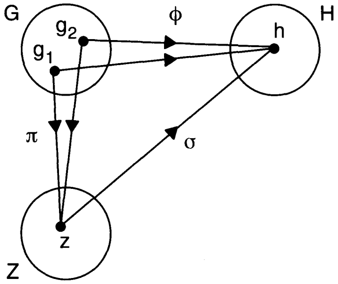
$$\begin{tikzcd}
G \arrow[r, "\phi"] \arrow[d, "\pi"'] & H \\
Z \arrow[ur, "\sigma"'] &
\end{tikzcd}$$

$\sigma : Z \to H$

$\phi = \sigma \circ \pi$

Proof : Induced homomorphism

(a) It is to be proved that $\sigma := \phi \circ \pi^{-1}$ is a mapping although an element $z \in Z$ may have more than one preimage in G which is mapped to H by ϕ. Since the mapping π is surjective, every element $z \in Z$ has at least one preimage in G. Let g_1, g_2 be two different preimages of z, that is $\pi(g_1) = \pi(g_2) = z$. Since $g_1 \circ g_2^{-1}$ is also an element of G and π is homomorphic, it follows that $\pi(g_1 \circ g_2^{-1}) = \pi(g_1) \circ \pi(g_2)^{-1} = \pi(g_1) \circ \pi(g_1)^{-1} = 1_Z$. Thus $g_1 \circ g_2^{-1}$ is an element of $\ker \pi$ and therefore by hypothesis also of $\ker \phi$. Thus $\phi(g_1 \circ g_2^{-1}) = 1_H$, and this implies $\phi(g_1) = \phi(g_2) =: h \in H$. Thus every element $z \in Z$ has a unique image $\sigma(z) = h$ in H. Hence the relation σ is a mapping.



$$\phi(g_1) = \phi(g_2) = h$$

$$\pi(g_1) = \pi(g_2) = z$$

$$\sigma(z) = h$$

(b) It is to be proved that the mapping $\sigma = \phi \circ \pi^{-1}$ satisfies the condition $\sigma \circ \pi = \phi$. Since π is a mapping, the relation $\pi^{-1} \circ \pi$ contains the identical mapping 1_G. Therefore $\sigma \circ \pi = \phi \circ \pi^{-1} \circ \pi \supseteq \phi \circ 1_G = \phi$, and thus $\sigma \circ \pi \supseteq \phi$. Since a mapping from G to H cannot be contained in another mapping from G to H, it follows that $\sigma \circ \pi = \phi$.

(c) It is to be proved that the mapping σ is homomorphic. For elements $z_1, z_2 \in Z$, there are elements $g_1, g_2 \in G$ with $z_1 = \pi(g_1)$, $z_2 = \pi(g_2)$. Since ϕ and π are homomorphic mappings and $\sigma \circ \pi = \phi$:

$$\begin{aligned}
\sigma(z_1 \circ z_2) &= \sigma(\pi(g_1) \circ \pi(g_2)) &&= \sigma(\pi(g_1 \circ g_2)) \\
&= (\sigma \circ \pi)(g_1 \circ g_2) &&= \phi(g_1 \circ g_2) \\
&= \phi(g_1) \circ \phi(g_2) &&= (\sigma \circ \pi)(g_1) \circ (\sigma \circ \pi)(g_2) \\
&= \sigma(\pi(g_1)) \circ \sigma(\pi(g_2)) &&= \sigma(z_1) \circ \sigma(z_2)
\end{aligned}$$

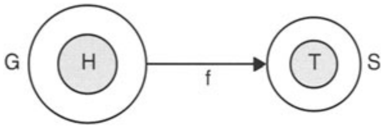
Since $\sigma(z_1 \circ z_2) = \sigma(z_1) \circ \sigma(z_2)$, the mapping σ is homomorphic.

(d) It is to be proved that for given mappings ϕ and π there is exactly one mapping σ with $\sigma \circ \pi = \phi$. For every element $z \in Z$ there is an element $g \in G$ with $\pi(g) = z$ and $h := \phi(g)$. For two mappings $\sigma_1 : Z \rightarrow H$ and $\sigma_2 : Z \rightarrow H$ with $\sigma_1 \circ \pi(g) = \phi(g)$ and $\sigma_2 \circ \pi(g) = \phi(g)$, this implies $\sigma_1(z) = h$ and $\sigma_2(z) = h$. Hence the mappings σ_1 and σ_2 are equal.

(e) If ϕ is surjective, then for every element $h \in H$ there is an element $g \in G$ with $\phi(g) = h$ and $z := \pi(g)$. Since $\sigma \circ \pi = \phi$, it follows that $\sigma \circ \pi(g) = \phi(g)$, and thus $\sigma(z) = h$. Hence the mapping σ is surjective : $\sigma(Z) = H$.

(f) It is to be proved that σ is injective under the assumption $\ker \pi = \ker \phi$. The homomorphism σ is injective if its kernel is trivial, that is if $\ker \sigma = \{1_Z\}$. For every element $z \in Z$ with $\sigma(z) = 1_H$ there is an element $g \in G$ with $\pi(g) = z$. Then $\phi(g) = \sigma \circ \pi(g) = \sigma(z) = 1_H$, and hence $g \in \ker \phi = \ker \pi$ and $\pi(g) = 1_Z$. Thus only the identity element 1_Z of Z is mapped to the identity element 1_H of H by σ, that is $\ker \sigma = \{1_Z\}$, and hence the mapping σ is injective.

Homomorphic mappings of subgroups : Let $(G\,;\circ)$ be a group with a subgroup H, and let $(S\,;\circ)$ be a group with a subgroup T. Let the mapping $f : G \rightarrow S$ be homomorphic. Then f(H) is a subgroup of S. The preimage $f^{-1}(T)$ is a subgroup of G.



Proof : Homomorphic mapping of subgroups

(1) If f(H) contains the elements b_1 and b_2, then there are elements $h_1, h_2 \in H$ with $b_1 = f(h_1)$ and $b_2 = f(h_2)$. Then H also contains the element $h_1 \circ h_2^{-1}$. Hence f(H) contains the element $f(h_1 \circ h_2^{-1}) = f(h_1) \circ f(h_2)^{-1} = b_1 \circ b_2^{-1}$. By property (E2) of subgroups in Section 7.2, f(H) is a group.

(2) If $f^{-1}(T)$ contains the elements c_1 and c_2, then T contains the elements $f(c_1)$ and $f(c_2)$. It follows that T contains the element $f(c_1) \circ f(c_2)^{-1} = f(c_1 \circ c_2^{-1})$, and hence $f^{-1}(T)$ contains the element $c_1 \circ c_2^{-1}$. By property (E2), $f^{-1}(T)$ is a group.

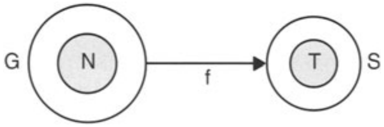
Homomorphic mappings of normal subgroups : Let $(G\,;\circ)$ be a group with a normal subgroup N, and let $(S\,;\circ)$ be a group with a normal subgroup T. The normal subgroups exhibit the following properties under a homomorphic mapping $f : G \rightarrow S$:

(1) The image f(N) is a normal subgroup in the image f(G).

(2) If the mapping f is surjective, then f(N) is a normal subgroup in S.

(3) The preimage $f^{-1}(T)$ is a normal subgroup in G.

(4) The kernel of f is a normal subgroup in G.



Proof : Homomorphic mapping of normal subgroups

(1) Let N be a normal subgroup of G. For arbitrary elements $x \in f(G)$ and $y \in f(N)$ there are elements $g \in G$ and $n \in N$ with $x = f(g)$ and $y = f(n)$. Since N is a normal subgroup of G, the product $g \circ n \circ g^{-1}$ is an element of N. Hence its image $f(g \circ n \circ g^{-1}) = f(g) \circ f(n) \circ f(g)^{-1} = x \circ y \circ x^{-1}$ is an element of f(N). Since y and $x \circ y \circ x^{-1}$ are elements of f(N), f(N) is a normal subgroup of f(G).

(2) If the homomorphism $f : G \rightarrow S$ is surjective, then by definition $f(G) = S$. It follows by (1) that $f(N)$ is a normal subgroup of S.

(3) Let T be a normal subgroup of S. For arbitrary elements $g \in G$ and $n \in f^{-1}(T)$, the image $f(g \circ n \circ g^{-1}) = f(g) \circ f(n) \circ f(g)^{-1}$ is an element of the normal subgroup T of S since $f(n)$ is an element of T. Hence $g \circ n \circ g^{-1}$ is an element of $f^{-1}(T)$. Since n and $g \circ n \circ g^{-1}$ are elements of $f^{-1}(T)$, $f^{-1}(T)$ is a normal subgroup of G.

(4) The kernel of f is the preimage of the normal subgroup $\{1_S\}$ of S. Hence (3) implies that the kernel is a normal subgroup in G.

Quotient set of a normal subgroup : The cosets of a normal subgroup N in a group $(G ; \circ)$ form a partition of the group. The class with the representative a is designated by [a]. In the quotient set G/N, an inner operation $\circ$ is defined by the rule $[a] \circ [b] = [a \circ b]$. This operation is well-defined since the result is independent of the choice of representatives.

$$[a] \circ [b] = [a \circ b] \quad \Rightarrow \quad (x \in [a] \quad \wedge \quad y \in [b] \quad \Rightarrow \quad x \circ y \in [a \circ b])$$

Proof : Inner operation in the quotient set of a normal subgroup

Let [a] and [b] be cosets of a normal subgroup N in the group $(G ; \circ)$. It is to be proved that for arbitrary elements $x \in [a]$ and $y \in [b]$ the product $x \circ y$ is an element of the class $[a \circ b]$. By definition, the normal subgroup N contains elements n_1 and n_2 such that $x = a \circ n_1$ and $y = b \circ n_2$. Hence :

$$x \circ y = a \circ n_1 \circ b \circ n_2 = a \circ b \circ (b^{-1} \circ n_1 \circ b) \circ n_2$$

Every transform $b^{-1} \circ n_1 \circ b$ of the element n_1 of the normal subgroup N is an element n_3 of N. The product $n_3 \circ n_2$ is an element n_4 of the group N. It follows from $x \circ y = a \circ b \circ n_4$ that $x \circ y$ is an element of the class $[a \circ b]$. Hence the operation is well-defined.

Natural homomorphism : The canonical mapping $k : G \rightarrow G/N$ from a group $(G ; \circ)$ to the quotient set G/N with respect to a normal subgroup N of G is homomorphic; it is called the natural homomorphism (the canonical projection) of G with respect to N. That the mapping k is homomorphic follows from the definition of the inner operation $\circ$ of the domain $(G/N ; \circ)$.

$$k : G \rightarrow G/N \quad \text{with} \quad k(a) = [a]$$

Quotient group with respect to a normal subgroup : Since the canonical mapping $k : G \rightarrow G/N$ from a group $(G ; \circ)$ to the quotient set G/N is homomorphic, the domain $(G/N ; \circ)$ is also a group. The group $(G/N ; \circ)$ is called the quotient group (factor group) of the group G with respect to the normal subgroup N.

Example 1 : Quotient set of a cyclic group

The cyclic group $A = \{1, a, a^2, a^3\}$ has the following quotient set :

normal subgroup : $N \quad = \{1, a^2\}$

cosets of N : $1 \circ N = N \circ 1 = a^2 \circ N = N \circ a^2 = \{1, a^2\} = [1_A]$
$a \circ N = N \circ a = a^3 \circ N = N \circ a^3 = \{a, a^3\} = [a]$

quotient set : $A/N = \{[1_A], [a]\}$

Example 2 : Quotient group of the symmetry group of the tetrahedron

The symmetry group of the regular tetrahedron in Example 2 of Section 7.3.2 has the following quotient group (see also Example 2 in Section 7.4.2) :

normal subgroup : $N \quad = \{a_0, a_9, a_{10}, a_{11}\}$

cosets of N : $[a_0] = \{a_0, a_9, a_{10}, a_{11}\}$
$[a_1] = \{a_1, a_4, a_5, a_8\}$
$[a_2] = \{a_2, a_3, a_6, a_7\}$

quotient set : $A/N = \{[a_0], [a_1], [a_2]\}$

The domain $(A/N ; \circ)$ is a group. For example, $[a_1] \circ [a_2] = [a_1 \circ a_2] = [a_0]$. This result is independent of the choice of representatives for the classes, as the following examples demonstrate :

$$
\begin{array}{lll}
a_1 \circ a_3 = a_{11} & \qquad & a_4 \circ a_2 = a_{11} \\
a_1 \circ a_6 = a_9 & & a_5 \circ a_2 = a_9 \\
a_1 \circ a_7 = a_{10} & & a_8 \circ a_2 = a_{10}
\end{array}
$$

The product of an element of $[a_1]$ with an element of $[a_2]$ is invariably an element of $[a_0]$.

7.5.3 ISOMORPHISM

Identical structures : Studying groups by comparing their subgroups requires a definition of the concept of ”groups with identical structure”. The existence of a homomorphic (structurally compatible) mapping between two groups does not imply that they possess the same structure. Two groups are identically structured (isomorphic) if there is a bijective mapping between the groups which is homomorphic in both directions.

In a homomorphic mapping $f : G \rightarrow H$, the number of elements in the group G is often far greater than the number of elements in the group H. It is therefore not fruitful to compare the groups using the mapping f. Instead one seeks a group isomorphic to H which can be derived from G. The first isomorphism theorem asserts that the quotient group G/N is isomorphic to H if the kernel of f is a normal subgroup N of G.

Many of the subsequent studies of group structures are based on the two remaining isomorphism theorems. The second isomorphism theorem applies to a group G with a subgroup H and a normal subgroup N : The groups $H/H \cap N$ and $H \circ N/N$ are isomorphic. The third isomorphism theorem applies to a group G with two normal subgroups H and N such that $N \subseteq H \subseteq G$: The groups G/H and $(G/N) / (H/N)$ are isomorphic.

Isomorphic mappings : Let a homomorphic mapping $f : G \rightarrow H$ from the group G to the group H be bijective. Then f is called an isomorphic mapping (an isomorphism). The groups G and H are said to be isomorphic (identically structured). The isomorphism of groups is designated by $G \cong H$ (G is isomorphic to H).

Finite isomorphic groups contain the same number of elements. However, finite groups with the same number of elements are not necessarily isomorphic (see Example 1). For a suitable choice of designations for the elements, the product tables of isomorphic groups are identical. Isomorphic groups are therefore often said to be identical.

Properties of isomorphic mappings :

(1) Let $(G ; \circ)$ and $(H ; \circ)$ be groups. Let the mapping $f : G \rightarrow H$ be an isomorphism. Then the inverse mapping $f^{-1} : H \rightarrow G$ is also an isomorphism. Thus $G \cong H$ implies $H \cong G$.

(2) Let the mappings $f_1 : G_1 \rightarrow G_2$ and $f_2 : G_2 \rightarrow G_3$ be isomorphisms. Then the composition $f_2 \circ f_1 : G_1 \rightarrow G_3$ is also an isomorphism. Thus from $G_1 \cong G_2$ and $G_2 \cong G_3$ it follows that $G_1 \cong G_3$.

Proof : Properties of isomorphic mappings

(1) For elements $h_1, h_2 \in H$, there are elements $g_1, g_2 \in G$ such that $h_1 = f(g_1)$, $h_2 = f(g_2)$ and $h_1 \circ h_2 = f(g_1 \circ g_2)$. For the inverse mapping f^{-1}, this implies :

$$f^{-1}(h_1 \circ h_2) = f^{-1} \circ f(g_1 \circ g_2) = g_1 \circ g_2 = f^{-1}(h_1) \circ f^{-1}(h_2)$$

(2) The composition of homomorphic mappings is a homomorphism. The composition of bijective mappings is bijective. Hence $f_2 \circ f_1$ is an isomorphism.

Example 1 : Non-isomorphic finite groups of the same order

It is difficult to determine the number of non-isomorphic groups of a given order. The following table shows the number of non-isomorphic groups of orders 1 to 15. The product tables for the non-isomorphic groups of order 4 are specified.

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

$\circ$	a_0	a_1	a_2	a_3
a_0	a_0	a_1	a_2	a_3
a_1	a_1	a_2	a_3	a_0
a_2	a_2	a_3	a_0	a_1
a_3	a_3	a_0	a_1	a_2

cyclic group

$\circ$	a_0	a_1	a_2	a_3
a_0	a_0	a_1	a_2	a_3
a_1	a_1	a_0	a_3	a_2
a_2	a_2	a_3	a_0	a_1
a_3	a_3	a_2	a_1	a_0

Klein's four-group

Kernel of a natural homomorphism : Let N be a normal subgroup in a group $(G ; \circ)$. Then $k : G \to G/N$ is a natural homomorphism. The subgroup N of G is exactly the set of elements which are mapped to the identity element $1_{G/N} = [1_G]$ of the quotient set G/N. Hence the normal subgroup N is the kernel of the natural homomorphism k.

$$n \in N \quad \Leftrightarrow \quad k(n) = 1_{G/N}$$

$$1_{G/N} = [1_G] = 1_G \circ N = N$$

N is a normal subgroup in G $\Leftrightarrow$ N is the kernel of $k : G \to G/N$.

First isomorphism theorem : Let the kernel of a surjective group homomorphism $f : G \to H$ be N. Then the quotient group G/N and the group H are isomorphic.

$$\begin{tikzcd}
G \arrow[r, "f"] \arrow[d, "k"'] & H = f(G) \cong G/N \\
G/N \arrow[ur, "i"'] &
\end{tikzcd}$$

homomorphic mapping	:	$f :$	$G \to H$	with	$\ker f$	$=$	N
canonical mapping	:	$k :$	$G \to G/N$	with	$k(a)$	$=$	$[a]$
isomorphic mapping	:	$i :$	$G/N \to H$	with	$i([a])$	$=$	$f(a)$
composition	:	$f = i \circ k$					

Proof : First isomorphism theorem

According to Section 7.5.2, the kernel of the homomorphism f is a normal subgroup in G. The natural homomorphism $k : G \to G/N$ is surjective. Hence f induces a homomorphism $i : G/N \to H$. Since f is surjective, i is also surjective. By hypothesis, $\ker k = \ker f = N$, so that i is injective (see Section 7.5.2). Since i is surjective and injective, i is an isomorphism. Hence the groups G/N and H are isomorphic.

Example 2 : Displacement of a body along the real axis

Let A be a body in a closed interval on the real axis $\mathbb{R}$. The linear mapping $s_{a,b} : A \to \mathbb{R}$ is a displacement of the body A along the axis $\mathbb{R}$. Let this displacement be given by the sum of a rigid motion b and a linear deformation ax of the body.

$$s_{a,b} : A \to \mathbb{R} \quad \text{with} \quad s_{a,b}(x) = ax + b \quad \text{and} \quad a \neq 0$$

The linear mappings $s_{a,b}$ with different parameters $a,b \in \mathbb{R}$ form a group with the identity element $s_{1,0}$. Applying the mapping $s_{c,d}$ after the mapping $s_{a,b}$ yields a mapping $s_{ac,bc+d}$. Every mapping $s_{a,b}$ has an inverse :

$$s_{c,d} \circ s_{a,b}(x) = c(ax+b)+d = (ac)x + (bc+d) = s_{ac,bc+d}(x)$$

$$s_{a,b}^{-1}(x) = \frac{1}{a}x - \frac{b}{a}$$

$$s_{a,b}^{-1} \circ s_{a,b}(x) = \frac{1}{a}(ax+b) - \frac{b}{a} = x = s_{1,0}(x)$$

The general relationship between displacement, rigid motion and deformation is determined by studying the structure of the group G of linear mappings. The mapping f from the displacement group G to the deformation group H is homomorphic:

$$f : G \rightarrow H \quad \text{with} \quad f(s_{a,b}) = s_{a,0}$$

$$f(s_{c,d} \circ s_{a,b}) = f(s_{ac,bc+d}) = s_{ac,0} = s_{c,0} \circ s_{a,0} = f(s_{c,d}) \circ f(s_{a,b})$$

The identity element 1_H of the deformation group is $f(s_{1,0}) = s_{1,0}$. The kernel N of the homomorphic mapping f contains the rigid motions. The image $H = f(G)$ contains the deformations.

$$N = \ker(f) = \{s_{1,b} \mid b \in \mathbb{R}\}$$

$$H = f(G) = \{s_{a,0} \mid a \in \mathbb{R} \wedge a \neq 0\}$$

A class $[s_{a,b}]$ in the factor group of the natural homomorphism $k : G \rightarrow G/N$ contains all displacements with the same expansion factor a :

$$[s_{a,b}] = N \circ s_{a,b} = \{s_{1,d} \mid d \in \mathbb{R}\} \circ s_{a,b} = \{s_{a,e} \mid e \in \mathbb{R}\}$$

The isomorphism $i : G/N \rightarrow H$ maps the displacement class with the expansion factor a and different motions e to a displacement with the same expansion factor a and the motion $e = 0$:

$$i([s_{a,b}]) = f(s_{a,b}) = s_{a,0}$$

$$\begin{tikzcd}
\boxed{\begin{array}{l}\text{displacement group G :}\\ \{s_{a,b} \mid a, b \in \mathbb{R} \wedge a \neq 0\}\end{array}} \arrow[r, "f"] \arrow[d, "k"'] & \boxed{\begin{array}{l}\text{deformation group H :}\\ \{s_{a,0} \mid a \in \mathbb{R} \wedge a \neq 0\}\end{array}} \\
\boxed{\begin{array}{l}\text{displacement classes :}\\ [s_{a,b}] = \{s_{a,e} \mid e \in \mathbb{R}\}\end{array}} \arrow[ur, "i"'] &
\end{tikzcd}$$

Second isomorphism theorem : Let $(G\ ;\circ)$ be a group, $H \subseteq G$ a subgroup and $N \subseteq G$ a normal subgroup. Then :

(1) $H \circ N$ is a subgroup of G with normal subgroup N.

(2) $H \cap N$ is a normal subgroup of H.

(3) The groups $H/H \cap N$ and $H \circ N/N$ are isomorphic.

Proof : Second isomorphism theorem

(1) Since N is a normal subgroup of G, $g \circ N = N \circ g$ holds for every element $g \in G$. For every element h of the subgroup H of G, $h \circ N = N \circ h$, and hence also $H \circ N = N \circ H$. Since the factors may be interchanged, $H \circ N$ is a subgroup of G by property (P2) of the products of subgroups (see Section 7.2).

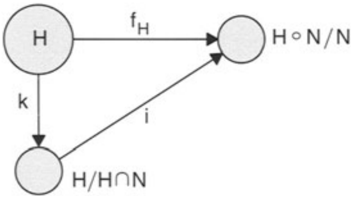
An arbitrary element $h \circ n \in H \circ N$ is an element of G. For the normal subgroup N in G, this implies $h \circ n \circ N = N \circ h \circ n$. Hence N is a normal subgroup in $H \circ N$.

(2) Since H is a subgroup and N is a normal subgroup of G, property (T3) of normal subgroups implies that $H \cap N$ is a normal subgroup of H.

(3) For the group $H \circ N$ with the normal subgroup N, there is a natural homomorphism $f : H \circ N \rightarrow H \circ N/N$. The images of the elements $h \circ n \in H \circ N$ and $h \in H$ are equal, since the normal subgroup N is mapped to the unit element of $H \circ N/N$:

$$f(h \circ n) \;=\; f(h) \circ f(n) \;=\; f(h)$$

Hence the restriction $f_H : H \rightarrow H \circ N/N$ is a surjective homomorphism with kernel $N \cap H$. For the group H with the normal subgroup $H \cap N$, there is a natural homomorphism $k : H \rightarrow H/H \cap N$ with the kernel $H \cap N$. Since the surjective homomorphisms f_H and k have the same kernel $H \cap N$, the induced homomorphism $i : H/H \cap N \rightarrow H \circ N/N$ is an isomorphism (see Section 7.5.2).



Third isomorphism theorem : Let $(G ; \circ)$ be a group, and let N, H be normal subgroups in G with $N \subseteq H \subseteq G$. Then :

(1) N is a normal subgroup of H.

(2) H/N is a normal subgroup of G/N.

(3) The groups (G/N) / (H/N) and G/H are isomorphic.

Proof : Third isomorphism theorem

(1) Since N is a normal subgroup of G, by definition $g \circ N = N \circ g$ for every element $g \in G$, in particular for every $g \in H$. Hence N is a normal subgroup of H.

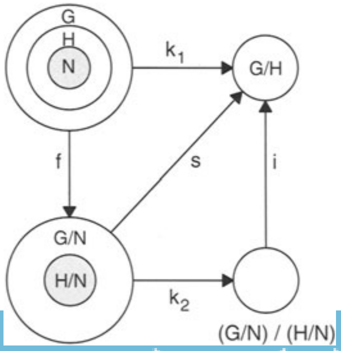
(2) Every element of the quotient group H/N is a coset $h \circ N$. Since $h \in G$, the coset $h \circ N$ is also an element of G / N. Every element of H / N is therefore an element of G / N.

The natural homomorphism $f : G \rightarrow G/N$ maps the group G to G/N and the normal subgroup $H \triangleleft G$ to $H/N \subseteq G/N$. According to Section 7.5.2 (homomorphic mappings of normal subgroups), H/N is a normal subgroup in G/N.

(3) The natural homomorphism $k_1 : G \rightarrow G/H$ is surjective, and its kernel contains the normal subgroup N, that is $N \subseteq \ker k_1$. The natural homomorphism $f : G \rightarrow G/N$ is surjective with kernel N. According to Section 7.5.2, the homomorphism k_1 induces a surjective homomorphism $s : G/N \rightarrow G/H$. The kernel of s is the normal subgroup H/N of G/N, since k_1 maps H to the identity element of G/H :

$$s^{-1}(1_{G/H}) \;=\; (f \circ k_1^{-1})(1_{G/H}) \;=\; f(k^{-1}(1_{G/H})) \;=\; f(N) \;=\; H/N$$

By the first isomorphism theorem, the surjective homomorphism s induces an isomorphism $i : (G/N)/(H/N) \rightarrow G/H$. The groups G /H and (G/ N) / (H / N) are therefore isomorphic.



$k_1 : \; G \rightarrow G/H$

$f \;\; : \; G \rightarrow G/N$

$s \;\; : \; G/N \rightarrow G/H$

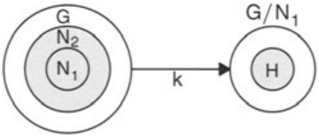
$k_2 : \; G/N \rightarrow (G/N) / (H/N)$

$i \;\; : \; (G/N)/(H/N) \rightarrow G/H$

Extended third isomorphism theorem : Let (G ; ∘) be a group, let N_1 be a normal subgroup of G, and let H be a normal subgroup of the quotient group G/N_1. Then for the natural homomorphism $k : G \to G/N_1$:

(1) The preimage $N_2 := k^{-1}(H)$ is a normal subgroup of G and contains N_1.

(2) H is the quotient group N_2/N_1.



Proof : Extended third isomorphism theorem

The natural homomorphism $k : G \to G/N_1$ is surjective. In Section 7.5.2, the preimage of the normal subgroup H of G/N_1 is shown to be a normal subgroup $N_2 := k^{-1}(H)$ of G. The group N_2 contains the group N_1, since the homomorphism k maps the group N_1 to the unit element 1_H in H. It follows from $N_1 \subset N_2 \subset G$ and $N_2 = k^{-1}(H)$ that $H = N_2/N_1$.

7.5.4 ISOMORPHIC TYPES OF GROUPS

Introduction : In this section, some frequently used isomorphisms are compiled and proved using the tools of the preceding section :

(1) Every infinite cyclic group is isomorphic to the group ($\mathbb{Z}$; +) of the integers.

(2) Every finite cyclic group of order n is isomorphic to the group $\mathbb{Z}_n$ of residue classes.

(3) Every finite group G is isomorphic to a group of permutations of the group G.

Isomorphic cyclic groups : Up to isomorphism, the additive group ($\mathbb{Z}$; +) of the integers is the only infinite cyclic group. Up to isomorphism, the group $\mathbb{Z}_n$ of residue classes in Section 7.4.3 is the only finite cyclic group of order n.

Proof : Isomorphic cyclic groups

Let a be the generating element of a cyclic group (G ; +). The mapping f from the integers $\mathbb{Z}$ to the group G is a surjective homomorphism :

$$f : \mathbb{Z} \rightarrow G \quad \text{with} \quad f(z) = za$$

$$f(z_1 + z_2) = (z_1 + z_2)\,a = z_1 a + z_2 a = f(z_1) + f(z_2)$$

If the group G is infinite, then the kernel of the mapping f is the set {0}. If the group G is finite of order n, then the kernel of f is the subgroup $\mathbb{Z}n$. By the first isomorphism theorem, the group G and the quotient group $\mathbb{Z}$ / ker f are isomorphic.

$$\begin{array}{llllllll} \text{G infinite} & : & \mathbb{Z}/\ker f & = & \mathbb{Z}/\{0\} & = & \mathbb{Z} & \Rightarrow \quad G \cong \mathbb{Z} \\ \text{G finite} & : & \mathbb{Z}/\ker f & = & \mathbb{Z}/\mathbb{Z}n & = & \mathbb{Z}_n & \Rightarrow \quad G \cong \mathbb{Z}_n \end{array}$$

Isomorphic permutation groups : Let a mapping $\phi : A \rightarrow B$ from a finite non-empty set A to a non-empty set B be bijective. Then there is an isomorphic mapping $f : S(B) \rightarrow S(A)$ between the complete permutation groups S(A) and S(B). Hence the group of all permutations on a set of n elements is unique up to isomorphism. This permutation group is called the symmetric group for a set of n elements and is designated by S_n.

Proof : Isomorphic permutation groups

(1) Every element ω of the permutation group S(B) is a permutation of B, that is $\omega : B \rightarrow B$. The image of the permutation ω under the mapping f is defined to be the ϕ-transform of ω :

$$f : S(B) \rightarrow S(A) \quad \text{with} \quad f(\omega) = \phi^{-1} \circ \omega \circ \phi \qquad a_1, a_2 \in A$$

$$\phi^{-1} \circ \omega \circ \phi(a_1) = \phi^{-1} \circ \omega(b_1) = \phi^{-1}(b_2) = a_2 \qquad b_1, b_2 \in B$$

(2) The mapping f is a homomorphism, since for permutations $\omega_1 \in S(B)$ and $\omega_2 \in S(B)$ with the images $\phi^{-1} \circ \omega_1 \circ \phi$ and $\phi^{-1} \circ \omega_2 \circ \phi$ in S(A) one has :

$$f(\omega_2 \circ \omega_1) = \phi^{-1} \circ (\omega_2 \circ \omega_1) \circ \phi = (\phi^{-1} \circ \omega_2 \circ \phi) \circ (\phi^{-1} \circ \omega_1 \circ \phi)$$
$$f(\omega_2 \circ \omega_1) = f(\omega_2) \circ f(\omega_1)$$

(3) The mapping f is bijective. If $f^{-1} : S(A) \rightarrow S(B)$ is defined by $f^{-1}(\gamma) = \phi \circ \gamma \circ \phi^{-1}$, then the compositions $f \circ f^{-1}$ and $f^{-1} \circ f$ satisfy :

$$f \circ f^{-1}(\gamma) = \phi^{-1} \circ (\phi \circ \gamma \circ \phi^{-1}) \circ \phi = \gamma \qquad \gamma \in S(A)$$
$$f^{-1} \circ f(\omega) = \phi \circ (\phi^{-1} \circ \omega \circ \phi) \circ \phi^{-1} = \omega \qquad \omega \in S(B)$$

(4) Since the mapping f is bijective and homomorphic, the permutation groups S(A) and S(B) are isomorphic.

Left translation of a group : Let every element of a group (G ; ∘) be multiplied from the left by the element $g \in G$. Since the equation $g \circ a = b$ has a unique solution $a = g^{-1} \circ b \in G$ for every $b \in G$, the mapping $\phi_g : G \rightarrow G$ with $\phi_g(a) = g \circ a$ is bijective. The mapping ϕ_g is called the left translation of the group G defined by g. For $g \neq 1$, the left translation is not homomorphic !

$$\phi_g : G \rightarrow G \quad \text{with} \quad \phi_g(a) = g \circ a$$
$$\phi_g(a \circ b) = g \circ a \circ b = \phi_g(a) \circ b$$
$$\phi_g(a \circ b) \neq \phi_g(a) \circ \phi_g(b) \quad \Rightarrow \quad \phi_g \text{ is not homomorphic}$$

Cayley's Theorem : Every finite group (G ; ∘) is isomorphic to a group of permutations of the set G. Thus, there is an injective homomorphism $f : G \rightarrow S(G)$. The image of the element g of G is the left translation ϕ_g. The image of the group G is a subgroup of the symmetric group S(G).

$$f : G \rightarrow S(G) \quad \text{with} \quad f(g) = \phi_g$$

Proof : Cayley's Theorem

(1) The relation f defined above is a mapping. Equal elements $g = h$ of G have equal images ϕ_g and ϕ_h in S(G). For $a \in G$:

$$\phi_g(a) = g \circ a = h \circ a = \phi_h(a) \quad \Rightarrow \quad \phi_g = \phi_h$$

(2) The mapping f is injective. Equal elements $\phi_g = \phi_h$ in f(G) have equal preimages g and h in G, since for $a \in G$, the element a^{-1} is also in G :

$$\phi_g = \phi_h \quad \Rightarrow \quad g \circ a = h \circ a \quad \Rightarrow \quad g \circ a \circ a^{-1} = h \circ a \circ a^{-1} \quad \Rightarrow \quad g = h$$

(3) The mapping f is homomorphic, since for $g, h \in G$:

$$\phi_{g \circ h}(a) = g \circ h \circ a = g \circ \phi_h(a) = \phi_g(\phi_h(a)) = \phi_g \circ \phi_h(a)$$
$$f(g \circ h) = \phi_{g \circ h} = \phi_g \circ \phi_h$$

(4) Since the homomorphism f is injective, the groups G and $f(G)$ are isomorphic.

Right translation : If in defining a permutation of a group $(G\,;\circ)$ the left translation $\phi_g(a) = g \circ a$ is replaced by the right translation $\sigma_g(a) = a \circ g$, the mapping $k : G \to S(G)$ with $k(g) = \sigma_g$ is generally not a homomorphism, since for $g, h \in G$:

$$\sigma_{g \circ h}(a) = a \circ g \circ h = \sigma_g(a) \circ h = \sigma_h(\sigma_g(a)) = \sigma_h \circ \sigma_g$$
$$k(g \circ a) = \sigma_{g \circ h} = \sigma_h \circ \sigma_g \neq \sigma_g \circ \sigma_h$$

For the definition of the composition $\sigma_g \circ \sigma_h$ used here (σ_g after σ_h), the mapping which maps the elements of a group to right translations is therefore homomorphic only if the group is commutative (abelian). For general groups, mappings to left and right translations must be distinguished.

Example : Cayley's Theorem

A group $G = \{a_1, a_2, a_3\}$ with the identity element a_1 has the following product table :

$\circ$	a_1	a_2	a_3
a_1	a_1	a_2	a_3
a_2	a_2	a_3	a_1
a_3	a_3	a_1	a_2

The permutation group $S = \{\sigma_1, \sigma_2, \sigma_3\}$ isomorphic to G contains the following permutations. The mapping $f : G \to S$ with $f(a_i) = \sigma_i$ is bijective and homomorphic. The group S is a subgroup of the permutation group S_3 .

$$\sigma_1 = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_1 \circ a_1 & a_1 \circ a_2 & a_1 \circ a_3 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_1 & a_2 & a_3 \end{pmatrix}$$

$$\sigma_2 = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_2 \circ a_1 & a_2 \circ a_2 & a_2 \circ a_3 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_2 & a_3 & a_1 \end{pmatrix}$$

$$\sigma_3 = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_3 \circ a_1 & a_3 \circ a_2 & a_3 \circ a_3 \end{pmatrix} = \begin{pmatrix} a_1 & a_2 & a_3 \\ a_3 & a_1 & a_2 \end{pmatrix}$$

7.5.5 AUTOMORPHISMS

Introduction : An isomorphic mapping from a group to itself is called an automorphism. Thus, an automorphism is a homomorphic permutation. The composition of automorphisms on a group G leads to the automorphism group of G. This group is a subgroup of the permutation group of G.

The mapping of the elements of a group G to their g-transforms is an automorphism on G and is called an inner automorphism. The inner automorphisms of G with composition as the inner operation form a subgroup of the automorphism group. The mapping from a group to its inner automorphisms is a group homomorphism. Not every automorphism is an inner automorphism.

The action of automorphic mappings on subgroups is important for the structure of a group. Subgroups which are mapped to themselves under all automorphisms are said to be characteristic; they are normal subgroups of the group. Only characteristic subgroups of a normal subgroup of a group are also normal subgroups of the group itself : This is not true for other normal subgroups of a normal subgroup.

Automorphic mappings : An isomorphic mapping $f : G \to G$ from a group $(G\ ;\circ)$ to itself is said to be automorphic (an automorphism of G). The set of automorphisms of G is designated by aut G.

$$\text{aut } G \ := \{ f : G \to G \mid f \text{ is isomorphic} \}$$

Properties of automorphic mappings

(A1) The composition $f_1 \circ f_2$ of automorphisms f_1 and f_2 is defined as an inner operation in the set aut G. The domain (aut G ; ∘) is a group; it is called the automorphism group of G.

(A2) If the sets G and H are isomorphic, then their automorphism groups are also isomorphic.

$$G \cong H \quad \Rightarrow \quad \text{aut } G \cong \text{aut } H$$

Proof A1 : The domain (aut G ; ∘) is a group.

The domain (aut G ; ∘) has the properties of a group :

(1) The identity element is the identity mapping i :

$$i : G \to G \quad \text{with} \quad i(g) = g$$

(2) Together with the automorphisms f_1 and f_2, the set aut G also contains their composition $f_1 \circ f_2$. In fact, every automorphism is an isomorphism. By property (2) of isomorphisms in Section 7.5.3, the composition $f_1 \circ f_2 : G \to G$ is an isomorphic mapping from G to itself, and hence by definition an element of aut G.

(3) If $f : G \to G$ is an automorphism, then by property (1) of isomorphisms the inverse $f^{-1} : G \to G$ is also an automorphism, and hence an element of aut G.

Proof A2 : The automorphism groups of isomorphic sets are isomorphic.

(1) Let the mapping $\phi : G \to H$ be an isomorphism with $\phi(g) = h$. Every element of aut H is an isomorphism $\omega : H \to H$. Let the image of ω under the mapping $f : \text{aut } H \to \text{aut } G$ be the ϕ-transform of ω :

$$f : \text{aut } H \to \text{aut } G \quad \text{with} \quad f(\omega) = \phi^{-1} \circ \omega \circ \phi \qquad g_1, g_2 \in G$$
$$\phi^{-1} \circ \omega \circ \phi(g_1) = \phi^{-1} \circ \omega(h_1) = \phi^{-1}(h_2) = g_2 \qquad h_1, h_2 \in H$$

(2) The mapping f is a homomorphism, since for the automorphisms $\omega_1, \omega_2 \in$ aut H with the images $\phi^{-1} \circ \omega_i \circ \phi$ in aut G :

$$f(\omega_2 \circ \omega_1) = \phi^{-1} \circ (\omega_2 \circ \omega_1) \circ \phi = (\phi^{-1} \circ \omega_2 \circ \phi) \circ (\phi^{-1} \circ \omega_1 \circ \phi)$$
$$f(\omega_2 \circ \omega_1) = f(\omega_2) \circ f(\omega_1)$$

(3) The mapping f is bijective. Using $f^{-1} : \text{aut } G \to \text{aut } H$ with $f^{-1}(\gamma) = \phi \circ \gamma \circ \phi^{-1}$ for $\gamma \in$ aut G one obtains :

$$f \circ f^{-1}(\gamma) = \phi^{-1} \circ (\phi \circ \gamma \circ \phi^{-1}) \circ \phi = \gamma \qquad \gamma \in \text{aut } G$$
$$f^{-1} \circ f(\omega) = \phi \circ (\phi^{-1} \circ \omega \circ \phi) \circ \phi^{-1} = \omega \qquad \omega \in \text{aut } H$$

(4) Since the mapping f is bijective and homomorphic, the automorphism groups aut G and aut H are isomorphic.

Inner automorphisms : For every element g of a group $(G ; \circ)$, the relation $\sigma_g : G \to G$ with $\sigma_g(a) = g \circ a \circ g^{-1}$ is an automorphism. It is called the inner automorphism of G with respect to g.

$$\sigma_g : G \to G \quad \text{with} \quad \sigma_g(a) = g \circ a \circ g^{-1} \quad \wedge \quad a, g \in G$$

Proof : The relation σ_g is an automorphism.

(1) The relation σ_g is a mapping. Every element a of G has an image $\sigma_g(a)$. Equal elements $a = b$ of G have the same image :

$$a = b \quad \Rightarrow \quad g \circ a \circ g^{-1} = g \circ b \circ g^{-1} \quad \Rightarrow \quad \sigma_g(a) = \sigma_g(b)$$

(2) The mapping σ_g is homomorphic, since for elements $a, b \in G$:

$$\sigma_g(a \circ b) = g \circ a \circ b \circ g^{-1} = (g \circ a \circ g^{-1}) \circ (g \circ b \circ g^{-1}) = \sigma_g(a) \circ \sigma_g(b)$$

(3) The kernel of the mapping σ_g is trivial, since $\sigma_g(a) = 1_G$ implies $g \circ a \circ g^{-1} = 1_G \Rightarrow a = g^{-1} \circ g = 1_G$. Hence the mapping σ_g is injective. Every element b of the image has a preimage $a = g^{-1} \circ b \circ g$. Thus σ_g is injective and surjective, and hence bijective. Since σ_g is bijective and homomorphic, σ_g is an automorphism.

Properties of inner automorphisms

(I1) The set of inner automorphisms of G with the composition $\sigma_{g_1} \circ \sigma_{g_2}$ as the inner operation on the elements σ_{g_1} and σ_{g_2} is a group. It is designated by int G.

(I2) The mapping from a group G to the group of its inner automorphisms is a homomorphism :

$$f : G \rightarrow \text{int } G \quad \text{with} \quad f(g) = \sigma_g$$

Proof : Properties of inner automorphisms

(I1) The domain (int G ; $\circ$) has the properties of a group :

- The identity element is the identity mapping σ_{1_G} :

  $$\sigma_{1_G}(a) = 1_G \circ a \circ 1_G^{-1} = a$$

- Together with σ_{g_1} and σ_{g_2}, the set int G also contains their composition $\sigma_{g_3} = \sigma_{g_1} \circ \sigma_{g_2}$, since together with g_1 and g_2 the group G also contains $g_3 = g_1 \circ g_2$:

  $$\sigma_{g_1} \circ \sigma_{g_2}(a) = g_1 \circ (g_2 \circ a \circ g_2^{-1}) \circ g_1^{-1} = g_3 \circ a \circ g_3^{-1} = \sigma_{g_3}(a)$$

- Together with σ_g, the set int G also contains the inverse $\sigma_g^{-1} = \sigma_{g^{-1}}$:

  $$\sigma_g \circ \sigma_{g^{-1}}(a) = g \circ (g^{-1} \circ a \circ g) \circ g^{-1} = 1_G \circ a \circ 1_G^{-1} = \sigma_{1_G}(a)$$

(I2) For elements $a, g, h \in G$ and the inner automorphism $\sigma_{g \circ h}$:

$$\sigma_{g \circ h}(a) = g \circ h \circ a \circ (g \circ h)^{-1} = g \circ (h \circ a \circ h^{-1}) \circ g^{-1}$$
$$\sigma_{g \circ h}(a) = g \circ \sigma_h(a) \circ g^{-1} = \sigma_g \circ \sigma_h(a)$$

For $f : G \rightarrow \text{int } G$ with $f(g) = \sigma_g$ it follows that $f(g \circ h) = f(g) \circ f(h)$.

Outer automorphisms : The group of inner automorphisms of an abelian group contains only the identity mapping. However, there are other automorphisms of abelian groups (see Example 2). An automorphism of G which is not an inner automorphism is called an outer automorphism of G.

Example 1 : Inner automorphisms of the tetrahedral symmetry group

The inner automorphisms of the symmetry group of a regular tetrahedron are shown in Example 1 of Section 7.4.4. Row k of the transformation table contains the images under the inner automorphism $f_k : G \rightarrow G$ with $f_k(a_m) = a_k^{-1} \circ a_m \circ a_k$. The entry in row k and column m is the image $a_k^{-1} \circ a_m \circ a_k$ of the element a_m. The subgroup $\{a_0, a_9, a_{10}, a_{11}\}$ is a normal subgroup of the symmetry group (see Example 2 in Section 7.4.2). Each of the 12 inner automorphisms of G maps every element of this subgroup to an element of the subgroup.

Example 2 : Outer automorphism of a cyclic group

The following table contains the products of a cyclic group of order 4. Since the group is abelian, the group of inner automorphisms consists only of the identity mapping : $a^{-1} \circ b \circ a = a^{-1} \circ a \circ b = b$ for all $a, b \in G$. The mapping $f : G \rightarrow G$ with $f(a) = a^3$ is a homomorphism, since f is bijective and $f(a \circ b) = (a \circ b)^3 = a^3 \circ b^3 = f(a) \circ f(b)$. The mapping f is an outer automorphism of G.

$\circ$	a_0	a_1	a_2	a_3
a_0	a_0	a_1	a_2	a_3
a_1	a_1	a_2	a_3	a_0
a_2	a_2	a_3	a_0	a_1
a_3	a_3	a_0	a_1	a_2

$f(a_0) = a_0$

$f(a_1) = a_3$

$f(a_2) = a_2$

$f(a_3) = a_1$

Characteristic subgroups : Let G be a group. A subgroup H of G is said to be characteristic in G if for every automorphism $\phi : G \rightarrow G$ the image of H is contained in H.

$$\text{H is characteristic in G} \quad :\Leftrightarrow \quad \bigwedge_{\phi \in \text{aut } G} (\phi(H) \subseteq H)$$

If the group G is finite, it follows that $\phi(H) = H$, since the mapping ϕ is isomorphic. If the group G is infinite, then $\phi(H) \subset H$ would imply that the condition $\phi^{-1}(H) \subseteq H$ for the inverse automorphism ϕ^{-1} is not satisfied. Hence the condition $\phi(H) = H$ applies to all groups :

$$\text{H is characteristic in G} \quad \Leftrightarrow \quad \bigwedge_{\phi \in \text{aut } G} (\phi(H) = H)$$

Properties of characteristic subgroups

(C1) Every characteristic subgroup of a group G is a normal subgroup of G.

(C2) Let N be a normal subgroup of the group G, and let H be a characteristic subgroup in N. Then H is a normal subgroup of G.

Proof : Properties of characteristic subgroups

(C1) Let a subgroup $H \subseteq G$ be characteristic in G. Then $\phi(H) = H$ holds for every automorphism ϕ on G, in particular for every inner automorphism on G. For every element $g \in G$ and $h_1 \in H$, therefore, $\sigma_g(h_1) = g \circ h_1 \circ g^{-1} = h_2 \in H$, so that $g \circ h_1 = h_2 \circ g$. Hence H is a normal subgroup of G with $g \circ H = H \circ g$.

(C2) Since N is a normal subgroup of G, $g \circ N = N \circ g$ for every $g \in G$. Hence there is an automorphism $\phi : N \rightarrow N$ with $\phi(h) = g^{-1} \circ n \circ g$ for every $g \in G$. Since the subgroup H is characteristic in N, the image $\phi(h)$ of every element $h \in H$ under the automorphism ϕ is an element of H : $g^{-1} \circ h_1 \circ g \circ h_2$ with $h_2 \in H$ for all $h_1 \in H, g \in G$. Thus $g \circ H = H \circ g$ for all $g \in G$. Hence H is a normal subgroup of G.

Notes :

(1) By property (C1), every characteristic subgroup of a group G is a normal subgroup of G. There are, however, also normal subgroups of a group which are not characteristic subgroups of this group (see Example 3).

(2) By property (C2), every characteristic subgroup H of a normal subgroup N of a group G is a normal subgroup of N and of G. There are, however, also normal subgroups H of a normal subgroup N of a group G which are not normal subgroups of G (see Example 3).

(3) If a subgroup H is characteristic in a normal subgroup N of G, then H is not necessarily a characteristic subgroup of G.

Properties of normal subgroups : By the definition in Section 7.4.2, a subgroup H of a group $(G ; \circ)$ is a normal subgroup of G if $g \circ H = H \circ g$ for every $g \in G$. Normal subgroups have the following properties :

(N1) A subgroup H of a group $(G ; \circ)$ is a normal subgroup of G if and only if H is mapped to itself under every inner automorphism of G.

(N2) Let N be a normal subgroup of a group $(G ; \circ)$, and let $\phi : G \rightarrow G$ be an inner automorphism of G. Then the automorphism may be restricted to $\phi_N : N \rightarrow N$.

Proof : Properties of normal subgroups

(N1) Let H be a subgroup of G which is mapped to itself under every inner automorphism σ_g. Then for every element $g \in G$ and every element $h_1 \in H$ there is an element $h_2 \in H$ such that $g \circ h_1 \circ g^{-1} = h_2$ and hence $g \circ h_1 = h_2 \circ g$. Thus $g \circ H \subseteq H \circ g$. Analogously, $g^{-1} \circ h_1 \circ g = h_3$ implies $H \circ g \subseteq g \circ H$. Together, the two inclusions imply $g \circ H = H \circ g$ for every $g \in G$: The subgroup H is a normal subgroup of G.

Let H be a normal subgroup in G. Then for every element $g \in G$ and every element $h_1 \in H$ there is an element $h_2 \in H$ with $g \circ h_1 = h_2 \circ g$. Together with g the group G also contains g^{-1}, and hence $h_2 = g \circ h_1 \circ g^{-1}$. Thus every inner automorphism σ_g of G maps the group H to itself.

(N2) Together with ϕ, the group of inner automorphisms of G also contains the inverse ϕ^{-1}. Hence $\phi \circ \phi^{-1}(N) = N = \phi^{-1} \circ \phi\,(N)$. Since N is a normal subgroup, also $\phi\,(N) \subseteq N$ and $\phi^{-1}(N) \subseteq N$. These conditions are simultaneously satisfied only if $\phi\,(N) = \phi^{-1}(N) = N$. Hence the automorphism ϕ may be restricted to N.

Example 3 : Subgroups of the symmetric group S_4

The subgroups and normal subgroups of the symmetric group S_4 are determined in Sections 7.7.7 and 7.7.8. In this example, a subset of the subgroups of S_4 is considered :

alternating group : $A_4 = \{a_1,a_3,a_6,a_9,a_{11},a_{12},a_{13},a_{14},a_{15},a_{16},a_{17},a_{18}\}$
Klein's four-group : $V_4 = \{a_1,a_3,a_6,a_9\}$
cyclic subgroup : $Z_1 = \{a_1,a_3\}$
trivial subgroup : $\mathrm{I} = \{a_1\}$

These groups form a chain $S_4 \supset A_4 \supset V_4 \supset Z_1 \supset \mathrm{I}$, in which each of the subgroups is a normal subgroup of its predecessor, that is $\mathrm{I} \triangleleft Z_1$, $Z_1 \triangleleft V_4$, $V_4 \triangleleft A_4$, $A_4 \triangleleft S_4$. The four-group V_4 is also a normal subgroup of its second predecessor S_4. This is due to the fact that A_4 is a normal subgroup of S_4 and V_4 is a characteristic subgroup in A_4.

The cyclic group Z_1 is not a normal subgroup of its second predecessor A_4. The four-group V_4 is a normal subgroup in A_4 and Z_1 is a normal subgroup in V_4, but Z_1 is not a characteristic subgroup of V_4. For example, for $a_3 \in Z_1$ and $a_{11} \in A_4$, the transform $a_{11} \circ a_3 \circ a_{11}^{-1} = a_6$ is not an element of Z_1.

7.6 ABELIAN GROUPS

7.6.1 INTRODUCTION

The inner operation of an abelian group is commutative. The order in which the operation is applied in expressions is therefore arbitrary. This property leads to essential simplifications in the structure of abelian groups compared to the structure of non-commutative groups. In particular, it turns out that every finitely generated abelian group may be represented as a direct sum of cyclic groups, albeit not necessarily uniquely. This property of abelian groups is widely applied, for instance in algebraic topology.

In analogy with real vector spaces, the concept of a linear combination of elements is introduced for abelian groups. In contrast to the case of real vector spaces, however, the elements of an abelian group can be uniquely represented as linear combinations of the elements of a basis only if every element of the group generates a subgroup of infinite order. A group with this property is called a free abelian group. The concept of a direct sum of subgroups of an abelian group is introduced in order to study the structure of abelian groups which contain elements of finite order.

The study of homomorphic mappings of direct sums shows that every finitely generated abelian group H is the image of a free abelian group A. The linear combinations of the basis of A whose image is the identity element in H form a normal subgroup N in A. The quotient group A/N is isomorphic to H. This isomorphism leads to a decomposition of H into a direct sum of cyclic groups (fundamental theorem for abelian groups).

The decomposition of an abelian group into cyclic groups is generally not unique. The decomposition considered in the fundamental theorem for abelian groups leads to cyclic groups whose orders are infinite or form a divisor chain. However, an abelian group may also be decomposed into cyclic groups whose orders are infinite or powers of primes. This decomposition is treated in Section 7.9. It is unique up to isomorphism and the order of the summands.

7.6.2 CLASSIFICATION OF ABELIAN GROUPS

Introduction : Due to the commutative property of an abelian group (A ; +), every subgroup H of A is a normal subgroup in A. The subgroup H may be used to partition the abelian group into cosets. The subgroup X of elements of finite order in A is a special normal subgroup. It is called the torsion group of A. The quotient group A/X is a group of infinite degree. This classification of A forms the basis for the analysis of the structure of abelian groups in Section 7.9.

Abelian groups : A group is said to be abelian if its inner operation is commutative. The sum notation (A ; +) is generally used for abelian groups. The identity element is designated by 0. The inverse of an element a of the group A is designated by –a, the n-fold sum of the element a by na. For n = 0, let na = 0.

$$\bigwedge_{a\in A}\ \bigwedge_{b\in A}\ (a+b\ =\ b+a)$$

Every cyclic group is abelian. Cyclic groups are treated in Sections 7.3.4 to 7.3.6. However, there are also abelian groups which are not cyclic.

Order of an element : An element a of a group (A ; +) is called an element of order n if it generates a subgroup gp(a) of order n (see Section 7.3.6). If the order of the element a is finite, then n is the least positive integer for which na = 0 holds. If the order of the element a is infinite, then na is non-zero for all positive integers. The identity element 0 is an element of finite order.

$$\text{a is an element of finite order} \quad :\Leftrightarrow \quad \bigvee_{k\in\mathbb{N}'} (ka = 0)$$

$$\text{a is an element of infinite order} \quad :\Leftrightarrow \quad \bigwedge_{k\in\mathbb{N}'} (ka \neq 0)$$

Abelian group of finite degree : An abelian group (A ; +) is called an abelian group of finite degree (torsion group) if every element a ∈ A is an element of finite order.

$$(A\,;+)\ \text{is an abelian group of finite degree} \quad :\Leftrightarrow \quad \bigwedge_{a\in A}\ \bigvee_{k\in\mathbb{N}'} (ka = 0)$$

Every finite abelian group (A ; +) is an abelian group of finite degree, since by Fermat's lesser theorem (F2) in Section 7.4.2 the n-fold sum of every element a of a group of finite order n is equal to the identity element 0. However, an abelian group of finite degree may be an infinite group (see Example 4) if the number of generating elements is infinite.

$$(A\,;+)\ \text{is finite} \quad \Rightarrow \quad (A\,;+)\ \text{is an abelian group of finite degree}$$

Abelian group of infinite degree : An abelian group $(A\,;+)$ is called an abelian group of infinite degree and is said to be torsion-free if every element $a \in A$ except for the identity element 0 is an element of infinite order.

$$(A\,;+) \text{ is an abelian group of infinite degree } :\Leftrightarrow \bigwedge_{\substack{a \in A \\ a \neq 0}} \bigwedge_{k \in \mathbb{N}'} (ka \neq 0)$$

Every abelian group of infinite degree is an infinite abelian group, but an infinite abelian group may contain elements of finite order besides the identity element 0.

$$(A\,;+) \text{ is an abelian group of infinite degree} \quad \Rightarrow \quad (A\,;+) \text{ is infinite}$$

Abelian group of mixed degree : An abelian group $(A\,;+)$ is called an abelian group of mixed degree if it contains elements of finite order other than the identity element 0 and also elements of infinite order. An abelian group of mixed degree is an infinite group.

Subgroups of abelian groups : Every subgroup $(X\,;+)$ of an abelian group $(A\,;+)$ is abelian. If X contains exactly the elements of finite order of an abelian group $(A\,;+)$, then $(X\,;+)$ is a subgroup of $(A\,;+)$. This subgroup is called the torsion subgroup of A and is designated by tor A.

Proof : Properties of subgroups of an abelian group

(1) Every subgroup $(X\,;+)$ of an abelian group $(A\,;+)$ is abelian, since the elements of X are also elements of A, and hence the commutativity of the addition $+$ of the group is inherited by the subgroup.

(2) If a subset X of A contains all elements of finite order in A, then X contains the identity element 0. For every element x of order n, X contains the inverse element $-x$, since $nx = 0$ implies $n(-x) = 0$. If X contains an element x_1 of order n_1 and an element x_2 of order n_2, then X contains the element $x_1 + x_2$, since $n_1 x_1 = 0$ and $n_2 x_2 = 0$ implies $n_1 n_2 (x_1 + x_2) = 0$. Hence the domain $(X\,;+)$ has the properties of a group.

Normal subgroups of abelian groups : A left coset of a subgroup $(X\,;+)$ in an abelian group $(A\,;+)$ with a representative a is designated by $a + X$. A right coset is designated by $X + a$. Every subgroup $(X\,;+)$ of an abelian group $(A\,;+)$ is a normal subgroup of A, so that $a + X = X + a = [a]$.

Proof : Every subgroup of an abelian group is a normal subgroup.

The left coset with the representative a is the set $a + X = \{a + x \mid x \in X\}$. The right coset with the same representative a is the set $X + a = \{x + a \mid x \in X\}$. Since the addition $+$ is commutative, the left and right coset are equal : $a + X = X + a$. By the definition of normal subgroups in Section 7.4.2, $(X\,;+)$ is therefore a normal subgroup in A.

Quotient groups : The cosets $[a_0]$, $[a_1]$,... of a normal subgroup (X ; +) in an abelian group (A ; +) form the quotient set $A/X = \{[a_0], [a_1],...\}$ defined in Section 7.5.2. The addition + of cosets in the quotient set A/X is defined by the rule $[a_0] + [a_1] = [a_0 + a_1]$. The quotient group (A/X ; +) is an abelian group with the set X acting as the identity element. If the subgroup (X ; +) contains all elements of finite order of an abelian group (A ; +) of mixed degree, then the quotient group A/X is an abelian group of infinite degree (torsion-free).

Proof : Properties of quotient groups of abelian groups

(1) The quotient group (A/X ; +) is abelian, since the addition + is commutative : $[a_0] + [a_1] = [a_0 + a_1] = [a_1 + a_0] = [a_1] + [a_0]$.

(2) For a representative $x \in X$ of the normal subgroup, $x + X = X = X + x$, so that the set X is the coset [0]. It follows from $[a] + [0] = [a + 0] = [a]$ that $X = [0]$ is the identity element of the quotient group.

(3) The quotient group (A/X ; +) is an abelian group of infinite degree if its identity element X is the only element of finite order. Let a coset [a] in A/X be of finite order n. Then $[na] = n[a] = X$. Hence na is an element of X.

Since the elements of X are of finite order by hypothesis, there is a number m such that $mna = 0$. Hence a is an element of finite order, and therefore an element of the group X. Hence the coset [a] coincides with X, and therefore X is the only element of finite order in the quotient group A/X.

Example 1 : Cyclic four-group

The following sum table defines the addition + in a cyclic group (A ; +) of order 4 with the generating element u and the set $A = gp(u) = \{0, u, 2u, 3u\}$. The sum table is symmetric with respect to the diagonal from the top left corner to the bottom right corner. This symmetry follows from the commutative property $x + y = y + x$ of the addition for $x, y \in A$. The cyclic four-group (A ; +) is abelian.

+	0	u	2u	3u
0	0	u	2u	3u
u	u	2u	3u	0
2u	2u	3u	0	u
3u	3u	0	u	2u

The group (N; +) with $N = \{0, 2u\}$ is a subgroup of (A ; +). Thus, since A is abelian, N is a normal subgroup of A. The cosets of N in A are $[a_0] = [0] = \{0, 2u\} = N$ and $[a_1] = [u] = \{u, 3u\}$. The quotient group (A/N ; +) with $A/N = \{[a_0], [a_1]\}$ and the identity element $N = [a_0]$ is abelian. If two elements of $[a_0]$ are added, the result is an element of $[a_0]$. If two elements of $[a_1]$ are added, the result is an element of $[a_0]$. If an element of $[a_0]$ and an element of $[a_1]$ are added in arbitrary order, the result is an element of $[a_1]$.

+	$[a_0]$	$[a_1]$
$[a_0]$	$[a_0]$	$[a_1]$
$[a_1]$	$[a_1]$	$[a_0]$

→

+	0	2u	u	3u
0	0	2u	u	3u
2u	2u	0	3u	u
u	u	3u	2u	0
3u	3u	u	0	2u

Example 2 : Klein's four-group

The following sum table defines the addition + in Klein's four-group (G ; +) with $G = \{0, a, b, c\}$. Klein's four-group is abelian but not cyclic.

+	0	a	b	c
0	0	a	b	c
a	a	0	c	b
b	b	c	0	a
c	c	b	a	0

The groups (A ; +), (B ; +), (C ; +) with $A = \{0, a\}$, $B = \{0, b\}$, $C = \{0, c\}$ are subgroups of (G ; +). Since G is abelian, they are normal subgroups of G. The cosets of A are $[a_0] = [0] = \{0, a\} = A$ and $[a_1] = [b] = \{b, c\}$. The quotient group (G / A; +) with $G/A = \{[a_0], [a_1]\}$ and the identity element $A = [a_0]$ is abelian. In contrast to the cyclic four-group (Example 1), Klein's four-group has several subgroups and hence several normal subgroups and quotient groups.

Example 3 : Addition of integers

The integers form an abelian group ($\mathbb{Z}$; +) of infinite degree with respect to addition, since the addition of integers is commutative and no integer except for the identity element 0 is of finite order. The group ($\mathbb{Z}$; +) is an infinite cyclic group with the generating element 1 and the set $\mathbb{Z} = \text{gp}(1)$.

The infinite cyclic group ($\mathbb{Z}4$; +) with the generating element 4 and $\mathbb{Z}4 = \text{gp}(4) = \{\ldots,-8,-4,0,4,8,\ldots\}$ is a subgroup of ($\mathbb{Z}$; +). Hence $\mathbb{Z}4$ is a normal subgroup in $\mathbb{Z}$. The cosets of $\mathbb{Z}4$ are the residue classes with respect to division by 4 :

$$\begin{array}{lcll} [0] & = & \{\ldots, -8, -4, 0, 4, 8,\ldots\} = \mathbb{Z}4 & \text{residue class with remainder 0} \\ [1] & = & \{\ldots, -7, -3, 1, 5, 9,\ldots\} & \text{residue class with remainder 1} \\ [2] & = & \{\ldots, -6, -2, 2, 6, 10,\ldots\} & \text{residue class with remainder 2} \\ [3] & = & \{\ldots, -5, -1, 3, 7, 11,\ldots\} & \text{residue class with remainder 3} \end{array}$$

The quotient group $\mathbb{Z}_4 = \mathbb{Z}/\mathbb{Z}4 = \{[0],[1],[2],[3]\}$ with the identity element $\mathbb{Z}4 = [0]$ is abelian. The set of residue classes has the following sum table :

+	[0]	[1]	[2]	[3]
[0]	[0]	[1]	[2]	[3]
[1]	[1]	[2]	[3]	[0]
[2]	[2]	[3]	[0]	[1]
[3]	[3]	[0]	[1]	[2]

The quotient group ($\mathbb{Z}_4$; +) is a cyclic group of order 4.

Example 4 : Addition of rational numbers

The integers $\mathbb{Z}$ are a normal subgroup of the group ($\mathbb{Q}$; +) of rational numbers. The element [q] of the quotient group $\mathbb{Q}/\mathbb{Z}$ is the infinite subset of $\mathbb{Q}$ which is obtained from q by adding arbitrary integers z.

$$[q] = \{h \in \mathbb{Q} \mid \bigvee_{z \in \mathbb{Z}} (h = q + z)\}$$

Let the normal form of the rational number q be $\frac{m}{n}$ with $m < n$. Then the element $[\frac{m}{n}]$ of the quotient group $\mathbb{Q}/\mathbb{Z}$ generates a finite cyclic group of order n :

$$n[\tfrac{m}{n}] = [n\tfrac{m}{n}] = [m] = [0]$$

Since there are infinitely many natural numbers, there is also an infinite number of cosets $[\frac{m}{n}]$ of different order n in $\mathbb{Q}/\mathbb{Z}$.

7.6.3 LINEAR COMBINATIONS

Introduction : The commutative property of abelian groups allows all expressions formed from elements of such groups to be represented as linear combinations. If an abelian group contains only elements of infinite order, then every element of the group is a unique linear combination of the elements of a subset of the group. In analogy with the real vector spaces treated in Section 3.5.2, the concept of a basis is therefore introduced, and the effect of basis transformations is studied. This leads to the concepts of free abelian group, rank of an abelian group, basis of minimal size with respect to a subgroup and minimal generating set of a subgroup. These concepts are important for the study of the structure of abelian groups in Section 7.6.5. However, the structure of abelian groups which contain elements of finite order cannot be studied using bases, since their elements are not unique linear combinations of the elements of an independent subset of the group. Such abelian groups are represented as direct sums.

Linear combination : A sum of elements x_i from a finite subset $X = \{x_1,...,x_s\}$ of an abelian group $(A\,;+)$ is called a linear combination of X. Every element x_i of X and its inverse $-x_i$ may occur more than once in the linear combination. Since by the commutativity of the abelian group the value of the sum is independent of the order of the summands, the elements x_i in the sum may be ordered according to the index i and combined into terms $n_i x_i$ with integers $n_i \in \mathbb{Z}$. A general linear combination of X is designated by $\Sigma\, n_i x_i$. The numbers n_i are called the coefficients of the linear combination. If the index range is not self-evident, it is explicitly specified above and below the sum symbol Σ.

$$\sum_{i=1}^{s} n_i x_i \;=\; n_1 x_1 + \ldots + n_s x_s \qquad\qquad n_i \in \mathbb{Z}\,,\; x_i \in X$$

Linearly independent subset : A subset X of an abelian group $(A\,;+)$ is said to be linearly independent if the identity element 0 of the group A can only be represented as a linear combination of the elements x_i of X using the coefficients $n_i = 0$.

$$n_i x_i + \ldots + n_s x_s = 0 \quad\Rightarrow\quad n_1 = \ldots = n_s = 0 \qquad n_i \in \mathbb{Z}$$

A linearly independent subset X of an abelian group A contains only elements of infinite order. By definition, an element $x_i \in A$ of finite order $n > 0$ cannot be contained in X, since $n x_i = 0$ for $n \neq 0$. The identity element 0 cannot be contained in X since $n\,0 = 0$ for every value of n, and hence also for $n \neq 0$.

The restriction of the linearly independent subsets of an abelian group to elements of infinite order marks an essential difference from the linear independence of real vectors. For every vector $\mathbf{u} \neq \mathbf{0}$ of a real vector space the equation $r\mathbf{u} = \mathbf{0}$ with $r \in \mathbb{R}$ holds only for $r = 0$.

Basis of an abelian group : A linearly independent subset $B = \{b_1,\dots,b_s\}$ of an abelian group $(A\,;+)$ is called a basis of A if every element a of A may be represented as a linear combination of a finite number of basis elements b_i. For a fixed element a, the coefficients n_i of this linear combination are unique.

$$\bigwedge_{a\in A} (a = n_1 b_1 + \dots + n_s b_s \quad \wedge \quad n_i \in \mathbb{Z} \quad \wedge \quad b_i \in B)$$

$$n_1 b_1 + \dots + n_s b_s = 0 \quad \Rightarrow \quad n_1 = \dots = n_s = 0$$

Proof : Uniqueness of the coefficients of an element in a basis

Let the linear combinations $\Sigma m_i b_i$ and $\Sigma n_i b_i$ be representations of the same element a of an abelian group A with the basis B. The coefficient of the basis element b_i in the difference of the two linear combinations is $m_i - n_i$. Since the basis is linearly independent, the condition $\Sigma(m_i - n_i)b_i = 0$ is only satisfied by the coefficients $m_i - n_i = 0$. The coefficients m_i and n_i of the element b_i in the two linear combinations are therefore equal, and hence the representation of the element a in the basis B is unique.

$$a = m_1 b_1 + \dots + m_s b_s$$

$$a = n_1 b_1 + \dots + n_s b_s$$

$$0 = (m_1 - n_1) b_1 + \dots + (m_s - n_s) b_s \quad \Rightarrow \quad m_i = n_i \quad \text{for} \quad i = 1,\dots,s$$

Free abelian group : An abelian group is called a free abelian group if it has a basis. Every element b_i of this basis B generates an infinite cyclic group. Every element a of a free abelian group A is a unique linear combination of the basis elements.

$$A = \{a \mid a = n_1 b_1 + \dots + n_s b_s \ \wedge \ n_i \in \mathbb{N}\,,\ b_i \in B\}$$

Finitely generated free abelian group : A basis of a free abelian group may be an infinite set. If a free abelian group $(A\,;+)$ has a finite basis B, then A is called a finitely generated free abelian group. The relationship between an element a of the group A and the elements b_i of the finite basis B may be represented as a scalar product in vector notation by combining the coordinates n_i of the linear combination and the basis elements b_i into vectors.

$$\bigwedge_{a\in A} (a = \mathbf{n}\cdot\mathbf{b} \quad \wedge \quad b_i \in B,\ n_i \in \mathbb{Z})$$

$$\mathbf{n} = \begin{bmatrix} n_1 \\ \vdots \\ n_s \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_s \end{bmatrix}$$

Transformation of the basis : Let the subsets B and C be bases of a finitely generated free abelian group (A ; +). Then every element c_i of the basis C is an element of the group A, and hence a unique linear combination of the basis B. The bases B and C are equipotent. The transformation of the basis may be represented in matrix notation :

$$\mathbf{c} = \mathbf{T}\,\mathbf{b}$$

$$\begin{array}{|c|}\hline c_1 \\ \hline \vdots \\ \hline c_s \\ \hline\end{array} = \begin{array}{|c|c|c|}\hline t_{11} & \cdots & t_{1s} \\ \hline \vdots & & \vdots \\ \hline t_{s1} & \cdots & t_{ss} \\ \hline\end{array} * \begin{array}{|c|}\hline b_1 \\ \hline \vdots \\ \hline b_s \\ \hline\end{array}$$

Proof : Equipotent bases

Let the basis B contain r elements, and let the basis C contain s > r elements. Each of the elements $c_1,...,c_s$ of the basis C is represented as a linear combination of the elements $b_1, ..., b_r$ of the basis B :

$$\begin{array}{l} n_{11}\, b_1 + n_{12}\, b_2 + ... + n_{1r}\, b_r \;=\; c_1 \\ \vdots \\ n_{s1}\, b_1 + n_{s2}\, b_2 + ... + n_{sr}\, b_r \;=\; c_s \end{array}$$

The first r equations are used to eliminate $b_1, ..., b_r$ in the (r + 1)-th equation. The first equation is used to eliminate b_1 in equations 2 to r + 1. To eliminate b_1 in equation m, the n_{m1}-fold multiple of the first equation is subtracted from the n_{11}-fold multiple of the m-th equation. If $n_{11} = 0$, the first equation is first exchanged with an equation k for which $n_{k1} \neq 0$. If $n_{k1} = 0$ for each of the rows 1,...,r + 1, then b_1 is already eliminated. The procedure is continued by eliminating b_2 in equations 3 to r + 1 using the second equation. In this way, since the basis B contains exactly r elements, each of the elements $b_1, ..., b_r$ can be eliminated in equation r + 1 using one of the first r equations. The (r + 1)-th equation takes the following form :

$$w_1\, c_1 + ... + w_r\, c_r + c_{r+1} = 0 \qquad\qquad w_i \in \mathbb{Z}$$

The coefficient 1 of c_{r+1} contradicts the definition of the basis C, since in every representation of the identity element 0 as a linear combination of the basis elements $c_1, ..., c_s$ all coefficients $w_1, ..., w_s$ are zero. It follows that C does not contain more elements than B. Analogously, B does not contain more elements than C. Hence the bases B and C are equipotent.

Transformation of the group : If B and C are bases of an abelian group (A ; +), then every element of the group A may alternatively be represented as a linear combination **m • b** of the basis B or as a linear combination **n • c** of the basis C. The coefficients **m** for the basis B are derived from the coefficients **n** for the basis C.

$$\bigwedge_{a \in A} (a = \mathbf{m} \cdot \mathbf{b} = \mathbf{n} \cdot \mathbf{c}) \quad \wedge \quad \mathbf{c} = \mathbf{R}\,\mathbf{b} \quad \Rightarrow \quad \mathbf{m} = \mathbf{R}^T \mathbf{n}$$

Rank of an abelian group : If an abelian group has a basis, then all bases obtained by transforming this basis contain the same number of elements. The number of elements is called the rank of the abelian group. If a group has no basis, it has no rank.

Size of a basis : Let H be a subgroup of a finitely generated free abelian group (A ; +) of rank s. For $s > 1$, the basis of the group A is not unique. If a certain basis $B = (b_1, ..., b_s)$ is chosen, then every element h of the subgroup H is a unique linear combination of the basis elements :

$$h = n_1 b_1 + ... + n_s b_s \qquad h \in H,\ n_i \in \mathbb{Z},\ b_i \in B$$

The non-zero coefficient of least magnitude among the coefficients $n_1, ..., n_s$ is determined for an arbitrary non-zero element $h \in H$. If this coefficient is negative, then the associated basis element b_i is replaced by $-b_i$. Thus there is always a least positive value among the coefficients for the element h.

The least positive coefficient in the representations of all non-zero elements $h \in H$ as linear combinations of a fixed basis B is determined. This value is called the size of the basis B with respect to the subgroup H and is designated by w_B.

Among the values w_B for all bases B, the least value w is determined. This value may occur for more than one basis. Let one of the bases of size w with respect to H be $Y = \{y_1, ..., y_s\}$. The basis Y is called a basis of minimal size with respect to the subgroup H.

Properties of a basis of minimal size : Let the minimal size of the bases of a finitely generated free abelian group (A ; +) with respect to a subgroup H be w. Let $Y = \{y_1, ..., y_s\}$ be a basis of A of minimal size with respect to H. Let h_{min} be the element of H which leads to the size w of the basis Y. Then every coefficient of the linear combination for h_{min} is divisible by the minimal size w.

$$h_{min} = w(y_1 + k_2 y_2 + ... + k_s y_s) \qquad k_i \in \mathbb{Z}$$

Proof : Properties of a basis of minimal size

The basis elements are numbered such that the minimal size w is the coefficient of y_1 in the linear combination for h_{min} :

$$h_{min} = w y_1 + n_2 y_2 + ... + n_s y_s \qquad w \in N',\ n_i \in \mathbb{Z}$$

An arbitrary coefficient n_m of this linear combination is represented in the remainder form $n_m = k_m w + r_m$ with $0 \le r_m < w$:

$$h_{min} = w(y_1 + k_m y_m) + n_2 y_2 + ... + r_m y_m + ... + n_s y_s$$

Since every element a of A can be represented as a linear combination of the basis Y, it can also be represented as a linear combination of $\bar{Y} = \{y_1 + k_m y_m, y_2, \ldots, y_s\}$:

$$\begin{aligned} a &= c_1 y_1 + c_2 y_2 + \ldots + c_m y_m + \ldots + c_s y_s \\ a &= c_1 (y_1 + k_m y_m) + c_2 y_2 + \ldots + (c_m - c_1 k_m) y_m + \ldots + c_s y_s \end{aligned}$$

If a linear combination of the elements of $\bar{Y}$ is zero, then a corresponding linear combination of the elements of Y is zero :

$$\begin{aligned} t_1 (y_1 + k_m y_m) + t_2 y_2 + \ldots + t_m y_m + \ldots + t_s y_s &= 0 \quad \Rightarrow \\ t_1 y_1 + t_2 y_2 + \ldots + (t_m + t_1 k_m) y_m + \ldots + t_s y_s &= 0 \end{aligned}$$

Since Y is a basis of A, it follows that $t_1 = t_2 = \ldots = t_m + t_1 k_m = \ldots = t_s = 0$. This implies $t_1 = t_2 = \ldots = t_s = 0$. Hence $\bar{Y}$ is a basis of A. The representation of h_{min} in this basis contains the coefficient r_m for y_m with $0 \le r_m < w$. Since w is the minimal size of a basis of A with respect to H, it follows that $r_m = 0$. Hence n_m is divisible by w. This property holds for m = 2,...,s. Hence each of the coefficients $n_2, \ldots, n_s$ is divisible by w.

Minimal generating set of a subgroup : Let H be a subgroup of a finitely generated free abelian group (A ; +) of rank s. Then every element of the subgroup H can be represented in an arbitrary basis $B = \{b_1, \ldots, b_s\}$ of the group A. However, for a special basis $X = \{x_1, \ldots, x_s\}$ of the group A there are numbers $c_1, \ldots, c_s$ and r such that :

(1) The set $E = \{c_1 x_1, \ldots, c_r x_r\}$ generates the subgroup H.

(2) For $r < i \le s$, the numbers c_i are zero. For $i \le r$, the numbers c_i form a divisor chain $c_1 \mid c_2 \mid \ldots \mid c_r$. Thus, the number c_i is a divisor of c_{i+1} .

The set E is called a minimal generating set of the subgroup H in the group A.

Proof : Minimal generating set of a subgroup

(1) A basis of minimal size w with respect to the subgroup H is determined for the group A. Let this basis be $\{y_1, \ldots, y_s\}$. Then, according to the preceding proof, for a suitable order of the indices of the basis elements there is an element $h_1 \in H$ with $h_1 = w(y_1 + k_2 y_2 + \ldots + k_s y_s)$. Let $x_1 := y_1 + k_2 y_2 + \ldots + k_s y_s$ and $c_1 := w$, so that $h_1 = c_1 x_1$. Then $(x_1, y_2, \ldots, y_s)$ is a basis of A. For an arbitrary element $h \in H$:

$$h = a_1 x_1 + a_2 y_2 + \ldots + a_s y_s \qquad a_i \in \mathbb{Z}$$

The coefficient a_1 is a multiple of c_1, since division with remainder yields $a_1 = p_1 c_1 + r_1$ with $0 \le r_1 < c_1$, and hence

$$\begin{aligned} h - p_1 h_1 &= (p_1 c_1 + r_1) x_1 - p_1 c_1 x_1 + a_2 y_2 + \ldots + a_s y_s \\ &= r_1 x_1 + a_2 y_2 + \ldots + a_s y_s \end{aligned}$$

Since c_1 is the minimal size of a basis of A with respect to H and $h - p_1 h_1$ is an element of H, it follows that $r_1 = 0$, and hence

$$h \;=\; p_1 c_1 x_1 + a_2 y_2 + \ldots + a_s y_s \qquad\qquad p_1 \in \mathbb{Z}$$

(2) The basis elements $y_2, \ldots, y_s$ generate a subgroup A' of A. The set H' of all elements $h' \in H$ which can be represented as linear combinations of the elements $y_2, \ldots, y_s$ is a subgroup of both H and A'. A basis of minimal size w' with respect to the subgroup H' is determined for the group A'. Let this basis be $\{y_2', \ldots, y_s'\}$. Then each element $h' \in H'$ has a representation

$$h' \;=\; a_2' y_2' + a_3' y_3' + \ldots + a_s' y_s' \qquad\qquad a_i' \in \mathbb{Z}$$

With $x_2 := y_2' + k_3' y_3' + \ldots + k_s' y_s'$ and $c_2 := w'$, it follows by analogy with (1) that the coefficient a_2' is a multiple of c_2 :

$$h' \;=\; p_2 c_2 x_2 + a_3' y_3' + \ldots + a_s' y_s' \qquad\qquad p_2 \in \mathbb{Z}$$

The elements of H may thus be represented in the basis $\{x_1, x_2, y_3', \ldots, y_s'\}$ by the following linear combinations :

$$h \;=\; p_1 c_1 x_1 + p_2 c_2 x_2 + a_3' y_3' + \ldots + a_s' y_s'$$

(3) The minimal size c_1 is a divisor of c_2. To prove this, consider the element $h = c_1 x_1 + c_2 x_2$ of H. Division with remainder yields $c_2 = p c_1 + r$ with $0 \le r < c_1$, and hence

$$h \;=\; c_1(x_1 + p x_2) + r x_2$$

Since c_1 is the minimal size of a basis of A with respect to H, it follows that $r = 0$, and hence $c_2 = p c_1$.

(4) The construction in step (2) is continued until the subgroup H' for the remaining basis elements $y_{r+1}', \ldots, y_s'$ is empty. These basis elements are not required for generating H. For $r < i \le s$, the basis element $x_i = y_i'$ is therefore chosen, and c_i is set to zero. Then for all $h \in H$:

$$\begin{aligned} h &= p_1 c_1 x_1 + \ldots + p_r c_r x_r && \text{with } c_i \mid c_{i+1} \quad \text{for } 1 \le i < r \\ c_i &= 0 && \text{for } \; i > r \end{aligned}$$

Rank of a subgroup : Let H be a subgroup of a finitely generated free abelian group (A ; +) of rank s. Then the rank r of H is less than or equal to the rank s of A. For if $E = \{c_1 x_1, \ldots, c_s x_s\}$ is the minimal generating set of H with $c_{r+1} = \ldots = c_s = 0$, then every element of H is a linear combination of the elements $c_1 x_1, \ldots, c_r x_r$ with $r \le s$.

7.6.4 DIRECT SUMS

Introduction : In the preceding sections, bases are shown to be unsuitable for studying the structure of general abelian groups : For an element a of finite order n, the coefficient in a linear combination is not unique, since $na = 2na = ... = 0$.

A unique representation of every element of an abelian group in terms of certain elements of the group is desired. This is achieved by forming the direct sum of elements of certain subgroups of the abelian group. The direct sum contains exactly one element from each of the subgroups under consideration. In contrast to a linear combination, in which an element of the basis may occur n times, each element occurs only once in the direct sum. By virtue of this rule, the representation of an element of the group as a direct sum becomes unique. However, the choice of subgroups is not unique.

Note : Elements of cyclic subgroups may occur in a direct sum. These elements are often represented as the m-fold multiple of the generating element a of the subgroup. Thus the m-th element of the subgroup is designated by ma instead of a_m. This element ma enters into the direct sum only once. For an element of order n it is assumed that $0 \le m < n$ holds, so that ma is a unique element of the subgroup $\{0, a, ..., (n-1)a\}$.

Outer direct sum : The cartesian (direct) product $T = A \times B$ of the abelian groups $(A ; +)$ and $(B ; +)$ contains the general element $t = (a, b)$ with $a \in A$ and $b \in B$. The inner operation $+$ on two elements t_1 and t_2 of the product T is defined as follows :

$$T := A \times B \qquad \text{with} \qquad t_i = (a_i, b_i)$$

$$t_1 + t_2 = (a_1, b_1) + (a_2, b_2) := (a_1 + a_2, b_1 + b_2)$$

The domain $(T ; +)$ has the properties of an abelian group and is called the outer direct sum of the groups A and B.

(1) The identity element of the group T is $(0_A, 0_B)$.

(2) If t_1 and t_2 are elements of T, then $t_1 + t_2$ is also an element of T, since $(a_1 + a_2) \in A$ and $(b_1 + b_2) \in B$.

(3) If T contains the element $t = (a, b)$, then T also contains the inverse element $-t = (-a, -b)$, since A contains the inverse $-a$ of a and B contains the inverse $-b$ of b.

(4) The group T is abelian, since the groups A and B are abelian :

$$t_1 + t_2 = (a_1 + a_2, b_1 + b_2) = (a_2 + a_1, b_2 + b_1) = t_2 + t_1$$

Inner direct sum : Let $(A\,;+)$ and $(B\,;+)$ be subgroups of an abelian group $(G\,;+)$. Let every element g of the group G be a unique sum of an element $a \in A$ and an element $b \in B$. Then the group G is called the inner direct sum of the subgroups A and B. The inner direct sum is designated by $G = A \oplus B$. The group $(G\,;+)$ and the outer direct sum $(A \times B\,;+)$ are isomorphic.

$$(g = a_1 + b_1 \quad \wedge \quad g = a_2 + b_2) \quad \Rightarrow \quad (a_1 = a_2 \quad \wedge \quad b_1 = b_2)$$

$$G = A \oplus B \cong A \times B = T$$

Proof : Isomorphism of the direct sums

The outer direct sum $(A \times B\,;+)$ and the inner direct sum $(A \oplus B\,;+)$ are isomorphic if the mapping between the groups is bijective and homomorphic in both directions.

$$\begin{array}{llllll} f & : A \times B & \rightarrow & A \oplus B & \text{with} & f((a,b)) = a + b \\ f^{-1} & : A \oplus B & \rightarrow & A \times B & \text{with} & f^{-1}(a+b) = (a,b) \end{array}$$

(1) The mapping f is bijective, since every element g of the group $G = A \oplus B$ has a unique representation (a, b) in $A \times B$ and a unique representation $a + b$ in $A \oplus B$.

(2) The mapping is homomorphic in both directions :

$$\begin{aligned} f((a_1,b_1) + (a_2,b_2)) &= f((a_1 + a_2, b_1 + b_2)) = a_1 + a_2 + b_1 + b_2 \\ &= (a_1 + b_1) + (a_2 + b_2) = f((a_1,b_1)) + f((a_2,b_2)) \end{aligned}$$

$$\begin{aligned} f^{-1}((a_1 + b_1) + (a_2 + b_2)) &= f^{-1}((a_1 + a_2) + (b_1 + b_2)) = (a_1 + a_2, b_1 + b_2) \\ &= (a_1,b_1) + (a_2,b_2) \\ &= f^{-1}(a_1 + b_1) + f^{-1}(a_2 + b_2) \end{aligned}$$

Example 1 : Isomorphic direct sums

The following sum table defines the addition + in an abelian group $(G\,;+)$ with $G = \{0, a, b, c\}$. The sum table of a cyclic group gp(u) of order 4 which is not isomorphic with G is also shown for comparison.

+	0	a	b	c
0	0	a	b	c
a	a	0	c	b
b	b	c	0	a
c	c	b	a	0

Klein's four-group G

+	0	u	2u	3u
0	0	u	2u	3u
u	u	2u	3u	0
2u	2u	3u	0	u
3u	3u	0	u	2u

cyclic four-group gp(u)

The group G contains the proper subgroups $A = \{0, a\}$, $B = \{0, b\}$ and $C = \{0, c\}$. The elements of G are unique sums of the elements of A and B. Hence G is the direct inner sum of A and B.

$$G = A \oplus B = \{0+0, a+0, 0+b, a+b\} = \{0, a, b, c\}$$

The set $A \times B$ of the outer direct sum $(A \times B; +)$ is given by $\{(0,0), (a,0), (0,b), (a,b)\}$. The mapping $f : A \times B \to A \oplus B$ maps the elements of G as follows :

$$f((0,0)) = 0+0 = 0 \qquad f((0,b)) = 0+b = b$$
$$f((a,0)) = a+0 = a \qquad f((a,b)) = a+b = c$$

The representation of G as an inner direct sum of A and B is not unique, since $G = A \oplus B = B \oplus C = C \oplus A$.

Multiple direct sums : The concept of a direct sum is extended to a finite family of n subgroups G_i of an abelian group $(G; +)$. The subgroup G_i contains the elements $\{g_{i(1)}, g_{i(2)}, \ldots\}$. Every element of the group G is a sum of one element $g_{i(m)}$ from each of the groups G_i. The intersection of the subsets G_i contains only the identity element 0 of the group G. The union of the subsets G_i is a subset of G.

outer direct sum : $T := G_1 \times \ldots \times G_n$ with $t_{(m)} = (g_{1(m)}, \ldots, g_{n(m)})$

$t_{(i)} + t_{(m)} := (g_{1(i)} + g_{1(m)}, \ldots, g_{n(i)} + g_{n(m)})$

inner direct sum : $G = G_1 \oplus \ldots \oplus G_n$ with $g_{(m)} = g_{1(m)} + \ldots + g_{n(m)}$

$i \neq m \quad \Rightarrow \quad G_i \cap G_m = \{0\}$

Infinite inner direct sum : The concept of an inner direct sum cannot be directly extended to an infinite number of summands, since an infinite sum of the form $g_{(m)} = g_{1(m)} + g_{2(m)} + \ldots$ is not defined. For an infinite family of subgroups G_i with $i \in I$, it is assumed that for all but a finite number of groups G_i the element 0 is chosen for the summation. These elements do not influence the value of the direct sum and are therefore omitted. Let the indices of the non-zero elements be $i_1, \ldots, i_s$. The resulting direct sum is designated by $\sum_i g_i$.

$$g_{(m)} = \sum_i g_{i(m)} = g_{i_1(m)} + \ldots + g_{i_s(m)}$$

An infinite inner direct sum is designated by $G_1 \oplus G_2 \oplus \ldots$. If the notation is meant to represent both finite and infinite inner direct sums, then the symbol $\oplus G_i$ with $i \in I$ is used.

$$G := G_1 \oplus G_2 \oplus \ldots \qquad \text{with} \qquad g_{i(m)} \in G_i$$
$$i \neq m \quad \Rightarrow \quad G_i \cap G_m = \{0\}$$

Properties of direct sums of abelian groups :

(D1) Let $G_1,...,G_n$ be subgroups of the abelian group G. Their sum $G_1 + ... + G_n$ is direct if and only if there is only one way to represent the identity element 0 of G as a sum of elements g_i from $G_1,...,G_n$, namely $g_1 = ... = g_n = 0$:

$$(g_i \in G_i \quad \wedge \quad g_1 + ... + g_n = 0) \quad \Rightarrow \quad g_1 = ... = g_n = 0$$

(D2) Let the abelian group G be the sum of its subgroups $G_1,...,G_n$ with $n \geq 2$. This sum is direct if and only if each of the subgroups G_i has only the identity element 0 in common with the sum S_i of the remaining subgroups.

$$S_i = G_1 + ... + G_{i-1} + G_{i+1} + ... + G_n$$

$$G = G_1 + ... + G_n \quad \wedge \quad \bigwedge_i \ (G_i \cap S_i = \{0\}) \quad \Leftrightarrow$$

$$G = G_1 \oplus ... \oplus G_n$$

(D3) Let the abelian group G be the sum of its subgroups $G_1,...,G_n$. Let each subgroup G_i with $i = 1,...,n$ be the sum of subgroups G_{ik} with $k = 1,...,m_i$. The group G is the direct sum of the subgroups G_{ik} if and only if the sum of $G_1,...,G_n$ and each of the sums of $G_{i1},...,G_{im_i}$ for $i = 1,...,n$ is direct.

Proof : Representation of abelian groups as direct sums

(D1) Let $g_1 = ... = g_n = 0$ be the only representation of the identity element 0 as a sum $g_1 + ... + g_n = 0$ with $g_i \in G_i$. Let an element $a \in G$ be represented by two sums of elements $a_i, b_i \in G_i$:

$$a = a_1 + ... + a_n$$

$$a = b_1 + ... + b_n$$

Since the order of the operations in abelian groups is irrelevant, the difference of these equations may be written as follows :

$$0 = (a_1 - b_1) + ... + (a_n - b_n)$$

Since the only representation of 0 as a sum is given by $g_i = a_i - b_i = 0$, it follows that $a_i = b_i$. The sum for the element a is therefore unique. Hence G is the direct sum of the subgroups $G_1,...,G_n$.

Conversely, let G be the direct sum of $G_1,...,G_n$. Then the representation $a = a_1 + ... + a_n$ of an arbitrary element $a \in G$ is unique. The subtraction $a - a$ shows that the identity element 0 of G can only be represented in the form $0 = 0 + ... + 0$.

(D2) Let the element $a \in G$ be contained in G_i and in S_i. Then the group S_i also contains the inverse element $-a$. Thus there is a representation $a + (-a) = 0$ with $a \neq 0$. This contradicts (D1). Hence the sum is not direct.

Assume that the sum is not direct. Then by (D1) there exists a representation $0 = a + b$ with $a \in G_i$ and $b \in S_i$ and $a, b \neq 0$. Together with a, the group G_i also contains the inverse $-a$. From $b = -a$ it follows that $b \in G_i$, and hence $b \in G_i \cap S_i$.

(D3) Let the group G be the direct sum of its subgroups G_{ik}. Then each of the sums $G_{i1} \oplus ... \oplus G_{im_i}$ is a direct sum. For the sum $G_1 + ... + G_n$, every term g_i in the equation $g_1 + ... + g_n = 0$ is decomposed into the unique sum $g_{i1} + ... + g_{im_i}$. Since the group G is the direct sum of the groups G_{ik}, the resulting equation $g_{11} + ... + g_{nm_n} = 0$ is satisfied only by $g_{ik} = 0$. Hence $g_i = 0$: The sum of $G_1, ..., G_n$ is direct by (D1).

Conversely, assume $G = G_1 \oplus ... \oplus G_n$ and $G_i = G_{i1} \oplus ... \oplus G_{im_i}$ for $i = 1, ..., n$. Then every element $g \in G$ is a unique sum $g_1 + ... + g_n$ of elements $g_i \in G_i$. But every element g_i is a unique sum $g_{i1} + ... + g_{im_i}$ of elements $g_{ik} \in G_{ik}$. It follows by substitution that $g = (g_{11} + ... + g_{1m_1}) + ... + (g_{n1} + ... + g_{nm_n})$. Since the addition $+$ in the group G is associative, it follows that $g = g_{11} + ... + g_{1m_1} + ... + g_{n1} + ... + g_{nm_n}$ is a unique sum of elements from each of the subgroup G_{ik}. Hence the group G is the direct sum of the subgroups G_{ik}.

Scaled groups : An abelian group (G ; +) is scaled by an integer n by mapping every element g of G to its n-fold multiple ng. The group G scaled by n is designated by nG ; it is a subgroup of G.

$$nG := \{ng \mid g \in G\} \qquad n \in \mathbb{Z}$$

Proof : Scaled groups are subgroups.

The set nG contains the identity element $n \cdot 0 = 0$. Since for any element g the group G contains an inverse element g^{-1}, it follows that nG contains the element ng^{-1} inverse to ng. For ng_1 and ng_2, the sum $ng_1 + ng_2 = n(g_1 + g_2)$ is also contained in nG, since together with g_1 and g_2 the sum $g_1 + g_2$ is contained in G. Thus nG has the properties of a group.

Direct sums of scaled groups : Let an abelian group (G ; +) be the direct sum of the groups A and B. Then the scaled group nG is the direct sum of the scaled groups nA and nB.

$$G = A \oplus B \quad \Rightarrow \quad nG = nA \oplus nB \qquad n \in \mathbb{Z}$$

Proof : Direct sums of scaled groups

(1) Since the groups A and B are disjoint except for the identity element, their subgroups nA and nB are also disjoint except for the identity element. Hence by (D2) there is an inner direct sum $nA \oplus nB$.

(2) Every element $g \in G$ has a unique representation $g = a + b$ with $a \in A$ and $b \in B$. Hence $ng = na + nb$ is an element of $nA \oplus nB$, that is $nG \subseteq nA \oplus nB$.

(3) For arbitrary elements $a \in A$ and $b \in B$, $na + nb = n(a + b)$ is an element of nG. Thus $nA \oplus nB \subseteq nG$.

(4) From $nA \oplus nB \subseteq nG \subseteq nA \oplus nB$ it follows that $nG = nA \oplus nB$.

Properties of cyclic subgroups of outer direct sums

(K1) Let $G_1, ..., G_s$ be general groups, and let $a_i \in G_i$ be an element of finite order m_i. Then the element $(a_1, ..., a_s)$ of the cartesian product $G_1 \times ... \times G_s$ generates a cyclic subgroup of order $\text{lcm}(m_1, ..., m_s)$.

(K2) The cartesian product $G := G_1 \times ... \times G_s$ of finite groups $G_1, ..., G_s$ is a finite cyclic group if and only if $G_1, ..., G_s$ are cyclic groups with pairwise mutually prime orders. In this case, an element $(a_1, ..., a_s)$ of G is a generating element of G if and only if the elements a_i generate the groups G_i.

$$G \; := \; G_1 \times ... \times G_s$$

$$\bigwedge_i \; G_i = gp(a_i) \quad \Leftrightarrow \quad G = gp(a_1, ..., a_s)$$

Proof : Properties of cyclic subgroups of outer direct sums

(K1) By property (Z2) of cyclic groups in Section 7.3.4, the r-th multiple of a_i is zero if and only if r contains the order m_i of a_i as a factor. Hence $r(a_1, ..., a_s) = (ra_1, ..., ra_s)$ is the identity element $(0,...,0)$ if and only if r is a common multiple of $m_1, ..., m_s$. The least such number is $\text{lcm}(m_1, ..., m_s)$. This is the order of the group generated by $(a_1, ..., a_s)$.

(K2) Let the cartesian product $G = G_1 \times ... \times G_s$ of finite groups G_i be cyclic. The elements of the form $(0,...,g_i,...,0) \in G$ for every $g_i \in G_i$ form a subgroup of G, which is cyclic by (U1) in Section 7.3.6. This subgroup is isomorphic to G_i, so that G_i is also cyclic.

Let the element $a = (a_1, ..., a_s)$ be a generating element of the cyclic group G. Let the order of the element a_i be m_i. By (K1), $\text{ord}\, G = \text{ord}\,(gp(a)) = \text{lcm}(m_1, ..., m_s)$. Since G is the cartesian product of $G_1,...,G_s$, it follows that $\text{ord}\, G_1 \cdots \text{ord}\, G_s = \text{ord}\, G = \text{lcm}(m_1, ..., m_s)$. Also $a_i \in G_i$ implies $m_i \le \text{ord}\, G_i$. Altogether, it follows that the orders $m_1, ..., m_s$ are mutually prime and that $\text{ord}\, G_i = m_i$. This further implies that a_i generates the group G_i.

Let the factors of a cartesian product $G_1 \times ... \times G_s$ be finite cyclic groups $G_i = gp(a_i)$ with mutually prime orders $m_i = \text{ord}\, G_i$. Then $\gcd(m_i, m_k) = 1$ implies $\text{lcm}(m_1, ..., m_s) = m_1 \cdots m_s$. By (K1), the element $(a_1, ..., a_s)$ generates a cyclic group of order $m_1 \cdots m_s$. Since $\text{ord}\,(G_1 \times ... \times G_s) = m_1 \cdots m_s$, the group G is cyclic.

Example 2 : Cyclic cartesian product

Let the order of the group $G_1 = gp(a)$ be 3, and let the order of the group $G_2 = gp(b)$ be 2. Since the orders of G_1 and G_2 are mutually prime, the cartesian product $G = G_1 \times G_2$ is a cyclic group of order $3 \cdot 2 = 6$. Since a and b are generating elements of G_1 and G_2, respectively, by property (K2) the pair (a,b) is a generating element of G. Since 2a is also a generating element of G_1, the pair (2a, b) is also a generating element of G.

+	(0,0)	(a,b)	(2a,0)	(0,b)	(a,0)	(2a,b)
(0,0)	(0,0)	(a,b)	(2a,0)	(0,b)	(a,0)	(2a,b)
(a,b)	(a,b)	(2a,0)	(0,b)	(a,0)	(2a,b)	(0,0)
(2a,0)	(2a,0)	(0,b)	(a,0)	(2a,b)	(0,0)	(a,b)
(0,b)	(0,b)	(a,0)	(2a,b)	(0,0)	(a,b)	(2a,0)
(a,0)	(a,0)	(2a,b)	(0,0)	(a,b)	(2a,0)	(0,b)
(2a,b)	(2a,b)	(0,0)	(a,b)	(2a,0)	(0,b)	(a,0)

sum table of the product group (G ;+)

$$G = G_1 \times G_2 \quad \text{mit} \quad G_1 = \{0, a, 2a\} \quad \text{und} \quad G_2 = \{0, b\}$$

Chinese remainder theorem : Let the positive natural numbers $n_1, ..., n_s$ be pairwise mutually prime. Then for arbitrary natural numbers $a_1, ..., a_s$ there is a natural number b such that b is congruent to a_i modulo n_i :

$$\bigwedge_{i \neq k} (\gcd(n_i, n_k) = 1) \;\Rightarrow\; \bigwedge_{a_1,...,a_s \in \mathbb{N}} \; \bigvee_{b \in \mathbb{N}} \; (b \equiv a_i \bmod n_i)$$

Every integer c which is congruent to b modulo $n_1 \cdots n_s$ is also congruent to a_i modulo n_i for $i = 1,...,s$:

$$c \equiv b \mod n_1 \cdots n_s \;\Rightarrow\; \bigwedge_{i \in \{1,...,s\}} (c \equiv a_i \bmod n_i)$$

Proof : Chinese remainder theorem

(1) By property (K2), a cartesian product $G := G_1 \times ... \times G_s$ of finite groups $G_1,...,G_s$ is a cyclic group $(G\,;+)$ if and only if $G_1,...,G_s$ are cyclic groups $(G_i\,;+)$ with pairwise mutually prime orders. Cyclic groups of order n_i with the generating elements x_i for $i = 1,...,s$ are chosen as factors G_i of the cartesian product. Since the orders $n_1,...,n_s$ of the factors $gp(x_i)$ are by hypothesis pairwise mutually prime, G is cyclic. By (K1), G is of order $n := n_1 \cdots n_s$. The general element of the cartesian product G is :

$$g \;=\; (a_1x_1,...,a_sx_s) \qquad\qquad a_i \in \mathbb{N}$$

The general element of the finite cyclic group G with the generating element $(x_1,...,x_s)$ is :

$$g \;=\; b(x_1,...,x_s) \;=\; (bx_1,...,bx_s) \qquad\qquad b \in \mathbb{N}$$

These are two different representations of the same element g of the group G. The uniqueness of an element $g \in G$ implies $a_ix_i = b\,x_i$ for $i = 1,...,s$. By property (Z2) of cyclic groups in Section 7.3.4, the elements a_ix_i and $b\,x_i$ of the group $gp(x_i)$ of order n_i are equal if and only if $b \equiv a_i \bmod n_i$ for $i = 1,...,s$. Hence there is such a number b.

(2) By property (Z2) of cyclic groups, $b \equiv c \bmod n$ implies that the elements $b(x_1,..., x_s)$ and $c(x_1,..., x_s)$ of the cyclic group $(G\,;+)$ of order n are equal. As before, $a_ix_i = c\,x_i$ for $i = 1,...,s$, and hence $c \equiv a_i \bmod n_i$.

Example 3 : Chinese remainder theorem

The numbers $(n_1,n_2,n_3) = (2,3,5)$ are mutually prime. Let the tuple $(a_1,a_2,a_3) = (103, 86, 69)$ be given. By the remainder theorem, there is a natural number b with $b = a_i \bmod n_i$. One such number is $b = 29$:

$$\begin{array}{rclcrcl} 103 \bmod 2 & = & 1 & \qquad & 29 \bmod 2 & = & 1 \\ 86 \bmod 3 & = & 2 & & 29 \bmod 3 & = & 2 \\ 69 \bmod 5 & = & 4 & & 29 \bmod 5 & = & 4 \end{array}$$

The numbers $b_1 = 29$ and $b_2 = 89$ are congruent modulo $2 \cdot 3 \cdot 5 = 30$, that is $29 \equiv 89 \bmod 30$. Hence also $89 \equiv a_i \bmod n_i$.

$$\begin{array}{rcl} 89 \bmod 2 & = & 1 \\ 89 \bmod 3 & = & 2 \\ 89 \bmod 5 & = & 4 \end{array}$$

7.6.5 CONSTRUCTIONS OF ABELIAN GROUPS

Introduction : An abelian group is to be constructed from a given set of abelian groups $(G_i \,;\, +)$ with $i \in I$. It is not assumed that the groups G_i are subgroups of the same group G. Thus no operation is defined for elements from different groups $G_i \neq G_m$. The inner direct sum of the groups G_i is therefore not defined either. In the following, it is shown that there is a group H with subgroups H_i such that H is the inner direct sum of the subgroups H_i and H_i is isomorphic to G_i.

$$H = \bigoplus_{i \in I} H_i \qquad \text{with } H_i \cong G_i$$

Thus a new abelian group H is constructed from the given groups G_i. Isomorphic groups are often regarded as identical, and the construction of the new group is designated by $G_1 \oplus G_2 \oplus \ldots$. The constructed group G is called the direct sum of the groups G_i. This terminology refers to the relationship described above.

The elements of the constructed set H are mappings from the index set I to the union of the given groups G_i. The value of a mapping $h : I \rightarrow \cup G_i$ for the index i is an element of the group G_i. The sum of the mappings is chosen as the inner operation on H. The set H of index mappings is thus equipped with a group structure.

A subset of the mappings in H maps the index i to an arbitrary element of G_i and every index $k \neq i$ to the identity element of the group G_k. This subset is designated by H_i. It is a subgroup of H. The group H is the inner direct sum of its subgroups H_i. The groups H_i and G_i are isomorphic. These relationships are proved in the following.

The index set I used in the construction of the group H is arbitrary; it may be finite or infinite. The construction of the group H as a direct sum of subgroups H_i which are isomorphic to given groups G_i is therefore an extension of the concept of a cartesian product.

Index mappings : Let $(G_i \,;\, +)$ with $i \in I$ be an abelian group with the general element $g_{i(s)}$ and the identity element 0_i. The index set I and each of the groups G_i may be infinite.

The index set I is mapped to the union $\cup G_i$ of the abelian groups G_i. A mapping $h : I \rightarrow \cup G_i$ is called an index mapping if the index i is mapped to an element $g_{i(s)}$ of G_i. Thus the image h(I) contains exactly one element from each group G_i. If the index set I is infinite, then it is assumed that $h(i) = 0_i$ for all but a finite number of $i \in I$. The set of index mappings is designated by H :

$$H := \{ h \mid h : I \rightarrow \cup G_i \text{ with } h(i) = g_{i(s)} \in G_i \}$$

The set H is equipped with a group structure (H ; +). For two index mappings h_1, $h_2 \in H$ the inner operation + yields the sum $h_1 + h_2$ of the mappings. The identity element h_0 maps the index i to the identity element 0_i. The inverse of h in the additive group H is $-h$. The images h(i) and $-h(i)$ are inverses of each other.

$$\begin{array}{llll} + & : H \times H \rightarrow H & \text{with} & (h_1 + h_2)(i) = h_1(i) + h_2(i) \\ h_0 & : I \rightarrow \cup G_i & \text{with} & h_0(i) = 0_i \\ -h & : I \rightarrow \cup G_i & \text{with} & -h(i) = -g_{i(s)} \end{array}$$

Proof : The domain (H ; +) is an abelian group.

The domain (H ; +) has the properties of a group :

$$\begin{array}{lllllll} (h_1 + h_2)(i) & = h_1(i) + h_2(i) & = g_{i(s_1)} + g_{i(s_2)} & = g_{i(s_3)} & = h_3(i) \\ (h_0 + h)(i) & = h_0(i) + h(i) & = 0_i + g_{i(s)} & = g_{i(s)} & = h(i) \\ (h - h)(i) & = h(i) - h(i) & = g_{i(s)} - g_{i(s)} & = 0_i & = h_0(i) \end{array}$$

The addition in the group (H ; +) is commutative :

$$h_1(i) + h_2(i) = g_{i(s_1)} + g_{i(s_2)} = g_{i(s_2)} + g_{i(s_1)} = h_2(i) + h_1(i)$$

Subgroups of the index mappings : In the group (H ; +) of index mappings, there is exactly one mapping which maps the index i to a given element $g_{i(s)}$ of G_i and every index $k \neq i$ to the identity element 0_k of the group G_k. This mapping is designated by $h_{i(s)}$. The set of mappings $h_{i(s)}$ for all elements $g_{i(s)} \in G_i$ is designated by H_i. The domain (H_i ; +) is a subgroup of H.

$$H_i := \{ h_{i(s)} \in H \mid h_{i(s)}(i) = g_{i(s)} \quad \wedge \quad \bigwedge_{k \neq i} h_{i(s)}(k) = 0_k \}$$

Proof : The domain (H_i ; +) is a subgroup of H.

(1) The set H_i contains the identity element h_0 of H, since $h_0(k) = 0_k$ for every $k \in I$.

(2) With the mappings $h_{i(r)}$ and $h_{i(s)}$, H_i also contains their sum $h_{i(r)} + h_{i(s)}$. In fact, for every $k \neq i$:

$$h_{i(r)}(k) + h_{i(s)}(k) = 0_k + 0_k = 0_k = h_0(k)$$

For the index i, $g_{i(r)} + g_{i(s)}$ is an element $g_{i(t)}$ of the group G_i, and hence $h_{i(t)}$ is a mapping in H_i :

$$h_{i(r)}(i) + h_{i(s)}(i) = g_{i(r)} + g_{i(s)} = g_{i(t)} = h_{i(t)}(i)$$

(3) The inverse $-h_{i(s)}$ of $h_{i(s)}$ is a mapping in H_i. In fact, $-h_{i(s)}(k) = -0_k = 0_k$ for every index $k \neq i$. For the index i, the group G_i contains $g_{i(s)}$ and therefore also contains $-g_{i(s)}$.

Direct sums of index mappings : The group H is the direct sum of the subgroups H_i. Every index mapping h in the group (H ; +) is the unique sum of mappings, one each from a finite number of subgroups H_i.

$$H = \bigoplus_{i \in I} H_i$$

Proof : The group H is the direct sum of the subgroups H_i.

(1) Every mapping $h \in H$ maps the index $i \in I$ to an element of the group G_i. By hypothesis, $h(i) = 0_i$ for all but a finite number of $i \in I$. Assume $h(i) \neq 0_i$ for $i = c_1,...,c_n$.

$$\begin{array}{lcllc} h(c_r) & = & g_{c_r(s)} & \text{for} & c_r \in \{c_1,...,c_n\} \\ h(i) & = & 0_i & \text{for} & i \notin \{c_1,...,c_n\} \end{array}$$

(2) In the group H_{c_r}, there is exactly one mapping $h_{c_r(s)}$ which also maps the index c_r to the element $g_{c_r(s)}$ of G_{c_r}. This mapping is designated by h_r :

$$\begin{array}{lcllc} h_r(c_r) & = & g_{c_r(s)} & \text{for} & r = 1,...,n \\ h_r(i) & = & 0_i & \text{for} & i \neq c_r \end{array}$$

(3) Every index $c_r \in \{c_1,...,c_n\}$ corresponds to a mapping $h_r \in H_{c_r}$. The image under the mapping h is the sum of the images under the mappings $h_1,...,h_n$:

$$\begin{array}{lcllll} h(i) & = & h_1(i) + ... + h_r(i) + ... + h_n(i) & & \\ g_{c_r(s)} & = & 0_{c_r} + ... + g_{c_r(s)} + ... + 0_{c_r} & \text{for} & c_r \in \{c_1,...,c_n\} \\ 0_i & = & 0_i + ... + 0_i + ... + 0_i & \text{for} & i \notin \{c_1,...,c_n\} \end{array}$$

It follows from the definition of the sum of mappings that

$$h = h_1 + ... + h_r + ... + h_n$$

Thus every $h \in H$ may be represented as a finite sum of mappings $h_r \in H_{c_r}$.

(4) It remains to be shown that different valuations of the sum lead to different elements of H. Two different valuations differ in at least one summand, for example in the summand from H_k. These different summands are different index mappings. For the index k, these mappings yield different elements of G_k. All remaining summands are index mappings from some H_j with $j \neq k$ and therefore yield the value 0_k for k. Altogether, the two valuations thus yield different elements of G_k for the index k. Hence different valuations of the sum lead to different index mappings in H.

(5) It follows from (3) and (4) that every element of the group H is a unique sum of one element each from a finite number of subgroups H_r of H.

Isomorphism of the summands : The subgroup H_i of H and the abelian group G_i are isomorphic, since the following mapping is bijective and homomorphic :

$$f_i : H_i \to G_i \quad \text{with} \quad f_i(h_{i(s)}) = g_{i(s)} \quad \text{and} \quad h_{i(s)}(i) = g_{i(s)}$$

(1) Let the images of the elements $h_{i(r)}$ and $h_{i(s)}$ of H_i be equal, that is $f_i(h_{i(r)}) = f_i(h_{i(s)})$. Then $g_{i(r)} = g_{i(s)}$, and hence also $h_{i(r)}(i) = h_{i(s)}(i)$. For every $k \neq i$, by definition $h_{i(r)}(k) = h_{i(s)}(k) = 0_k$. Hence the elements $h_{i(r)}$ and $h_{i(s)}$ are also equal. The mapping f_i is surjective, since by definition H_i contains exactly the mappings for all elements of G_i. Hence the mapping f_i is bijective.

(2) The mapping f_i is homomorphic, since for elements $h_{i(r)}$ and $h_{i(s)}$:

$$f_i(h_{i(r)} + h_{i(s)}) = (h_{i(r)} + h_{i(s)})(i) = h_{i(r)}(i) + h_{i(s)}(i) = f_i(h_{i(r)}) + f_i(h_{i(s)})$$

Construction : Let a set of abelian groups $(G_i\,;+)$ with $i \in I$ be given. Then for every group G_i the subgroup H_i of H may be constructed. Since G_i and H_i are isomorphic, the sum table for H_i is the sum table for G_i with relabeled elements. The sum table for the group $H = \oplus H_i$ is constructed from the sum tables of the subgroups H_i.

$$H = \oplus H_i \quad \wedge \quad H_i \cong G_i$$

$$H_i = \{ h_{i(s)} \mid h_{i(s)}(i) = g_{i(s)} \ \wedge \ \bigwedge_{k \neq i} h_{i(s)}(k) = 0_k \}$$

Example 1 : Construction of Klein's four-group

The cyclic groups $G_1 = \{0_1, a\}$ and $G_2 = \{0_2, b\}$ are used to construct a new group $H = H_1 \oplus H_2$ with $H_i \cong G_i$. The groups G_i are abelian and have the following sum tables :

G_1

+	0_1	a
0_1	0_1	a
a	a	0_1

G_2

+	0_2	b
0_2	0_2	b
b	b	0_2

The subgroups $H_1 = \{h_0, h_{1(1)}\}$ and $H_2 = \{h_0, h_{2(1)}\}$ contain the following mappings of the index set $I = \{1, 2\}$:

$$h_0(1) = 0_1 \qquad h_{1(1)}(1) = a \qquad h_{2(1)}(1) = 0_1$$

$$h_0(2) = 0_2 \qquad h_{1(1)}(2) = 0_2 \qquad h_{2(1)}(2) = b$$

The subgroups H_1 and H_2 are isomorphic to G_1 and G_2. The sum tables of H_i and G_i differ only in the designations of the elements :

H_1

$+$	h_0	$h_{1(1)}$
h_0	h_0	$h_{1(1)}$
$h_{1(1)}$	$h_{1(1)}$	h_0

H_2

$+$	h_0	$h_{2(1)}$
h_0	h_0	$h_{2(1)}$
$h_{2(1)}$	$h_{2(1)}$	h_0

The elements of $H = H_1 \oplus H_2$ are determined as sums :

$$\begin{aligned} h_0 &= h_0 + h_0 \\ h_1 &= h_{1(1)} + h_0 \\ h_2 &= h_0 + h_{2(1)} \\ h_3 &= h_{1(1)} + h_{2(1)} \end{aligned}$$

The sum table for H is constructed from the sum tables of H_1 and H_2, for example :

$$\begin{aligned} h_1 + h_2 &= h_2 + h_1 = h_{1(1)} + h_{2(1)} = h_3 \\ h_1 + h_3 &= h_3 + h_1 = h_{1(1)} + h_{1(1)} + h_{2(1)} = h_0 + h_{2(1)} = h_2 \\ h_2 + h_3 &= h_3 + h_2 = h_{1(1)} + h_{2(1)} + h_{2(1)} = h_{1(1)} + h_0 = h_1 \end{aligned}$$

H

$+$	h_0	h_1	h_2	h_3
h_0	h_0	h_1	h_2	h_3
h_1	h_1	h_0	h_3	h_2
h_2	h_2	h_3	h_0	h_1
h_3	h_3	h_2	h_1	h_0

Direct sums of infinite cyclic groups : The construction of an abelian group H from given abelian groups $(G_i \, ; +)$ with $i \in I$ is specialized to infinite cyclic groups $G_i = gp(b_i)$. To simplify the notation, a finite index set $I = \{1,...,n\}$ is considered. In analogy with the general case, the group H of index mappings is represented as a direct sum of subgroups H_i. Every subgroup H_i is isomorphic to a given group G_i and is uniquely determined by the generating element b_i of G_i. The group H is isomorphic to a free abelian group F with the basis $B := \{b_i \mid i \in I\}$.

$$\begin{array}{lll} H = \oplus H_i & \text{with} & H_i \cong G_i = gp(b_i) \\ H \cong F = gp(B) & \text{with} & B = \{b_i \mid i \in I\} \end{array}$$

Proof : Direct sums of infinite cyclic groups

(1) Every cyclic group is abelian, so that $G_i = gp(b_i)$ is abelian. An index mapping $h : I \to \cup G_i$ maps every index $i \in I$ to an image G_i. The index mapping which maps the index i to the image $s_i\, b_i$ with $s_i \in \mathbb{Z}$ is designated by $h_{s_1 \ldots s_n}$. The index mappings form a group H :

$$H := \{ h_{s_1 \ldots s_n} \mid h_{s_1 \ldots s_n} : I \to \cup G_i \quad \text{with} \quad h_{s_1 \ldots s_n}(i) = s_i\, b_i \}$$

(2) In the group (H ; +) of index mappings, there is exactly one mapping which maps the index i to the element sb_i of G_i and every index $k \neq i$ to the identity element 0_k of the group G_k. This mapping is designated by $h_{i(s)}$. The set of mappings $h_{i(s)}$ for all elements $sb_i \in G_i$ is designated by H_i. Then, as was already proved for abelian groups in general, $(H_i ; +)$ is a subgroup of H. The elements of H_i are added by adding their indices.

$$H_i := \{ h_{i(s)} \in H \mid h_{i(s)}(i) = sb_i \quad \wedge \quad \bigwedge_{k \neq i} h_{i(s)}(k) = 0_k \}$$

$$h_{i(s)}(i) + h_{i(u)}(i) = sb_i + ub_i = (s+u)b_i = h_{i(s+u)}(i)$$

$$h_{i(s)}(k) + h_{i(u)}(k) = 0_k = h_{i(s+u)}(k) \quad \text{for} \quad k \neq i$$

(3) For the general construction, it was proved that every element of the group H is a unique valuation of the direct sum $\oplus H_i$. The representation of the element $h_{s_1 \ldots s_n}$ as a sum contains the summand $h_{i(s_i)}$ from the subgroup H_i :

$$h_{s_1 \ldots s_n} = h_{1(s_1)} + \ldots + h_{n(s_n)}$$

The rule for adding elements in H is obtained by adding their representations as direct sums :

$$\begin{aligned} h_{s_1 \ldots s_n} + h_{u_1 \ldots u_n} &= h_{1(s_1)} + \ldots + h_{n(s_n)} + h_{1(u_1)} + \ldots + h_{n(u_n)} \\ &= h_{1(s_1)} + h_{1(u_1)} + \ldots + h_{n(s_n)} + h_{n(u_n)} \\ &= h_{1(s_1+u_1)} + \ldots + h_{n(s_n+u_n)} \\ h_{s_1 \ldots s_n} + h_{u_1 \ldots u_n} &= h_{(s_1+u_1) \ldots (s_n+u_n)} \end{aligned}$$

(4) To prove the isomorphism of the group H to a free abelian group F with the basis $B = \{ b_i \mid i \in I \}$, the following mapping from H to F is considered :

$$f : H \to F \qquad \text{with} \qquad f(h_{s_1 \ldots s_n}) = s_1\, b_1 + \ldots + s_n\, b_n$$

The mapping f is bijective. In fact, if two images $f(h_{s_1 \ldots s_n})$ and $f(h_{u_1 \ldots u_n})$ are equal, this implies $s_1\, b_1 + \ldots + s_n\, b_n = u_1\, b_1 + \ldots + u_n\, b_n$, and hence $s_i = u_i$, so that the preimages $h_{s_1 \ldots s_n}$ and $h_{u_1 \ldots u_n}$ are equal. Thus the mapping f is injective. The mapping f is also surjective, since every element of F may be represented in the form $s_1\, b_1 + \ldots + s_n\, b_n$ and therefore possesses a preimage $h_{s_1 \ldots s_n} \in H$. Thus f is bijective. The mapping f is also homomorphic :

$$\begin{aligned} f(h_{s_1 \dots s_n} + h_{u_1 \dots u_n}) &= f(h_{(s_1+u_1)\dots(s_n+u_n)}) \\ &= (s_1+u_1)b_1 + \dots + (s_n+u_n)b_n \\ &= (s_1 b_1 + \dots + s_n b_n) + \dots + (u_1 b_1 + \dots + u_n b_n) \\ f(h_{s_1 \dots s_n} + h_{u_1 \dots u_n}) &= f(h_{s_1 \dots s_n}) + f(h_{u_1 \dots u_n}) \end{aligned}$$

Since the mapping f is bijective and homomorphic, the groups F and H are isomorphic.

Example 2 : Construction of an infinite abelian group

Let the cyclic groups $G_1 = gp(a)$ and $G_2 = gp(b)$ be infinite. These groups are used to construct a new group $H = H_1 \oplus H_2$ with $H_i \cong G_i$. The group H is shown to be isomorphic to the infinite abelian group F with the basis $\{a, b\}$.

$$\begin{aligned} H_1 &= \{h_{1(s)} \mid h_{1(s)}(1) = sa \quad \wedge \quad h_{1(s)}(2) = 0_2 \quad \wedge \quad s \in \mathbb{Z}\} \\ H_2 &= \{h_{2(s)} \mid h_{2(s)}(1) = 0_1 \quad \wedge \quad h_{2(s)}(2) = sb \quad \wedge \quad s \in \mathbb{Z}\} \end{aligned}$$

The groups H_1 and H_2 are isomorphic to the known groups G_1 and G_2. Hence they possess the sum tables of infinite cyclic groups :

H_1

+	⋯	$h_{1(-1)}$	$h_{1(0)}$	$h_{1(1)}$	⋯
⋮	⋯				
$h_{1(-1)}$		$h_{1(-2)}$	$h_{1(-1)}$	$h_{1(0)}$	
$h_{1(0)}$		$h_{1(-1)}$	$h_{1(0)}$	$h_{1(1)}$	
$h_{1(1)}$		$h_{1(0)}$	$h_{1(1)}$	$h_{1(2)}$	
⋮					⋯

H_2

+	⋯	$h_{2(-1)}$	$h_{2(0)}$	$h_{2(1)}$	⋯
⋮	⋯				
$h_{2(-1)}$		$h_{2(-2)}$	$h_{2(-1)}$	$h_{2(0)}$	
$h_{2(0)}$		$h_{2(-1)}$	$h_{2(0)}$	$h_{2(1)}$	
$h_{2(1)}$		$h_{2(0)}$	$h_{2(1)}$	$h_{2(2)}$	
⋮					⋯

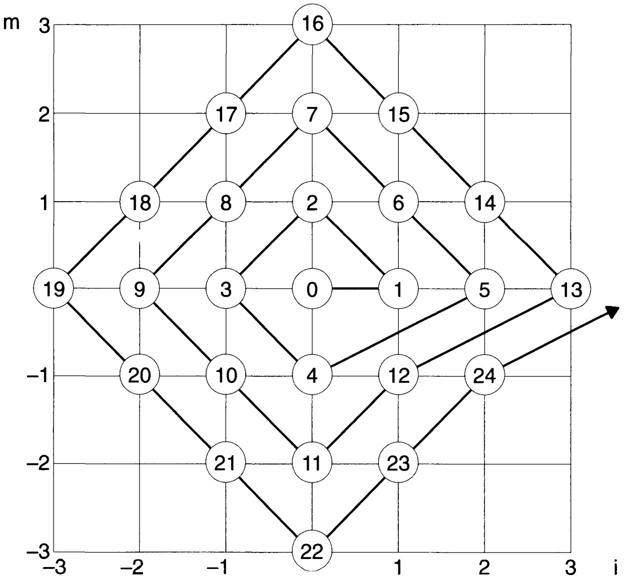
The sum table for the elements $h_{im} = h_{1(i)} + h_{2(m)}$ of H is determined by adding the indices of the elements (for notational reasons only non-negative indices are shown) :

$$h_{im} + h_{kn} = h_{(i+k)\,(m+n)}$$

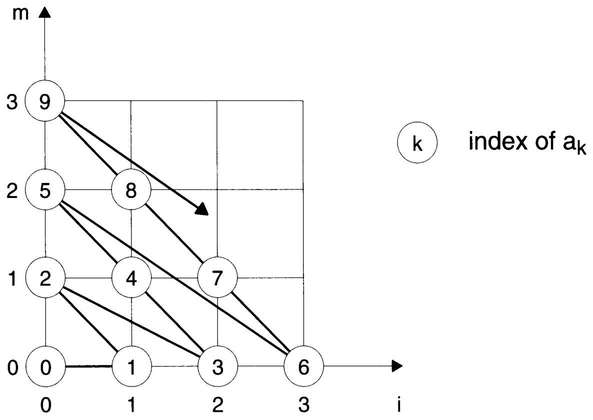
H

+	···	h_{00}	h_{10}	h_{01}	h_{20}	h_{11}	h_{02}	h_{30}	h_{21}	···
⋮	···									
h_{00}		h_{00}	h_{10}	h_{01}	h_{20}	h_{11}	h_{02}	h_{30}	h_{21}	
h_{10}		h_{10}	h_{20}	h_{11}	h_{30}	h_{21}	h_{12}	h_{40}	h_{31}	
h_{01}		h_{01}	h_{11}	h_{02}	h_{21}	h_{12}	h_{03}	h_{31}	h_{22}	
h_{20}		h_{20}	h_{30}	h_{21}	h_{40}	h_{31}	h_{22}	h_{50}	h_{41}	
h_{11}		h_{11}	h_{21}	h_{12}	h_{31}	h_{22}	h_{13}	h_{41}	h_{32}	
h_{02}		h_{02}	h_{12}	h_{03}	h_{22}	h_{13}	h_{04}	h_{32}	h_{23}	
h_{30}		h_{30}	h_{40}	h_{31}	h_{50}	h_{41}	h_{32}	h_{60}	h_{51}	
h_{21}		h_{21}	h_{31}	h_{22}	h_{41}	h_{32}	h_{23}	h_{51}	h_{42}	
⋮										···

The mapping $f : H \rightarrow F$ with $f(h_{im}) = ia + mb$ isomorphically maps the group H to the free abelian group F with the basis {a, b}. The elements of F may be numbered according to the following scheme :



If the indices i and m are restricted to non-negative values instead, the following numbering may be chosen :



The elements of F are designated by a_k. The indices i and m are mapped to the index k according to the numbering. Then the free abelian group F has the following sum table :

F

+	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...
a_0	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	...
a_1	a_1	a_3	a_4	a_6	a_7	a_8	a_{10}	a_{11}	
a_2	a_2	a_4	a_5	a_7	a_8	a_9	a_{11}	a_{12}	
a_3	a_3	a_6	a_7	a_{10}	a_{11}	a_{12}	a_{15}	a_{16}	
a_4	a_4	a_7	a_8	a_{11}	a_{12}	a_{13}	a_{16}	a_{17}	
a_5	a_5	a_8	a_9	a_{12}	a_{13}	a_{14}	a_{17}	a_{18}	
a_6	a_6	a_{10}	a_{11}	a_{15}	a_{16}	a_{17}	a_{21}	a_{22}	
a_7	a_7	a_{11}	a_{12}	a_{16}	a_{17}	a_{18}	a_{22}	a_{23}	
⋮	⋮								...

7.6.6 DECOMPOSITIONS OF ABELIAN GROUPS

Introduction : The preceding section deals with the construction of an abelian group from given abelian groups. The question arises whether, conversely, a given abelian group can always be represented as a direct sum of subgroups.

To decompose an abelian group into a direct sum of subgroups, a mapping from a free abelian group to the given abelian group is studied. According to the definition in Section 7.6.3, every free abelian group F has a basis B. Every element of the basis B generates an infinite cyclic group. Every element of F is a linear combination of the basis elements. A free abelian group is an abelian group of infinite degree (see Section 7.6.2).

If the abelian group G to be decomposed is a free group, the decomposition is achieved by determining a basis B of G. If the group G is of mixed degree, a free group F with a basis B and a bijective mapping $\theta : B \to Y$ with $\theta(b_i) = y_i$ to a generating set Y of G is chosen. The mapping θ leads to a homomorphic mapping $f : F \to G$ with kernel N; the group G and the quotient set F/N are isomorphic. This property is used in the proof of the fundamental theorem for abelian groups, which asserts that every finitely generated abelian group is a direct sum of a finite number of cyclic groups. However, this decomposition is not unique. Unique decompositions of abelian groups are treated in Section 7.9.

The last part of this section treats the decomposition of scaled abelian groups. This decomposition is required for the study of abelian groups in Section 7.9.

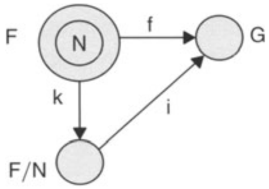
Homomorphic mapping of a free abelian group : Let (F ; +) be a free abelian group, and let $B = \{b_i \mid i \in I\}$ be a basis of F. To define a homomorphic mapping $f : F \to G$ from the free abelian group F to an abelian group G, it suffices to define a mapping $\theta : B \to Y$ from the basis B to a generating set Y of G. Every element a of F is by definition a unique linear combination of basis elements b_i. The image f(a) is defined to be the corresponding linear combination of the images y_i of the basis elements.

$$\begin{array}{llrcll}
\theta : B \to Y & \text{with} & \theta(b_i) & = & y_i & \\
f : F \to G & \text{with} & a & = & n_1 b_1 + \ldots + n_s b_s & n_i \in \mathbb{Z} \\
 & & f(a) & = & n_1 y_1 + \ldots + n_s y_s &
\end{array}$$

The mapping f is homomorphic, since for elements $a = a_1 b_1 + \ldots + a_s b_s$ and $c = c_1 b_1 + \ldots + c_s b_s$ of F :

$$\begin{aligned} f(a+c) &= f((a_1+c_1)b_1 + \ldots + (a_s+c_s)b_s) \\ &= (a_1+c_1)y_1 + \ldots + (a_s+c_s)y_s \\ &= (a_1 y_1 + \ldots + a_s y_s) + (c_1 y_1 + \ldots + c_s y_s) \\ &= f(a_1 b_1 + \ldots + a_s b_s) + f(c_1 b_1 + \ldots + c_s b_s) \\ f(a+c) &= f(a) + f(c) \end{aligned}$$

First isomorphism theorem for abelian groups : Let $(G;+)$ be an abelian group of mixed degree with a generating set $Y = \{y_i \mid i \in I\}$, and let $(F;+)$ be a free abelian group with a basis $B = \{b_i \mid i \in I\}$ for the same index set I. Then the bijective mapping $\theta : B \rightarrow Y$ with $\theta(b_i) = y_i$ induces a homomorphic mapping $f : F \rightarrow G$ with kernel N ; the group G and the quotient group F/N are isomorphic.



$$\begin{aligned} &\theta : B \rightarrow Y \quad \text{with} \quad \theta(b_i) = y_i \\ &f : F \rightarrow G \quad \text{with} \quad f(a) = g \quad \text{and} \quad a = n_1 b_1 + \ldots + n_s b_s \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad\; g = n_1 y_1 + \ldots + n_s y_s \end{aligned}$$

Proof : First isomorphism theorem for abelian groups

Since the group G is of mixed degree, it also contains elements of finite order, for example the elements $g_1, g_2, \ldots$ of order $s_1, s_2, \ldots$. Then by definition $s_1 g_1 = s_2 g_2 = \ldots = 0_G$. Let the preimage of g_i be a_i. Since the mapping f is homomorphic, every linear combination of $s_1 a_1, s_2 a_2, \ldots$ is also mapped to the identity element 0_G. These linear combinations form the kernel N of the mapping f. According to Section 7.5.2, the kernel of the homomorphic mapping f is a normal subgroup in G. Since f is surjective, the general first isomorphism theorem in Section 7.5.3 shows that the group G and the quotient group F/N are isomorphic.

Fundamental theorem for abelian groups : Let $(G;+)$ be a finitely generated abelian group. Then there are elements $t_1, \ldots, t_m$ and $u_1, \ldots, u_n$ in the group G such that G is the direct sum of the cyclic subgroups generated by $t_1, \ldots, t_m$ and $u_1, \ldots, u_n$. The orders of the finite summands $T_i = gp(t_i)$ form a divisor chain. The summands $U_i = gp(u_i)$ are infinite.

$$G = T_1 \oplus \ldots \oplus T_m \oplus U_1 \oplus \ldots \oplus U_n \qquad m,n \in \mathbb{N}$$

$$\text{ord } T_i \mid \text{ord } T_{i+1} \qquad 1 \le i < m$$

m number of summands $T_i = gp(t_i)$ of finite order

n number of summands $U_i = gp(u_i)$ of infinite order

If the direct sum contains no finite summands, then G is a free abelian group. If the group G is finite, then the direct sum contains only finite summands. In that case, its order is the product of the orders of the summands.

$m = 0 \Rightarrow$ G is a free abelian group

$n = 0 \Rightarrow \text{ord } G = \text{ord } T_1 \cdot \ldots \cdot \text{ord } T_m$

The decomposition of a finitely generated abelian group into a direct sum of cyclic subgroups is not unique. The decomposition in the fundamental theorem for abelian groups contains the least possible number of cyclic summands. Other decompositions are treated in Section 7.9.

Proof : Fundamental theorem for abelian groups

(1) By the first isomorphism theorem for abelian groups, the abelian group (G ; +) is isomorphic to the quotient group F/N of a free abelian group (F ; +). Since G is finitely generated and its generating set is bijectively mapped to the basis B of F, it follows that $B = \{b_1,...,b_s\}$ is finite. By Section 7.6.3, the basis B may be chosen such that the subgroup N of F has a minimal generating set $E = \{c_1 b_1,...,c_s b_s\}$ with $c_i \in \mathbb{N}$ and $c_{r+1} = ... = c_s = 0$ with $r \le s$.

$$G \cong F/N$$
$$F = gp(b_1) \oplus \ldots \oplus gp(b_s) = F_1 \oplus \ldots \oplus F_s$$
$$N = gp(c_1 b_1) \oplus \ldots \oplus gp(c_s b_s) = N_1 \oplus \ldots \oplus N_s$$

(2) The subgroup $N_i = gp(c_i b_i)$ is a normal subgroup of the abelian group $F_i = gp(b_i)$. Hence there is a natural homomorphism $f_i : F_i \to F_i / N_i$ with the kernel N_i.

(3) A mapping $f : \oplus F_i \to \oplus(F_i / N_i)$ is defined for the group $F = \oplus F_i$. The element $a = n_1 b_1 + ... + n_s b_s$ in $\oplus F_i$ is mapped to the element $f_1(n_1 b_1) + ... + f_s(n_s b_s)$. The image is unique, since by definition every element of the direct sum $\oplus F_i$ corresponds to a unique sum $n_1 b_1 + ... + n_s b_s$ and the image $f_i(n_i b_i)$ of every term is also unique.

(4) The mapping $f : \oplus F_i \rightarrow \oplus(F_i / N_i)$ is homomorphic, since for elements $a = a_1 b_1 + ... + a_s b_s$ and $u = u_1 b_1 + ... + u_s b_s$ of $\oplus F_i$:

$$\begin{aligned} f(a+u) &= f\,((a_1 + u_1)b_1 + ... + (a_s + u_s)b_s) \\ &= f_1((a_1 + u_1)b_1) + ... + f_s((a_s + u_s)b_s) \\ &= f_1(a_1 b_1) + ... + f_s(a_s b_s) + f_1(u_1 b_1) + ... + f_s(u_s b_s) \\ &= f\,(a_1 b_1 + ... + a_s b_s) + f(u_1 b_1 + ... + u_s b_s) \\ f(a+u) &= f(a) + f(u) \end{aligned}$$

(5) The kernel of the mapping $f : \oplus F_i \rightarrow \oplus(F_i / N_i)$ is $N = \oplus N_i$. Every element $n \in N$ must satisfy the condition $f(n) = 0$. Since the element n is a unique direct sum $n = n_1 b_1 + ... + n_s b_s$, this condition takes the form $f(n_1 b_1 + ... + n_s b_s) = f_1(n_1 b_1) + ... + f_s(n_s b_s) = 0$. Since f_i has kernel N_i and the representation of the identity element 0 is unique, this condition is satisfied only for $n_i b_i \in N_i$, and hence $\ker f \subseteq \oplus N_i$. But $\oplus N_i \subseteq \ker f$ also holds, since N_i is the kernel of f_i. Hence $\ker f = \oplus N_i = N$.

(6) By the general first isomorphism theorem in Section 7.5.3, the surjective homomorphic mapping $f : \oplus F_i \rightarrow \oplus(F_i / N_i)$ with kernel N yields the isomorphism $\oplus(F_i / N_i) \cong (\oplus F_i)/N = F/N$. Thus the group $G = F/N$ is isomorphic to the direct sum of the quotient groups F_i / N_i :

$$G \cong \oplus(F_i / N_i) \quad \text{with} \quad F_i / N_i = gp(b_i) / gp(c_i\, b_i)$$

(7) For $c_i \neq 0$, the quotient group F_i / N_i is isomorphic to the group of residue classes modulo c_i (see Section 7.5.4). For $c_{r+1} = ... = c_s = 0$, the quotient group F_i / N_i is isomorphic to the additive group $(\mathbb{Z}\,;+)$ of the integers. Hence the finitely generated abelian group G is isomorphic to the direct sum of $m = r$ cyclic groups $T_i = gp(t_i)$ of order c_i and $n = s - r$ cyclic groups $U_i = gp(u_i)$ of infinite order.

(8) According to Section 7.6.3, the numbers $c_1,...,c_r$ in the minimal generating set $\{c_1 b_1,...,c_s b_s\}$ of the subgroup N form a divisor chain. The numbers c_i are also the orders of the groups T_i. Hence the orders of the subgroups T_i form a divisor chain.

(9) If the direct sum contains no summands of infinite order, that is if $G = T_1 \oplus ... \oplus T_m$, then every element of G is a unique sum of one element each from $T_1,...,T_m$. Each such combination of elements from $T_1,...,T_m$ is an element of G. Hence the order of G is the product of the orders of the summands T_i.

Decomposition of scaled abelian groups : Let an abelian group (G ; +) be the direct sum of m groups $G_1,...,G_m$. Then the group $nG = \{ng \mid g \in G\}$ scaled by the natural number n is the direct sum of m groups $nG_1,...,nG_m$. The quotient group G/nG is isomorphic to the direct sum of m groups $H_1,...,H_m$ with $H_i = G_i / nG_i$.

$$\begin{array}{rclcrcl}
G & = & G_1 \oplus ... \oplus G_m & \text{with} & G_i & = & gp(a_i) \\
nG & = & nG_1 \oplus ... \oplus nG_m & \text{with} & nG_i & = & gp(na_i) \\
G/nG & \cong & H_1 \oplus ... \oplus H_m & \text{with} & H_i & = & G_i/nG_i
\end{array}$$

Proof : Decomposition of scaled abelian groups

(1) Since G is the direct sum of the groups $G_1,...,G_m$, every element $g \in G$ has a unique representation as a sum $g = g_1 + ... + g_m$ with $g_i \in G_i$. Hence also $ng = ng_1 + ... + ng_m$. This representation of ng is unique, since ng is an element of G and ng_i is an element of G_i. Hence nG is the direct sum of the groups $nG_1,...,nG_m$.

(2) The subgroup nG_i is a normal subgroup in G_i. Hence there is a natural homomorphism $f_i : G_i \rightarrow G_i / nG_i$ with kernel nG_i. As in points (3) to (5) of the proof of the fundamental theorem for abelian groups, it follows that the mapping $f : \oplus G_i \rightarrow \oplus (G_i / nG_i)$ is homomorphic with kernel $nG = \oplus nG_i$. As in point (6) of the proof of the fundamental theorem, it follows that the group H := G/nG is isomorphic to the direct sum of the quotient groups $H_i = G_i/nG_i$.

Example 1 : Decomposition of Klein's four-group

Klein's four-group (G ; +) is abelian and is generated by two elements, for example $G = gp(g_1,g_2)$. Its sum table is determined in Example 1 of Section 7.6.5.

+	g_0	g_1	g_2	g_3
g_0	g_0	g_1	g_2	g_3
g_1	g_1	g_0	g_3	g_2
g_2	g_2	g_3	g_0	g_1
g_3	g_3	g_2	g_1	g_0

G

G is to be decomposed into cyclic subgroups. Each of the generating elements g_1,g_2 of G is associated with an element of the basis $B = \{b_1,b_2\}$ of a free abelian group F. Then the mapping $f : F \rightarrow G$ defined in the proof of the fundamental theorem for abelian groups is homomorphic and has the following kernel :

$$N = \{n_1 b_1 + n_2 b_2 \mid g = n_1 g_1 + n_2 g_2 = 0_G\}$$

The sum table of the four-group shows that the condition $g = 0$ is satisfied for even values of n_1 and n_2. Hence the kernel of f is

$$N = \{2m_1b_1 + 2m_2b_2 \mid m_1, m_2 \in \mathbb{Z}\}$$

The sum table of F is shown in Example 2 of Section 7.6.5. If $b_1 = a_1$ and $b_2 = a_2$ are chosen to form a basis, then the groups F and N are defined as follows :

$$F = \{n_1a_1 + n_2a_2 \mid n_i \in \mathbb{Z}\}$$

$$N = \{2n_1a_1 + 2n_2a_2 \mid n_i \in \mathbb{Z}\}$$

The minimal generating set of N is $\{2a_1, 2a_2\}$. The subgroup $N_1 = gp(2a_1)$ is a normal subgroup of the subgroup $F_1 = gp(a_1)$. The cosets of N_1 in F_1 are $[a_0]$ and $[a_1]$. The quotient group F_1 / N_1 is a cyclic group of order 2 generated by $[a_1]$.

+	$[a_0]$	$[a_1]$
$[a_0]$	$[a_0]$	$[a_1]$
$[a_1]$	$[a_1]$	$[a_0]$

F_1/N_1

$$[a_0] = \{ 2na_1 \mid n \in \mathbb{N} \}$$

$$[a_1] = \{(2n+1)a_1 \mid n \in \mathbb{N}\}$$

Analogous statements hold for the subgroups $N_2 = gp(2a_2)$ and $F_2 = gp(a_2)$. By the proof of the fundamental theorem for abelian groups, there is a homomorphic mapping $f : F \to F/N$ with

$$F = F_1 \oplus F_2 = gp(a_1) \oplus gp(a_2)$$

$$F/N \cong (F_1/N_1) \oplus (F_2/N_2) = \{ [a_0], [a_1]\} \oplus \{[a_0], [a_2] \}$$

$$F/N = \{ [a_0], [a_1], [a_2], [a_4] \}$$

The sum table for the group F/N follows from the rule $[a_i] + [a_m] = [a_i + a_m]$ and the sum table for F; for example :

$$[a_1] + [a_2] = [a_1 + a_2] = [a_4]$$

$$[a_1] + [a_4] = [a_1 + a_1 + a_2] = [a_0 + a_2] = [a_2]$$

+	$[a_0]$	$[a_1]$	$[a_2]$	$[a_4]$
$[a_0]$	$[a_0]$	$[a_1]$	$[a_2]$	$[a_4]$
$[a_1]$	$[a_1]$	$[a_0]$	$[a_4]$	$[a_2]$
$[a_2]$	$[a_2]$	$[a_4]$	$[a_0]$	$[a_1]$
$[a_4]$	$[a_4]$	$[a_2]$	$[a_1]$	$[a_0]$

F/N

The sum tables show that the groups G and F/N are isomorphic. The group F/N is the inner direct sum of the cyclic subgroups F_1 / N_1 and F_2 / N_2.

7.7 PERMUTATIONS

7.7.1 INTRODUCTION

Every finite group is isomorphic to a group of permutations. The structure of permutations is therefore studied in detail. The concept of direct sums for abelian groups is replaced by the concept of cycles for permutations. A cycle leaves a part of the permuted set invariant and maps each of the remaining elements to its neighbor (endless belt). It turns out that every permutation can be reduced to a product of disjoint cycles (canonical decomposition of a permutation).

The canonical decomposition of permutations may be used to determine conjugate elements in permutation groups. The transform of each cycle in the decomposition with respect to a permutation ω is determined by replacing its elements a_i by their images $\omega(a_i)$. In a symmetric group S_n, conjugate permutations form an equivalence class. This property is used for the further study of the structure of general groups in Section 7.8.

A decomposition of a permutation into transpositions is obtained from its decomposition into cycles. Every transposition interchanges two elements of the permuted set. The number of transpositions determines the sign of the permutation. This property is used for instance to determine the coordinates of ε-tensors in Chapter 9.

The subgroup of even permutations (alternating group) in a symmetric group determines whether an equation can be solved by radicals and whether a shape can be constructed using only compass and straightedge (Galois theory). The properties of alternating groups are therefore studied and proved in detail.

Finally, the structure of the symmetric group of degree 4 is studied. First all subgroups of S_4 are analyzed. There is only one complete chain $I \triangleleft V_4 \triangleleft A_4 \triangleleft S_4$ of normal subgroups with the alternating group A_4 and Klein's four-group V_4. The conjugate elements and the commutators of S_4 are also determined. This prepares the ground for the general study of the structure of finite groups in Section 7.8.

7.7.2 SYMMETRIC GROUPS

Introduction : Some definitions and properties of permutation groups from preceding sections are summarized in this section. The definition of a permutation is adopted from the introductory example in Section 7.3.1. The isomorphism of permutation groups of equal degree and the isomorphism of every finite group to a permutation group are proved in Section 7.5.4. The geometric interpretation of the permutation groups alluded to in the examples of preceding sections is deepened with the definition of symmetry groups. In contrast to the preceding examples, the isometries of the space $\mathbb{R}^n$ are permutations on infinite sets. Symmetry groups are subgroups of symmetric groups and must not be confused with the latter.

Permutations : For every natural number $n \geq 1$, let $X_n = \{1,...,n\}$ be the set of numbers from 1 to n. A permutation ϕ on the set X_n is a bijective mapping from X_n to itself. The permutation $\phi : X_n \rightarrow X_n$ is represented using the scheme defined in Section 7.3.1. The order of the columns in the scheme is arbitrary.

$$\phi \quad : \quad \begin{bmatrix} 1 & 2 & \dots & n \\ \phi(1) & \phi(2) & \dots & \phi(n) \end{bmatrix}$$

Identity permutation : The identity permutation $i : X_n \rightarrow X_n$ maps every element of X_n to itself. Thus it is represented by the following scheme :

$$i \quad : \quad \begin{bmatrix} 1 & 2 & \dots & n \\ 1 & 2 & \dots & n \end{bmatrix}$$

Inverse permutation : The product of a permutation ϕ with its inverse ϕ^{-1} yields the identity permutation i. The inverse permutation ϕ^{-1} is obtained from ϕ by interchanging the rows of the scheme.

$$\phi^{-1} : \begin{bmatrix} \phi(1) & \phi(2) & \dots & \phi(n) \\ 1 & 2 & \dots & n \end{bmatrix}$$

$$\phi \circ \phi^{-1} = \phi^{-1} \circ \phi = i$$

Symmetric groups : The set of all permutations ϕ on a set X_n with n elements is designated by S_n. Let the composition $\phi_i \circ \phi_m$ of permutations be defined as an inner operation in S_n. The domain $(S_n ; \circ)$ is a group (see Section 7.3.1). The group S_n is called the symmetric group of degree n and is also designated by $S(X_n)$. In Section 7.5.4, the symmetric groups on sets with the same number of elements are shown to be isomorphic.

$$S_n := \{\phi \mid \phi : X_n \rightarrow X_n \quad \wedge \quad \phi \circ \phi^{-1} = i\}$$

Cayley's Theorem : By Cayley's Theorem (see Section 7.5.4) every finite group is isomorphic to a group $(A\,;\circ)$ of permutations. However, the group A is generally not a complete symmetric group. Since for a set X_n with n elements there are n! permutations, the orders of the symmetric groups are 1, 2, 6, 24, 120,... . While there is for instance no symmetric group of order 4, by Cayley's Theorem for every group of order 4 there is an isomorphic group of permutations $(A\,;\circ)$. This group is a subgroup of a symmetric group. For example, the group S_4 of order 24 contains subgroups of order 2, 3, 4, 6, 8 and 12, which are compiled in Section 7.7.6.

Isometries : The symmetric group S_R on the euclidean space $(\mathbb{R}^n\,;d)$ is the set of bijective mappings ϕ from $\mathbb{R}^n$ to itself :

$$S_R := \{\phi \mid \phi : \mathbb{R}^n \to \mathbb{R}^n \quad \wedge \quad \phi \circ \phi^{-1} = i\}$$

A general permutation changes the distance $d(x, y)$ between two points $x, y \in \mathbb{R}^n$. A special permutation which preserves the distance $d(x, y)$ between all pairs of points $x, y \in \mathbb{R}^n$ is called an isometry and is designated by σ. The set I_R of isometries on $\mathbb{R}^n$ is a subgroup of the symmetric group S_R on $\mathbb{R}^n$.

$$\sigma \text{ is an isometry} \quad :\Leftrightarrow \quad \bigwedge_{x,y \in \mathbb{R}^n} (d(x, y) \ = \ d(\sigma(x),\, \sigma(y)))$$

Proof : The isometries I_R on $\mathbb{R}^n$ form a subgroup of the symmetric group S_R. The domain $(I_R\,;\circ)$ has the properties of a group :

(1) The identity permutation i is an element of I_R, since it preserves all distances in $\mathbb{R}^n$.

(2) If I_R contains the isometries ω and σ, then I_R also contains the isometry $\omega \circ \sigma$. In fact, for the isometries ω and σ :

$$\bigwedge_{x,y \in \mathbb{R}^n} (d(x, y) \ = \ d(\sigma(x),\, \sigma(y)))$$

$$\bigwedge_{a,b \in \mathbb{R}^n} (d(a, b) \ = \ d(\omega(a),\, \omega(b)))$$

Choosing $a = \sigma(x)$ and $b = \sigma(y)$, it follows by substitution that

$$\bigwedge_{x,y \in \mathbb{R}^n} (d(x, y) \ = \ d(\omega\circ\sigma(x),\, \omega\circ\sigma(y))) \quad \Rightarrow \quad \omega\circ\sigma \in I_R$$

(3) For every isometry σ, the inverse σ^{-1} is also contained in I_R. In fact, since every permutation σ is bijective, for all $a, b \in \mathbb{R}^n$ there are points $x, y \in \mathbb{R}^n$ such that $\sigma(x) = a$ and $\sigma(y) = b$. Since σ is an isometry, it follows that $d(x, y) = d(\sigma(x), \sigma(y)) = d(a, b)$. With $i = \sigma^{-1} \circ \sigma$, this implies $d(x, y) = d(i(x), i(y)) = d(\sigma^{-1}\circ\sigma(x),\, \sigma^{-1}\circ\sigma(y)) = d(\sigma^{-1}(a),\, \sigma^{-1}(b))$. Combining these results yields $d(a, b) = d(\sigma^{-1}(a), \sigma^{-1}(b))$. Hence σ^{-1} is an element of I_R.

Symmetry groups : Let $A \subset \mathbb{R}^n$ be a shape in the euclidean space $(\mathbb{R}^n ; d)$. In general, an isometry $\sigma \in I_R$ does not map the shape A to itself. The set I_A of isometries which map every point of A to a point of A (and every point of $\mathbb{R}^n - A$ to a point of $\mathbb{R}^n - A$) is called the symmetry group of this shape. The symmetry group I_A of the shape A is a subgroup of the isometry group of the space $\mathbb{R}^n$, and hence also a subgroup of the symmetric group S_R on the space $\mathbb{R}^n$:

$$I_A := \{\sigma \in I_R \mid \bigwedge_{x\in A} (\sigma(x) \in A \quad \wedge \quad \sigma^{-1}(x) \in A)\}$$

$$I_A \subset I_R \subset S_R$$

Proof : Group properties of the symmetry group I_A of a shape A

(1) The identity permutation i is an element of I_A, since it leaves all points of A invariant.

(2) If I_A contains the isometries ω and σ, then I_A also contains the isometry $\omega \circ \sigma$. In fact, for the isometries ω and σ :

$$\bigwedge_{x\in A} (\sigma(x) \in A \quad \wedge \quad \sigma^{-1}(x) \in A) \quad \wedge \quad \bigwedge_{y\in A} (\omega(y) \in A \quad \wedge \quad \omega^{-1}(y) \in A)$$

Setting $y = \sigma(x)$ and $x = \omega^{-1}(y)$, it follows from the fact that σ and ω are bijective that the quantifier $\bigwedge$ for $x \in A$ and $y \in A$ applies to the same set of elements. Hence by substituting for y and x, respectively, one obtains :

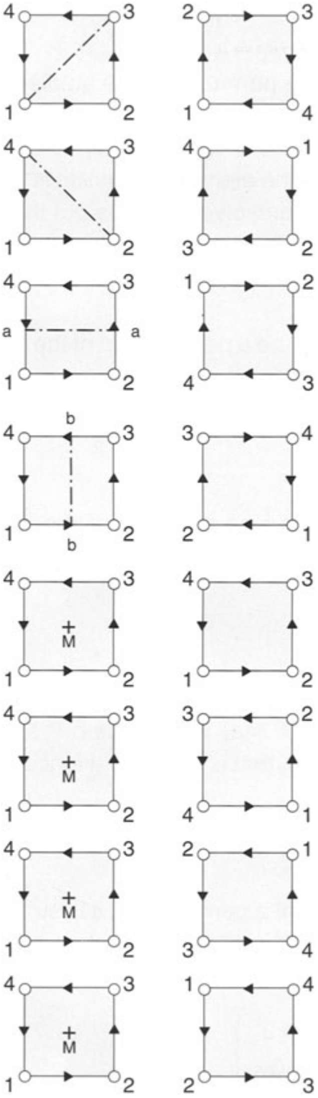
$$\bigwedge_{y\in A} (\omega(y) \in A) \quad \Rightarrow \quad \bigwedge_{x\in A} (\omega \circ \sigma(x) \in A)$$

$$\bigwedge_{x\in A} (\sigma^{-1}(x) \in A) \quad \Rightarrow \quad \bigwedge_{y\in A} (\sigma^{-1} \circ \omega^{-1}(y) \in A)$$

(3) Since it was assumed that the bijective isometry σ maps no points of $\mathbb{R}^n - A$ to A, the inverse σ^{-1} maps no points of A to $\mathbb{R}^n - A$. Hence σ^{-1} is an element of I_A.

Example : Symmetry group of a square in the plane

The covering operations of a square are isometries $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. Such an isometry σ is completely described by the images of three corners of the square. The representation becomes clearer if the image of the fourth corner is also explicitly specified. Then the isometries σ are replaced by permutations $\phi : X_4 \rightarrow X_4$ of the four corners. The symmetry group of a square in the plane contains eight permutations $\{\phi_1, ..., \phi_8\}$. The reflections ϕ_1 to ϕ_4 in $\mathbb{R}^2$ may also be regarded as rotations in $\mathbb{R}^3$. They change the orientation of the boundary of the square.



$\phi_1 = \begin{pmatrix}1 & 2 & 3 & 4\\ 1 & 4 & 3 & 2\end{pmatrix}$ reflection in 1 – 3

$\phi_2 = \begin{pmatrix}1 & 2 & 3 & 4\\ 3 & 2 & 1 & 4\end{pmatrix}$ reflection in 2 – 4

$\phi_3 = \begin{pmatrix}1 & 2 & 3 & 4\\ 4 & 3 & 2 & 1\end{pmatrix}$ reflection in a – a

$\phi_4 = \begin{pmatrix}1 & 2 & 3 & 4\\ 2 & 1 & 4 & 3\end{pmatrix}$ reflection in b – b

$\phi_5 = \begin{pmatrix}1 & 2 & 3 & 4\\ 1 & 2 & 3 & 4\end{pmatrix}$ rotation, 0 degrees

$\phi_6 = \begin{pmatrix}1 & 2 & 3 & 4\\ 4 & 1 & 2 & 3\end{pmatrix}$ rotation, 90 degrees

$\phi_7 = \begin{pmatrix}1 & 2 & 3 & 4\\ 3 & 4 & 1 & 2\end{pmatrix}$ rotation, 180 degrees

$\phi_8 = \begin{pmatrix}1 & 2 & 3 & 4\\ 2 & 3 & 4 & 1\end{pmatrix}$ rotation, 270 degrees

7.7.3 CYCLES

Introduction : The structure of individual finite permutations is studied in this section. The permutations are decomposed into products of disjoint cycles. Each cycle is a permutation which maps each element a_i in a subset $\{a_1,...,a_s\}$ of the underlying set X_n to the element a_{i+1}, except for the element a_s, which is mapped to a_1. All other elements of X_n are mapped to themselves. It turns out that every permutation may be represented as a product of disjoint cycles, which are unique up to their order.

Fixed point of a permutation : Let $\phi : X_n \to X_n$ be a permutation of the finite set $X_n = \{1,...,n\}$. Then an element a of X_n is called a fixed point of the permutation ϕ if it is mapped to itself, that is if $\phi(a) = a$.

$$\text{a is a fixed point of } \phi \quad :\Leftrightarrow \quad \phi(a) = a$$

Range of a permutation : Let $\phi : X_n \to X_n$ be a permutation of the finite set $X_n = \{1,...,n\}$. The subset of elements of X_n which are not fixed points of ϕ is called the range of ϕ and is designated by $W[\phi]$.

$$W[\phi] \ := \ \{a \in X_n \mid \phi(a) \neq a\}$$

The permutation ϕ maps the elements of the range $W[\phi]$ to the range $W[\phi]$. In fact, for an arbitrary element a of $W[\phi]$, the image $b := \phi(a) \neq a$. Since $b \neq a$ and the permutation is injective, it follows that $\phi(b) \neq \phi(a)$, that is $\phi(b) \neq b$. Hence b is not a fixed point of ϕ.

$$\phi(W[\phi]) = W[\phi]$$

Representation of permutations : To represent a permutation, it is sufficient to represent it on its range. Only the elements of the range will therefore be shown in the scheme of a permutation in the following.

$$\phi : \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ \phi(1) & 2 & \phi(3) & \phi(4) & 5 \end{bmatrix} = \begin{bmatrix} 1 & 3 & 4 \\ \phi(1) & \phi(3) & \phi(4) \end{bmatrix}$$

Commuting permutations : The composition $\phi_i \circ \phi_m$ of permutations $\phi_i, \phi_m \in S_n$ is generally not commutative. The order of two permutations may, however, be changed if their ranges are disjoint.

$$W[\phi_i] \cap W[\phi_m] = \emptyset \quad \Rightarrow \quad \phi_i \circ \phi_m = \phi_m \circ \phi_i$$

Proof : Commuting permutations

By hypothesis, every element a of the range $W[\phi_i]$ of the permutation ϕ_i is a fixed point of the permutation ϕ_m, that is $\phi_m(a) = a$. The image $b := \phi_i(a)$ is an element of the range $W[\phi_i]$. Hence the order of the permutations may be changed for $a \in W[\phi_i]$:

$$\begin{aligned} \phi_i \circ \phi_m(a) &= \phi_i\,(a) = b \\ \phi_m \circ \phi_i(a) &= \phi_m(b) = b \end{aligned}$$

By hypothesis, every element b of the range $W[\phi_m]$ of the permutation ϕ_m is a fixed point of the permutation ϕ_i, that is $\phi_i(b) = b$. The image $a := \phi_m(b)$ is by definition an element of the range $W[\phi_m]$. Hence the order of the permutations may be changed for $b \in W[\phi_m]$:

$$\begin{aligned} \phi_i \circ \phi_m(b) &= \phi_i\,(a) = a \\ \phi_m \circ \phi_i(b) &= \phi_m(b) = a \end{aligned}$$

Elements which lie in none of the two ranges are fixed points of both permutations. Thus in all cases the permutations are seen to commute : $\phi_i \circ \phi_m = \phi_m \circ \phi_i$.

Cycles : Let $S_n = S(X_n)$ be a symmetric group. A permutation ϕ in S_n is called a cycle if there is a set $\{a_1, ..., a_s\}$ of s elements in X_n with the following properties :

(1) Every element of X_n which does not belong to $\{a_1, ..., a_s\}$ is a fixed point of ϕ.

(2) Every element of $\{a_1, ..., a_{s-1}\}$ is mapped to the element with the next higher index, while a_s is mapped to a_1.

$$\begin{aligned} \phi(a_i) &= a_{i+1} \quad \text{for} \quad i = 1,...,s-1 \\ \phi(a_s) &= a_1 \end{aligned}$$

The short notation $<a_1, ..., a_s>$ is introduced for a cycle ϕ. The identity mapping is designated by $<a_1>$. The notations $<a_1,a_2,a_3>$, $<a_2,a_3,a_1>$, $<a_3,a_1,a_2>$ all describe the same cycle.

Length of a cycle : The cycle $\phi = <a_1, ..., a_s>$ with s elements is said to be of length $s-1$. The length of a cycle ϕ is designated by $L(\phi)$. A cycle of length 1 is called a transposition. A cycle of length 2 is called a three-cycle, as it contains three elements.

Inverse of a cycle : The inverse of the cycle $<a_1, ..., a_s>$ is given by the cycle $<a_s, ..., a_1>$, since $<a_1, ..., a_s> \circ <a_s, ..., a_1>$ is the identity mapping.

$$\begin{aligned} \phi(a_i) &= a_{i+1} \quad \wedge \quad \phi^{-1}(a_{i+1}) = a_i \quad \text{for} \quad i = 1,...,s-1 \\ \phi(a_s) &= a_1 \quad \wedge \quad \phi^{-1}(a_1) = a_s \end{aligned}$$

Equivalent elements of a permutation : The elements a, b in the range $W[\phi]$ of a permutation ϕ are said to be equivalent if a multiple composition of the permutation ϕ maps the element a to the element b. Multiple compositions are considered to include non-positive powers of ϕ. The equivalence of elements is designated by $a \sim b$. The relation $\sim$ is an equivalence relation.

$$a \sim b \quad \Rightarrow \quad \bigvee_{r \in \mathbb{Z}} \phi^r(a) = b$$

Proof : $a \sim b$ is an equivalence relation.

(1) Reflexive : For every element a in the range, $\phi^0(a) = a$, and hence $a \sim a$.

(2) Symmetric : For every equivalent pair of elements $a \sim b$ there is a number r such that $\phi^r(a) = b$. The corresponding composition using the inverse permutation yields $a = \phi^{-r}(b)$, and hence $b \sim a$. Thus for every pair $a \sim b$ the relation also contains the pair $b \sim a$.

(3) Transitive : For equivalent elements $a \sim b$ and $b \sim c$ there are numbers r and s such that $\phi^r(a) = b$ and $\phi^s(b) = c$, hence $\phi^{r+s}(a) = c$ and therefore $a \sim c$.

Orbit of an element : The range $W[\phi]$ of a permutation ϕ is partitioned into classes of equivalent elements. An element $a \in K$ is chosen as the representative of a class K. Since the permutation is finite, there is a least positive number r such that $\phi^r(a) = a$. In the following, the set $H = \{\phi(a), \phi^2(a), ..., \phi^r(a)\}$ is shown to contain exactly the elements of the class K.

Every element of H is an element of K, since it is obtained from a by multiple composition of ϕ. For an arbitrary element $b \in K$, there is by definition a number s such that $b = \phi^s(a)$. For $1 \le s \le r$, b is contained in H. For $s \le 0$ or $s > r$, the number s is represented in remainder form modulo r, that is $s = qr + t$ with $1 \le t \le r$. Then $b = \phi^{qr+t}(a) = \phi^t(a)$. Hence b is an element of H.

The class K is called the orbit of the element a. With the designations $a_i = \phi^i(a)$, the orbit may be written as the set $K = \{a_1, ..., a_r\}$. It is the range of a cycle ϕ_k. This cycle leaves the elements of X_n which do not belong to K invariant. The images of the elements in K for the permutation ϕ and the cycle ϕ_k coincide.

$$\begin{aligned}
\phi_k(a_i) &= a_{i+1} && \text{for} \quad i = 1,...,r-1 \\
\phi_k(a_r) &= a_1 \\
\phi\,(a_i) &= \phi \circ \phi^i(a) = \phi^{i+1}(a) = a_{i+1} \\
\phi\,(a_r) &= \phi \circ \phi^r(a) = \phi(a) \quad = a_1
\end{aligned}$$

Canonical decomposition of a permutation into cycles : Every permutation ϕ of a symmetric group S_n may be represented as a product of disjoint cycles ϕ_i. Each of these cycles is the orbit of an element of X_n. This representation is called the canonical representation of ϕ :

$$\phi \; = \; \phi_1 \circ \phi_2 \circ \ldots \circ \phi_m$$

ϕ_i factor of the canonical representation of ϕ

The canonical representation of ϕ has the following properties, with $< a_1 >$ designating the identity mapping :

(1) The factors ϕ_i are unique, but their order may be changed.

(2) All factors of a permutation $\phi \neq < a_1 >$ are different from $< a_1 >$.

(3) Any two factors have disjoint ranges.

(4) The canonical decomposition of the permutation $\phi = < a_1 >$ is defined to be $< a_1 >$.

Proof : Canonical decomposition of a permutation

(1) The equivalence relation $\sim$ partitions the range $W[\phi]$ of the permutation ϕ into disjoint equivalence classes $K_1, \ldots, K_m$. The equivalence class K_i is the orbit of a representative $a_i \in K_i$ and thus corresponds to a cycle ϕ_i with range $W[\phi_i] = K_i$. The permutation ϕ is therefore the product of cycles $\phi_1, \ldots, \phi_m$. The cycles commute since their ranges are disjoint.

$$\phi \; = \; \phi_1 \circ \phi_2 \circ \ldots \circ \phi_m \qquad \wedge \qquad \phi_i \circ \phi_k \; = \; \phi_k \circ \phi_i$$

(2) The partition of the range $W[\phi]$ into the equivalence classes $K_1, \ldots, K_m$ is unique. If there are two decompositions $\phi = \alpha_1 \circ \ldots \circ \alpha_j$ and $\phi = \beta_1 \circ \ldots \circ \beta_k$ for the permutation ϕ, then one of the cycles α_i and one of the cycles β_i must coincide with ϕ on the orbit K_i. Thus $j = k = m$, and for a suitable numbering also $\alpha_i = \beta_i$. Hence the decomposition of ϕ is unique.

Determination of the canonical decomposition : Let a permutation ϕ of a set X_n be given. To determine the canonical decomposition of ϕ into cycles, choose an arbitrary element $a \in X_n$ and determine the least number $r \in \mathbb{N}'$ for which $\phi^r(a) = a$. If the orbit $K_1 := \{ \phi(a), \phi^2(a), \ldots, \phi^r(a) \}$ does not contain all elements of X_n, choose an arbitrary element b of the remaining set $X_n - K_1$ and determine its orbit $K_2 := \{ \phi(b), \phi^2(b), \ldots, \phi^s(b) \}$. Fixed points of ϕ lead to one-element cycles $<c>$ which correspond to the identity mapping and are not included in the decomposition. The process is continued until all elements of X_n have been taken into account. For example, if the permutation ϕ possesses two disjoint cycles ϕ_1 and ϕ_2, then its canonical decomposition is

$$\phi \; = \; \phi_1 \circ \phi_2 \; = \; < \phi(a), \ldots, \phi^r(a) > \circ < \phi(b), \ldots, \phi^s(b) >$$

Length of a permutation : Let the canonical decomposition of a permutation ϕ of the symmetric group S_n be $\phi = \phi_1 \circ ... \circ \phi_m$. The sum of the lengths of the cycles ϕ_i is called the length of the permutation ϕ and is designated by $L(\phi)$.

$$L(\phi) = \sum_{i=1}^{m} L(\phi_i)$$

Example 1 : Canonical decomposition of a permutation

Let the permutation ϕ on the set $X_6 = \{1,2,...,6\}$ with fixed point 4 be given :

$$\phi : \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 3 & 2 & 4 & 1 & 5 \end{bmatrix}$$

The element 1 of X_6 is chosen and the orbit $\{6,5,1\}$ is determined. In the remaining set $\{2,3,4\}$, the element 2 is chosen and the orbit $\{3,2\}$ is determined. The remaining set $\{4\}$ contains only the fixed point 4; this leads to the identity mapping, which is not included in the decomposition.

$$\phi(1) = 6 \qquad \phi^2(1) = \phi(6) = 5 \qquad \phi^3(1) = \phi(5) = 1$$
$$\phi(2) = 3 \qquad \phi^2(2) = \phi(3) = 2$$

The permutation ϕ may therefore be represented as the product of the cycles ϕ_1 and ϕ_2 :

$$\phi = \phi_1 \circ \phi_2 = <1,5,6> \circ <2,3>$$

The length of the permutation ϕ is $L(\phi) = 2 + 1 = 3$.

Products of cycles : The decomposition of a permutation into a product of cycles is not unique. If arbitrary cycles ω_1 and ω_2 are chosen, their product is a permutation $\phi = \omega_1 \circ \omega_2$. This permutation ϕ has a canonical decomposition $\phi = \phi_1 \circ ... \circ \phi_s$, which is unique up to the order of the cycles. However, its factors are generally different from ω_1 and ω_2.

Example 2 : Canonical decomposition of a product of cycles

Let the cycles $\omega_1 = < 2, 5, 1, 3 >$ and $\omega_2 = < 6, 3, 4, 1 >$ in the set $X_6 = \{ 1,2,...,6 \}$ be given. Their product is a permutation $\phi = \omega_1 \circ \omega_2$, which may be decomposed by the method described above :

$$\phi = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 5 & 2 & 4 & 1 & 6 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 2 & 4 & 1 & 5 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 6 & 5 & 4 & 3 & 1 & 2 \end{bmatrix}$$

The element 1 leads to the orbit $\{6,2,5,1\}$, leaving the set $\{3,4\}$. The element 3 leads to the orbit $\{4, 3\}$. The canonical decomposition is therefore

$$\phi = \phi_1 \circ \phi_2 = \{6,2,5,1\} \circ \{4, 3\}$$

The canonical decomposition can also be carried out without explicitly determining the permutation ϕ, namely by determining the images with the product $\omega_1 \circ \omega_2$:

$$\begin{aligned}
\phi(1) &= \omega_1(\omega_2(1)) = \omega_1(6) = 6 & \qquad \phi(3) &= \omega_1(\omega_2(3)) = \omega_1(4) = 4 \\
\phi(6) &= \omega_1(\omega_2(6)) = \omega_1(3) = 2 & \qquad \phi(4) &= \omega_1(\omega_2(4)) = \omega_1(1) = 3 \\
\phi(2) &= \omega_1(\omega_2(2)) = \omega_1(2) = 5 \\
\phi(5) &= \omega_1(\omega_2(5)) = \omega_1(5) = 1
\end{aligned}$$

7.7.4 CONJUGATE PERMUTATIONS

Introduction : Two permutations are said to be similar if the number of cycles in their canonical decompositions is equal and the cycles have the same length when suitably numbered. In this section, similar permutations are shown to be conjugate in S_n. Conjugate elements of groups are defined in Section 7.4.4. Every permutation group can be partitioned into classes of conjugate elements. This is demonstrated for the symmetric group S_4 in Section 7.7.7.

Cycle form of conjugation : The ω-transform $\phi = \omega \circ \sigma \circ \omega^{-1}$ of a permutation $\sigma \in S_n$ is to be determined. Let the canonical decomposition of σ be $\sigma = \sigma_1 \circ \ldots \circ \sigma_m$. Then the canonical decomposition of ϕ is $\phi = \phi_1 \circ \ldots \circ \phi_m$ with the same number of cycles and $\phi_i = \omega \circ \sigma_i \circ \omega^{-1}$. If σ_i is the cycle $<a_1, \ldots, a_s>$, then ϕ_i is the cycle $<\omega(a_1),\ldots,\omega(a_s)>$.

$$\begin{aligned} \sigma &= \sigma_1 \circ \ldots \circ \sigma_m \\ \phi &= \phi_1 \circ \ldots \circ \phi_m \\ \sigma_i &= <a_1, \ldots, a_s> \quad \Rightarrow \quad \phi_i = <\omega(a_1),\ldots,\omega(a_s)> \end{aligned}$$

Proof : Cycle form of conjugation

(1) The transform $\phi_i = \omega \circ \sigma_i \circ \omega^{-1}$ of the cycle $\sigma_i = <a_1, \ldots, a_s>$ is given by the cycle $\phi_i = <\omega(a_1),\ldots,\omega(a_s)>$. In fact, for every element $a_k \in \sigma_i$ it follows from $\sigma_i(a_k) = a_{k+1}$ and $\sigma_i(a_s) = a_1$ that ϕ_i maps the element $\omega(a_k)$ to $\omega(a_{k+1})$ and $\omega(a_s)$ to $\omega(a_1)$:

$$\begin{aligned} k \neq s : \ \phi_i(\omega(a_k)) &= \omega \circ \sigma_i \circ \omega^{-1}(\omega(a_k)) = \omega \circ \sigma_i(a_k) = \omega(a_{k+1}) \\ k = s : \ \phi_i(\omega(a_s)) &= \omega \circ \sigma_i \circ \omega^{-1}(\omega(a_s)) = \omega \circ \sigma_i(a_s) = \omega(a_1) \end{aligned}$$

Every element $x \in X_n$ which is not contained in the cycle σ_i is a fixed point of σ_i with $\sigma_i(x) = x$. This implies that ϕ_i leaves the element $\omega(x)$ invariant, so that $\omega(x)$ is a fixed point of ϕ_i :

$$\phi_i(\omega(x)) = \omega \circ \sigma_i \circ \omega^{-1}(\omega(x)) = \omega \circ \sigma_i(x) = \omega(x)$$

The permutation ϕ_i therefore maps the elements $\omega(a_1),\ldots,\omega(a_s)$ of the permuted set $\omega(X_n)$ cyclically, while all other elements of $\omega(X_n)$ are fixed points. Hence ϕ_i is a cycle $<\omega(a_1),\ldots,\omega(a_s)>$ in S_n.

(2) The ω-transform $\phi = \omega \circ \sigma \circ \omega^{-1}$ of the permutation σ with the canonical decomposition $\sigma = \sigma_1 \circ \ldots \circ \sigma_m$ is expressed as a product of the ω-transforms $\phi_i = \omega \circ \sigma_i \circ \omega^{-1}$ of the cycles σ_i :

$$\begin{aligned} \phi &= \omega \circ \sigma \circ \omega^{-1} = \omega \circ \sigma_1 \circ \ldots \circ \sigma_m \circ \omega^{-1} \\ \phi &= (\omega \circ \sigma_1 \circ \omega^{-1}) \circ (\omega \circ \sigma_2 \circ \omega^{-1}) \circ \ldots \circ (\omega \circ \sigma_m \circ \omega^{-1}) \\ \phi &= \phi_1 \circ \phi_2 \circ \ldots \circ \phi_m \end{aligned}$$

(3) Since the cycles σ_i of the canonical decomposition of σ are disjoint and the mapping $\omega : X_n \rightarrow X_n$ is bijective, the cycles ϕ_i are also disjoint. Hence $\phi = \phi_1 \circ \phi_2 \circ \ldots \circ \phi_m$ is the canonical decomposition of the permutation ϕ.

Example 1 : Cycle form of conjugate permutations

Let the permutations ω and σ of the set $X_5 = \{1, 2, 3, 4, 5\}$ be given.

$$\omega = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 1 & 3 \end{bmatrix} \qquad \sigma = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 5 & 1 & 3 \end{bmatrix}$$

The canonical decomposition of the permutation σ is $\phi_1 \circ \phi_2 = <1,4> \circ <3,5>$. The ω-transform of σ is determined in cycle form.

$$\begin{aligned} \omega \circ \sigma \circ \omega^{-1} &= (\omega \circ \phi_1 \circ \omega^{-1}) \circ (\omega \circ \phi_2 \circ \omega^{-1}) \\ &= <\omega(1), \omega(4)> \circ <\omega(3), \omega(5)> \\ &= <2,1> \circ <4,3> \end{aligned}$$

To check the result, the transform $\omega \circ \sigma \circ \omega^{-1}$ is calculated directly and compared with the product $<2,1> \circ <4,3>$.

$$\begin{aligned} \omega \circ \sigma \circ \omega^{-1} &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 5 & 4 & 1 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 2 & 5 & 1 & 3 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 4 & 1 & 5 & 3 & 2 \end{bmatrix} \\ &= \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 3 & 5 \end{bmatrix} \end{aligned}$$

$$<2,1> \circ <4,3> = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 3 & 4 & 5 \end{bmatrix} \circ \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 4 & 3 & 5 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 2 & 1 & 4 & 3 & 5 \end{bmatrix}$$

Similar permutations : Let the canonical decompositions of the permutations σ and ϕ in S_n be $\sigma = \sigma_1 \circ \ldots \circ \sigma_k$ and $\phi = \phi_1 \circ \ldots \circ \phi_m$, respectively. The permutations σ and ϕ are said to be similar if their canonical decompositions contain the same number of cycles, that is $k = m$, and the cycles have the same length when suitably numbered, that is $L(\sigma_i) = L(\phi_i)$.

$$\begin{aligned} \sigma, \phi \in S_n \text{ are similar} \quad :\Leftrightarrow \quad \sigma &= \sigma_1 \circ \ldots \circ \sigma_m \quad \wedge \\ \phi &= \phi_1 \circ \ldots \circ \phi_m \quad \wedge \\ L(\sigma_i) &= L(\phi_i) \end{aligned}$$

Classes of conjugate permutations : The permutations ϕ and σ are conjugate in the symmetric group S_n if and only if they are similar. The relation "conjugate" is an equivalence relation and partitions S_n into disjoint classes. Hence similar permutations form a class of conjugate elements of S_n.

Proof : Permutations are conjugate in S_n if and only if they are similar.

(1) Let the permutations ϕ and σ be conjugate. Then there is a permutation $\omega \in S_n$ such that $\phi = \omega \circ \sigma \circ \omega^{-1}$. Let the canonical decomposition of σ be $\sigma_1 \circ \ldots \circ \sigma_m$. Then the canonical decomposition of ϕ is the product $\phi_1 \circ \ldots \circ \phi_m$ with $\phi_i = \omega \circ \sigma_i \circ \omega^{-1}$ (see cycle form of conjugation). The number m of cycles in the decompositions of ϕ and σ is equal, and the lengths of $\sigma_i = \langle a_1, \ldots, a_s \rangle$ and $\phi_i = \langle \omega(a_1), \ldots, \omega(a_s) \rangle$ are also equal. Hence the permutations ϕ and σ are similar.

(2) Let the permutations ϕ and σ be similar. Let their canonical decompositions be $\phi = \phi_1 \circ \ldots \circ \phi_m$ and $\sigma = \sigma_1 \circ \ldots \circ \sigma_m$ with $L(\phi_i) = L(\sigma_i)$. Then there are bijective mappings $\omega_i : W[\sigma_i] \to W[\phi_i]$ for $i = 1, \ldots, m$, and hence there is a bijective mapping $\omega_W : W[\sigma] \to W[\phi]$ between the ranges. Since the ranges $W[\sigma]$ and $W[\phi]$ are equipotent, there is a bijective mapping from the fixed points of σ to the fixed points of ϕ. Altogether, there is therefore a permutation $\omega : X_n \to X_n$ which transforms $\sigma_i = \langle a_1, \ldots, a_s \rangle$ into $\phi_i = \langle \omega(a_1), \ldots, \omega(a_s) \rangle$, and it follows from the cycle form of conjugation that $\phi = \omega \circ \sigma \circ \omega^{-1}$ is the ω-transform of σ.

Example 2 : Classes of conjugate elements of $S_3 = \{\phi_1, \ldots, \phi_6\}$

$$\phi_1 : \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} \qquad \phi_2 : \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{bmatrix} \qquad \phi_3 : \begin{bmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{bmatrix}$$

$$\phi_4 : \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{bmatrix} \qquad \phi_5 : \begin{bmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix} \qquad \phi_6 : \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$$

The cycle decompositions of the permutations ϕ_i are :

$$\begin{array}{llllllll}
\phi_1(1) = 1 & & & : & \phi_1 & = & \langle 1 \rangle \\
\phi_2(1) = 2 & \phi_2(2) = 3 & \phi_2(3) = 1 & : & \phi_2 & = & \langle 1,2,3 \rangle \\
\phi_3(1) = 3 & \phi_3(3) = 2 & \phi_3(2) = 1 & : & \phi_3 & = & \langle 3,2,1 \rangle \\
\phi_4(1) = 2 & \phi_4(2) = 1 & & : & \phi_4 & = & \langle 1,2 \rangle \\
\phi_5(2) = 3 & \phi_5(3) = 2 & & : & \phi_5 & = & \langle 2,3 \rangle \\
\phi_6(1) = 3 & \phi_6(3) = 1 & & : & \phi_6 & = & \langle 3,1 \rangle
\end{array}$$

The three classes of conjugate elements of S_3 are therefore given by the sets $\{\langle 1 \rangle\}$, $\{\langle 1,2 \rangle, \langle 2,3 \rangle, \langle 3,1 \rangle\}$ and $\{\langle 1,2,3 \rangle, \langle 3,2,1 \rangle\}$.

7.7.5 TRANSPOSITIONS

Introduction : In Section 7.7.3, transpositions are defined as cycles of length 1. Every element of a symmetric group can be represented as a product of transpositions. The sign of a permutation is conveniently determined using transpositions. This sign is often required in applications, for example in determining the coordinates of the ε-tensor in Chapter 9.

Decomposition into transpositions : Every element of a symmetric group S_n with $n \geq 2$ may be represented as a product of transpositions. The permutation ϕ is first canonically decomposed into cycles, that is $\phi = \phi_1 \circ ... \circ \phi_m$. Then each cycle is decomposed into transpositions as follows : $\phi_i = <a_1, a_2, ..., a_s> = <a_1, a_2> \circ <a_2, a_3, ..., a_s> = <a_1, a_2> \circ <a_2, a_3> \circ ... \circ <a_{s-1}, a_s>$.

Multiplication by a transposition : The permutation ϕ in the symmetric group $S_n = S(X_n)$ is multiplied by the transposition $<a_1, b_1>$ of the elements a_1 and b_1 of X_n. The difference of the lengths of the resulting permutation $\gamma = <a_1, b_1> \circ \phi$ and of the permutation ϕ is either 1 or –1.

$$\gamma = <a_1, b_1> \circ \phi \quad \Rightarrow \quad L(\gamma) = L(\phi) \pm 1$$

Proof : Multiplication by a transposition

The permutation ϕ is canonically decomposed into cycles : $\phi = \phi_1 \circ ... \circ \phi_m$. The following cases are distinguished according to the position of the elements a_1, b_1 in the cycles :

(1) The elements lie in the same cycle $\phi_i = <a_1, ..., a_r, b_1, ..., b_s>$.

$$\gamma_i = <a_1, b_1> \circ \phi_i = \begin{pmatrix} a_1 & b_1 \\ b_1 & a_1 \end{pmatrix} \circ \begin{pmatrix} a_1 & ... & a_{r-1} & a_r & b_1 & ... & b_{s-1} & b_s \\ a_2 & ... & a_r & b_1 & b_2 & ... & b_s & a_1 \end{pmatrix}$$

$$= \begin{pmatrix} a_1 & ... & a_{r-1} & a_r & b_1 & ... & b_{s-1} & b_s \\ a_2 & ... & a_r & a_1 & b_2 & ... & b_s & b_1 \end{pmatrix}$$

$$= <a_1, ..., a_r> \circ <b_1, ..., b_s>$$

The length of the cycles which do not contain a_1 and b_1 remains the same. The length $L(\phi_i) = r + s - 1$ is replaced by the length $L(\gamma_i) = (r-1) + (s-1)$, so that $L(\gamma) = L(\phi) - 1$.

(2) The elements a_1 and b_1 lie in two different cycles $\phi_1 = <a_1,...,a_r>$ and $\phi_2 = <b_1,...,b_s>$. As demonstrated in case (1), the product $\phi_1 \circ \phi_2$ may be replaced by $<a_1,b_1> \circ <a_1,..., a_r, b_1,...,b_s>$. Then multiplying $\phi_1 \circ \phi_2$ by the transposition $<a_1, b_1>$ yields :

$$\begin{aligned} <a_1, b_1> \circ \phi_1 \circ \phi_2 &= <a_1, b_1> \circ <a_1, b_1> \circ <a_1,...,a_r, b_1,...,b_s> \\ &= <a_1,...,a_r, b_1,...,b_s> \end{aligned}$$

The length of the cycles which do not contain a_1 and b_1 remains the same. $L(\phi_1 \circ \phi_2) = r+s-2$ and $L(<a_1,b_1> \circ \phi_1 \circ \phi_2) = r+s-1$ together yields $L(\gamma) = L(\phi)+1$.

(3) The element a_1 lies in cycle ϕ_1 , the element b_1 is a fixed point. A cycle of length 0 is formed for the fixed point, that is $\phi_2 = <b_1>$. This case is included in case (2) :

$$\begin{aligned} <a_1,b_1> \circ \phi_1 \circ \phi_2 &= <a_1,b_1> \circ <a_1,b_1> \circ <a_1,...,a_r, b_1> \\ &= <a_1,...,a_r, b_1> \end{aligned}$$

(4) Both elements are fixed points. In this case the cycles $\phi_1 = <a_1>$ and $\phi_2 = <b_1>$ of length 0 are formed. This case is also included in case (2).

$$\begin{aligned} <a_1, b_1> \circ <a_1> \circ <b_1> &= <a_1,b_1> \circ <a_1,b_1> \circ <a_1,b_1> \\ &= <a_1,b_1> \end{aligned}$$

Sign of a permutation : A permutation ϕ is said to be even if its length $L(\phi)$ is an even number. Otherwise the permutation ϕ is said to be odd. The number $(-1)^{L(\phi)}$ is called the sign (signum, signature, parity) of the permutation ϕ and is designated by $\operatorname{sgn} \phi$.

$$\phi = \phi_1 \circ ... \circ \phi_m \quad \wedge \quad L(\phi) = \sum_{i=1}^{m} L(\phi_i) \quad \Rightarrow \quad \operatorname{sgn} \phi = (-1)^{L(\phi)}$$

Determination of the sign using transpositions : Let a permutation ϕ be the product of s transpositions $<a_i, b_i>$. Then the permutation is even if the number s is even. Otherwise, the permutation is odd.

$$\phi = <a_1,b_1> \circ ... \circ <a_s, b_s> \quad \Rightarrow \quad \operatorname{sgn} \phi = (-1)^s$$

Proof : Even and odd permutations

(1) If the permutation ϕ is a transposition $<a,b>$, then $L(\phi) = 1$. The permutation ϕ is therefore odd, and the statement holds for $s = 1$.

(2) Let a permutation ϕ be the product of s transpositions. Multiplying ϕ by an additional transposition leads to a permutation σ which is the product of $(s+1)$ transpositions. Since $L(\sigma) = L(\phi) \pm 1$, it follows that $\operatorname{sgn} \sigma = -\operatorname{sgn} \phi$.

(3) Since a permutation with $s = 1$ is odd and the sign alternates for consecutive values of s, all permutations which are the product of an odd number of transpositions are odd. All other permutations are even.

Homomorphism of the symmetric group : The mapping sgn from the symmetric group S_n to the set $\{-1,1\}$ defined by the sign of the permutations in S_n is a homomorphism.

$$\text{sgn} : S_n \rightarrow \{-1,1\}$$

$$\text{sgn}(\phi_1 \circ \phi_2) = \text{sgn}(\phi_1) \circ \text{sgn}(\phi_2)$$

Proof : sgn is a homomorphism.

The permutations are decomposed into transpositions : $\phi_1 = \alpha_1 \circ \ldots \circ \alpha_k$ and $\phi_2 = \beta_1 \circ \ldots \circ \beta_m$. Then $\phi_1 \circ \phi_2 = \alpha_1 \circ \ldots \circ \alpha_k \circ \beta_1 \circ \ldots \circ \beta_m$.

$$\text{sgn}(\phi_1 \circ \phi_2) = (-1)^{k+m} = (-1)^k \cdot (-1)^m = \text{sgn}\,\phi_1 \cdot \text{sgn}\,\phi_2$$

Example : Determination of the sign of a permutation

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

The sign of the permutation ϕ is to be determined. First the orbit of the element 1 is determined :

$$\phi(1) = 7 \quad \phi(7) = 2 \quad \phi(2) = 3 \quad \phi(3) = 4 \quad \phi(4) = 6 \quad \phi(6) = 5 \quad \phi(5) = 1$$

The canonical decomposition of ϕ is the cycle $< 7, 2, 3, 4, 6, 5, 1 >$. Hence a decomposition of ϕ into transpositions is given by $\phi = <7,2> \circ <2,3> \circ <3,4> \circ <4,6> \circ <6,5> \circ <5,1>$. Since the number 6 of transpositions is even, the sign of ϕ is 1. The permutation ϕ is even.

7.7.6 SUBGROUPS OF A SYMMETRIC GROUP

Subgroups of a symmetric group : Let $S_m = S(X_m)$ be a symmetric group. For a subset $X_n \subseteq X_m$, a subset $S_n \subseteq S_m$ is formed. The subset S_n contains every permutation ϕ from S_m for which every element of the difference $X_m - X_n$ is a fixed point. The permutations in S_n form a subgroup of S_m.

Order of permutation groups : The order of the symmetric group S_n is given by the factorial $n!$. The index of the group S_n in the group S_{n+1} is $n+1$.

Proof : The order of permutation groups

(1) The left cosets of the subgroup S_n in the permutation group S_{n+1} are :

$$\phi_a \circ S_n = \{\phi \in S_{n+1} \mid \phi_a \circ \phi_n = \phi \quad \wedge \quad \phi_n \in S_n\}$$

To determine these cosets, consider permutations $\phi_a \in S_{n+1}$ which interchange an element $a \in \{1,...,n+1\}$ with the element $n+1$. In the special case $a = n+1$, the permutation ϕ_a is the identity mapping in S_{n+1}.

$$\phi_a := \begin{pmatrix} a & n+1 \\ n+1 & a \end{pmatrix} \qquad\qquad a = 1,...,n+1$$

The coset $\phi_a \circ S_n$ contains exactly those permutations $\phi \in S_{n+1}$ which map the element $n+1$ to the element a. In fact, if ϕ is an element of $\phi_a \circ S_n$, then $\phi_n \in S_n$ implies :

$$\phi(n+1) = \phi_a \circ \phi_n(n+1) = \phi_a(n+1) = a$$

Conversely, $\phi(n+1) = a$ implies :

$$\phi_a \circ \phi(n+1) = \phi_a(a) = n+1 \quad \Rightarrow \quad \phi_a \circ \phi \in S_n$$

Thus $\phi_a \circ \phi = \phi_n \in S_n$, and hence $\phi = \phi_a^{-1} \circ \phi_n = \phi_a \circ \phi_n$. Thus the $n+1$ left cosets $\phi_a \circ S_n$ form a partition of S_{n+1}. The index of S_n in S_{n+1} is therefore $n+1$.

(2) Lagrange's Theorem yields ord $S_{n+1} = (n+1) \cdot$ ord S_n. It follows by induction that ord $S_n = n!$.

Alternating groups : The even permutations in a symmetric group S_n form a subgroup of S_n. This subgroup is called the alternating group of degree n and is designated by A_n or by $A(X_n)$.

Proof : Group properties of A_n

(1) The identity mapping is even and is therefore contained in A_n.

(2) The product of two even permutations is an even permutation and is therefore contained in A_n.

(3) The inverse of an even permutation is an even permutation and is therefore contained in A_n.

$$\begin{aligned} \phi \circ \phi^{-1} = 1_A \quad &\Rightarrow \quad \text{sgn } \phi \cdot \text{sgn } \phi^{-1} = \text{sgn } 1_A = 1 \\ &\Rightarrow \quad \text{sgn } \phi = \text{sgn } \phi^{-1} \end{aligned}$$

Properties of the alternating groups :

(A1) The alternating group A_n is a normal subgroup in S_n.

(A2) For $n \geq 2$, the index of the alternating group A_n in the symmetric group S_n is 2 and the order of A_n is $\frac{1}{2} n!$.

(A3) For $n \geq 4$, the group A_n is non-abelian.

(A4) For $n \geq 3$, every element of A_n may be represented as a product of three-cycles.

(A5) The alternating group A_n is the commutator subgroup of the symmetric group S_n defined in Section 7.8.2. Hence A_n is a characteristic subgroup of S_n.

(A6) For $n \geq 5$, the three-cycles in A_n form a class of conjugate elements.

(A7) For $n \geq 5$, the group A_n is simple : It contains no proper subgroup which is a normal subgroup of A_n.

Proof : Properties of the alternating groups

(A1) Since sgn is a homomorphism, $\text{sgn}(s \circ a \circ s^{-1}) = \text{sgn } s \circ \text{sgn } a \circ \text{sgn } s^{-1} = \text{sgn } a$ for arbitrary elements $a \in A_n$ and $s \in S_n$. The transformed element $s \circ a \circ s^{-1}$ is therefore an element of A_n. Hence A_n is a normal subgroup in S_n.

(A2) Every even permutation is an element of A_n. Multiplying by the identity mapping <1> does not change the sign of a permutation. Hence the left coset $<1> \circ A_n$ contains all even permutations in S_n.

Multiplication by the transposition <1,2> is admissible for $n \geq 2$ and changes the sign of every permutation. Let the permutation ω in S_n be odd. Then $\phi = <1,2> \circ \omega$ is an element of A_n. But $\phi = <1,2> \circ \omega$ implies $\omega = <1,2> \circ \phi$. Hence the left coset $<1,2> \circ A_n$ contains all odd permutations in S_n.

Every permutation in S_n is either even or odd. Thus the classes $<1> \circ A_n$ and $<1,2> \circ A_n$ form a partition of S_n. Hence the index of A_n in S_n is 2. Since $\text{ord } S_n = n!$, it follows that $\text{ord } A_n = \frac{1}{2} n!$.

(A3) If $n \geq 4$, the group A_n contains the cycles $\phi = <1,2,3>$, $\omega = <1,2,4>$ and $\gamma = <2,4,3>$. The transformation $\omega \circ \phi \circ \omega^{-1} = <\omega(1), \omega(2), \omega(3)> = <2,4,3> = \gamma$ yields $\omega \circ \phi = <1,2,4> \circ <1,2,3> = <2,4,3> \circ <1,2,4> \neq \phi \circ \omega$. Hence A_n is non-abelian.

(A4) Every element ϕ of A_n is an even permutation and therefore the product of an even number of transpositions ω_i. For any combination of elements $a \neq b \neq c \neq d$ from X_n, the product of two transpositions is equal to a product of three-cycles. Hence ϕ is a product of three-cycles :

$$\phi \ = \ (\omega_1 \circ \omega_2) \circ (\omega_3 \circ \omega_4) \circ \ldots$$

$$\begin{array}{lllll}
\omega_m \circ \omega_{m+1} & = & <a,b> \circ <a,b> & = & <a,b,c>^3 \ = \ <1> \\
\omega_m \circ \omega_{m+1} & = & <a,b> \circ <b,c> & = & <b,c,a> \ = \ <a,b,c> \\
\omega_m \circ \omega_{m+1} & = & <a,b> \circ <c,d> & = & <a,b> \circ <b,c> \circ <b,c> \circ <c,d> \\
 & & & = & <a,b,c> \circ <b,c,d>
\end{array}$$

(A5) In the symmetric groups S_1 and S_2, the commutators $\phi \circ \omega \circ \phi^{-1} \circ \omega^{-1}$ yield only the identity element $<1>$: The assertion is true for $n = 1, 2$. For $n \geq 3$, every three-cycle is a commutator in S_n. To prove this, one makes use of the equation $<a,b>^2 = <1>$, that is $<a,b> = <a,b>^{-1}$:

$$\begin{array}{lll}
<a,b,c> & = & <a,c,b> \circ <a,c,b> \\
 & = & <a,c> \circ <c,b> \circ <a,c> \circ <c,b> \\
 & = & <a,c> \circ <c,b> \circ <a,c>^{-1} \circ <c,b>^{-1}
\end{array}$$

Since every permutation $\phi \in A_n$ may be represented as a product of three-cycles and every three-cycle may be represented as a commutator, every permutation $\phi \in A_n$ is a product of commutators and hence an element of the commutator group K of S_n. Thus $A_n \subseteq K$. Since A_n is a normal subgroup in S_n and S_n / A_n is abelian, by property (K2) of commutator subgroups in Section 7.8.2 K is a subset of A_n, that is $K \subseteq A_n$. Hence $A_n \subseteq K \subseteq A_n$, and therefore $K = A_n$.

Proof A6 : For $n \geq 5$, the three-cycles in A_n form a class of conjugate elements.

(1) In Section 7.7.4, it is proved that two permutations are conjugate in S_n if and only if they are similar. The three-cycles $\phi = <a_1, a_2, a_3>$ and $\gamma = <b_1, b_2, b_3>$ are similar and therefore conjugate in S_n. It is to be shown that ϕ and γ are also conjugate in A_n.

For the elements ϕ and γ conjugate in S_n, there is a permutation $\omega_1 \in S_n$ such that $\omega_1 \circ \phi \circ \omega_1^{-1} = \gamma$. For $n \geq 5$, there is a transposition $<c,d>$ in S_n such that the ranges $\{a_1, a_2, a_3\}$ of ϕ and $\{c,d\}$ of $<c,d>$ are disjoint.

Hence the permutations ϕ and $<c,d>$ commute. For the permutation defined by $\omega_2 := \omega_1 \circ <c,d>$, this leads to $\omega_2 \circ \phi \circ \omega_2^{-1} = \omega_1 \circ <c,d> \circ \phi \circ <c,d>^{-1} \circ \omega_1^{-1} = \omega_1 \circ \phi \circ \omega_1^{-1} = \gamma$. Since either ω_1 or ω_2 is an even permutation, ϕ and γ are conjugate elements of A_n.

(2) If the permutation γ is a three-cycle and the permutation ϕ is not, then the permutations are not similar, and hence not conjugate in S_n. Thus ϕ is not an element of the class $[\gamma]$. The class $[\gamma]$ contains only three-cycles.

Proof A7 : The alternating group A_n with $n \geq 5$ is simple.

Property (A4) of A_n shows that every element of A_n with $n \geq 5$ can be represented as a product of three-cycles. In the following it is shown that a normal subgroup $N \neq \{1\}$ of A_n contains all three-cycles of S_n. Hence N contains all elements of A_n. Since $N = A_n$, the group A_n is simple.

(1) The commutator $k = \phi \circ \omega \circ \phi^{-1} \circ \omega^{-1}$ of a three-cycle $\omega \in A_n$ and a permutation $\phi \in N$ is an element of N. In fact, together with ϕ the group N also contains the inverse ϕ^{-1}. It follows from $\omega \circ N = N \circ \omega$ that $\omega \circ \phi^{-1} \circ \omega^{-1} \in N$. Since the group N contains the elements ϕ and $\omega \circ \phi^{-1} \circ \omega^{-1}$, it also contains their product $k = \phi \circ \omega \circ \phi^{-1} \circ \omega^{-1}$.

(2) N is shown to contain a three-cycle. Every permutation $\phi \in N$ with $\phi \neq i$ is even and therefore the product of at least two different transpositions. Hence at least three elements a,b,c are not fixed points of ϕ.

If ϕ contains exactly three elements which are not fixed points, then ϕ is itself a three-cycle $<a,b,c>$. If ϕ contains more than three elements which are not fixed points, ϕ is canonically decomposed into cycles. For each of the possible arrangements of the elements a,b,c in the cycles of the canonical decomposition of ϕ it is shown that for a suitable choice of the three-cycle ω the commutator $k = \phi \circ \omega \circ \phi^{-1} \circ \omega^{-1}$ is either a three-cycle or may be used to construct a commutator which is a three-cycle. By (1), this commutator is an element of N.

(2a) A cycle with more than three elements : $\phi = <a,b,c,d...>$
Choose $\omega = <a,b,c>$, so that $\phi \circ \omega \circ \phi^{-1} = <b,c,d>$
$k = <b,c,d> \circ <c,b,a> = <d,b> \circ <b,c> \circ <c,b> \circ <b,a> = <d,b,a>$
Thus k is a three-cycle in N.

(2b) A cycle with three elements : $\phi = <a,b,c> \circ <d,e,...>$
Choose $\omega = <a,b,d>$, so that $\phi \circ \omega \circ \phi^{-1} = <b,c,e>$
$k = <b,c,e> \circ <d,b,a> = <c,e,b,a,d> \in N$
A three-cycle in N is constructed from k using (2a).

(2c) There are only cycles with two elements: $\phi = \langle a,b \rangle \circ \langle c,d \rangle \circ \ldots$
Case 1 : There is a fixed point $\phi(e) = e$
Choose $\omega = \langle a,c,e \rangle$, so that $\phi \circ \omega \circ \phi^{-1} = \langle b,d,e \rangle$
$k = \langle b,d,e \rangle \circ \langle e,c,a \rangle = \langle b,d,e,c,a \rangle \in N$
A three-cycle in N is constructed from k using (2a).

Case 2 : There is an element e with $\phi(e) \neq e$ and $e \neq a,b,c,d$
Choose $\omega = \langle a,c,e \rangle$, so that $\phi \circ \omega \circ \phi^{-1} = \langle b,d,\phi(e) \rangle$
$k = \langle b,d,\phi(e) \rangle \circ \langle e,c,a \rangle \in N$
A three-cycle in N is constructed from k using (2b).

(3) Let the three-cycle in N determined in (2) be γ. By property (A5), A_n is a characteristic subgroup of S_n. The normal subgroup N of A_n is therefore also a normal subgroup of S_n. Since γ is an element of N, it follows that N contains all elements of S_n conjugate to γ :

$$N = \{\beta \circ \gamma \circ \beta^{-1} \mid \beta \in S_n\}$$

By property (A6), the three-cycles in A_n form a class of conjugate elements for $n \geq 5$. Since $\gamma \in N$ is a three-cycle, the class N of the elements conjugate to γ contains all three-cycles of S_n and hence by (A4) all elements of A_n.

Example 5 : Isomorphism of the symmetry group of the triangle with S_3

The symmetry group of the equilateral triangle described in Example 1 of Section 7.3.2 is isomorphic to the symmetric group S_3. The elements $\{a_0,...,a_5\}$ of the symmetry group are mapped to the permutations $\{p_0,...,p_5\}$ of Example 1 in Section 7.3.1. The product tables of the symmetry group and the symmetric group coincide if the covering operation a_i is mapped to the permutation p_i.

Example 6 : Isomorphism of the symmetry group of the tetrahedron with A_4

The symmetry group of the regular tetrahedron described in Example 2 of Section 7.3.2 is isomorphic to the alternating group A_4. Hence the symmetry group contains only even permutations. This is due to the fact that the odd permutations in the symmetry group S_4 cannot be carried out physically with the three-dimensional tetrahedron as a solid in three-dimensional space.

7.7.7 GROUP STRUCTURE OF THE SYMMETRIC GROUP S_4

Introduction : The symmetric group S_4 contains the 24 permutations of the set of numbers $\{1, 2, 3, 4\}$. All subgroups of S_4 are determined in this section. For this purpose, the product table of S_4 is constructed. Then the cyclic subgroups of S_4 generated by one element are determined. With the generating elements of these subgroups, subgroups of S_4 with two generating elements are determined. Finally, the chains formed by the subgroups are studied.

Construction of the product table : In the following, the permutations in the symmetric group S_4 are represented in a simplified form; the upper row of the scheme defined in Section 7.3.1 is omitted and replaced by a grid.

ϕ_1 =	1 2 3 4	ϕ_7 =	3 1 4 2	ϕ_{13} =	3 2 4 1	ϕ_{19} =	2 1 3 4
ϕ_2 =	2 3 4 1	ϕ_8 =	3 4 2 1	ϕ_{14} =	4 2 1 3	ϕ_{20} =	3 2 1 4
ϕ_3 =	3 4 1 2	ϕ_9 =	2 1 4 3	ϕ_{15} =	2 4 3 1	ϕ_{21} =	4 2 3 1
ϕ_4 =	4 1 2 3	ϕ_{10} =	4 3 1 2	ϕ_{16} =	4 1 3 2	ϕ_{22} =	1 3 2 4
ϕ_5 =	2 4 1 3	ϕ_{11} =	1 3 4 2	ϕ_{17} =	2 3 1 4	ϕ_{23} =	1 4 3 2
ϕ_6 =	4 3 2 1	ϕ_{12} =	1 4 2 3	ϕ_{18} =	3 1 2 4	ϕ_{24} =	1 2 4 3

The inner operation $\phi_i \circ \phi_m$ on the permutations is composition. It leads to the product table for the group $(S_4 ; \circ)$ shown below. Often only the indices of the permutations are shown in the following, such as i instead of ϕ_i.

Determination of the cyclic subgroups : For every element ϕ_i of S_4, the cyclic subgroup is determined using the product table, for example :

$$\text{gp}(\phi_1) = \{\phi_1\} \qquad \text{gp}(\phi_3) = \{\phi_1, \phi_3\}$$

$$\text{gp}(\phi_2) = \{\phi_1, \phi_2, \phi_3, \phi_4\} \qquad \text{gp}(\phi_4) = \{\phi_1, \phi_2, \phi_3, \phi_4\}$$

The example shows that different elements (in this case ϕ_2 and ϕ_4) may generate the same subgroup. The generating element of a subgroup may also occur in a different group (in this case ϕ_3), but in the group $\{\phi_1, \phi_2, \phi_3, \phi_4\}$ it is not a generating element ! In the following, the subgroups are compiled according to their order and identified by designations such as E_1, Z_i, D_m and V_n.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	2	3	4	1	18	23	14	16	20	12	5	19	10	22	7	24	8	21	13	6	11	15	9	17
3	3	4	1	2	21	9	22	24	6	19	18	13	12	15	14	17	16	11	10	23	5	7	20	8
4	4	1	2	3	11	20	15	17	23	13	21	10	19	7	22	8	24	5	12	9	18	14	6	16
5	5	16	22	13	6	7	1	11	21	18	19	2	23	8	10	20	4	24	14	12	3	9	17	15
6	6	20	9	23	7	1	5	19	3	24	14	16	17	11	18	12	13	15	8	2	22	21	4	10
7	7	12	21	17	1	5	6	14	22	15	8	20	4	19	24	2	23	10	11	16	9	3	13	18
8	8	14	19	11	16	24	17	9	10	1	20	7	5	23	4	22	21	2	6	15	12	13	18	3
9	9	23	6	20	22	3	21	10	1	8	15	17	16	18	11	13	12	14	24	4	7	5	2	19
10	10	18	24	15	13	19	12	1	8	9	4	21	22	2	20	5	7	23	3	11	17	16	14	6
11	11	8	14	19	20	15	4	21	18	5	12	1	6	17	13	9	3	16	7	10	2	23	24	22
12	12	21	17	7	10	13	19	2	16	20	1	11	15	3	6	18	14	9	4	5	8	24	22	23
13	13	5	16	22	19	12	10	4	17	23	3	18	14	1	9	11	15	6	2	21	24	8	7	20
14	14	19	11	8	2	18	23	22	15	7	16	6	1	13	17	3	9	12	5	24	20	4	10	21
15	15	10	18	24	4	11	20	7	14	22	17	9	3	12	16	1	6	13	21	8	23	2	19	5
16	16	22	13	5	24	17	8	20	12	2	6	14	18	9	1	15	11	3	23	7	19	10	21	4
17	17	7	12	21	8	16	24	23	13	4	9	15	11	6	3	14	18	1	20	22	10	19	5	2
18	18	24	15	10	23	14	2	5	11	21	13	3	9	16	12	6	1	17	22	19	4	20	8	7
19	19	11	8	14	12	10	13	3	24	6	2	5	7	4	23	21	22	20	1	18	16	17	15	9
20	20	9	23	6	15	4	11	12	2	16	7	8	24	21	5	10	19	22	17	1	14	18	3	13
21	21	17	7	12	9	22	3	18	5	11	10	4	20	24	19	23	2	8	15	13	1	6	16	14
22	22	13	5	16	3	21	9	15	7	14	24	23	2	10	8	4	20	19	18	17	6	1	12	11
23	23	6	20	9	14	2	18	13	4	17	22	24	8	5	21	19	10	7	16	3	15	11	1	12
24	24	15	10	18	17	8	16	6	19	3	23	22	21	20	2	7	5	4	9	14	13	12	11	1

product table for the symmetric group S_4
(row ϕ_i, column ϕ_m ; result $\phi_i \circ \phi_m$)

Cyclic subgroups of S_4 :

order 1 : $I = gp(1)$

	1

I

order 2 : $Z_1 = gp(3)$, $Z_2 = gp(6)$, $Z_3 = gp(9)$, $Z_4 = gp(19)$, $Z_5 = gp(20)$, $Z_6 = gp(21)$, $Z_7 = gp(22)$, $Z_8 = gp(23)$, $Z_9 = gp(24)$

	1	3
1	1	3
3	3	1

Z_1

	1	6
1	1	6
6	6	1

Z_2

	1	9
1	1	9
9	9	1

Z_3

	1	19
1	1	19
19	19	1

Z_4

	1	20
1	1	20
20	20	1

Z_5

	1	21
1	1	21
21	21	1

Z_6

	1	22
1	1	22
22	22	1

Z_7

	1	23
1	1	23
23	23	1

Z_8

	1	24
1	1	24
24	24	1

Z_9

order 3 : $D_1 = gp(11)$, $D_2 = gp(13)$, $D_3 = gp(15)$, $D_4 = gp(17)$

	1	11	12
1	1	11	12
11	11	12	1
12	12	1	11

D_1

	1	13	14
1	1	13	14
13	13	14	1
14	14	1	13

D_2

	1	15	16
1	1	15	16
15	15	16	1
16	16	1	15

D_3

	1	17	18
1	1	17	18
17	17	18	1
18	18	1	17

D_4

order 4 : $V_1 = gp(2)$, $V_2 = gp(5)$, $V_3 = gp(8)$

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

V_1

	1	5	6	7
1	1	5	6	7
5	5	6	7	1
6	6	7	1	5
7	7	1	5	6

V_2

	1	8	9	10
1	1	8	9	10
8	8	9	10	1
9	9	10	1	8
10	10	1	8	9

V_3

Determination of the subgroups with two generating elements : None of the elements of S_4 generates the entire group S_4 : The symmetric group of degree 4 is not cyclic. Each of the cyclic subgroups of S_4 has a generating element. This is generally not unique. For example, V_1 is generated by ϕ_2 and by ϕ_4. The elements $E = \{\phi_2, \phi_3, \phi_5, \phi_6, \phi_8, \phi_9, \phi_{11}, \phi_{13}, \phi_{15}, \phi_{17}, \phi_{19}$ to $\phi_{24}\}$ are chosen for the following investigation.

The generating elements of a group $gp(\phi_i, \phi_m)$ with two generating elements belong to different cyclic subgroups. In order to determine the subgroups with two generating elements, all pairs $(\phi_i, \phi_m) \in E \times E$ with $i \neq m$ are considered. Some of these pairs generate the entire group S_4. Other pairs generate subgroups of S_4. The designations of the generated subgroups are shown in the matrix below.

Some generating elements of cyclic subgroups are also non-generating elements of other cyclic subgroups. For example, the generating element ϕ_3 of $Z_1 = gp(3)$ is also a non-generating element of $V_1 = gp(2)$. The groups V_1, V_2, V_3 therefore occur in the matrix, even though they are generated by a single element.

		(4)		(7)		(10)		(12)	(14)	(16)	(18)						
		2	3	5	6	8	9	11	13	15	17	19	20	21	22	23	24
(4)	2		V_1	●	H_1	●	H_1	●	●	●	●	●	H_1	●	●	H_1	●
	3	V_1		H_2	V_4	H_3	V_4	A_4	A_4	A_4	A_4	H_3	V_5	H_2	H_2	V_5	H_3
(7)	5	●	H_2		V_2	●	H_2	●	●	●	●	●	●	H_2	H_2	●	●
	6	H_1	V_4	V_2		H_3	V_4	A_4	A_4	A_4	A_4	H_3	H_1	V_6	V_6	H_1	H_3
(10)	8	●	H_3	●	H_3		V_3	●	●	●	●	H_3	●	●	●	●	H_3
	9	H_1	V_4	H_2	V_4	V_3		A_4	A_4	A_4	A_4	V_7	H_1	H_2	H_2	H_1	V_7
(12)	11	●	A_4	●	A_4	●	A_4		A_4	A_4	A_4	●	●	●	X_1	X_1	X_1
(14)	13	●	A_4	●	A_4	●	A_4	A_4		A_4	A_4	●	X_2	X_2	●	●	X_2
(16)	15	●	A_4	●	A_4	●	A_4	A_4	A_4		A_4	X_3	●	X_3	●	X_3	●
(18)	17	●	A_4	●	A_4	●	A_4	A_4	A_4	A_4		X_4	X_4	●	X_4	●	●
	19	●	H_3	●	H_3	H_3	V_7	●	●	X_3	X_4		X_4	X_3	X_4	X_3	V_7
	20	H_1	V_5	●	H_1	●	H_1	●	X_2	●	X_4	X_4		X_2	X_4	V_5	X_2
	21	●	H_2	H_2	V_6	●	H_2	●	X_2	X_3	●	X_3	X_2		V_6	X_3	X_2
	22	●	H_2	H_2	V_6	●	H_2	X_1	●	●	X_4	X_4	X_4	V_6		X_1	X_1
	23	H_1	V_5	●	H_1	●	H_1	X_1	●	X_3	●	X_3	V_5	X_3	X_1		X_1
	24	●	H_3	●	H_3	H_3	V_7	X_1	X_2	●	●	V_7	X_2	X_2	X_1	X_1	

designations of the generated groups $gp(\phi_i, \phi_m)$

● complete permutation group S_4 generated

Subgroups of S_4 with two generating elements:

order 12 : $A_4 = gp(11,13)$

	1	3	6	9	11	12	13	14	15	16	17	18
1	1	3	6	9	11	12	13	14	15	16	17	18
3	3	1	9	6	18	13	12	15	14	17	16	11
6	6	9	1	3	14	16	17	11	18	12	13	15
9	9	6	3	1	15	17	16	18	11	13	12	14
11	11	14	15	18	12	1	6	17	13	9	3	16
12	12	17	13	16	1	11	15	3	6	18	14	9
13	13	16	12	17	3	18	14	1	9	11	15	6
14	14	11	18	15	16	6	1	13	17	3	9	12
15	15	18	11	14	17	9	3	12	16	1	6	13
16	16	13	17	12	6	14	18	9	1	15	11	3
17	17	12	16	13	9	15	11	6	3	14	18	1
18	18	15	14	11	13	3	9	16	12	6	1	17

A_4

order 8 : $H_1 = gp(2,6)$ $H_2 = gp(5,9)$ $H_3 = gp(3,8)$

	1	2	3	4	6	9	20	23
1	1	2	3	4	6	9	20	23
2	2	3	4	1	23	20	6	9
3	3	4	1	2	9	6	23	20
4	4	1	2	3	20	23	9	6
6	6	20	9	23	1	3	2	4
9	9	23	6	20	3	1	4	2
20	20	9	23	6	4	2	1	3
23	23	6	20	9	2	4	3	1

H_1

	1	3	5	6	7	9	21	22
1	1	3	5	6	7	9	21	22
3	3	1	21	9	22	6	5	7
5	5	22	6	7	1	21	3	9
6	6	9	7	1	5	3	22	21
7	7	21	1	5	6	22	9	3
9	9	6	22	3	21	1	7	5
21	21	7	9	22	3	5	1	6
22	22	5	3	21	9	7	6	1

H_2

	1	3	6	8	9	10	19	24
1	1	3	6	8	9	10	19	24
3	3	1	9	24	6	19	10	8
6	6	9	1	19	3	24	8	10
8	8	19	24	9	10	1	6	3
9	9	6	3	10	1	8	24	19
10	10	24	19	1	8	9	3	6
19	19	8	10	3	24	6	1	9
24	24	10	8	6	19	3	9	1

H_3

order 6 : X_1 = gp(11,22) X_2 = gp(13,24)

X_3 = gp(15,19) X_4 = gp(17,19)

	1	11	12	22	23	24
1	1	11	12	22	23	24
11	11	12	1	23	24	22
12	12	1	11	24	22	23
22	22	24	23	1	12	11
23	23	22	24	11	1	12
24	24	23	22	12	11	1

X_1

	1	13	14	20	21	24
1	1	13	14	20	21	24
13	13	14	1	21	24	20
14	14	1	13	24	20	21
20	20	24	21	1	14	13
21	21	20	24	13	1	14
24	24	21	20	14	13	1

X_2

	1	15	16	19	21	23
1	1	15	16	19	21	23
15	15	16	1	21	23	19
16	16	1	15	23	19	21
19	19	23	21	1	16	15
21	21	19	23	15	1	16
23	23	21	19	16	15	1

X_3

	1	17	18	19	20	22
1	1	17	18	19	20	22
17	17	18	1	20	22	19
18	18	1	17	22	19	20
19	19	22	20	1	18	17
20	20	19	22	17	1	18
22	22	20	19	18	17	1

X_4

order 4 : V_4 = gp(3,6) V_5 = gp(3,20)

V_6 = gp(6,21) V_7 = gp(9,19)

	1	3	6	9
1	1	3	6	9
3	3	1	9	6
6	6	9	1	3
9	9	6	3	1

V_4

	1	3	20	23
1	1	3	20	23
3	3	1	23	20
20	20	23	1	3
23	23	20	3	1

V_5

	1	6	21	22
1	1	6	21	22
6	6	1	22	21
21	21	22	1	6
22	22	21	6	1

V_6

	1	9	19	24
1	1	9	19	24
9	9	1	24	19
19	19	24	1	9
24	24	19	9	1

V_7

More than two generating elements : Groups with three generating elements contain the generating elements of one of the groups gp(x,y) with two generating elements which have already been determined, and an additional generating element of a cyclic group which is not contained in gp(x,y). If one of the groups A_4, H_i or X_i is chosen for gp(x,y) then the entire symmetric group S_4 is generated. If one of the groups $V_4,...,V_7$ is chosen for gp(x,y), then one of the groups A_4 or H_i is generated. These can, however, already be generated with two elements. Thus all subgroups of S_4 can be generated with two elements.

Isomorphism between the subgroups :

The subgroups of order 8 are isomorphic. The following isomorphisms serve to prove this :

$f_1 :=$

1	2	3	4	6	9	20	23
1	5	6	7	3	9	21	22

$H_1 \cong H_2$

$f_2 :=$

1	3	5	6	7	9	21	22
1	6	8	9	10	3	19	24

$H_2 \cong H_3$

The isomorphism of the other groups which are generated by at least 2 elements is demonstrated analogously :

$$X_1 \cong X_2 \cong X_3 \cong X_4$$

$$V_4 \cong V_5 \cong V_6 \cong V_7$$

According to Section 7.5.4, the cyclic subgroups of equal order are isomorphic :

$$V_1 \cong V_2 \cong V_3$$

$$D_1 \cong D_2 \cong D_3 \cong D_4$$

$$Z_1 \cong ... \cong Z_9$$

The cyclic groups of order 4 and the groups of order 4 generated by two elements are, however, not isomorphic. This is shown in Example 1 of Section 7.5.3.

Canonical decomposition of the permutations in S_4 : The relationships between the group S_4 and its subgroups are studied by canonical decomposition of the permutations in S_4. The decomposition leads to the following subsets of S_4 :

(a) The identity permutation $\phi_1 = <1>$

(b) Six two-cycles can be formed on the set $X_4 = \{1, 2, 3, 4\}$. Each of these cycles is the canonical decomposition of an element of S_4 :

ϕ_{19}	=	$<1,2>$	ϕ_{21}	=	$<1,4>$	ϕ_{23}	=	$<2,4>$
ϕ_{20}	=	$<1,3>$	ϕ_{22}	=	$<2,3>$	ϕ_{24}	=	$<3,4>$

(c) Three pairs of two-cycles with disjoint ranges can be formed on the set X_4. Each of these pairs is the canonical decomposition of an element of S_4 :

$$\phi_3 = <1,3> \circ <2,4>$$
$$\phi_6 = <1,4> \circ <2,3>$$
$$\phi_9 = <1,2> \circ <3,4>$$

(d) Eight three-cycles can be formed on the set X_4. Each of these three-cycles is the canonical decomposition of an element of S_4. The elements form pairs of inverses.

ϕ_{11}	=	$<2,3,4>$	ϕ_{12}	=	$<2,4,3>$	$\phi_{11} \circ \phi_{12}$	=	ϕ_1
ϕ_{13}	=	$<1,3,4>$	ϕ_{14}	=	$<1,4,3>$	$\phi_{13} \circ \phi_{14}$	=	ϕ_1
ϕ_{15}	=	$<1,2,4>$	ϕ_{16}	=	$<1,4,2>$	$\phi_{15} \circ \phi_{16}$	=	ϕ_1
ϕ_{17}	=	$<1,2,3>$	ϕ_{18}	=	$<1,3,2>$	$\phi_{17} \circ \phi_{18}$	=	ϕ_1

(e) The canonical decompositions of the remaining six elements of S_4 are four-cycles. They form pairs of inverses.

ϕ_2	=	$<1,2,3,4>$	ϕ_4	=	$<1,4,3,2>$	$\phi_2 \circ \phi_4$	=	ϕ_1
ϕ_5	=	$<2,4,3,1>$	ϕ_7	=	$<1,3,4,2>$	$\phi_5 \circ \phi_7$	=	ϕ_1
ϕ_8	=	$<1,3,2,4>$	ϕ_{10}	=	$<4,2,3,1>$	$\phi_8 \circ \phi_{10}$	=	ϕ_1

The subgroups of S_4 are formed from these subsets as follows :

$Z_1 - Z_3$: one element from (c) and ϕ_1

$Z_4 - Z_9$: one element from (b) and ϕ_1

$D_1 - D_4$: two mutually inverse elements from (d) and ϕ_1

$V_1 - V_3$: two mutually inverse elements from (e), the square of one of these elements (the squares of the elements are equal) and ϕ_1

V_4 : the elements in (c) and ϕ_1

$V_5 - V_7$: all compositions of the cycles which occur as a pair of (c)

$$
\begin{array}{llll}
V_5 : & \phi_1 = \langle 1,3\rangle \circ \langle 1,3\rangle = \langle 2,4\rangle \circ \langle 2,4\rangle & & \\
& \phi_3 = \langle 1,3\rangle \circ \langle 2,4\rangle & \phi_{20} = \langle 1,3\rangle & \phi_{23} = \langle 2,4\rangle \\
V_6 : & \phi_1 = \langle 1,4\rangle \circ \langle 1,4\rangle = \langle 2,3\rangle \circ \langle 2,3\rangle & & \\
& \phi_6 = \langle 1,4\rangle \circ \langle 2,3\rangle & \phi_{21} = \langle 1,4\rangle & \phi_{22} = \langle 2,3\rangle \\
V_7 : & \phi_1 = \langle 1,2\rangle \circ \langle 1,2\rangle = \langle 3,4\rangle \circ \langle 3,4\rangle & & \\
& \phi_9 = \langle 1,2\rangle \circ \langle 3,4\rangle & \phi_{19} = \langle 1,2\rangle & \phi_{24} = \langle 3,4\rangle
\end{array}
$$

$H_1 - H_3$: the group $V_4 = \{\phi_1, \phi_3, \phi_6, \phi_9\}$, two mutually inverse elements from (e) and both of the two-cycles from (b) which do not occur as transpositions in the elements from (e) :

$$
\begin{array}{lll}
H_1 : & \phi_2 = \phi_4^{-1} = \langle 1,2,3,4\rangle = \langle 1,2\rangle \circ \langle 2,3\rangle \circ \langle 3,4\rangle \circ \langle 4,1\rangle & \\
& \phi_{20} = \langle 1,3\rangle & \phi_{23} = \langle 2,4\rangle \\
H_2 : & \phi_5 = \phi_7^{-1} = \langle 1,2,4,3\rangle = \langle 1,2\rangle \circ \langle 2,4\rangle \circ \langle 4,3\rangle \circ \langle 3,1\rangle & \\
& \phi_{21} = \langle 1,4\rangle & \phi_{22} = \langle 2,3\rangle \\
H_3 : & \phi_8 = \phi_{10}^{-1} = \langle 1,3,2,4\rangle = \langle 1,3\rangle \circ \langle 3,2\rangle \circ \langle 2,4\rangle \circ \langle 4,1\rangle & \\
& \phi_{19} = \langle 1,2\rangle & \phi_{24} = \langle 3,4\rangle
\end{array}
$$

A_4 : All elements from (c) and (d) and ϕ_1. These are the even permutations.

The analysis of the permutations by canonical decomposition shows that isomorphic subgroups may bear different relationships to S_4. Although V_4 is isomorphic to V_5, V_6 and V_7, the role played by V_4 in S_4 differs from the role played by V_5, V_6 and V_7. The significance of this difference becomes apparent in the study of the normal subgroups in the following section : The four-group V_4 is a normal subgroup in S_4, while the groups V_5, V_6 and V_7 are not.

Compilation of the subgroups : All subgroups of the symmetric group S_4 are compiled in the following. The symbol • indicates that the element associated with that column is contained in the group.

Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
S_4	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●
A_4	●		●			●			●		●	●	●	●	●	●	●	●						
H_1	●	●	●	●		●			●											●			●	
H_2	●		●		●	●	●		●												●	●		
H_3	●		●			●		●	●	●									●					●
X_1	●										●	●										●	●	●
X_2	●												●	●						●	●			●
X_3	●														●	●			●		●		●	
X_4	●																●	●	●	●		●		
V_4	●		●			●			●															
V_5	●		●																	●			●	
V_6	●					●															●	●		
V_7	●								●										●					●
V_1	●	●	●	●																				
V_2	●				●	●	●																	
V_3	●							●	●	●														
D_1	●										●	●												
D_2	●												●	●										
D_3	●														●	●								
D_4	●																●	●						
Z_1	●		●																					
Z_2	●					●																		
Z_3	●								●															
Z_4	●																		●					
Z_5	●																			●				
Z_6	●																				●			
Z_7	●																					●		
Z_8	●																						●	
Z_9	●																							●
I	●																							

Chains of subgroups in S_4 : Some of the subgroups of S_4 are contained in other subgroups. The chains of subgroups in S_4 are listed in the following :

$S_4 \supset A_4 \supset V_4 \supset (Z_1 \cong Z_2 \cong Z_3) \supset \mathrm{I}$
$S_4 \supset A_4 \supset (D_1 \cong D_2 \cong D_3 \cong D_4) \supset \mathrm{I}$

$S_4 \supset H_1 \supset V_1 \supset Z_1 \supset \mathrm{I}$
$S_4 \supset H_2 \supset V_2 \supset Z_2 \supset \mathrm{I}$
$S_4 \supset H_3 \supset V_3 \supset Z_3 \supset \mathrm{I}$
$S_4 \supset (H_1 \cong H_2 \cong H_3) \supset V_4 \supset (Z_1 \cong Z_2 \cong Z_3) \supset \mathrm{I}$

$S_4 \supset H_1 \supset V_5 \supset Z_1 \supset \mathrm{I}$
$S_4 \supset H_2 \supset V_6 \supset Z_2 \supset \mathrm{I}$
$S_4 \supset H_3 \supset V_7 \supset Z_3 \supset \mathrm{I}$

$S_4 \supset X_1 \supset D_1 \supset \mathrm{I}$
$S_4 \supset X_2 \supset D_2 \supset \mathrm{I}$
$S_4 \supset X_3 \supset D_3 \supset \mathrm{I}$
$S_4 \supset X_4 \supset D_4 \supset \mathrm{I}$

$S_4 \supset X_1 \supset (Z_7 \cong Z_8 \cong Z_9) \supset \mathrm{I}$
$S_4 \supset X_2 \supset (Z_5 \cong Z_6 \cong Z_9) \supset \mathrm{I}$
$S_4 \supset X_3 \supset (Z_4 \cong Z_6 \cong Z_8) \supset \mathrm{I}$
$S_4 \supset X_4 \supset (Z_4 \cong Z_5 \cong Z_7) \supset \mathrm{I}$

Notes :

(1) The intersection of two subgroups of S_4 is a subgroup of S_4, for example :

$A_4 \cap H_1 = V_4$	$H_1 \cap H_2 = V_4$
$A_4 \cap X_1 = D_1$	$H_1 \cap X_1 = Z_8$
$A_4 \cap V_1 = Z_1$	$H_1 \cap V_2 = Z_2$

(2) The order 24 of S_4 contains the proper divisors 2, 3, 4, 6, 8, 12, and the group S_4 contains subgroups of these orders. The order 12 of A_4 contains the proper divisors 2, 3, 4, 6, but the group A_4 does not contain a subgroup of order 6.

7.7.8 CLASS STRUCTURE OF THE SYMMETRIC GROUP S_4

Normal subgroups : The group structure of the symmetric group S_4 shows that there are subgroups of S_4 which are also subgroups of other subgroups of S_4. Examples are furnished by the chains $S_4 \supset A_4 \supset V_4 \supset Z_1 \supset \mathrm{I}$, $S_4 \supset H_1 \supset V_5 \supset Z_1 \supset \mathrm{I}$, $S_4 \supset X_1 \supset Z_7 \supset \mathrm{I}$. For each subgroup U, the left and right cosets may be formed in every subgroup of S_4 which contains U. The normal subgroups (subgroups with identical left and right cosets) are of special importance for the class structure of S_4. Some of the normal subgroups are determined in the following. All normal subgroups are compiled in the subsequent table. The subgroups which are not normal are also shown.

$$\begin{array}{lcrcl}
A_4 \lhd S_4 & : & 1 \circ A_4 & = & \{1,3,6,9,11,12,13,14,15,16,17,18\} = A_4 \circ 1 = [1] \\
 & & 2 \circ A_4 & = & \{2,4,5,7,8,10,19,20,21,22,23,24\} = A_4 \circ 2 = [2]
\end{array}$$

$$\begin{array}{lcrcl}
V_4 \lhd S_4 & : & 1 \circ V_4 & = & \{1,3,6,9\} = V_4 \circ 1 = [1] \\
 & & 2 \circ V_4 & = & \{2,4,20,23\} = V_4 \circ 2 = [2] \\
 & & 5 \circ V_4 & = & \{5,7,21,22\} = V_4 \circ 5 = [5] \\
 & & 8 \circ V_4 & = & \{8,10,19,24\} = V_4 \circ 8 = [8] \\
 & & 11 \circ V_4 & = & \{11,14,15,18\} = V_4 \circ 11 = [11] \\
 & & 12 \circ V_4 & = & \{12,13,16,17\} = V_4 \circ 12 = [12]
\end{array}$$

$$\begin{array}{lcrcl}
V_4 \lhd A_4 & : & 1 \circ V_4 & = & \{1,3,6,9\} = V_4 \circ 1 = [1] \\
 & & 11 \circ V_4 & = & \{11,14,15,18\} = V_4 \circ 11 = [11] \\
 & & 12 \circ V_4 & = & \{12,13,16,17\} = V_4 \circ 12 = [12]
\end{array}$$

$$\begin{array}{lcrcl}
Z_1 \lhd H_1 & : & 1 \circ Z_1 & = & \{1,3\} = Z_1 \circ 1 = [1] \\
 & & 2 \circ Z_1 & = & \{2,4\} = Z_1 \circ 20 = [2] \\
 & & 6 \circ Z_1 & = & \{6,9\} = Z_1 \circ 6 = [6] \\
 & & 20 \circ Z_1 & = & \{20,23\} = Z_1 \circ 20 = [20]
\end{array}$$

$$\begin{array}{lcrcl}
Z_1 \lhd V_4 & : & 1 \circ Z_1 & = & \{1,3\} = Z_1 \circ 1 = [1] \\
 & & 6 \circ Z_1 & = & \{6,9\} = Z_1 \circ 6 = [6]
\end{array}$$

$$\begin{array}{lcrcl}
Z_1 \lhd V_1 & : & 1 \circ Z_1 & = & \{1,3\} = Z_1 \circ 1 = [1] \\
 & & 2 \circ Z_1 & = & \{2,4\} = Z_1 \circ 2 = [2]
\end{array}$$

$$\begin{array}{lcrcl}
V_5 \lhd H_1 & : & 1 \circ V_5 & = & \{1,3,20,23\} = V_5 \circ 1 = [1] \\
 & & 2 \circ V_5 & = & \{2,4,6,9\} = V_5 \circ 2 = [2]
\end{array}$$

$$\begin{array}{lcrcl}
D_1 \lhd X_1 & : & 1 \circ D_1 & = & \{1,11,12\} = D_1 \circ 1 = [1] \\
 & & 22 \circ D_1 & = & \{22,23,24\} = D_1 \circ 22 = [22]
\end{array}$$

Subgroups form a chain of normal subgroups if every subgroup in the chain is a normal subgroup in every subgroup to its right. The following chains of normal subgroups are contained in the table of normal subgroups :

$I \triangleleft V_4 \triangleleft A_4 \triangleleft S_4$

$I \triangleleft Z_1 \triangleleft V_5 \triangleleft H_1$

$I \triangleleft Z_2 \triangleleft V_6 \triangleleft H_2$

$I \triangleleft Z_3 \triangleleft V_7 \triangleleft H_3$

$I \triangleleft V_1 \triangleleft H_1$

$I \triangleleft V_2 \triangleleft H_2$

$I \triangleleft V_3 \triangleleft H_3$

$I \triangleleft V_4 \triangleleft H_1$

$I \triangleleft V_4 \triangleleft H_2$

$I \triangleleft V_4 \triangleleft H_3$

$I \triangleleft Z_1 \triangleleft V_1$

$I \triangleleft Z_2 \triangleleft V_2$

$I \triangleleft Z_3 \triangleleft V_3$

$I \triangleleft Z_1 \triangleleft V_4$

$I \triangleleft Z_2 \triangleleft V_4$

$I \triangleleft Z_3 \triangleleft V_4$

$I \triangleleft Z_5 \triangleleft V_5$

$I \triangleleft Z_8 \triangleleft V_5$

$I \triangleleft Z_6 \triangleleft V_6$

$I \triangleleft Z_7 \triangleleft V_6$

$I \triangleleft Z_4 \triangleleft V_7$

$I \triangleleft Z_9 \triangleleft V_7$

$I \triangleleft D_1 \triangleleft X_1$

$I \triangleleft D_2 \triangleleft X_2$

$I \triangleleft D_3 \triangleleft X_3$

$I \triangleleft D_4 \triangleleft X_4$

Notes :

(1) The chain $I \triangleleft V_4 \triangleleft A_4 \triangleleft S_4$ of normal subgroups does not contain the normal subgroup Z_1 of V_4, since Z_1 is not a normal subgroup of A_4 and of S_4. This is explained in Section 7.5.4 (Example 3).

(2) The chain $I \triangleleft V_4 \triangleleft A_4 \triangleleft S_4$ of normal subgroups ends with the entire group S_4. The group A_4 is called the alternating subgroup of S_4. The subgroup V_4 is called Klein's four-group. The significance of the chain $I \triangleleft V_4 \triangleleft A_4 \triangleleft S_4$ of normal subgroups is studied in Section 7.8.

(3) The chains of normal subgroups ending in V_4 cannot be continued to A_4 or S_4 since Z_i is not a normal subgroup of A_4 or S_4. The other chains of normal subgroups end with H_i , X_i , V_{1-3} or V_{5-7} because these subgroups are not normal subgroups of S_4.

(4) The intersection of two normal subgroups of a group is a normal subgroup of the group. For example, H_1 contains the normal subgroups V_1, V_4, V_5, Z_1 and I. The intersections of these normal subgroups yield :

$$V_1 \cap V_4 = V_1 \cap V_5 = V_1 \cap Z_1 = V_4 \cap V_5 = V_4 \cap Z_1 = V_5 \cap Z_1 = Z_1 \triangleleft H_1$$

(5) The intersection of a normal subgroup of S_4 and a subgroup U of S_4 is a normal subgroup of U, for example :

$$A_4 \cap X_1 = D_1 \quad \Rightarrow \quad D_1 \triangleleft X_1$$
$$A_4 \cap V_5 = Z_1 \quad \Rightarrow \quad Z_1 \triangleleft V_5$$
$$A_4 \cap V_1 = Z_1 \quad \Rightarrow \quad Z_1 \triangleleft V_1$$
$$V_4 \cap V_5 = Z_1 \quad \Rightarrow \quad Z_1 \triangleleft V_5$$

◁		S	A	H			X				V			V	V			D				Z			Z						I
		4	4	1	2	3	1	2	3	4	1	2	3	4	5	6	7	1	2	3	4	1	2	3	4	5	6	7	8	9	
S	4	●																													
A	4	●	●																												
H	1	□		●																											
	2	□			●																										
	3	□				●																									
X	1	□					●																								
	2	□						●																							
	3	□							●																						
	4	□								●																					
V	1	□		●							●																				
	2	□			●							●																			
	3	□				●							●																		
V	4	●	●	●	●	●								●																	
V	5	□		●											●																
	6	□			●											●															
	7	□				●											●														
D	1	□	□				●											●													
	2	□	□					●											●												
	3	□	□						●											●											
	4	□	□							●											●										
Z	1	□	□	●	□	□					●			●	●							●									
	2	□	□	□	●	□						●		●		●							●								
	3	□	□	□	□	●							●	●			●							●							
Z	4	□				□			□	□							●								●						
	5	□		□				□		□					●											●					
	6	□			□			□	□							●											●				
	7	□			□		□			□						●												●			
	8	□		□			□		□						●														●		
	9	□				□	□	□									●													●	
I		●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●	●

normal subgroups of subgroups of S_4

● row group is a normal subgroup in column group

□ row group is a subgroup in column group, but not a normal subgroup

Conjugate elements of S_4 : The elements $a, b \in S_4$ are conjugate if there is an element $g \in S_4$ for which $a = g^{-1} \circ b \circ g$. Conjugate elements form an equivalence class. The equivalence classes of conjugate elements form a partition of S_4. The following table shows the conjugate elements of S_4.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
2	1	2	3	4	8	9	10	7	6	5	17	18	11	12	13	14	15	16	21	23	24	19	20	22
3	1	2	3	4	7	6	5	10	9	8	15	16	17	18	11	12	13	14	24	20	22	21	23	19
4	1	2	3	4	10	9	8	5	6	7	13	14	15	16	17	18	11	12	22	23	19	24	20	21
5	1	10	9	8	5	6	7	2	3	4	16	15	11	12	17	18	14	13	20	24	22	21	19	23
6	1	4	3	2	5	6	7	10	9	8	18	17	16	15	14	13	12	11	24	23	21	22	20	19
7	1	8	9	10	5	6	7	4	3	2	13	14	18	17	12	11	15	16	23	19	22	21	24	20
8	1	5	6	7	4	3	2	8	9	10	17	18	15	16	12	11	14	13	24	21	23	20	22	19
9	1	4	3	2	7	6	5	8	9	10	14	13	12	11	18	17	16	15	19	23	22	21	20	24
10	1	7	6	5	2	3	4	8	9	10	16	15	18	17	13	14	11	12	24	22	20	23	21	19
11	1	10	9	8	4	3	2	5	6	7	11	12	17	18	14	13	16	15	21	19	20	23	24	22
12	1	7	6	5	8	9	10	4	3	2	11	12	16	15	18	17	13	14	20	21	19	24	22	23
13	1	7	6	5	10	9	8	2	3	4	18	17	13	14	11	12	16	15	23	21	24	19	22	20
14	1	8	9	10	4	3	2	7	6	5	15	16	13	14	18	17	12	11	22	24	20	23	19	21
15	1	8	9	10	2	3	4	5	6	7	18	17	12	11	15	16	13	14	21	24	23	20	19	22
16	1	5	6	7	8	9	10	2	3	4	14	13	17	18	15	16	12	11	23	22	19	24	21	20
17	1	5	6	7	10	9	8	4	3	2	15	16	12	11	14	13	17	18	20	22	24	19	21	23
18	1	10	9	8	2	3	4	7	6	5	14	13	16	15	11	12	17	18	22	19	23	20	24	21
19	1	7	6	5	4	3	2	10	9	8	13	14	11	12	16	15	18	17	19	22	23	20	21	24
20	1	4	3	2	8	9	10	5	6	7	16	15	14	13	12	11	18	17	22	20	24	19	23	21
21	1	10	9	8	7	6	5	4	3	2	17	18	14	13	16	15	11	12	23	24	21	22	19	20
22	1	8	9	10	7	6	5	2	3	4	12	11	15	16	13	14	18	17	20	19	21	22	24	23
23	1	4	3	2	10	9	8	7	6	5	12	11	18	17	16	15	14	13	21	20	19	24	23	22
24	1	5	6	7	2	3	4	10	9	8	12	11	14	13	17	18	15	16	19	21	20	23	22	24

conjugation table : $\phi_i = \phi_k^{-1} \circ \phi_m \circ \phi_k$ (row k, column m)

Each of the equivalence classes contains only elements of equal order. The elements of order 2 in the classes [3] and [19] differ in that the groups generated by the elements in the class [3] are each contained in one of the groups generated by elements in the class [2].

[1]	=	{1}	element of order 1
[3]	=	{3, 6, 9}	element of order 2
[19]	=	{19, 20, 21, 22, 23, 24}	element of order 2
[11]	=	{11, 12, 13, 14, 15, 16, 17, 18}	element of order 3
[2]	=	{2, 4, 5, 7, 8, 10}	element of order 4

Commuting elements of S_4 : The elements $a,g \in S_4$ are said to commute (see Section 3.2) if $a \circ g = g \circ a$, that is $a \circ g \circ a^{-1} \circ g^{-1} = 1$. The following table shows the values of the commutator $a \circ g \circ a^{-1} \circ g^{-1}$ for all pairs of elements of $S_4 \times S_4$.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	12	3	16	18	3	14	18	16	12	18	14	12	16	14	18	3	16	12	3	14
3	1	1	1	1	6	1	6	9	1	9	9	6	6	9	9	6	6	9	9	1	6	6	1	9
4	1	1	1	1	17	3	13	15	3	11	15	13	17	15	11	17	13	11	15	3	13	17	3	11
5	1	11	6	18	1	1	1	13	6	16	18	16	16	11	18	13	13	11	16	18	6	6	11	13
6	1	3	1	3	1	1	1	9	1	9	3	9	9	3	3	9	9	3	9	3	1	1	3	9
7	1	15	6	14	1	1	1	12	6	17	14	17	17	15	14	12	12	15	17	14	6	6	15	12
8	1	17	9	16	14	9	11	1	1	1	14	17	16	11	11	17	16	14	9	17	14	11	16	9
9	1	3	1	3	6	1	6	1	1	1	6	3	3	6	6	3	3	6	1	3	6	6	3	1
10	1	13	9	12	15	9	18	1	1	1	15	13	12	18	18	13	12	15	9	13	15	18	12	9
11	1	17	9	16	17	3	13	13	6	16	1	1	9	3	6	3	6	9	17	13	16	12	12	12
12	1	15	6	14	15	9	18	18	3	14	1	1	6	9	3	9	3	6	15	18	14	11	11	11
13	1	11	6	18	15	9	18	15	3	11	9	6	1	1	6	3	9	3	18	14	14	11	15	14
14	1	17	9	16	12	3	16	12	6	17	3	9	1	1	9	6	3	6	16	13	13	17	12	13
15	1	13	9	12	17	3	13	12	6	17	6	3	6	9	1	1	9	3	16	17	16	12	16	13
16	1	11	6	18	14	9	11	18	3	14	3	9	3	6	1	1	6	9	15	14	15	18	15	11
17	1	15	6	14	14	9	11	15	3	11	6	3	9	3	9	6	1	1	18	18	15	18	11	14
18	1	13	9	12	12	3	16	13	6	16	9	6	3	6	3	9	1	1	17	17	13	17	16	12
19	1	17	9	16	15	9	18	9	1	9	18	16	17	15	15	16	17	18	1	17	15	18	16	1
20	1	3	1	3	17	3	13	18	3	14	14	17	13	14	18	13	17	18	18	1	13	17	1	14
21	1	15	6	14	6	1	6	13	6	16	15	13	13	14	15	16	16	14	16	14	1	1	15	13
22	1	11	6	18	6	1	6	12	6	17	11	12	12	18	11	17	17	18	17	18	1	1	11	12
23	1	3	1	3	12	3	16	15	3	11	11	12	16	11	15	16	12	15	15	1	16	12	1	11
24	1	13	9	12	14	9	11	9	1	9	11	12	13	14	14	12	13	11	1	13	14	11	12	1

table of commutators of S_4 : $\phi_i = \phi_k \circ \phi_m \circ \phi_k^{-1} \circ \phi_m^{-1}$ (row k, column m)

The commutator formed with the elements a, g is an element of S_4 ; it is designated by k or by [a,g]. Ordered pairs $(a,g) \in S_4 \times S_4$ with the same value of the commutator k form an equivalence class, which is designated by [k]. Commutators are treated in Section 7.8.2.

7.8 GENERAL GROUPS

7.8.1 INTRODUCTION

In contrast to the inner operation of an abelian group, the inner operation of a general group is generally not commutative. The order of the operands in an expression is therefore relevant. This property significantly complicates the study of the properties of non-abelian groups. In particular, expressions can generally not be simplified to linear combinations.

The structure of non-abelian groups is studied by classifying their elements, using the classification methods developed in Section 7.4. Their effectiveness relies on the fact that subsets of elements of general groups may be conjugate or commutative. Normalizers are introduced for a classification using conjugate subsets, the center of the group is defined for a classification using commutative subsets, and the commutator group is defined for a classification using non-commutative subsets. The center and the commutator group are normal subgroups of the group; they are used in forming quotient groups.

Every finite group is isomorphic to a group of permutations. The diverse group and class structure of permutation groups is illustrated in Sections 7.7.7 and 7.7.8 for the group S_4. The existence of subgroups in a symmetric group S_n may be studied using Cauchy's Theorem and the theorems of Sylow. If a group contains no proper subgroups, then the group is cyclic of prime order. If the group is abelian, then for every divisor m of the order of the group there is a subgroup of order m. If the prime power p^n is a divisor of the order of a finite group, then the group contains a subgroup of order p^n.

A further aim in studying the structure of general groups is to find a simple subgroup which has no proper normal subgroups. In Section 7.7.6, the alternating group A_n is shown to be simple for $n \geq 5$, while in Section 7.7.8 the group A_4 is shown to contain the normal subgroup V_4 (Klein's four-group).

Starting from a constructible normal subgroup (center, commutator group), a nested chain of quotient groups is determined, for instance the central series or the derived series of the group. Groups whose derived series ends with the trivial subgroup {1} are called soluble and have special properties. The generalization of this concept is the normal series, which is refined to a composition series. Any two composition series of a finite group are similar. These series are used in Galois theory.

7.8.2 CLASSES IN GENERAL GROUPS

Introduction : The partition of a group $(G\,;\circ)$ into the cosets of a subgroup of G is defined in Section 7.4. This requires a subgroup of G to be known. In the following the transformation of the elements of a subset A of G is shown to determine a subgroup of G, which is called the normalizer of A in G. The left and right cosets of the normalizer are generally different : The normalizer is therefore generally not a normal subgroup of G.

The center of $(G\,;\circ)$ is a special subgroup of G. The center contains the elements of G which commute with every element of G. The center is a normal subgroup in G and may therefore be used to construct a quotient group. The properties of the quotient group and the relationships between the structure of the group and the structure of the quotient group provided by the isomorphism theorems lead to important properties of non-abelian groups.

A further special subgroup of $(G\,;\circ)$ is the commutator group D(G), also called the derived group of G or the first derivative of G. To determine the commutator group, the elements of the cartesian product $G \times G$ are classified according to the value of the commutator $a \circ b \circ a^{-1} \circ b^{-1}$ of the pair (a,b). The derived group is a normal subgroup in G and may therefore also be used to construct a quotient group. The quotient group G/D(G) is abelian; it is called the abelianized group G. The structure of G/D(G) may therefore be studied using the methods in Section 7.6.

Normalizer in a subgroup : Let H be a subgroup of a group $(G\,;\circ)$, and let A be a non-empty subset of G. The subset of H whose elements transform the subset A into itself is called the normalizer of A in H and is designated by $N_H(A)$. As a special case, the groups H and G may coincide.

$$N_H(A) := \{h \in H \mid A = h^{-1} \circ A \circ h\}$$

Properties of normalizers

(N1) The normalizer $N_H(A)$ is a subgroup of H. The normalizer may therefore be used to partition H into left and right cosets. The left and right cosets may, however, be different : The normalizer $N_H(A)$ is generally not a normal subgroup in H. For $H = G$, the normalizer $N_G(A)$ of every subset A of G is a subgroup of G.

(N2) The normalizer $N_H(A)$ is the intersection of the normalizer $N_G(A)$ with the subgroup H : $N_H(A) = H \cap N_G(A)$.

(N3) If the subset A is a subgroup of G, then A is a normal subgroup in the normalizer $N_G(A)$.

(N4) A subgroup A of the subgroup H is a normal subgroup in the subgroup H if and only if $N_H(A) = H$.

(N5) The number of H-conjugates of the subset A is equal to the index of $N_H(A)$ in H, that is $[H : N_H(A)]$.

Proof : Properties of normalizers

(N1) The normalizer $N_H(A)$ contains the identity element 1_G, the inverse element h^{-1} for each element h and the product $h_1 \circ h_2$ for any two elements h_1, h_2. Hence $N_H(A)$ is a subgroup of H.

$$\bigwedge_{a\in A} (a \circ 1_G \;=\; 1_G \circ a) \;\Rightarrow\; 1_G \in N_H(A)$$

$$\bigwedge_{h\in H} (A \circ h \;=\; h \circ A \;\Rightarrow\; A \circ h^{-1} = h^{-1} \circ A)$$

$$\bigwedge_{h_i\in H} (A \circ h_1 \;=\; h_1 \circ A \;\wedge\; A \circ h_2 = h_2 \circ A \;\Rightarrow\; A \circ h_1 \circ h_2 \;=\; h_1 \circ h_2 \circ A)$$

(N2) The normalizer $N_G(A)$ contains every element $g \in G$ for which $g \circ A = A \circ g$ holds. The normalizer $N_H(A)$ contains every element $h \in H \subseteq G$ for which $h \circ A = A \circ h$ holds. Hence $N_H(A) = H \cap N_G(A)$.

(N3) With every element a, the subgroup A also contains the inverse element a^{-1}. It follows from $a = a^{-1} \circ a \circ a$ and $a \in G$ that every element $a \in A$ is an element of the normalizer $N_G(A)$. For every element n of the normalizer $N_G(A)$, by definition $n \circ A = A \circ n$. Hence A is a normal subgroup of $N_G(A)$.

(N4) The statements "A is a normal subgroup in H" and "$N_H(A) = H$" are equivalent :

$$A \lhd H \;\Leftrightarrow\; \bigwedge_{h\in H} (A \circ h = h \circ A) \;\Leftrightarrow\; \bigwedge_{h\in H} (h \in N_H(A)) \;\Leftrightarrow\; N_H(A) = H$$

(N5) By the definition in Section 7.4.4, the set $B \subseteq G$ is an H-conjugate of $A \subseteq G$ if there is an element $h \in H$ for which $B = h^{-1} \circ A \circ h$ holds. The number of H-conjugates of A is equal to the index of $N_H(A)$ in H if there is a bijective mapping of the right cosets of $N_H(A)$ in A to the H-conjugates of A. This is the case if elements $h_1, h_2 \in H$ which belong to the same coset of $N_H(A)$ also lead to the same H-conjugate of A and vice versa :

$$N_H(A) \circ h_1 \;=\; N_H(A) \circ h_2 \;\Leftrightarrow\; h_1^{-1} \circ A \circ h_1 \;=\; h_2^{-1} \circ A \circ h_2$$

$N_H(A) \circ h_1 = N_H(A) \circ h_2$ implies $n_1 \circ h_1 = n_2 \circ h_2$ with $n_1, n_2 \in N_H(A)$. Along with n_1 and n_2, the group $N_H(A)$ contains the element $n = n_1^{-1} \circ n_2$, so that $h_1 = n \circ h_2$. With $n^{-1} \circ A \circ n = A$ one obtains :

$$h_1^{-1} \circ A \circ h_1 \;=\; h_2^{-1} \circ n^{-1} \circ A \circ n \circ h_2 \;=\; h_2^{-1} \circ A \circ h_2$$

Conversely, it follows from $h_1^{-1} \circ A \circ h_1 \;=\; h_2^{-1} \circ A \circ h_2$ by multiplication with h_1 from the left and with h_1^{-1} from the right that $h_2 \circ h_1^{-1}$ is an element of $N_H(A)$:

$$h_1 \circ h_1^{-1} \circ A \circ h_1 \circ h_1^{-1} \;=\; A \;=\; h_1 \circ h_2^{-1} \circ A \circ h_2 \circ h_1^{-1}$$

It follows from $h_1 = n \circ h_2$ that h_1 is an element of the right coset $N_H(A) \circ h_2$. Since any two right cosets of a subgroup are either identical or disjoint, this implies $N_H(A) \circ h_1 = N_H(A) \circ h_2$.

Center of a group : In a non-commutative group $(G\,;\circ)$, the commutative law $a\circ b = b\circ a$ may be satisfied for a subset of $G\times G$. The subset of G whose elements commute with all elements of G is called the center of the group G and is designated by Z(G).

$$Z(G) := \{b\in G \mid \bigwedge_{a\in G} a\circ b = b\circ a\}$$

The center Z(G) of a group G has the following properties :

(Z1) The center Z(G) is a characteristic subgroup of G.

(Z2) Every subgroup of the center Z(G) is a normal subgroup in G.

(Z3) Let $A=\{a\}$ be a one-element subset of the group G. Then the normalizer $N_G(A)$ is the entire group G if a is an element of the center Z(G). Otherwise $N_G(A)$ is a proper subgroup of G.

(Z4) The center of every symmetric group S_n with $n\geq 3$ contains only the identity permutation i.

Proof : Properties of a center

(Z1) The center Z(G) of a group $(G\,;\circ)$ has the properties of a group, since it contains the identity element 1_G, the inverse element b^{-1} for every element b and the element $b_1\circ b_2$ for any two elements b_1,b_2 :

$$\bigwedge_{a\in G} (a\circ 1_G = 1_G\circ a) \quad\Rightarrow\quad 1_G \in Z(G)$$

$$\bigwedge_{b\in Z} (\bigwedge_{a\in G} (a\circ b = b\circ a) \Rightarrow \bigwedge_{a\in G} (b^{-1}\circ a = a\circ b^{-1}))$$

$$\bigwedge_{b_1 b_2\in Z} (\bigwedge_{a\in G} (a\circ b_1 = b_1\circ a \;\wedge\; a\circ b_2 = b_2\circ a) \quad\Rightarrow$$

$$\bigwedge_{a\in G} (a\circ b_1\circ b_2 = b_1\circ b_2\circ a))$$

By Section 7.5.5, the subgroup Z(G) is characteristic in the group G if Z(G) is invariant under every automorphism $\phi : G\to G$. Thus for every element g of G and every element a of Z(G), the equation $g\circ\phi(a) = \phi(a)\circ g$ must hold. Since automorphisms are surjective, there is an element $b\in G$ such that $\phi(b) = g$, and hence

$$g\circ\phi(a) = \phi(b)\circ\phi(a) = \phi(b\circ a) = \phi(a\circ b) = \phi(a)\circ\phi(b) = \phi(a)\circ g$$

(Z2) For every element $a\in H\subseteq Z_G$ and every element $g\in G$, $a\circ g = g\circ a$. Thus $g\circ H = H\circ g$ for every $g\in G$, and hence H is a normal subgroup of G.

(Z3) For a one-element set $A=\{a\}$, the normalizer in the group G takes the form

$$N_G(A) = \{g\in G \mid g^{-1}\circ a\circ g = a\}$$

If a is an element of the center Z(G), then the condition $g^{-1} \circ a \circ g = a$ is satisfied for every element g of G. Hence in this case $N_G(A) = G$. If a is not an element of the center Z(G), then by definition there is at least one element in G for which the condition $g^{-1} \circ a \circ g = a$ is not satisfied. Thus $N_G(A) \subset G$. By property (N1) of normalizers, $N_G(A)$ is a proper subgroup of G.

(Z4) For every permutation $\phi \in S_n$ which is not the identity mapping, there is an element $a \in X_n$ which is not mapped to itself but to $\phi(a) = b \neq a$. For $n \geq 3$, there is an element $c \in X_n$ other than a and b. In Section 7.7.4, the ϕ-transform of the cycle <a,c> is shown to be the cycle $<\phi(a), \phi(c)>$. Since $\phi(a) = b$ is by hypothesis different from a and c, the cycle <a,c> does not commute with ϕ. Hence $\phi \neq i$ does not belong to the center of S_n, so that $Z(S_n) = \{i\}$ for $n \geq 3$.

$$\begin{aligned} \phi \neq i \quad & \Rightarrow \quad \bigvee_{a \in X_n} (\phi(a) = b \neq a) \\ n \geq 3 \quad & \Rightarrow \quad \bigvee_{c \in X_n} (c \neq a \ \wedge \ c \neq b \ \wedge \ \phi \circ <a,c> \circ \phi^{-1} = <b, \phi(c)>) \\ & \Rightarrow \quad \phi \circ <a,c> = <b, \phi(c)> \circ \phi \ \neq \ <a,c> \circ \phi \end{aligned}$$

Class equation of a group : Let (G ; ∘) be a finite group with center Z(G). The group is partitioned into conjugacy classes. Let $t_1, ..., t_m$ be those representatives of these classes which are not contained in Z(G). The normalizers $N_G(T_k)$ of the one-element sets $T_k = \{t_k\}$ in G possess the indices $[G : N_G(T_k)]$. The order of the group G is the sum of the order of the center Z(G) and the indices $[G : N_G(T_k)]$ for $k = 1,...,m$. This relationship is called the class equation of the group G.

$$\begin{aligned} \text{ord } G \quad & = \quad \text{ord } Z(G) \ + \ \sum_{k=1}^{m} [G : N_G(T_k)] \\ N_G(T_k) \quad & = \quad \{g \in G \mid t_k = g^{-1} \circ t_k \circ g\} \end{aligned}$$

Proof : Class equation of a group

The group G is partitioned into s disjoint conjugacy classes R_i according to Section 7.4.4. Then the order of G is equal to the sum of the orders of the classes R_i :

$$\text{ord } G = \sum_{i=1}^{s} \text{ord } R_i \qquad \text{(i)}$$

If the representative of R_i is an element a of the center Z(G), then this element commutes with every element $g \in G$, that is $a = g^{-1} \circ a \circ g$. The class R_i therefore contains only the element a.

If the representative of R_i is not an element of the center Z(G), it is designated by t_k. In this case, the order of R_i is equal to the number of elements conjugate to t_k. By property (N5) of normalizers, this number is equal to the index of the normalizer $N_G(T_k)$ of the one-element set $T_k = \{t_k\}$ in G.

Substituting ord $R_i = 1$ in equation (i) for all classes with representative in Z(G) and ord $R_i = [G : N_G(T_k)]$ for all classes whose representative is not contained in Z(G) yields the class equation of the group G.

Example 1 : Class equation of the tetrahedral symmetry group

The symmetry group $G=\{a_0,...,a_{11}\}$ of a regular tetrahedron is introduced in Example 2 of Section 7.3.2. In Example 1 of Section 7.4.4, the group G is partitioned into conjugacy classes :

$$
\begin{aligned}
R_i &= [a_i] = \{a_k \in G \mid \bigvee_{g \in G} (a_k = g^{-1} \circ a_i \circ g)\} \\
R_1 &= [a_0] = \{a_0\} \\
R_2 &= [a_1] = \{a_1,a_4,a_5,a_8\} \\
R_3 &= [a_2] = \{a_2,a_3,a_6,a_7\} \\
R_4 &= [a_9] = \{a_9,a_{10},a_{11}\}
\end{aligned}
$$

The center Z(G) contains only the identity element a_0. The representatives not contained in Z(G) are a_1,a_2,a_9. The normalizers $N_G(T_k)$ for the sets $T_1=\{a_1\}$, $T_2=\{a_2\}$, $T_3=\{a_9\}$ are determined :

$$
\begin{aligned}
N_G(T_k) &= \{g \in G \mid t_k = g^{-1} \circ t_k \circ g\} \\
N_G(T_1) &= \{a_0,a_1,a_2\} \\
N_G(T_2) &= \{a_0,a_1,a_2\} \\
N_G(T_3) &= \{a_0,a_9,a_{10},a_{11}\}
\end{aligned}
$$

The right cosets of the normalizers $N_G(T_i)$ are

$$
\begin{aligned}
N_G(T_k) \circ a_i &= \{g \in G \mid n \circ a_i = g \quad \wedge \quad n \in N_G(T_k)\} \\
N_G(T_1) \circ a_0 &= \{a_0,a_1,a_2\} \\
N_G(T_1) \circ a_3 &= \{a_3,a_5,a_{11}\} \\
N_G(T_1) \circ a_4 &= \{a_4,a_7,a_{10}\} \\
N_G(T_1) \circ a_6 &= \{a_6,a_8,a_9\}
\end{aligned}
$$

$$
\begin{aligned}
N_G(T_3) \circ a_0 &= \{a_0,a_9,a_{10},a_{11}\} \\
N_G(T_3) \circ a_1 &= \{a_1,a_4,a_5,a_8\} \\
N_G(T_3) \circ a_2 &= \{a_2,a_3,a_6,a_7\}
\end{aligned}
$$

The right cosets of the normalizers $N_G(T_1)$ and $N_G(T_2)$ are identical. The number $[G : N_G(T_k)]$ of cosets of $N_G(T_k)$ is counted. The class equation of the symmetry group G is satisfied :

$$
\begin{aligned}
&[G : N_G(T_1)] = [G : N_G(T_2)] = 4 \\
&[G : N_G(T_3)] = 3 \\
&\text{ord } G = 12 \\
&\text{ord } Z(G) + \sum_1^3 [G : N_G(T_k)] = 1+4+4+3 = 12
\end{aligned}
$$

Commutators : An expression is sought whose value indicates whether two elements a,b of a group $(G\,;\circ)$ commute. The expression $a\circ b = k\circ b\circ a$ with $k\in G$ is suitable for this purpose. For commuting elements, $a\circ b = b\circ a$, and hence $k = 1_G$. For non-commuting elements, $a\circ b\circ a^{-1}\circ b^{-1} \neq 1_G$.

The element $k = a\circ b\circ a^{-1}\circ b^{-1}$ of G is called the commutator of the pair $(a,b)\in G\times G$ and is designated by [a,b]. The name commutator indicates that the order of the elements in the product $a\circ b$ is reversed if it is multiplied by the commutator [a,b] :

$$[a,b]\circ b\circ a \;=\; a\circ b\circ a^{-1}\circ b^{-1}\circ b\circ a \;=\; a\circ b$$

Commutator group : Generally, not every element g of a group $(G\,;\circ)$ can be represented as a commutator [a,b] with $(a,b)\in G\times G$. The set of all commutators and their finite products is designated by [G,G]. The domain $([G,G]\,;\circ)$ is a subgroup of G; it is called the commutator group of G :

(1) The set [G,G] contains the unit element, since $[a,a] = a\circ a\circ a^{-1}\circ a^{-1} = 1_G$.

(2) By definition, the set [G,G] contains the product of any two elements of [G,G].

(3) Together with the element [a,b], the set [G,G] also contains the inverse [b,a] :

$$[a,b]\circ[b,a] \;=\; a\circ b\circ a^{-1}\circ b^{-1}\circ b\circ a\circ b^{-1}\circ a^{-1} \;=\; 1_G$$

Properties of the commutator group

(K1) The commutator group [G,G] of a group $(G\,;\circ)$ is a characteristic subgroup of G : Every automorphism ϕ of G maps the elements of [G,G] to [G,G].

(K2) Let N be a normal subgroup of the group G. The quotient group G/N is abelian if and only if [G,G] is a subgroup of N.

(K3) [G,G] is the least normal subgroup of G which renders the quotient group G/[G,G] abelian. G/[G,G] is called the abelianized group G.

Proof : Properties of the commutator group

(K1) An automorphism ϕ is applied to an arbitrary commutator $[a, b]$. This yields $\phi(a \circ b \circ a^{-1} \circ b^{-1}) = \phi(a) \circ \phi(b) \circ \phi(a)^{-1} \circ \phi(b)^{-1}$, since the automorphism is by definition homomorphic. As $\phi(a)$ and $\phi(b)$ and their inverses $\phi(a)^{-1}$ and $\phi(b)^{-1}$ are elements of the group G, the image $k = \phi(a) \circ \phi(b) \circ \phi(a)^{-1} \circ \phi(b)^{-1}$ is by definition a commutator. Hence an arbitrary automorphism on G maps the elements of $[G,G]$ to $[G,G]$.

(K2) The homomorphism $f : G \to G/N$ maps the group to the quotient group G/N. The identity element of the quotient group is the normal subgroup N. For the product of arbitrary elements f(x) and f(y) of the quotient group :

$$
\begin{aligned}
f(x) \circ f(y) &= f(x \circ y) = f(x \circ y \circ (y \circ x)^{-1} \circ y \circ x) \\
f(x) \circ f(y) &= f(x \circ y \circ x^{-1} \circ y^{-1}) \circ f(y) \circ f(x)
\end{aligned}
$$

Let the commutator group $[G,G]$ be a subgroup of N. Then the commutator $x \circ y \circ x^{-1} \circ y^{-1}$ is an element of the normal subgroup N, so that its image $f(x \circ y \circ x^{-1} \circ y^{-1})$ is the identity element $f(1_G)$ of the quotient group G/N. But $f(x \circ y \circ x^{-1} \circ y^{-1}) = f(1_G)$ implies $f(x) \circ f(y) = f(y) \circ f(x)$: The quotient group G/N is abelian.

Conversely, let G/N be abelian. Then it follows from $f(x) \circ f(y) = f(y) \circ f(x)$ that $f(x \circ y \circ x^{-1} \circ y^{-1}) = f(1_G)$, and hence the commutator $x \circ y \circ x^{-1} \circ y^{-1}$ is an element of the normal subgroup N : $[G,G]$ is a subgroup of N.

(K3) If $N \subseteq G$ is a normal subgroup which renders G/N abelian, then by (K2) the inclusion $[G,G] \subseteq N$ holds. Since the characteristic subgroup $[G,G]$ of G is also a normal subgroup of G, the commutator group $[G,G]$ is the least normal subgroup of G which renders $G/[G,G]$ abelian.

Derivatives of a group : The commutator group $[G, G]$ is called the first derivative of G and is designated by $D^1 G$. G is correspondingly designated by $D^0 G$. The i-th derivative of the group G is defined inductively and is designated by $D^i G$:

$$
\begin{aligned}
D^0 G &= G \\
D^{i+1} G &= [D^i G, D^i G] \qquad i \in \mathbb{N}
\end{aligned}
$$

Since $D^{i+1}G$ is the commutator group of $D^i G$, $D^{i+1}G$ is always a normal subgroup of $D^i G$. The quotient set $D^i G / D^{i+1} G$ is abelian.

Example 2 : Commutator group of the tetrahedral symmetry group

The group table of the symmetry group of regular tetrahedra is shown in Example 2 of Section 7.3.2. This symmetry group is isomorphic to the alternating group A_4. The following table shows the cartesian product $A_4 \times A_4$ with the values of the commutators [a,b] of the symmetry group. The commutator group $[A_4, A_4]$ is $\{a_0, a_9, a_{10}, a_{11}\}$.

[ ]	a_0	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}
a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0	a_0
a_1	a_0	a_0	a_0	a_9	a_{10}	a_{11}	a_{10}	a_{11}	a_9	a_{11}	a_9	a_{10}
a_2	a_0	a_0	a_0	a_{11}	a_9	a_{10}	a_9	a_{10}	a_{11}	a_{10}	a_{11}	a_9
a_3	a_0	a_9	a_{11}	a_0	a_0	a_{11}	a_{10}	a_9	a_{10}	a_{10}	a_{11}	a_9
a_4	a_0	a_{10}	a_9	a_0	a_0	a_9	a_{11}	a_{10}	a_{11}	a_{11}	a_9	a_{10}
a_5	a_0	a_{11}	a_{10}	a_{11}	a_9	a_0	a_0	a_9	a_{10}	a_{11}	a_9	a_{10}
a_6	a_0	a_{10}	a_9	a_{10}	a_{11}	a_0	a_0	a_{11}	a_9	a_{10}	a_{11}	a_9
a_7	a_0	a_{11}	a_{10}	a_9	a_{10}	a_9	a_{11}	a_0	a_0	a_{10}	a_{11}	a_9
a_8	a_0	a_9	a_{11}	a_{10}	a_{11}	a_{10}	a_9	a_0	a_0	a_{11}	a_9	a_{10}
a_9	a_0	a_{11}	a_{10}	a_{10}	a_{11}	a_{11}	a_{10}	a_{10}	a_{11}	a_0	a_0	a_0
a_{10}	a_0	a_9	a_{11}	a_{11}	a_9	a_9	a_{11}	a_{11}	a_9	a_0	a_0	a_0
a_{11}	a_0	a_{10}	a_9	a_9	a_{10}	a_{10}	a_9	a_9	a_{10}	a_0	a_0	a_0

table of commutators : $[a,b] = a \circ b \circ a^{-1} \circ b^{-1}$ with $a,b \in \{a_0, ..., a_{11}\}$

Example 3 : Abelianized tetrahedral symmetry group

The first derivative $D^1 A_4 = \{a_0, a_9, a_{10}, a_{11}\}$ of the tetrahedral symmetry group A_4 is determined in Example 2. The group $N := D^1(A_4)$ is a normal subgroup in A_4; it leads to the following classification of A_4 :

$$a_0 \circ N = \{a_0, a_9, a_{10}, a_{11}\} = [a_0]$$

$$a_1 \circ N = \{a_1, a_4, a_5, a_8\} = [a_1]$$

$$a_2 \circ N = \{a_2, a_3, a_6, a_7\} = [a_2]$$

The abelianized group A_4 is the quotient group $\{[a_0], [a_1], [a_2]\}$. For any two elements, the product of their images under the mapping $f : G \to G/N$ is commutative:

$$a_9 \circ a_5 = a_1 : \quad f(a_9) \circ f(a_5) = f(a_1) = [a_1]$$

$$a_5 \circ a_9 = a_4 : \quad f(a_5) \circ f(a_9) = f(a_4) = [a_1]$$

$$\text{Thus} : \quad f(a_9) \circ f(a_5) = f(a_5) \circ f(a_9)$$

7.8.3 GROUPS OF PRIME-POWER ORDER

Introduction : Lagrange's Theorem shows that the order of every subgroup of a finite group $(G\,;\circ)$ is a divisor of the order of G. The question arises whether conversely every group whose order has the divisor m contains a subgroup of order m. If the group G is abelian, then it contains such a subgroup. If the group G is not commutative, then G does not necessarily contain such a subgroup. For example, the alternating group A_4 does not contain a subgroup of order 6, although 6 is a divisor of the order 12 of A_4 (see Example 1).

The theorems of Sylow show that a group whose order is divisible by a prime power p^a always contains a subgroup of order p^a. A subgroup of order p^s with $s < a$ is always a subgroup of a group of order p^a. A group with maximal exponent a is called a Sylow p-subgroup of G. The number of Sylow p-subgroups of G is of the form $1 + kp$ and is a divisor of the order of G.

Groups without proper subgroups : Let $(G\,;\circ)$ be a group other than the trivial group $\{1\}$. Let G be a group without proper subgroups; that is, there are no subgroups of G other than $\{1\}$ and G itself. Then G is a cyclic group of prime order.

Proof : Prime order of groups without proper subgroups

(1) The group G contains an element $a \neq 1$. The subgroup $gp(a)$ generated by the element a is different from $\{1\}$. Therefore $gp(a) = G$, and hence the group G is cyclic.

(2) The cyclic group G is not isomorphic to the infinite cyclic group $(\mathbb{Z}\,;+)$, since the group $\mathbb{Z}$ contains the subgroups $\mathbb{Z}n$. Hence the group G is finite.

(3) If there were a divisor n of the order of G, then by property (U2) of cyclic groups in Section 7.3.6 there would be a cyclic subgroup of order n in G. Since G does not contain any proper subgroups, the order of G is a prime.

Existence of subgroups of finite abelian groups : Let $(G\,;\circ)$ be a finite abelian group. Let the positive integer m be a divisor of the order of G. Then there is a subgroup of order m in G.

Proof : Existence of subgroups of finite abelian groups

The proof is carried out by induction. The statement is true for the group $\{1\}$. Let it be true for all groups of order less than ord G. It is proved that in this case the statement is also true for a group of order ord G.

(1) Let the divisor m of ord G be a prime p. If G itself is of order p, then the statement is true. For $p <$ ord G, there is an element $a \neq 1$ of G. The cyclic group $N = gp(a)$ is a normal subgroup of the abelian group G. It is used to

form the quotient group $H = G/N$. By Lagrange's Theorem in Section 7.4.2, $\text{ord}\, G = \text{ord}\, H \cdot \text{ord}\, N$. Since the prime p is a divisor of ord G, it is also a divisor of ord H or of ord N.

If the prime p is a divisor of ord N, then by property (U2) of cyclic subgroups the cyclic group N contains a subgroup of order p. Since N is a subgroup of G, G contains a subgroup of order p.

If the prime p is a divisor of $\text{ord}\, H < \text{ord}\, G$, then by the induction hypothesis $H = G/N$ contains a subgroup H' of order p. Since the group H' is of prime order p, it is cyclic by corollary (F3) to Lagrange's Theorem, and hence $H' = \text{gp}(h)$ with $h^p = 1_H$.

It is now to be proved that G contains a subgroup of order p if $H = G/N$ contains a subgroup of order p. Let the generating element h of H' be the class [y] of G/N. Let the order of the element y in G be s. Then $[y] \circ [y] = [y \circ y]$ leads to $h^s = [y^s] = [1_G] = 1_H$. Thus the prime p is a divisor of s. By property (U2) of cyclic subgroups, the subgroup gp(y) of G contains a subgroup of order p. Hence G also contains a subgroup of order p.

(2) Now assume that the divisor m of ord G is not a prime. Then there is a prime factor p of m, and by part (1) of the proof the group G contains a subgroup N of order p. This subgroup is a normal subgroup in the abelian group G. Hence there is a natural homomorphism $k : G \to H$ with $H = G/N$. Thus by Lagrange's Theorem $\text{ord}\, G = p \cdot \text{ord}\, H$.

Since the order of G contains the factor m, the order of H contains the factor $\frac{m}{p}$. By the induction hypothesis, $\text{ord}\, H < \text{ord}\, G$ implies that H contains a subgroup H' of order $\frac{m}{p}$. Thus by the extended third isomorphism theorem there is a subgroup $G' \subseteq G$ with $H' = G'/N$. Then Lagrange's Theorem shows that G' is of order $\text{ord}\, G' = \text{ord}\, H' \cdot \text{ord}\, N = m$, and hence the group G contains a subgroup of order m.

First theorem of Sylow : Let $(G ; \circ)$ be a finite group, p a prime, m a natural number and p^m a divisor of the order of G. Then G contains a subgroup of order p^m.

Proof : First theorem of Sylow

The proof is carried out by induction. The statement is true for a group of order p. Let it be true for all groups of order less than ord G. It is proved that in this case the statement is also true for a group of order ord G.

(1) Let the prime p be a divisor of the order of the center Z(G). Since Z is abelian, by the preceding theorem Z contains a subgroup N of order p. By property (Z2) of centers, the subgroup N is a normal subgroup of G, and hence there is a natural homomorphism $k : G \to H$ with $H = G/N$. By Lagrange's Theorem $\text{ord}\, G = \text{ord}\, N \cdot \text{ord}\, H = p \cdot \text{ord}\, H$.

Since ord G is divisible by p^m, it follows that ord H is divisible by p^{m-1}. Since ord H < ord G, by the induction hypothesis H contains a subgroup H' of order p^{m-1}. By the third isomorphism theorem, G contains a subgroup G' such that $H' = G'/N$. By Lagrange's Theorem, ord G' = ord $H' \cdot$ ord $N = p^m$, and hence the group G contains a subgroup G' of order p^m.

(2) Assume that the prime p is not a divisor of the order of the center Z(G). The class equation of G is $\text{ord}\, G = \text{ord}\, Z(G) + \Sigma\,[G : N_G\,(T_n)]$. Since the prime p divides the order of G but not the order of Z(G), it follows that p does not divide the sum $\Sigma\,[G : N_G\,(T_n)]$. Hence there is at least one representative T_i for which $[G : N_G\,(T_i)]$ is not divisible by p.

By Lagrange's Theorem, $\text{ord}\, G = [G : N_G\,(T_i)] \cdot \text{ord}\, N_G(T_i)$. Since p^m divides the order of G but p does not divide the index $[G : N_G(T_i)]$, the order of the normalizer $N_G(T_i)$ is divisible by p^m. By property (Z3) of centers, $N_G\,(T_i)$ is a proper subgroup of G, since the representative T_i is not an element of the center.

Since the order of $N_G(T_i)$ is divisible by p^m and less than ord G, the induction hypothesis implies that the normalizer $N_G(T_i)$ contains a subgroup H of order p^m. Since $N_G(T_i) \subset G$, the group G also contains the subgroup H of order p^m.

Cauchy's Theorem : In a finite group $(G\,;\circ)$, there is a cyclic subgroup of prime order p if and only if p divides the order of G.

Proof : Cauchy's Theorem

(1) If the prime p divides the order of G, then by the first theorem of Sylow G contains a subgroup of order p, and by corollary (F3) to Lagrange's Theorem every subgroup of prime order is cyclic.

(2) Conversely, if the group G contains a subgroup H of order p, then by Lagrange's Theorem the order p of H divides the order of G.

p-group : Let $(G\,;\circ)$ be a group, and let p be a prime. G is called a p-group if the order of every element of G is a power of p. The exponents m of the orders p^m of different elements of G may be different. A p-group is designated by G(p). The trivial group {1} is a p-group for every prime p.

$$\bigwedge_{a \in G(p)} (\text{ord}\, a = p^m \;\wedge\; m \in \mathbb{N}')$$

Properties of p-groups :

(P1) If a group $(G\,;\circ)$ is a p-group, then its order is a power n of the prime p : $\text{ord}\,G = p^n$.

(P2) If $(G\,;\circ)$ is a finite p-group other than $\{1\}$, then the prime p is a divisor of the order of the center of G, and hence $Z(G) \neq \{1\}$.

(P3) Let the order of a p-group $(G\,;\circ)$ be p^n with $n \geq 1$. Then G contains a normal subgroup of order p^{n-1}.

Proof : Properties of p-groups

(P1) If G is a p-group, then the order of every element of G is a power of p. By Cauchy's Theorem, the prime factorization of the order of G cannot contain primes other than p. Hence $\text{ord}\,G = p^n$.

(P2) For the p-group G and a normalizer $N_G(T_i)$ whose representative is not an element of the center Z(G), the class equation and Lagrange's Theorem yield :

$$\begin{aligned} \text{ord}\,G &= [G : N_G(T_i)] \cdot \text{ord}\,N_G(T_i) \\ &= \text{ord}\,Z(G) + \sum_{k=1}^{m} [G : N_G(T_k)] \end{aligned}$$

That T_i is not an element of the center Z(G) implies that $\text{ord}\,N_G(T_i) \neq \text{ord}\,G$, and hence $[G : N_G(T_i)] \neq 1$. Since G is a p-group, ord G and $\text{ord}\,N_G(T_i)$ are powers of p. Thus every index $[G : N_G(T_i)]$, and hence also the sum of these indices, is a multiple of p. Because ord G and the sum of the indices are divisible by p, ord Z(G) is divisible by p.

(P3) The proof is carried out by induction. Let $(G\,;\circ)$ be a group of order p^n. The statement is true for $n = 1$. Assume that the statement holds for all exponents of p less than n. By Cauchy's Theorem, G contains a cyclic subgroup gp(z) of order p. By (P2), the group $N_1 = gp(z)$ is a subgroup of the center Z(G), and hence by property (Z2) of the center it is a normal subgroup in G. Then Lagrange's Theorem yields $\text{ord}\,G = \text{ord}\,N_1 \cdot \text{ord}\,G/N_1$.

Since the order p^{n-1} of G/N_1 is less than p^n, by the induction hypothesis G/N_1 contains a normal subgroup N_2 of order p^{n-2}. By the extended third isomorphism theorem this implies that G contains a normal subgroup H with $N_2 = H/N_1$. Then Lagrange's Theorem yields $\text{ord}\,H = [H : N_1] \cdot \text{ord}\,N_1 = \text{ord}\,N_2 \cdot \text{ord}\,N_1 = p^{n-1}$. Thus G contains a normal subgroup H of order p^{n-1}.

Sylow p-subgroup : A subgroup S of a group $(G ; \circ)$ is called a Sylow p-subgroup of G if S is a p-group and every p-subgroup of G which contains S is identical with S.

Properties of Sylow p-subgroups : Let $(G ; \circ)$ be a group, p a prime, S a Sylow p-subgroup of G and H a p-subgroup of G.

(S1) If p^m is the highest power of the prime p which divides ord G, then there is a Sylow p-subgroup of order p^m in G.

(S2) The Sylow p-subgroup S is not a proper subgroup of the p-group H.

(S3) The normalizer of the Sylow p-subgroup S in H is the intersection of the groups S and H, that is $N_H(S) = H \cap S$.

(S4) Every G-conjugate of the Sylow p-subgroup S is a Sylow p-subgroup in G.

Proof : Properties of Sylow p-subgroups

(S1) By the first theorem of Sylow there is a subgroup S of order p^m in G. S is a Sylow p-subgroup if every p-subgroup H of G which contains S is identical with S. Let S be a subgroup of a p-group H. Then by Lagrange's Theorem :

$$\begin{aligned} \text{ord } G &= \text{ord } S \cdot [G : S] \\ \text{ord } H &= \text{ord } S \cdot [H : S] \\ \text{ord } G &= \text{ord } H \cdot [G : H] \;=\; \text{ord } S \cdot [H : S] \cdot [G : H] \end{aligned}$$

Since ord $S = p^m$ is the greatest power of p which divides ord G, it follows that p does not divide $[H : S] \cdot [G : H]$. Hence neither does p divide $[H : S]$. But since H is a p-group, it follows that $[H : S] = 1$ and ord S = ord H. Then $S \subseteq H$ and ord S = ord H imply $S = H$. Hence S is a Sylow p-subgroup.

(S2) Let the Sylow p-subgroup S be a subgroup of the p-subgroup H. Then by the definition of a Sylow p-subgroup $S = H$. Thus S is not a proper subgroup of H.

(S3) The normalizer $N_H(S)$ contains those elements $h \in H$ which transform every element $a \in S$ into an element $b = h^{-1} \circ a \circ h$ of S. For arbitrary elements $a, h \in H \cap S$, the transform $b = h^{-1} \circ a \circ h$ is an element of $H \cap S$, since $H \cap S$ is a group. Hence the normalizer $N_H(S)$ contains all elements of $H \cap S$, that is $H \cap S \subseteq N_H(S)$. It remains to be shown that $N_H(S) \subseteq H \cap S$. By property (N1) of normalizers, $N_H(S) \subseteq H$, so that only $N_H(S) \subseteq S$ remains to be shown.

By property (N2) of normalizers, $N_H(S)$ is a subgroup of $N_G(S)$. By property (N3) of normalizers, S is a normal subgroup of $N_G(S)$. Then the second isomorphism theorem implies :

S is a normal subgroup of $N_H(S) \circ S$. (i)

$S \cap N_H(S)$ is a normal subgroup of $N_H(S)$. (ii)

$$[N_H(S) \circ S : S] = [N_H(S) : N_H(S) \cap S] \tag{iii}$$

Lagrange's Theorem provides the relationships between the orders of the groups :

$$\text{ord } N_H(S)\circ S \;=\; [N_H(S)\circ S : S] \;\cdot\; \text{ord } S \qquad \text{(iv)}$$

$$\text{ord } N_H(S) \;=\; [N_H(S) : N_H(S)\cap S] \cdot \text{ord } N_H(S)\cap S \qquad \text{(v)}$$

Now (v) implies that $[N_H(S) : N_H(S)\cap S]$ is a power of p, since $N_H(S)$ and $N_H(S)\cap S$ are subgroups of the p-group H. By (iii), the index $[N_H(S)\circ S : S]$ is also a power of p. Since S is a p-group, (iv) implies that $N_H(S)\circ S$ is a p-group. By property (S2) of Sylow p-subgroups, S is not a proper subgroup of $N_H(S)\circ S$, so that $S = N_H(S)\circ S$. Hence $N_H(S) \subseteq N_H(S)\circ S = S$.

(S4) Let S_2 be a G-conjugate of the Sylow p-subgroup S_1, that is $S_2 = g^{-1}\circ S_1\circ g$ with $g\in G$. Let T_2 be a p-subgroup of G containing S_2. Then by Section 7.4.4 $T_1 := g\circ T_2\circ g^{-1}$ is also a p-subgroup. Since $S_1 = g\circ S_2\circ g^{-1}$, the p-subgroup T_1 contains S_1. Since S_1 is a Sylow p-subgroup, this implies $T_1 = S_1$. Thus $S_2 = g^{-1}\circ(g\circ T_2\circ g^{-1})\circ g = T_2$, and hence S_2 is a Sylow p-subgroup.

Second theorem of Sylow : Let (G ; ∘) be a finite group of order sp^m, where the number s is not divisible by the prime p. Let S be a subgroup of order p^m in G. Then every p-subgroup of G is contained in a subgroup of G conjugate to S.

Proof : Second theorem of Sylow

(1) Since p^m is the highest power of p which divides ord G, S is not properly contained in a p-subgroup of G. Hence S is a Sylow p-subgroup of G. The group S is not unique, since by property (S4) of Sylow p-subgroups every G-conjugate of S is also a Sylow p-subgroup of order p^m in G. The G-conjugates of S form a set M :

$$M \;=\; \{M_g \;\mid\; M_g = g^{-1}\circ S\circ g \quad\wedge\quad g\in G\}$$

(2) It is to be proved that every p-subgroup H of G is contained in one of the sets M_g. For this purpose the set M is partitioned into classes of H-conjugate sets with the system of representatives $\{M_1,...,M_s\}$ according to Section 7.4.4.

$$[M_i] \;=\; \{M_h\in M \;\mid\; \bigvee_{h\in H} (M_h = h^{-1}\circ M_i\circ h\,)\}$$

(3) If the condition $H\cap M_i = H$ is satisfied for the representative M_i, then it is also satisfied for every other element of the class $[M_i]$. It is therefore sufficient to determine whether H is a subgroup of one of the representatives M_i.

$$\begin{aligned}(a\in H \;\;\Rightarrow\;\; a\in M_i) \;\;\Rightarrow\;\; (b,h\in H \;\;&\Rightarrow\;\; c = h\circ b\circ h^{-1}\in H \;\;\Rightarrow\\ c\in M_i \;\;&\Rightarrow\;\; b = h^{-1}\circ c\circ h\in M_h)\end{aligned}$$

By property (S3) of Sylow p-subgroups, the intersection $H\cap M_i$ coincides with the normalizer $N_H(M_i)$. Hence H is a subgroup of M_i if $N_H(M_i) = H$, that is if $[H : N_H(M_i)] = 1$.

(4) By property (N5) of normalizers, the index $[H : N_H(M_i)]$ is equal to the number of H-conjugates of M_i, and the index $[G : N_G(S)]$ is equal to the number n of G-conjugates of S. Since the H-conjugates form a partition of M, summation over the classes yields :

$$n = \text{ord } M = [G : N_G(S)] = \sum_{i=1}^{s} [H : N_H(M_i)] \qquad \text{(i)}$$

(5) Lagrange's Theorem provides the relationships between the orders of the groups :

$$\text{ord } G = n \cdot \text{ord } N_G(S) \qquad \text{(ii)}$$

$$\text{ord } H = [H : N_H(M_i)] \cdot \text{ord } N_H(M_i) \qquad i = 1,\ldots,s \qquad \text{(iii)}$$

Property (N3) of normalizers shows that the Sylow p-subgroup S of order p^m is a subgroup of $N_G(S)$. Hence the number r in the order rp^m of $N_G(S)$ is not divisible by the prime p. Substituting ord $G = sp^m$ and ord $N_G(S) = rp^m$ into (ii) yields $n = \frac{s}{r}$. Since s is not divisible by p, n is not divisible by p. Hence at least one index $[H : N_H(M_j)]$ in (i) is not divisible by p. But in (iii) the orders of the p-group H and its subgroup $N_H(M_j)$ are powers of p. Hence the index $[H : N_H(M_j)]$ is equal to 1 : The Sylow p-subgroup M_j contains the p-group H as a subgroup.

Corollary to the second theorem of Sylow : The Sylow p-subgroups of a group G are G-conjugate.

Proof : Corollary to the second theorem of Sylow

Let S and P be Sylow p-subgroups of a group G for the same prime p. By the second theorem of Sylow, the p-subgroup P is contained in a subgroup of G which is conjugate to S. By definition, every p-subgroup of G which contains P coincides with P. Hence P is conjugate to S.

Third theorem of Sylow : Let $(G\,;\circ)$ be a finite group. Let the prime p be a divisor of the order of G. Then the number n_p of Sylow p-subgroups of G is a divisor of the order of G. The divisor has the form $n_p = 1 + kp$ with $k \in \mathbb{N}$.

Proof : Third theorem of Sylow

(1) Let the order of the group G be sp^m. Assume that s is not divisible by the prime p. Let S be a subgroup of order p^m in G. By the corollary to the second theorem of Sylow, the Sylow p-subgroups of G form a set M of G-conjugates of the group S. The set M is partitioned into classes of S-conjugate groups with the representatives $\{M_1,...,M_t\}$ according to Section 7.4.4.

$$M \;\; = \{M_g \mid M_g = g^{-1}\circ S \circ g \;\;\wedge\;\; g \in G\}$$
$$[M_i] = \{M_s \in M \mid \bigvee_{s \in S} (M_s = s^{-1}\circ M_i \circ s)\}$$

(2) The number $n_p = \text{ord } M$ of G-conjugates of S is to be determined. Since the classes of S-conjugates form a partition of M, the number n_p is equal to the sum of the numbers of S-conjugates over the classes. By property (N5) of normalizers, the number of S-conjugates of a representative M_i is equal to the index $[S : N_S(M_i)]$.

$$n_p \;\; = \sum_{i=1}^{t} [S : N_S(M_i)] \qquad \text{(i)}$$

(3) The reference group S provides a contribution $[S : N_S(S)] = [S : S] = 1$. For all other representatives, $M_i \neq S$. By property (S3) of Sylow p-subgroups, $N_S(M_i) = S \cap M_i$, so that $[S : N_S(M_i)] \neq 1$. Lagrange's Theorem provides the relationship between the orders of the groups :

$$\text{ord } S \;=\; [S : N_S(M_i)] \cdot \text{ord } N_S(M_i) \qquad \text{(ii)}$$

The orders of the Sylow p-subgroup S and its subgroup $N_S(M_i)$ are powers of the prime p. Since $[S : N_S(M_i)] \neq 1$, it follows from (ii) that the index is divisible by p, that is $[S : N_S(M_i)] = k_i\,p$ with $k_i \in \mathbb{N}$.

(4) The expressions for the indices determined in (3) are substituted into (i). With $k = \Sigma k_i$, the formula for n_p becomes

$$n_p \;\; = 1 + \sum_{M_i \neq S} k_i\,p \;=\; 1 + kp$$

(5) By property (N5) of normalizers, the number of G-conjugates of S is equal to the index $[G : N_G(S)]$ of the normalizer $N_G(S)$ in G, that is $n_p = [G : N_G(S)]$. By Lagrange's Theorem $\text{ord } G = \text{ord } N_G(S) \cdot [G : N_G(S)]$, and substitution yields $\text{ord } G = n_p \cdot \text{ord } N_G(S)$. Hence n_p is a divisor of the order of G.

Example 1 : Subgroups of the alternating group A_4

The alternating group A_4 and the symmetry group of a regular tetrahedron in Example 2 of Section 7.3.2 are isomorphic. In Section 7.7.7, the subgroups of the alternating group are obtained by enumeration as follows :

(a) Klein's four-group V_4 of order 4

(b) Four isomorphic groups $D_1 \cong D_2 \cong D_3 \cong D_4$ of order 3

(c) Three isomorphic groups $Z_1 \cong Z_2 \cong Z_3$ of order 2

In the following this result is obtained without enumeration using the theorems of Sylow. The prime factorization of the order of A_4 is $12 = 2^2 \cdot 3$. By property (S1) of Sylow p-subgroups, A_4 therefore contains Sylow p-subgroups of orders $2^2 = 4$ and 3. For the number n_p of Sylow p-subgroups the third theorem of Sylow yields :

$$\begin{aligned} n_2 &= 2k+1 \quad \wedge \quad n_2 \mid 12 \\ n_3 &= 3k+1 \quad \wedge \quad n_3 \mid 12 \end{aligned}$$

The orders of the elements are obtained using the product table in Section 7.3.2 :

$$\begin{aligned}
gp(a_1) &= \{a_0, a_1, a_2\} : & \text{ord } a_1 &= \text{ord } a_2 = 3 \\
gp(a_3) &= \{a_0, a_3, a_4\} : & \text{ord } a_3 &= \text{ord } a_4 = 3 \\
gp(a_5) &= \{a_0, a_5, a_6\} : & \text{ord } a_5 &= \text{ord } a_6 = 3 \\
gp(a_7) &= \{a_0, a_7, a_8\} : & \text{ord } a_7 &= \text{ord } a_8 = 3 \\
gp(a_9) &= \{a_0, a_9\} : & \text{ord } a_9 &= 2 \\
gp(a_{10}) &= \{a_0, a_{10}\} : & \text{ord } a_{10} &= 2 \\
gp(a_{11}) &= \{a_0, a_{11}\} : & \text{ord } a_{11} &= 2
\end{aligned}$$

The number of Sylow 3-subgroups is 4, since $k > 0$ by inspection, and since n_3 is not a divisor of ord $G = 12$ for $k > 1$. Hence $k = 1$ and $n_3 = 4$. The order of a Sylow 2-subgroup is 4. By definition, a 2-group cannot contain elements of order 3. The remaining elements $\{a_0, a_9, a_{10}, a_{11}\}$ form a subgroup of order 4. There are no further subgroups whose order divides 12. Thus there is only one Sylow 2-subgroup. It contains the 2-subgroups $gp(a_9)$, $gp(a_{10})$ and $gp(a_{11})$. In the third theorem of Sylow, $k = 0$ and $n_2 = 1$.

The group A_4 does not contain a subgroup of order 6, although the order of A_4 is divisible by 6. For if there were a subgroup H of order 6, then by the first theorem of Sylow it would have to contain subgroups of order 2 and 3, and thus for example the elements a_0, a_1, a_2, a_9. However, operations using these elements lead to more than 6 elements in the group H, for example $a_1 \circ a_9 = a_8$, $a_2 \circ a_9 = a_6$, $a_8 \circ a_1 = a_3$, etc. The same is true for the other combinations of elements from subgroups of orders 2 and 3. Hence A_4 does not contain any subgroups of order 6.

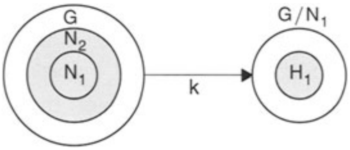
7.8.4 NORMAL SERIES

Introduction : If a group contains a normal subgroup, the quotient group with respect to this normal subgroup may be formed. The quotient group may contain a normal subgroup, and the quotient group of the quotient group with respect to this normal subgroup may be formed. By the extended third isomorphism theorem, the normal subgroups in the quotient group are associated with normal subgroups in the original group. Thus a nested chain of normal subgroups is formed. However, this chain generally does not contain all normal subgroups of the group.

If the trivial group $\{1\}$ is chosen as the first normal subgroup and the center of the group as the second normal subgroup, the continuation of this process yields the central series of the group. If the central series ends with the group itself, the group is said to be nilpotent. If the group itself is chosen as the first term and the commutator group as the second term, the continued formation of commutator groups leads to the derived series of the group. If the derived series ends with the trivial group $\{1\}$, the group is said to be soluble. Every nilpotent group is soluble, but the converse is not true.

A chain of normal subgroups which begins with the group itself and ends with the trivial group $\{1\}$ is called a normal series of the group. A second, longer chain of normal subgroups which contains the first chain is a refinement of the normal series. A normal series which cannot be refined is called a composition series. The quotients of a composition series are simple groups, that is groups without proper normal subgroups. Every finite group possesses at least one composition series. Any two composition series of a finite group have the same length. Their quotients are pairwise isomorphic.

Central series : The center $Z(G)$ of a group $(G\,;\circ)$ is by construction a normal subgroup of G (see Section 7.8.2). The extended third isomorphism theorem in Section 7.5.3 is used to construct further normal subgroups of G from $Z(G)$.



Only the group G is given for constructing the central series. The trivial group $\{1\}$ is chosen as a normal subgroup N_0 of G, and the quotient group $G/\{1\} \cong G$ is constructed. Its center $H_0 := Z(G/N_0) \cong Z(G)$ is by construction a normal subgroup of $G/\{1\}$. By the extended third isomorphism theorem, the preimage $N_1 := k^{-1}(H_0)$ is a normal subgroup of G and contains N_0. Also $H_0 = N_1/N_0$. Using G and N_1, the quotient group G/N_1 is constructed. Let its center be H_1, that is $H_1 = Z(G/N_1)$.

Then by construction H_1 is a normal subgroup of G/N_1. By the extended third isomorphism theorem, the preimage $N_2 := k^{-1}(H_1)$ is a normal subgroup of G and contains N_1. Also $H_1 = N_2/N_1$.

The construction of normal subgroups is continued. Generally, let N_i be a normal subgroup of G, and let H_i be the center of G/N_i. Then there is a normal subgroup N_{i+1} of G such that $H_i = N_{i+1}/N_i$. The chain $\{1\} = N_0 \subseteq N_1 \subseteq N_2 \subseteq \ldots$ is called the (ascending) central series of the group G. The centers $Z(G/N_i) = H_i = N_{i+1}/N_i$ are called the quotients (factors) of the central series.

$$\begin{aligned} \{1\} &= N_0 \subseteq N_1 \subseteq N_2 \subseteq \ldots \\ Z(G/N_i) &= N_{i+1}/N_i \quad \wedge \quad N_{i+1} \lhd G \end{aligned}$$

Nilpotent groups : If the center of a quotient group G/N_i for $N_i \neq G$ consists only of the identity element, it follows that $N_i = N_{i+1} = \ldots$, so that the central series does not end with the group G. For example, the alternating group A_4 has the center $N_1 = \{1\}$, so that $N_0 = N_1 = \ldots = \{1\}$.

A group $(G\,;\circ)$ is said to be nilpotent if there is a natural number i for which $N_i = G$ holds in the central series. If G is nilpotent and n is the least number for which $N_n = G$ holds, then the group G is said to be nilpotent of class n. Abelian groups are nilpotent of a class $n \leq 1$. The subgroup N_i of a nilpotent group is a proper subgroup of N_{i+1} for $i < n$.

$$\{1\} = N_0 \subset N_1 \subset \ldots \subset N_n = G$$

Derived series : The commutator group $[G, G]$ of a group $(G\,;\circ)$ is defined in Section 7.8.2. The commutator group is also called the first derivative of G and is designated by $D_1 := [G, G]$ or by $D^1 G$. By property (K3) of commutator groups, D_1 is the least normal subgroup of G which renders the quotient group G/D_1 abelian.

The process of forming derivatives of the group G may be continued. The commutator group $D_2 := [D_1, D_1]$ is the least normal subgroup in D_1 which renders the quotient group D_1/D_2 abelian. The group D_2 is called the second derivative of G.

Generally, let $D_{i+1} = [D_i, D_i]$ be the commutator group of D_i. Then D_{i+1} is the least normal subgroup of D_i which renders the quotient group D_i/D_{i+1} abelian. The group D_{i+1} is called the (i + 1)-th derivative of G. If G is designated by D_0, the commutator groups $D_0 \supseteq D_1 \supseteq D_2 \supseteq \ldots$ form a chain of normal subgroups. This chain is called the derived series of the group G. The quotient groups D_i/D_{i+1} are called the quotients (factors) of the derived series.

$$\begin{aligned} G &= D_0 \supseteq D_1 \supseteq D_2 \supseteq \ldots \\ D_{i+1} &= [D_i, D_i] \quad \wedge \quad D_{i+1} \lhd G \end{aligned}$$

Soluble groups : If the groups D_i and D_{i+1} in a derived series are equal, the series does not end with the trivial group $\{1\}$. For example, for $n \le 5$ all derivatives $D^k S_n$ of the symmetric group S_n are equal to the alternating group A_n, and likewise all derivatives $D^k A_n$ are equal to A_n. Hence $D_0 = S_n$ and $D_1 = D_2 = ... = A_n$.

A group $(G ; \circ)$ is said to be soluble if there is a natural number i for which $D_i = \{1\}$ in the derived series. If G is soluble and n is the least number for which $D_n = \{1\}$, the group G is said to be soluble of length n. The subgroup D_{i+1} of a soluble group is a proper subgroup of D_i for $i < n$.

$$G = D_0 \supset D_1 \supset ... \supset D_n = \{1\}$$

Normal series : The central series of a nilpotent group and the derived series of a soluble group are examples of normal series of a group $(G ; \circ)$. A chain of subgroups G_i of G which begins with G and ends with $\{1\}$ is called a normal series of G of length n if for $i = 0,...,n-1$ the subgroup G_{i+1} is a normal subgroup of G_i. The quotient groups G_i / G_{i+1} are called the quotients (factors) of the normal series. A group is simple if it does does not possess a normal series containing proper subgroups.

$$G = G_0 \supset G_1 \supset ... \supset G_n = \{1\}$$

Properties of soluble groups :

(A1) A group $(G ; \circ)$ is soluble if and only if there is a normal series for G with abelian quotients.

(A2) Every subgroup of a soluble group is soluble.

(A3) The image of a soluble group under a homomorphic mapping is soluble. In particular, every quotient group of a soluble group is soluble.

(A4) If a normal subgroup N in a group G and the quotient group G/N are soluble, then the group G is soluble.

(A5) The cartesian product $G_1 \times ... \times G_n$ is soluble if the groups G_i are all soluble.

(A6) Every nilpotent group is soluble.

(A7) Not every soluble group is nilpotent.

(A8) Every p-group is nilpotent and hence soluble.

Proof : Properties of soluble groups

(A1) Let the group G be soluble. Then G has a derived series with abelian quotients D_i / D_{i+1}, and hence G has a normal series with abelian quotients.

Conversely, let $G = G_0 \supset ... \supset G_n = \{1\}$ be a normal series with abelian quotients G_i / G_{i+1}. The normal subgroup G_0 contains the derivative D_0, since $G_0 = D_0 = G$. It is now assumed that the normal subgroup G_i contains the derivative D_i of G, and this is shown to imply that G_{i+1} also contains the derivative D_{i+1}.

The least normal subgroup of G_i which renders the quotient group G_i / G_{i+1} abelian is the commutator group $[G_i, G_i]$. Since G_i / G_{i+1} is abelian by hypothesis, it follows that $[G_i, G_i] \subseteq G_{i+1}$. Then $D_i \subseteq G_i$ implies $D_{i+1} = [D_i, D_i] \subseteq [G_i, G_i] \subseteq G_{i+1}$. By induction $D_n \subseteq G_n = \{1\}$, and hence G is soluble.

(A2) The soluble group G has a derived series $G = D_0 \supset ... \supset D_n = \{1\}$. For a subgroup H of G, this implies :

$$H = G \cap H = (D_0 \cap H) \supseteq (D_1 \cap H) \supseteq ... \supseteq (D_n \cap H) = \{1\}$$

The group D_i contains the subgroup $D_i \cap H$ and the normal subgroup D_{i+1}. By the second isomorphism theorem, $D_i \cap H \cap D_{i+1} = D_{i+1} \cap H$ is a normal subgroup in $D_i \cap H$. For the canonical homomorphic mapping $f : D_i \cap H \to (D_i \cap H) / (D_{i+1} \cap H)$ and the elements x, y of $D_i \cap H$:

$$f(x) \circ f(y) = f(x \circ y) = f(x \circ y \circ x^{-1} \circ y^{-1}) \circ f(y) \circ f(x)$$

Since $x, y \in D_i$, the commutator $x \circ y \circ x^{-1} \circ y^{-1}$ is an element of D_{i+1}. Since $x, y \in H$, the commutator is an element of H. Thus the commutator is an element of the normal subgroup $D_{i+1} \cap H$, and hence its image is the identity element of the quotient group. Thus the quotient group is abelian : $f(x) \circ f(y) = f(y) \circ f(x)$. Eliminating repeated groups therefore yields a normal series with abelian quotients for the group H. Hence by property (A1) the subgroup H is soluble.

(A3) Let the group G be soluble with the derived series $G = D^0 G \supset ... \supset D^n G = \{1\}$. The general generating element of the derivative $D^1 H$ of an arbitrary subgroup H of G is $a \circ b \circ a^{-1} \circ b^{-1}$ with $a, b \in H$. The image of this element under a homomorphism f is given by $f(a \circ b \circ a^{-1} \circ b^{-1}) = f(a) \circ f(b) \circ f(a^{-1}) \circ f(b^{-1}) = f(a) \circ f(b) \circ f(a)^{-1} \circ f(b)^{-1}$. This is the general generating element of the derivative $D^1 f(H)$. Hence $f(D^1 H) = D^1 f(H)$.

Since H is an arbitrary subgroup of G, it follows by induction that $f(D^i H) = D^i f(H)$, and hence in particular $f(D^i G) = D^i f(G)$. Thus applying the homomorphism f to the derived series for G yields $f(G) = D^0 f(G) \supseteq ... \supseteq D^n f(G) = \{1\}$. If any of these inclusions is an equality, the inclusions to its right are also equalities; omitting them yields the derived series $f(G) = D^0 f(G) \supset ... \supset D^k f(G) = \{1\}$ for f(G). Hence f(G) is soluble.

Every quotient group G/N of G is the image of G under the corresponding canonical homomorphism $k : G \to G/N$. Hence G/N is soluble.

(A4) By property (A1), the soluble quotient group G/N possesses a normal series $G/N = H = H_0 \subset ... \subset H_s = \{1_H\}$ with the abelian quotients H_i / H_{i+1}. By the extended third isomorphism theorem, for every normal subgroup H_i in H there is a normal subgroup G_i in G such that $H_i = G_i /N$. Since G_{i+1} is a normal subgroup in G_i and by the third isomorphism theorem G_i /G_{i+1} is isomorphic to H_i / H_{i+1}, there is a normal series $G = G_0 \supset ... \supset G_s = N$ with abelian quotients G_i /G_{i+1}. For the soluble group N there is a derived series $N = N_0 \supset N_1 \supset ... \supset N_t = \{1_G\}$ with the abelian quotients N_i / N_{i+1}. Hence G is a soluble group with the following normal series :

$$G = G_0 \supset ... \supset G_s = N = N_0 \supset ... \supset N_t = \{1_G\}$$

(A5) The projection $p : G_1 \times G_2 \to G_2$ is a homomorphic mapping with the kernel $(G_1 ,1) := \{(a,1) \mid a \in G_1\}$. By hypothesis, the quotient group G_2 is soluble. The kernel $(G_1 ,1)$ is a normal subgroup in $G_1 \times G_2$; it is isomorphic to the soluble group G_1. Thus by (A4) the cartesian product $G_1 \times G_2$ is also soluble. It follows by induction that the cartesian product $G_1 \times ... \times G_n$ is soluble.

(A6) The nilpotent group G possesses a central series $G = N_n \supset ... \supset N_0 = \{1_G\}$. The quotient N_{i+1} /N_i is the center of G/N_i and is therefore abelian. Thus the central series is a normal series with abelian quotients. Hence by virtue of property (A1) G is soluble.

(A7) This property is proved by the following example.

(A8) The proof is performed by induction. Let the order of the group G be p^n with a prime p. The statement holds for $n = 0$, since in this case $G = \{1\}$. Let the statement be true for groups of order $< p^n$. Let the center of a group G of order p^n be $N := Z(G)$. By Lagrange's Theorem, ord G = ord $N \cdot$ ord G/N.

By property (P2) in Section 7.8.3, the order of the center of G is divisible by p. Since G is of order p^n, G/N is of order p^k with $k < n$. By the induction hypothesis, G/N is soluble. The center N is abelian and therefore soluble. By property (A4), since G/N and N are soluble the group G is also soluble.

Example : Normal series of upper triangular matrices

Let the elements of the group $(G\,;\circ)$ be the regular upper triangular matrices G_i of dimension n. Let the operation $\circ$ on elements be matrix multiplication. The identity element of the group is the identity matrix I. If G_i contains the diagonal element a in row m, then G_i^{-1} contains the diagonal element a^{-1} in row m. The commutator group D_1 contains the commutators $G_i \circ G_s \circ G_i^{-1} \circ G_s^{-1}$. These are upper triangular matrices with diagonal elements 1; they are designated by H_r.

$$\begin{array}{ccccc}
G_i \circ G_s \circ G_i^{-1} \circ G_s^{-1} & & \begin{bmatrix} x & & \\ & y & \\ & & z \end{bmatrix} & \begin{bmatrix} \frac{1}{a} & & \\ & \frac{1}{b} & \\ & & \frac{1}{c} \end{bmatrix} & \begin{bmatrix} \frac{1}{x} & & \\ & \frac{1}{y} & \\ & & \frac{1}{z} \end{bmatrix} \\
G_i & \begin{bmatrix} a & & \\ & b & \\ & & c \end{bmatrix} & \begin{bmatrix} ax & & \\ & by & \\ & & cz \end{bmatrix} & \begin{bmatrix} x & & \\ & y & \\ & & z \end{bmatrix} & \begin{bmatrix} 1 & & \\ & 1 & \\ & & 1 \end{bmatrix}
\end{array}$$

If H_r contains the element a in row m of the codiagonal, then H_r^{-1} contains the element $-a$ in row m of the codiagonal. The commutator group D_2 contains the commutators $H_r \circ H_s \circ H_r^{-1} \circ H_s^{-1}$; these are upper triangular matrices with diagonal elements 1 and codiagonal elements 0.

$$\begin{array}{ccccc}
H_r \circ H_s \circ H_r^{-1} \circ H_s^{-1} & & \begin{bmatrix} 1 & x & \\ & 1 & y \\ & & 1 \end{bmatrix} & \begin{bmatrix} 1 & -a & \\ & 1 & -b \\ & & 1 \end{bmatrix} & \begin{bmatrix} 1 & -x & \\ & 1 & -y \\ & & 1 \end{bmatrix} \\
H_r & \begin{bmatrix} 1 & a & \\ & 1 & b \\ & & 1 \end{bmatrix} & \begin{bmatrix} 1 & a+x & \\ & 1 & y+b \\ & & 1 \end{bmatrix} & \begin{bmatrix} 1 & x & \\ & 1 & y \\ & & 1 \end{bmatrix} & \begin{bmatrix} 1 & 0 & \\ & 1 & 0 \\ & & 1 \end{bmatrix}
\end{array}$$

The k-th derivative of G contains triangular matrices with diagonal elements 1 and elements 0 on the codiagonals 1,...,k – 1. The n-th derivative contains only the identity matrix I. Hence G is a soluble group and possesses the derived series $G = D_0 \supset D_1 \supset D_2 \supset \ldots \supset D_n = \{I\}$.

The center of the group $(G\,;\circ)$ contains the special upper triangular matrices A with $A \circ B = B \circ A$ for every B in G. If the coefficients of A are designated by a_{ik} and the coefficients of B are designated by b_{km}, then the values of a_{ik} must satisfy the following conditions for arbitrary values of b_{km} :

$$\sum_{k=i}^{m} a_{ik}\, b_{km} = \sum_{k=i}^{m} b_{ik}\, a_{km} \qquad i, m = 1,...,n$$

Comparison of coefficients shows that this condition is satisfied only for $A = sI$ with $s \in \mathbb{R}$. The soluble group $(G\,;\circ)$ possesses the central series $\{I\} \subset S \subseteq S \subseteq ...$ with $S = \{sI \mid s \in \mathbb{R}\}$. Thus G is not nilpotent.

Composition series : A normal series $G = H_0 \supset ... \supset H_n = \{1\}$ is called a refinement of the normal series $G = G_0 \supset ... \supset G_m = \{1\}$ of a group $(G\,;\circ)$ if each of the groups G_i occurs in the normal series $H_0 \supset ... \supset H_n$ and $n > m$. A normal series $G_0 \supset ... \supset G_n$ of a group G is called a composition series of G if there is no refinement of the series.

Similar composition series : Two composition series $G_0 \supset ... \supset G_r$ and $H_0 \supset ... \supset H_s$ are said to be similar if their lengths r and s are equal and the quotients may be arranged so that they are pairwise isomorphic. Thus for similar composition series there is an index permutation p such that

$$G_i / G_{i+1} \cong H_{p(i)} \,/\, H_{p(i+1)}$$

Properties of composition series :

(K1) A normal series of a non-trivial group is a composition series if and only if its quotients are simple groups.

(K2) Let a finite group $(G\,;\circ)$ be soluble. Then any normal series of G with abelian quotients may be refined to a composition series whose quotients are cyclic groups of prime order.

(K3) Every finite group possesses at least one composition series.

(K4) Any two composition series of a finite group are similar (Jordan-Hölder Theorem).

Proof : Properties of composition series

(K1) Let $G_0 \supset ... \supset G_n$ be a normal series of G. By the extended third isomorphism theorem, its quotients are simple if and only if there is no normal subgroup of G_i which contains G_{i+1} and is different from both G_i and G_{i+1}. This is the case if and only if the normal series cannot be refined and hence is a composition series.

(K2) Let $G = G_0 \supset \ldots \supset G_n = \{1\}$ be a normal series with abelian quotients. Assume that the order of the quotients G_i / G_{i+1} is not a prime. Since the order of the group G_i / G_{i+1} is a product of primes, by the first theorem of Sylow in Section 7.8.3 the group G_i / G_{i+1} contains a subgroup U of prime order p. Due to its prime order, the group U is cyclic. It is a normal subgroup in the abelian group G_i / G_{i+1}.

Since the quotient group G_i / G_{i+1} contains the proper normal subgroup U, by the extended third isomorphism theorem the group G_i contains a normal subgroup H with $G_i \supset H \supset G_{i+1}$ and $H / G_{i+1} = U$. Since G_i / G_{i+1} is abelian, by property (K2) of commutator groups G_{i+1} contains the derivative $[G_i, G_i]$. Then $H \supset G_{i+1}$ implies $H \supset [G_i, G_i]$, so that by (K2) the quotient G_i / H is also abelian. The quotient $H / G_{i+1} = U$ is also abelian. Hence $G_0 \supset \ldots \supset G_i \supset H \supset G_{i+1} \supset \ldots \supset G_n$ is a normal series with abelian quotients. Since G is a finite group, the normal series may be refined until the order of each quotient is a prime.

(K3) The statement is true for trivial groups with the composition series $G = G_0 = \{1\}$ and simple groups with the composition series $G = G_0 \supset G_1 = \{1\}$. Assume that the statement is true for groups of order $< s$. It is to be shown that in this case the statement is also true for non-simple groups of order s.

Let G be a group of order s. By hypothesis G contains a proper normal subgroup N of order $< s$ with the composition series $N = N_0 \supset \ldots \supset N_m = \{1\}$. Let N be a maximal proper normal subgroup in G. Then the quotient group G/N is simple. By property (K1), the group G of order s possesses at least one composition series $G \supset N \supset N_1 \supset \ldots \supset N_m = \{1\}$.

(K4) The statement is true for trivial and simple groups. Assume that the statement is true for groups of order $< s$. It is to be shown that in this case the statement is also true for non-simple groups of order s. Let two composition series be given for a non-simple group G of order s :

$$G = G_0 \supset G_1 \supset \ldots \supset G_r = \{1\} \qquad (1)$$

$$G = H_0 \supset H_1 \supset \ldots \supset H_t = \{1\} \qquad (2)$$

If $G_1 = H_1$, then since ord $G_1 <$ ord G the induction hypothesis implies that the lengths r and t are equal and that the quotients may be arranged so that they are isomorphic : $G_i / G_{i+1} \cong H_{p(i)} / H_{p(i+1)}$ for $0 \le i < r$. Hence the statement holds for the group G of order s.

If $G_1 \neq H_1$, the groups G_1 and H_1 are normal subgroups of G. By the second isomorphism theorem, $G_1 \circ H_1$ is a subgroup of G with normal subgroup $G_1 \lhd G_1 \circ H_1$. The subgroup $G_1 \circ H_1$ is a normal subgroup of G :

$$G_1 \lhd G \quad \Leftrightarrow \quad \bigwedge_{a \in G} \ \bigwedge_{g_1 \in G_1} \ \bigvee_{g_2 \in G_1} (a \circ g_1 \ = \ g_2 \circ a)$$

$$H_1 \lhd G \quad \Leftrightarrow \quad \bigwedge_{a \in G} \ \bigwedge_{h_1 \in H_1} \ \bigvee_{h_2 \in H_1} (a \circ h_1 \ = \ h_2 \circ a)$$

$$\bigwedge_{a \in G} \ \bigwedge_{\substack{g_1 \in G_1 \\ h_1 \in H_1}} \ \bigvee_{\substack{g_2 \in G_1 \\ h_2 \in H_1}} (a \circ g_1 \circ h_1 = g_2 \circ a \circ h_1 = g_2 \circ h_2 \circ a) \quad \Leftrightarrow \quad G_1 \circ H_1 \lhd G$$

By the third isomorphism theorem, $G_1 \circ H_1 / G_1$ is a normal subgroup in G/G_1. Since G_1 is a proper subgroup of $G_1 \circ H_1$, the group $G_1 \circ H_1 / G_1$ is non-trivial. But by the definition of a composition series the quotient group G/G_1 is simple. Hence $G/G_1 = (G_1 \circ H_1)/G_1$, and therefore $G = G_1 \circ H_1$. For the group $G = G_1 \circ H_1$ with the normal subgroups G_1 and H_1, the second isomorphism theorem implies :

- The group $F := G_1 \cap H_1$ is a normal subgroup of G_1 and of H_1.
- $G/G_1 \cong H_1/F$ and $G/H_1 \cong G_1/F$ (3)

By (K3), the finite group F possesses a composition series $F = F_0 \supset ... \supset F_m = \{1\}$. By the definition of a composition series, the quotient groups G/G_1 and G/H_1 are simple. Therefore by the isomorphism (3) the quotient groups H_1/F and G_1/F are simple groups. This yields the following composition series :

$$G = G_0 \supset G_1 \supset F \supset F_1 \supset ... \supset F_m = \{1\} \qquad (4)$$

$$G = H_0 \supset H_1 \supset F \supset F_1 \supset ... \supset F_m = \{1\} \qquad (5)$$

The quotients of the series (4) are G/G_1, G_1/F and F_i/F_{i+1}, the quotients of the series (5) are $G/H_1 \cong G_1/F$, $H_1/F \cong G/G_1$ and F_i/F_{i+1}. These quotients may be arranged so that they are pairwise isomorphic. Hence the composition series (4) and (5) are similar.

The composition series (1) and (4) contain composition series for the group G_1 with ord $G_1 <$ ord G. By the induction hypothesis, the series for G_1 are similar, and hence the series (1) and (4) for G are also similar. Likewise, the composition series (2) and (5) are similar. Hence the similarity of (4) and (5) implies the similarity of (1) and (2).

7.9 UNIQUE DECOMPOSITION OF ABELIAN GROUPS

Introduction : By the fundamental theorem for abelian groups in Section 7.6.6, every finitely generated abelian group may be represented as a direct sum of cyclic subgroups. In the representation provided by the fundamental theorem for abelian groups, the orders of the finite summands form a divisor chain. There is, however, also another representation for abelian groups. Example 1 of Section 7.6.4 shows that for the same number of summands different (isomorphic) subgroups may be chosen as summands. The number of summands may also vary. The question arises whether there is a systematic relationship between the different decompositions of an abelian group.

The search for the possible decompositions of a finitely generated abelian group begins with the decomposition of the cyclic groups into direct sums of indecomposable subgroups whose order is either infinite or a prime power. Using this result, the decomposition of the abelian group into cyclic summands determined in the fundamental theorem for abelian groups is refined by decomposing each of these summands into indecomposable cyclic groups. The finite summands of the refined decomposition are combined into the torsion subgroup, the infinite summands are combined into a torsion-free subgroup. The abelian group is the direct sum of the torsion subgroup and the torsion-free subgroup.

The decomposition of the torsion subgroup into summands of prime-power order is unique up to isomorphism. The number of summands of infinite order is also unique. Using this unique decomposition, every finitely generated abelian group may be described by a small number of invariants, which define the type of the group. Finitely generated abelian groups are isomorphic if and only if they are of the same type.

Representation as a direct sum : The representation of abelian groups as direct sums is treated in Section 7.6.4 with properties (D1) to (D3). Finite abelian groups have a further property :

(D4) Every finite abelian group is the direct sum of its Sylow p-subgroups.

Proof : Abelian group as a direct sum of its Sylow p-subgroups.

Let the order of the abelian group G be m. The natural number m is decomposed into prime powers : $m = m_1 \cdots m_s$ with $m_i = p_i^{n_i}$ and $p_i \neq p_k$ for $i \neq k$. By property (S1) of Sylow p-subgroups in Section 7.8.3, for every factor m_i the group G contains a Sylow p_i-subgroup H_i of order m_i. Since every G-conjugate of the abelian group H_i coincides with H_i, the corollary to the second theorem of Sylow shows that H_i is the only Sylow p_i-subgroup of G.

In the sum $H_1 + ... + H_s$, let $a_1 + ... + a_s = 0$ with $a_i \in H_i$ be a representation of the identity element. The r-th multiple of this equation is formed with $r = m_2 \cdots m_s$. Then $m_i a_i = 0$ implies $ra_2 = ... = ra_s = 0$, and it follows that $ra_1 = 0$. Since the order of H_1 is prime to r, this implies $a_1 = 0$. Analogously, it follows that $a_2 = ... = a_s = 0$. By (D1), the sum is therefore direct.

The order of the direct sum $H_1 \oplus ... \oplus H_s$ is the product $m_1 \cdots m_s$ of the orders of its summands, and is therefore by hypothesis equal to the order of G. Hence $H_1 \oplus ... \oplus H_s = G$.

Decomposable abelian groups : An abelian group $(G ; +)$ is said to be decomposable if it is the direct sum of at least two proper subgroups. Otherwise, the group is said to be indecomposable.

$$G \text{ is decomposable} \quad :\Leftrightarrow \quad G = G_1 \oplus ... \oplus G_n \quad \wedge \quad n \geq 2$$

By property (D4), every indecomposable abelian group is a p-group. There are, however, p-groups which are decomposable. For example, Klein's four-group in Example 1 of Section 7.6.4 is of order 2^2, but by Example 1 in Section 7.6.6 it is decomposable.

Decomposability of cyclic groups : The infinite cyclic group is indecomposable. A cyclic group of finite order m is indecomposable if and only if m is a prime power p^n.

Proof : Decomposability of cyclic groups

(1) Let $\mathbb{Z}a$ and $\mathbb{Z}b$ be arbitrary subgroups of the group $(\mathbb{Z} ; +)$ of the integers. With $c = \text{lcm}(a, b)$, the intersection $\mathbb{Z}a \cap \mathbb{Z}b$ is the subgroup $\mathbb{Z}c \neq \{0\}$. Hence by property (D2) in Section 7.6.4 the group $\mathbb{Z}$ is indecomposable. All infinite cyclic groups are isomorphic with $\mathbb{Z}$, and hence also indecomposable.

(2) Let G be a cyclic group of order p^n with the prime p and the exponent $n \geq 1$. Let H_1 and H_2 be subgroups of G with ord $H_1 = p^r$ and ord $H_2 = p^{r+s}$ and $r, s \geq 0$. Since p^r is a divisor of p^{r+s}, by property (U2) of cyclic groups in Section 7.3.6 H_2 contains exactly one subgroup S of order p^r. Again by (U2), G does not contain more than one subgroup of order p^r. Hence $S = H_1$, and therefore $H_1 \subseteq H_2$. By property (D2) in Section 7.6.4, the sum $H_1 + H_2$ is not direct, since $H_1 \cap H_2 \neq \{0\}$ for $H_1 \neq \{0\}$. Hence the group G is indecomposable.

Let a cyclic group G of order m be indecomposable. Then a direct sum for G contains only G itself. Since every cyclic group is abelian, by (D4) the group G is the direct sum of its Sylow p-subgroups. Hence the order m of G is a prime power p^n.

Decomposability of abelian groups : Every non-trivial, finitely generated abelian group is the direct sum of a finite number of indecomposable cyclic groups.

Proof : Decomposition of abelian groups into indecomposable cyclic groups

By the fundamental theorem for abelian groups, every finitely generated abelian group (G ; +) may be represented as a direct sum of a finite number of cyclic subgroups. The finite cyclic summands may in turn be decomposed into their Sylow p-subgroups T_i. These cyclic subgroups are indecomposable. The infinite cyclic summands U_k are also indecomposable. Hence every finitely generated abelian group is a direct sum of a finite number of indecomposable cyclic groups.

$$G = T_1 \oplus \ldots \oplus T_m \oplus U_1 \oplus \ldots \oplus U_n$$

T_i finite cyclic Sylow p-subgroups
U_k infinite cyclic groups

Notes :

(1) Klein's four-group of prime-power order 2^2 is decomposable but not cyclic.

(2) The indecomposability of a cyclic group of prime-power order does not imply that the group contains no subgroups, but rather that it cannot be represented as a direct sum of these subgroups.

(3) If an abelian group is decomposed into its Sylow subgroups, then for every prime p there is exactly one Sylow p-subgroup. This subgroup is, however, generally not cyclic.

(4) If an abelian group G is decomposed into cyclic subgroups, different summands T_i and T_m of the direct sum may each contain a Sylow p-subgroup H_i and H_m, respectively, for the same prime p. These Sylow p-subgroups are necessarily cyclic. The subgroup H_i is a Sylow p-subgroup in the summand T_i, but not in the group G. The subgroup H_m is a Sylow p-subgroup in the summand T_m, but not in the group G.

Torsion group : Let (G ; +) be a finitely generated abelian group. The elements of finite order in G form the torsion subgroup tor G of G (see Section 7.6.2). The abelian group G is said to be torsion-free if $\text{tor}\, G = \{0\}$. The group G is called a torsion group if $\text{tor}\, G = G$. As a subgroup of the abelian group G, the torsion subgroup tor G is a normal subgroup in G. According to Section 7.6.2, the quotient group $G / \text{tor}\, G$ is torsion-free.

Unique decomposition of an abelian p-group : Every finite abelian group $(G\,;+)$ whose order is a prime power p^m is the direct sum of indecomposable cyclic subgroups G_i whose orders p^{m_i} are uniquely determined. Hence the decomposition of the p-group G into indecomposable cyclic subgroups is unique up to isomorphism and the order of the summands.

$$\begin{array}{rclcrcl} G & = & G_1 \oplus ... \oplus G_s & \text{with} & G_i & = & gp(a_i) \\ \text{ord } G & = & p^m = p^{m_1 + ... + m_s} & \text{with} & \text{ord } G_i & = & p^{m_i} \end{array}$$

Proof : Unique decomposition of an abelian p-group

(1) It was already shown that every non-trivial, finitely generated abelian group is the direct sum $G_1 \oplus ... \oplus G_s$ of indecomposable cyclic groups. The proof that this decomposition is unique for p-groups is carried out by induction. It is to be shown that G is the direct sum of a unique number s of groups $G_1, ..., G_s$ with unique orders ord $G_i = m_i$.

(2) It was already proved that a cyclic group of prime order p is indecomposable. The statement is therefore true for ord $G = p$. For groups with ord $G = p^m$ with $m \geq 2$, the statement is assumed to be true for groups of order less than p^m. To prove that in this case the statement also holds for groups with ord $G = p^m$, the scaled group $p^k G$ with $k \in \mathbb{N}$ is considered.

(3) Since the cyclic group $G_i = gp(a_i)$ of order p^{m_i} satisfies $p^k G_i = gp(p^k a_i)$, m_i is the least exponent for which $p^{m_i} G_i = \{0\}$ holds. By Section 7.6.6 :

$$p^k G \quad = \quad p^k G_1 \oplus ... \oplus p^k G_s$$

Thus $r := \max\{m_1, ..., m_s\}$ is the least exponent with the property $p^r G = \{0\}$. Since this property depends only on the group G, r is the same for every decomposition of G. Hence every decomposition contains at least one summand of order p^r. The remaining summands form a decomposition of a finitely generated abelian group whose order is less than p^m into indecomposable cyclic subgroups. By the induction hypothesis, they are therefore unique up to isomorphism and order. Altogether, it follows that the decomposition of G is also unique up to isomorphism and the order of the summands.

Unique decomposition of a finite torsion group : Every finitely generated torsion group $G \neq \{0\}$ is the direct sum $G_1 \oplus ... \oplus G_m$ of indecomposable cyclic subgroups, whose orders $q_i :=$ ord G_i are prime powers. The orders $q_1, ..., q_m$ are uniquely determined. The decomposition of G into indecomposable cyclic groups is thus uniquely determined up to isomorphism and the order of the summands.

Proof : Unique decomposition of a finite torsion group

(1) Every finitely generated abelian group is the direct sum of a finite number of indecomposable cyclic groups G_i. Since G is a torsion group, the summands G_i are finite. But the order of an indecomposable finite cyclic group is a prime power. In the direct sum, summands whose orders are powers of the same prime p are combined into the Sylow p-subgroup S_k of G.

$$G = G_1 \oplus ... \oplus G_m = S_1 \oplus ... \oplus S_n$$

(2) The order of each Sylow p-subgroup S_k is uniquely determined. In the decomposition of each p-group S_k into indecomposable cyclic groups, the orders of these groups are uniquely determined. Thus the orders q_i of the summands $G_1, ..., G_m$ are uniquely determined.

(3) Cyclic groups of equal order are isomorphic. The orders $q_1, ..., q_m$ of the cyclic groups G_i in $G = G_1 \oplus ... \oplus G_m$ are uniquely determined. Hence this decomposition is uniquely determined up to isomorphism and the order of the summands.

Unique decomposition of a torsion-free group : Every finitely generated torsion-free abelian group is the direct sum of a finite number of indecomposable infinite cyclic subgroups whose number is uniquely determined. Hence the decomposition of G into indecomposable infinite cyclic groups is uniquely determined up to isomorphism and the order the summands.

Proof : Unique decomposition of a torsion-free group

(1) Every finitely generated abelian group is the direct sum of a finite number of indecomposable cyclic groups G_i. Since the group G is torsion-free, all summands G_i are infinite groups.

$$G = G_1 \oplus ... \oplus G_r$$

(2) The scaled group 2G is a normal subgroup in G. According to Section 7.6.6, the quotient group G/2G is the direct sum of groups H_i which are isomorphic to the quotient groups $G_i/2G_i$ of order 2. The order 2^r of the quotient group G/2G depends only on G. Hence r is uniquely determined.

$$G/2G = H_1 \oplus ... \oplus H_r \quad \text{with} \quad H_i \cong G_i/2G_i \quad \text{and} \quad \text{ord } H_i = 2$$
$$\text{ord } G/2G = 2^r$$

(3) Every infinite cyclic group is isomorphic to the additive group ($\mathbb{Z}$; +) of the integers. The number of infinite cyclic groups is uniquely determined. Hence the decomposition is uniquely determined up to isomorphism and the order of the summands.

Unique decomposition of an abelian group : Every finitely generated abelian group $(G\,;+)$ is the direct sum of its torsion subgroup T and a torsion-free subgroup U. The torsion subgroup T is the direct sum of a finite number of indecomposable cyclic groups $T_1,\ldots,T_m$ of unique prime-power order. The torsion-free subgroup U is the direct sum of indecomposable infinite cyclic groups $U_1,\ldots,U_n$ whose finite number n is uniquely determined. The decomposition of the group G into its torsion subgroup and a torsion-free subgroup is unique up to isomorphism and the order of the summands.

$G = T \oplus U$	finitely generated abelian group
$T = T_1 \oplus \ldots \oplus T_m$	torsion subgroup tor G
$U = U_1 \oplus \ldots \oplus U_n$	torsion-free subgroup

Proof : Unique decomposition of an abelian group

Every finitely generated abelian group $(G\,;+)$ is a direct sum of indecomposable cyclic groups. Let G be the direct sum of the finite cyclic groups $H_1,\ldots,H_r$ and the infinite cyclic groups $W_1,\ldots,W_s$:

$$G = H_1 \oplus \ldots \oplus H_r \oplus W_1 \oplus \ldots \oplus W_s$$

The direct sums T and U are formed from these summands. If all summands are finite, let $U := \{0\}$. If all summands are infinite, let $T := \{0\}$.

$$T = H_1 \oplus \ldots \oplus H_r$$
$$U = W_1 \oplus \ldots \oplus W_s$$

(1) The group T is the torsion subgroup of G if it contains exactly the elements of G of finite order. For an arbitrary element $g \in G = T \oplus U$ of finite order k, there is a unique sum $g = t + u$ with $t \in T$ and $u \in U$. Then $kg = kt + ku = 0$ implies $ku = 0$, and hence $u = 0$. Thus g is an element of T. Conversely, since T is a finite subgroup of G, every $t \in T$ is an element of G of finite order.

(2) The group U is torsion-free if for an arbitrary element $u \in U$ the equation $ku = 0$ is only satisfied for $k = 0$. Since $U = W_1 \oplus \ldots \oplus W_s$ is a direct sum, u is a unique sum $u_1 + \ldots + u_s$ of elements $u_i \in W_i$. By definition of the direct sum, the condition $ku = ku_1 + \ldots + ku_s = 0$ is only satisfied if each of the terms is zero, that is if $ku_i = 0$ for $i = 1,\ldots,s$. Since every element u_i is of infinite order, this implies $k = 0$. Hence the group U is torsion-free.

(3) The torsion group T contains exactly the elements of G of finite order and is therefore uniquely determined by the group G. The subgroup T is a normal subgroup of the abelian group G. The second isomorphism theorem yields $G/T = (U+T)/T \cong U/(U \cap T)$. Since G is the direct sum of the subgroups U and T, $U \cap T = \{0\}$. Together this yields $G/T \cong U/\{0\} \cong U$. Thus, since T is uniquely determined by G, U is uniquely determined by G up to isomorphism.

By the preceding proofs, the groups T and U have unique decompositions. The torsion group T is the direct sum $T_1 \oplus ... \oplus T_m$ of indecomposable cyclic groups T_i of prime-power order. The torsion-free group U is the direct sum $U_1 \oplus ... \oplus U_n$ of indecomposable infinite cyclic groups U_i. The decompositions of T and U are unique up to isomorphism and the order of the summands.

Type of a finitely generated abelian group : Every finitely generated abelian group may be described by a small number of invariants using its decomposition into indecomposable cyclic subgroups : The prime powers of the orders of the finite summands and the number of infinite summands. The tuple of the invariants of a finitely generated abelian group (G ; +) is called the type of the abelian group and is designated by type G.

$$\text{type } G \ := \ (p_1^{n_1}, ..., p_m^{n_m} ; n)$$

p_i primes $p_1 \le ... \le p_m$ with $m \ge 0$

n_i positive integer exponent of p_i : $p_i = p_{i+1} \Rightarrow n_i \le n_{i+1}$

n number of infinite summands

Two finitely generated abelian groups are isomorphic if and only if they are of the same type. The invariants of the type are named as follows :

$p_i^{n_i}$ i-th torsion coefficient of G

n Betti number of G

Proof : Isomorphism of abelian groups of the same type

(1) Let the groups G and H be of the same type. Then the summands G_i in $G = G_1 \oplus ... \oplus G_m$ and H_i in $H = H_1 \oplus ... \oplus H_m$ are pairwise isomorphic to a group of residue classes $\mathbb{Z}_{q_i}$ with $q_i = p_i^{n_i}$ or to the group $\mathbb{Z}$ of integers. The isomorphism of the summands implies the isomorphism of the direct sums.

(2) Let the groups G and H be isomorphic. Then there is a bijective homomorphic mapping $f : G \to H$ with $f(g) = h$. The group G has a unique decomposition $G = G_1 \oplus ... \oplus G_m$, which is described by type G. Hence for every element $g \in G$ there is a unique sum $g = g_1 + ... + g_m$ with $g_i \in G_i$. Since f is homomorphic, the image of the element g is :

$$h = f(g) = f(g_1 + ... + g_m) = f(g_1) + ... + f(g_m) = h_1 + ... + h_m$$

Since the representation $g_1 + ... + g_n$ of g is unique and the mapping f is bijective, the representation $h_1 + ... + h_m$ of h is also unique. Hence H may be represented as the direct sum $H_1 \oplus ... \oplus H_m$ with $H_i = f(G_i)$. Since f is bijective, ord H_i = ord G_i. Altogether, it follows that type G = type H.

8 GRAPHS

8.1 INTRODUCTION

The structure of mathematics is based on relations between the elements of sets. The sets contain given elements. The relations are also sets and contain tuples which are formed from the elements of sets according to given rules of operation. In this manner, relationships between selected elements of sets are described symbolically.

The relations between elements may be visualized diagrammatically. In the diagram, the elements are represented as vertices, whereas the relationships are represented as edges. Simple relations can thus be represented visually in a plane. A relation for which such a visualization exists is called a graph. The diagram of vertices and edges is often also called a graph.

Since a graph is a visualizable relation, the algebra of relations forms the basis of graph theory. The algebra of relations for finite sets may be transformed into a boolean vector and matrix algebra. Basic definitions and rules of the algebra of relations for finite graphs are treated in Section 8.2.

Various applications require graphs with specific properties. Simple graphs, directed graphs, bipartite graphs, multigraphs and hypergraphs are distinguished with respect to these properties. The nomenclature for graphs varies considerably in the literature. For instance, simple and directed graphs are often also called ordinary graphs and digraphs, respectively. The different classes of graphs, their properties and their relationships are treated in Section 8.3.

Graphs may be classified with respect to their structural properties. For this purpose, the graph is considered as a domain consisting of vertices and edges. An edge sequence in the graph is a chain of connected edges which form either a path or a cycle. The study of paths and cycles leads to a definition of different forms of connectedness of graphs. The removal of some of the vertices and edges of a graph (a cut in the graph) leads to subgraphs; their connectedness is an essential structural property of the graph. The fundamentals of the structural analysis and further classification of simple and directed graphs are treated in Section 8.4.

The determination of paths in networks with specific properties is a basic problem of graph theory. A network is represented as a weighted graph in which the edges are weighted according to the properties under consideration. The various path problems are unified by abstraction. This leads to a path algebra for weighted graphs. The fundamentals of the path algebra and the algebraic methods of solution are treated in Section 8.5.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

The determination of flows in networks is a problem in graph theory related to the theory of optimization. As in the case of path problems, the networks for flow problems are represented by weighted graphs. The flows in the network must satisfy the law of conservation of mass and may be bounded by given capacities. To determine optimal flows, the principles of optimization are applied to graph theory. The fundamentals of flows in networks are treated in Section 8.6.

Graph theory has a large spectrum of applications. In computer science, for example, graph theory is applied in the theory of automata, in the theory of networks and in connection with formal languages and data structures. In engineering, it is applied to object-oriented modelling in the study of communications, transport and supply systems and of planning, decision and production processes. Some applications are shown as examples in connection with the theoretical foundations. The theoretical foundations treated here, as well as their applications, are restricted to finite graphs.

8.2 ALGEBRA OF RELATIONS

8.2.1 INTRODUCTION

The algebra of relations is based on boolean algebra and the algebra of sets. The basic definitions of these algebras are treated in Chapters 2 and 3. Their relationship is indicated in the following.

Boolean algebra : Boolean algebra is based on the two truth values "false" and "true". Unary and binary operations are defined for these truth values. The negation $\neg$ is a unary operation. The conjunction $\wedge$ and the disjunction $\vee$ are binary operations.

Algebra of sets : The algebra of sets is based on definitions and rules of set theory. If the elements of a set are taken from a given reference set, then the set may be regarded as a unary relation in the reference set. Unary and binary operations for sets are defined on the basis of boolean operations. The complement $\overline{}$ is a unary operation. The intersection $\cap$ and the union $\cup$ are binary operations.

Algebra of relations : The algebra of relations is based on the definition of relations. A binary relation is a set of ordered pairs of elements. It is a subset of a cartesian product of two sets. Since every relation is a set, all operations on sets may also be applied to relations. The complement $\overline{}$ is a unary operation. The intersection $\sqcap$ and the union $\sqcup$ are binary operations. In addition to these operations on sets, the transposition ^{T} is defined as a unary operation for the dual relation and the multiplication $\circ$ is defined as a binary operation for the composition of relations.

Boolean vectors and matrices : For finite sets, the algebra of relations is conveniently represented in vector and matrix form. Unary relations are boolean vectors, binary relations are boolean matrices with the truth values "false" and "true". The rules of boolean vector and matrix algebra are similar to the rules of vector and matrix algebra for real numbers.

Notation : Unary relations are represented by lowercase letters, binary relations by uppercase letters. The corresponding boolean vectors and matrices appear in boldface.

8.2.2 UNARY RELATIONS

Introduction : A unary relation is a subset of a set. Thus all rules of the algebra of sets hold for unary relations. Unary relations are specified by boolean vectors. This leads to a boolean vector algebra for unary relations.

Definition : Let a non-empty set M of elements and a unary operation on these elements be given. The value of the unary operation Rx for an element x is true or false. The corresponding unary relation u is the set of all elements x for which the unary operation Rx is true. It is a subset of M. The number of elements in the unary relation u is designated by I u I.

$$u := \{x \in M \mid Rx\} \subseteq M$$

Vector representation : Let M be a set with n elements. The elements of M are indexed by a mapping f : N → M with $f(i) = x_i$ and $1 \le i \le n$, so that $M = \{x_1, ..., x_n\}$. A unary relation $u \subseteq M$ is a subset of M. The elements of M which belong to the relation u are specified by a boolean vector **u** of dimension n. Every element $x_i \in M$ is bijectively associated with an element $u_i \in \mathbf{u}$. If the relation u contains the element x_i, then u_i has the value true (1) ; otherwise u_i has the value false (0).

A boolean vector **u** is an n-tuple of the truth values $W = \{0,1\}$, and hence an element of the n-fold cartesian product W^n. The elements of a vector **u** are usually arranged in a column scheme by regarding the index of the element u_i as a row index. In formulations of general properties and rules, a vector **u** is represented by a general element u_i in square brackets.

$$\mathbf{u} = [u_i] = \begin{array}{|c|}\hline u_1 \\ \hline \vdots \\ \hline u_i \\ \hline \vdots \\ \hline u_n \\ \hline \end{array} \quad : \quad \begin{array}{l} W = \{0,1\} \\ \mathbf{u} \in W^n \end{array}$$

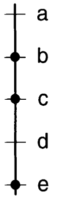
Graphical representation : The elements of a set may be represented in a one-dimensional grid. Each grid point corresponds to an element. The grid points contained in a unary relation are marked. The grid diagram is a graphical image of the boolean vector.

Example 1 : Representation of unary relations

Let a set M of elements and a unary relation $u \subseteq M$ be given. The grid diagram and the boolean vector **u** are shown :

$$M = \{a, b, c, d, e\}$$

$$u = \{b, c, e\} \subseteq M$$



$$\mathbf{u} = \begin{array}{|c|}\hline 0 \\ \hline 1 \\ \hline 1 \\ \hline 0 \\ \hline 1 \\ \hline \end{array} \begin{array}{c} a \\ b \\ c \\ d \\ e \end{array}$$

Special relations : The empty relation $\emptyset$ and the universal relation $e = M$ are special unary relations in the set M. They are also called the null relation and the all (complete, total) relation. The null relation corresponds to the boolean zero vector **0**, the all relation e corresponds to the boolean one vector **e**. A unary relation with exactly one element $x \in M$ is called a point relation or a point. A point relation is represented by a boolean unit vector. It is often designated by the name of its element.

null relation	$\emptyset := \{\ \}$
point relation	$x := \{x\}$
all relation	$e := M$

null relation $\emptyset$	point relation x	all relation e
$\mathbf{0} = \begin{array}{\|c\|}\hline 0 \\ \hline 0 \\ \hline 0 \\ \hline \vdots \\ \hline 0 \\ \hline \end{array}$	$\mathbf{x} = \begin{array}{\|c\|}\hline 0 \\ \hline 1 \\ \hline 0 \\ \hline \vdots \\ \hline 0 \\ \hline \end{array}$	$\mathbf{e} = \begin{array}{\|c\|}\hline 1 \\ \hline 1 \\ \hline 1 \\ \hline \vdots \\ \hline 1 \\ \hline \end{array}$

Equality and inclusion : The operations of equality $u = v$ and inclusion $u \sqsubseteq v$ on the relations u, v yield the logical constant true or false. If $u = v$ is true, then u and v are equal. If $u \sqsubseteq v$ is true, then u is contained in v.

equality	$u = v \quad :\Leftrightarrow \quad \bigwedge_{x} (x \in u \Leftrightarrow x \in v)$
inclusion	$u \sqsubseteq v \quad :\Leftrightarrow \quad \bigwedge_{x} (x \in u \Rightarrow x \in v)$
equality	$\mathbf{u} = \mathbf{v} \quad :\Leftrightarrow \quad \bigwedge_{i} (u_i \Leftrightarrow v_i)$
inclusion	$\mathbf{u} \sqsubseteq \mathbf{v} \quad :\Leftrightarrow \quad \bigwedge_{i} (u_i \Rightarrow v_i)$

Unary operation : The complement $\bar{u}$ is a unary operation on the relation u in the set M. It contains all elements $x \in M$ which are not contained in u.

complement	$\bar{u} := \{x \in M \mid x \notin u\}$
complement	$\bar{\mathbf{u}} := [\neg u_i]$

Binary operations : The intersection $u \sqcap v$ and the union $u \sqcup v$ are binary operations on the relations u and v. They are defined according to set theory :

intersection	$u \sqcap v := \{x \mid x \in u \wedge x \in v\}$
union	$u \sqcup v := \{x \mid x \in u \vee x \in v\}$
intersection	$\mathbf{u} \sqcap \mathbf{v} := [u_i \wedge v_i]$
union	$\mathbf{u} \sqcup \mathbf{v} := [u_i \vee v_i]$

Example 2 : Operations on unary relations

Let two boolean vectors $\mathbf{u}, \mathbf{v} \in W^4$ be given. The complement $\bar{\mathbf{u}}$, the intersection $\mathbf{u} \sqcap \mathbf{v}$ and the union $\mathbf{u} \sqcup \mathbf{v}$ are determined.

$$\mathbf{u} = \begin{array}{|c|}\hline 0 \\ \hline 1 \\ \hline 0 \\ \hline 1 \\ \hline\end{array} \quad \mathbf{v} = \begin{array}{|c|}\hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 0 \\ \hline\end{array} \quad \bar{\mathbf{u}} = \begin{array}{|c|}\hline 1 \\ \hline 0 \\ \hline 1 \\ \hline 0 \\ \hline\end{array} \quad \mathbf{u} \sqcap \mathbf{v} = \begin{array}{|c|}\hline 0 \\ \hline 1 \\ \hline 0 \\ \hline 0 \\ \hline\end{array} \quad \mathbf{u} \sqcup \mathbf{v} = \begin{array}{|c|}\hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline\end{array}$$

Algebraic structure : The unary relations in a reference set M are subsets of M. All of the different subsets of M are collected in the power set P(M). According to Section 3.4.3, the domain $(P(M); \sqcap, \sqcup, \bar{\ })$ with the power set P(M) and the operations $\sqcap, \sqcup, \bar{\ }$ is a boolean lattice. The following laws hold for the operations on elements u, v, w of the power set P(M) :

Property	Intersection $\sqcap$	Union $\sqcup$
associative	$(u \sqcap v) \sqcap w = u \sqcap (v \sqcap w)$	$(u \sqcup v) \sqcup w = u \sqcup (v \sqcup w)$
commutative	$u \sqcap v = v \sqcap u$	$u \sqcup v = v \sqcup u$
adjunctive	$u \sqcap (u \sqcup v) = u$	$u \sqcup (u \sqcap v) = u$
distributive	$u \sqcap (v \sqcup w) = (u \sqcap v) \sqcup (u \sqcap w)$	$u \sqcup (v \sqcap w) = (u \sqcup v) \sqcap (u \sqcup w)$
zero element	$u \sqcap \emptyset = \emptyset$	$u \sqcup \emptyset = u$
unit element	$u \sqcap e = u$	$u \sqcup e = e$
complement	$u \sqcap \bar{u} = \emptyset$	$u \sqcup \bar{u} = e$

In a boolean lattice, the inclusion $u \sqsubseteq v$ may be defined as follows in terms of the intersection, the union and the complements of the relations :

$$\begin{array}{llllllll} \text{inclusion :} & u \sqsubseteq v & :\Leftrightarrow & u \sqcap v = u & \Leftrightarrow & u \sqcup v = v \\ & & \Leftrightarrow & u \sqcap \overline{v} = \emptyset & \Leftrightarrow & \overline{u} \sqcup v = e \end{array}$$

The inclusion $u \sqsubseteq v$ is reflexive, antisymmetric and transitive. Hence it is a partial order relation. In the power set P(M), the null relation $\emptyset$ is the least relation, since it is contained in all relations, and the all relation e is the greatest relation, since it contains all relations.

Property	Inclusion
reflexive	$u \sqsubseteq u$
antisymmetric	$u \sqsubseteq v \quad \wedge \quad v \sqsubseteq u \quad \Rightarrow \quad u = v$
transitive	$u \sqsubseteq v \quad \wedge \quad v \sqsubseteq w \quad \Rightarrow \quad u \sqsubseteq w$
extreme	$\emptyset \sqsubseteq u \qquad u \sqsubseteq e$

8.2.3 HOMOGENEOUS BINARY RELATIONS

Introduction : A binary relation is a subset of the cartesian product of two sets. The relation is said to be homogeneous if the two factors of the product coincide. Thus a homogeneous binary relation contains ordered pairs $(x, y) \in M \times M$. Since every relation is a set, the rules of the algebra of sets also hold for homogeneous binary relations. Additional properties and rules result from the duality and composition of relations. Homogeneous binary relations are specified by quadratic boolean matrices. This leads to a boolean matrix algebra for binary relations.

Definition : Let a non-empty set M of elements and a binary operation for a relation R on M be given. The value of the binary operation xRy on the elements $x \in M$ and $y \in M$ is true or false. The corresponding homogeneous relation is the set of all ordered pairs (x, y) for which the binary operation xRy is true. It is a subset of the homogeneous cartesian product $M \times M$. The number of pairs in the binary relation R is designated by $|R|$.

$$R := \{(x, y) \in M \times M \mid xRy\} \subseteq M \times M$$

Matrix representation : Let M be a set with n elements. The elements of M are indexed by a mapping $f : N \rightarrow M$ with $f(i) = x_i$ and $1 \le i \le n$, so that $M = \{x_1, ..., x_n\}$. A homogeneous binary relation $R \subseteq M \times M$ is a subset of $M \times M$. The elements of $M \times M$ which belong to the relation are specified by a boolean matrix **R** of dimension $n \times n$. Every element $(x_i, x_j) \in M \times M$ is bijectively associated with an element $r_{ij} \in \mathbf{R}$. If the relation R contains the element (x_i, x_j), then r_{ij} has the value true (1); otherwise r_{ij} has the value false (0).

A boolean matrix **R** of a homogeneous relation R is an n^2-tuple of the truth values $W = \{0,1\}$, and hence an element of the n^2-fold cartesian product $W^{n \cdot n}$. The elements of a matrix **R** are usually arranged in a row and column scheme by regarding the indices i, j of the element r_{ij} as row and column indices, respectively. In formulations of general properties and rules, a matrix **R** is represented by a general element r_{ij} in square brackets.

$$\mathbf{R} = [r_{ij}] = \begin{array}{|c|c|c|c|c|} \hline r_{11} & \cdots & r_{1j} & \cdots & r_{1n} \\ \hline \vdots & & \vdots & & \vdots \\ \hline r_{i1} & \cdots & r_{ij} & \cdots & r_{in} \\ \hline \vdots & & \vdots & & \vdots \\ \hline r_{n1} & \cdots & r_{nj} & \cdots & r_{nn} \\ \hline \end{array} \qquad \begin{array}{l} W = \{0,1\} \\ \mathbf{R} \in W^{n \cdot n} \end{array}$$

Graphical representation : A homogeneous binary relation R on a set M is visually represented in a grid diagram, a relational diagram or a graph diagram. In the following, the different representations are described and illustrated by examples.

The grid diagram is a two-dimensional orthogonal grid with horizontal and vertical grid lines for the elements of the set M. Every grid point corresponds to a pair (x, y) $\in$ M $\times$ M. The grid points of the pairs (x, y) $\in$ R contained in the homogeneous relation R are marked. The grid diagram is a graphical image of the boolean matrix **R**.

The relational diagram consists of two point sets, each of which represents the set M of elements with their designations. If an element x is related to an element y, an arrow is drawn from the point x of the first point set to the point y of the second point set. The homogeneous relation R corresponds to the resulting set of arrows.

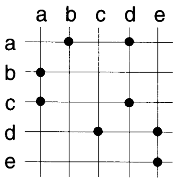
The graph diagram consist of a point set which represents the set M of elements with their designations. If an element x is related to an element y, an arrow is drawn from the point x to the point y. The homogeneous relation R corresponds to the resulting set of arrows. The graph diagram shows the elements of the set M and the relationships in a network-like structure. It is the representation used in graph theory. The points used to represent the elements are called vertices, the arrows are called directed edges.

Example 1 : Representation of a homogeneous relation

Let a set M of elements be given, and let a homogeneous relation R be given as a set of pairs of elements from M $\times$ M. The boolean matrix **R** and the various graphical representations are shown.

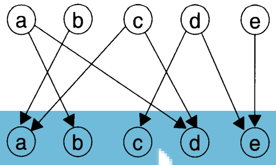
$$\begin{aligned} M &= \{a, b, c, d, e\} \\ R &= \{(a,b), (a,d), (b,a)\ (c,a), (c,d), (d,c), (d,e), (e,e)\} \end{aligned}$$

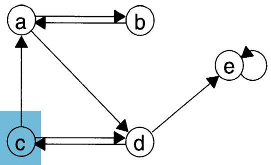
grid diagram and boolean matrix



$$\mathbf{R} = \begin{array}{|c|c|c|c|c|c} \multicolumn{1}{c}{a} & \multicolumn{1}{c}{b} & \multicolumn{1}{c}{c} & \multicolumn{1}{c}{d} & \multicolumn{1}{c}{e} & \\ \hline 0 & 1 & 0 & 1 & 0 & a \\ \hline 1 & 0 & 0 & 0 & 0 & b \\ \hline 1 & 0 & 0 & 1 & 0 & c \\ \hline 0 & 0 & 1 & 0 & 1 & d \\ \hline 0 & 0 & 0 & 0 & 1 & e \\ \hline \end{array}$$

relational diagram and graph diagram





Special relations : The null relation (empty relation) $\emptyset$, the identity relation I and the all relation (universal relation) E are special homogeneous binary relations on a set M. The corresponding boolean matrices are the zero matrix $\mathbf{0}$, the identity matrix $\mathbf{I}$ and the one matrix $\mathbf{E}$, respectively.

null relation $\emptyset = \{\ \}$

identity relation $\mathrm{I} = \{(x,x) \mid x \in M\}$

all relation $E = M \times M$

null relation $\emptyset$ identity relation I all relation E

$$\mathbf{0} = \begin{array}{|c|c|c|c|c|}\hline 0 & 0 & \cdots & & 0 \\ \hline 0 & 0 & & & \\ \hline \vdots & & \ddots & & \vdots \\ \hline & & & 0 & 0 \\ \hline 0 & & \cdots & 0 & 0 \\ \hline \end{array} \qquad \mathbf{I} = \begin{array}{|c|c|c|c|c|}\hline 1 & 0 & \cdots & \cdots & 0 \\ \hline 0 & 1 & \ddots & & \vdots \\ \hline \vdots & \ddots & \ddots & \ddots & \vdots \\ \hline \vdots & & \ddots & 1 & 0 \\ \hline 0 & \cdots & \cdots & 0 & 1 \\ \hline \end{array} \qquad \mathbf{E} = \begin{array}{|c|c|c|c|c|}\hline 1 & 1 & \cdots & & 1 \\ \hline 1 & 1 & & & \\ \hline \vdots & & \ddots & & \vdots \\ \hline & & & 1 & 1 \\ \hline 1 & & \cdots & 1 & 1 \\ \hline \end{array}$$

Equality and inclusion : The operations of equality $R = S$ and inclusion $R \sqsubseteq S$ on homogeneous relations R, S yield the logical constant true or false. If $R = S$ is true, then R and S are equal. If $R \sqsubseteq S$ is true, then R is contained in S.

equality $R = S \;:\Leftrightarrow\; \bigwedge_x \bigwedge_y ((x,y) \in R \;\Leftrightarrow\; (x,y) \in S)$

inclusion $R \sqsubseteq S \;:\Leftrightarrow\; \bigwedge_x \bigwedge_y ((x,y) \in R \;\Rightarrow\; (x,y) \in S)$

equality $\mathbf{R} = \mathbf{S} \;:\Leftrightarrow\; \bigwedge_i \bigwedge_j (r_{ij} \;\Leftrightarrow\; s_{ij})$

inclusion $\mathbf{R} \sqsubseteq \mathbf{S} \;:\Leftrightarrow\; \bigwedge_i \bigwedge_j (r_{ij} \;\Rightarrow\; s_{ij})$

Unary operations : The complement $\overline{R}$ and the transpose R^T are unary operations on a homogeneous binary relation R. The complement $\overline{R}$ contains all pairs of elements (x, y) of the cartesian product $M \times M$ which are not contained in R. The transpose R^T contains the dual pair (y, x) for every pair (x, y) of elements of R. It is therefore also called the dual relation.

complement $\overline{R} := \{(x,y) \in M \times M \mid (x,y) \notin R\}$

transpose $R^T := \{(y,x) \in M \times M \mid (x,y) \in R\}$

complement $\overline{\mathbf{R}} := [\neg r_{ij}]$

transpose $\mathbf{R}^T := [\ r_{ji}]$

A boolean matrix **R** is transposed according to the usual rules of matrix algebra, namely by interchanging the rows and columns of **R**.

Binary operations : The intersection $R \sqcap S$, the union $R \sqcup S$ and the product $R \circ S$ are binary operations on the homogeneous relations R and S. The intersection and the union are defined as in set theory. The product corresponds to the composition of two relations; the operation of forming products is called multiplication. In the algebra of relations it is convenient to define the composition $R \circ S$ of the relations in the order "first R, then S". This definition allows a direct transfer to boolean matrix algebra. However, it differs from the definition of the composition of relations in Section 2.4, since the order of R and S is reversed.

$$\begin{array}{lll} \text{intersection} & R \sqcap S := & \{(x,y) \mid \quad (x,y) \in R \;\wedge\; (x,y) \in S\ \} \\ \text{union} & R \sqcup S := & \{(x,y) \mid \quad (x,y) \in R \;\vee\; (x,y) \in S\ \} \\ \text{product} & R \circ S := & \{(x,y) \mid \bigvee_{z} ((x,z) \in R \;\wedge\; (z,y) \in S)\} \end{array}$$

$$\begin{array}{lll} \text{intersection} & \mathbf{R} \sqcap \mathbf{S} := & [r_{ij} \wedge s_{ij}] \\ \text{union} & \mathbf{R} \sqcup \mathbf{S} := & [r_{ij} \vee s_{ij}] \\ \text{product} & \mathbf{R} \circ \mathbf{S} := & [\bigvee_{k} r_{ik} \wedge s_{kj}] \end{array}$$

The operations on boolean matrices are similar to the operations on real matrices. The addition and the product of real matrices correspond to the union and the product of boolean matrices. The arithmetic operators + and * correspond to the logical operators $\vee$ and $\wedge$. Matrix multiplication is conveniently represented in a graphical scheme.

Example 2 : Operations on boolean matrices

Let boolean matrices **R, S** of the cartesian product $W^{4 \cdot 4}$ of the truth values W be given. The complement $\bar{\mathbf{R}}$, the transpose $\mathbf{S}^T$, the intersection $\mathbf{R} \sqcap \mathbf{S}$, the union $\mathbf{R} \sqcup \mathbf{S}$ and the product $\mathbf{R} \circ \mathbf{S}$ are determined.

relations **R** and **S**

$$\mathbf{R} = \begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 0 \\ \hline 0 & 1 & 1 & 0 \\ \hline 1 & 1 & 0 & 1 \\ \hline 0 & 0 & 1 & 1 \\ \hline \end{array} \qquad \mathbf{S} = \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline \end{array}$$

complement $\bar{\mathbf{R}}$ and transpose $\mathbf{S}^T$

$$\bar{\mathbf{R}} = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 1 \\ \hline 0 & 0 & 1 & 0 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array} \qquad \mathbf{S}^T = \begin{array}{|c|c|c|c|}\hline 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 \\ \hline 1 & 1 & 0 & 0 \\ \hline \end{array}$$

intersection $\mathbf{R} \sqcap \mathbf{S}$ and union $\mathbf{R} \sqcup \mathbf{S}$

$$\mathbf{R} \sqcap \mathbf{S} = \begin{array}{|c|c|c|c|}\hline 1&0&1&0\\\hline 0&1&1&0\\\hline 1&1&0&1\\\hline 0&0&1&1\\\hline\end{array} \sqcap \begin{array}{|c|c|c|c|}\hline 0&0&0&1\\\hline 1&0&0&1\\\hline 1&1&0&0\\\hline 0&0&1&0\\\hline\end{array} = \begin{array}{|c|c|c|c|}\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 1&1&0&0\\\hline 0&0&1&0\\\hline\end{array}$$

$$\mathbf{R} \sqcup \mathbf{S} = \begin{array}{|c|c|c|c|}\hline 1&0&1&0\\\hline 0&1&1&0\\\hline 1&1&0&1\\\hline 0&0&1&1\\\hline\end{array} \sqcup \begin{array}{|c|c|c|c|}\hline 0&0&0&1\\\hline 1&0&0&1\\\hline 1&1&0&0\\\hline 0&0&1&0\\\hline\end{array} = \begin{array}{|c|c|c|c|}\hline 1&0&1&1\\\hline 1&1&1&1\\\hline 1&1&0&1\\\hline 0&0&1&1\\\hline\end{array}$$

product $\mathbf{R} \circ \mathbf{S}$

$$\mathbf{R} \circ \mathbf{S} = \mathbf{T} \qquad \begin{array}{c|cccc|} & 1&2&3&4\\\hline 1&0&0&0&1\\ 2&1&0&0&1\\ 3&1&1&0&0\\ 4&0&0&1&0\\\hline\end{array}\ \mathbf{S}$$

$$\mathbf{R}\ \begin{array}{c|cccc|} & 1&2&3&4\\\hline 1&1&0&1&0\\ 2&0&1&1&0\\ 3&1&1&0&1\\ 4&0&0&1&1\\\hline\end{array} \quad \begin{array}{|cccc|}\hline 1&1&0&1\\ 1&1&0&1\\ 1&0&1&1\\ 1&1&1&0\\\hline\end{array}\ \mathbf{T}$$

calculation for t_{32} :

$$t_{32} = \bigvee_{k=1}^{4} (r_{3k} \wedge s_{k2})$$

$$t_{32} = (1 \wedge 0) \vee (1 \wedge 0) \vee (0 \wedge 1) \vee (1 \wedge 0)$$

$$t_{32} = 0 \vee 0 \vee 0 \vee 0 = 0$$

Algebraic structure : Homogeneous binary relations on a set M are subsets of the cartesian product $M \times M$. All of the different subsets of $M \times M$ are collected in the power set $P(M \times M)$. The power set $P(M \times M)$ is equipped with an algebraic structure by the operations $\sqcap$, $\sqcup$, $^{-}$, $\circ$, ^{T} defined above.

The domain $(P(M \times M) ; \sqcap, \sqcup, ^{-}, \circ, ^{T})$ is called an algebra of homogeneous relations; it has the following properties :

(1) The domain $(P(M \times M) ; \sqcap, \sqcup, ^{-})$ is a boolean lattice with the inclusion $\sqsubseteq$ as a partial order relation.

(2) The domain $(P(M \times M) ; \circ)$ is a semigroup with the identity relation acting as the identity element.

(3) The compatibility of the operations $\sqsubseteq$, $^{-}$, $\circ$, ^{T} is guaranteed by the following equivalence for $R, S, T \in P(M \times M)$.

$$R \circ S \sqsubseteq T \quad \Leftrightarrow \quad \overline{T} \circ S^{T} \sqsubseteq \overline{R} \quad \Leftrightarrow \quad R^{T} \circ \overline{T} \sqsubseteq \overline{S}$$

The properties of the algebra of homogeneous relations and the resulting rules of calculation are treated in the following.

Boolean lattice : The domain $(P(M \times M)\ ;\ \sqcap, \sqcup, ^{-})$ is a boolean lattice with the following properties for the relations $R, S, T \in P(M \times M)$:

Property	Intersection $\sqcap$	Union $\sqcup$
associative	$(R \sqcap S) \sqcap T = R \sqcap (S \sqcap T)$	$(R \sqcup S) \sqcup T = R \sqcup (S \sqcup T)$
commutative	$R \sqcap S = S \sqcap R$	$R \sqcup S = S \sqcup R$
adjunctive	$R \sqcap (R \sqcup S) = R$	$R \sqcup (R \sqcap S) = R$
distributive	$R \sqcap (S \sqcup T) = (R \sqcap S) \sqcup (R \sqcap T)$	$R \sqcup (S \sqcap T) = (R \sqcup S) \sqcap (R \sqcup T)$
zero element	$R \sqcap \emptyset = \emptyset$	$R \sqcup \emptyset = R$
unit element	$R \sqcap E = R$	$R \sqcup E = E$
complement	$R \sqcap \overline{R} = \emptyset$	$R \sqcup \overline{R} = E$

In a boolean lattice, the inclusion $\sqsubseteq$ can be defined as follows in terms of the intersection, the union and the complements of relations :

$$\text{inclusion :} \quad R \sqsubseteq S \ :\Leftrightarrow\ R \sqcap S = R \ \Leftrightarrow\ R \sqcup S = S$$
$$\Leftrightarrow\ R \sqcap \overline{S} = \emptyset \ \Leftrightarrow\ \overline{R} \sqcup S = E$$

As in the case of unary relations, the inclusion $R \sqsubseteq S$ of homogeneous binary relations is a partial order relation with the least relation $\emptyset$ and the greatest relation E.

Semigroup : The domain $(P(M \times M)\ ;\ \circ)$ is a semigroup with the identity relation acting as the identity element. The semigroup is not commutative, so that generally $R \circ S \neq S \circ R$. The following properties hold for $R, S, T \in P(M \times M)$:

Property	Multiplication $\circ$
associative	$R \circ (S \circ T) = (R \circ S) \circ T$
identity	$R \circ I = I \circ R = R$

Compatibility : The compatibility of the operations $\sqsubseteq, ^{-}, \circ, ^{T}$ requires that the following equivalence holds for $R, S, T \in P(M \times M)$:

$$R \circ S \sqsubseteq T \ \Leftrightarrow\ \overline{T} \circ S^T \sqsubseteq \overline{R} \ \Leftrightarrow\ R^T \circ \overline{T} \sqsubseteq \overline{S}$$

This equivalence is proved by transforming the following logical expression :

$$\bigwedge_x \bigwedge_y \bigwedge_z [(x,y) \in R \quad \wedge \quad (y,z) \in S \quad \Rightarrow \quad (x,z) \in T] \tag{1}$$

The logical expression is transformed according to the rules of formal logic and the definitions of the operations on relations. The rule $(a \Rightarrow b) \Leftrightarrow ((\neg a) \vee b)$ for implication as well as De Morgan's rules $\neg (a \wedge b) \Leftrightarrow ((\neg a) \vee (\neg b))$ and $\neg (a \vee b) \Leftrightarrow ((\neg a) \wedge (\neg b))$ are used.

$$\bigwedge_x \bigwedge_y \bigwedge_z [(x,y) \notin R \quad \vee \quad (y,z) \notin S \quad \vee \quad (x,z) \in T] \tag{2}$$

The logical expression (2) yields three equivalent expressions :

$$\begin{aligned}
&\bigwedge_z \bigwedge_x [\bigwedge_y ((x,y) \notin R \quad \vee \quad (y,z) \notin S) \quad \vee \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_x \bigwedge_y [\bigwedge_z ((y,z) \notin S \quad \vee \quad (x,z) \in T) \quad \vee \quad (x,y) \notin R] \quad \Leftrightarrow \\
&\bigwedge_y \bigwedge_z [\bigwedge_x ((x,y) \notin R \quad \vee \quad (x,z) \in T) \quad \vee \quad (y,z) \notin S]
\end{aligned} \tag{3}$$

Each of the three expressions in (3) is again transformed. The transformation for the first expression is shown.

$$\begin{aligned}
&\bigwedge_z \bigwedge_x [\bigwedge_y ((x,y) \notin R \quad \vee \quad (y,z) \notin S) \quad \vee \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_z \bigwedge_x [\bigwedge_y \neg ((x,y) \in R \quad \wedge \quad (y,z) \in S) \quad \vee \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_z \bigwedge_x [\neg \bigvee_y ((x,y) \in R \quad \wedge \quad (y,z) \in S) \quad \vee \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_z \bigwedge_x [\bigvee_y ((x,y) \in R \quad \wedge \quad (y,z) \in S) \quad \Rightarrow \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_z \bigwedge_x [(x,z) \in R \circ S \quad \Rightarrow \quad (x,z) \in T]
\end{aligned}$$

The transformation of all three expressions in (3) leads to the following result :

$$\begin{aligned}
&\bigwedge_z \bigwedge_x [(x,z) \in R \circ S \quad \Rightarrow \quad (x,z) \in T] \quad \Leftrightarrow \\
&\bigwedge_x \bigwedge_y [(x,y) \in \overline{T} \circ S^T \quad \Rightarrow \quad (x,y) \in \overline{R}] \quad \Leftrightarrow \\
&\bigwedge_y \bigwedge_z [(y,z) \in R^T \circ \overline{T} \quad \Rightarrow \quad (y,z) \in \overline{S}]
\end{aligned} \tag{4}$$

With the definition of inclusion, (4) implies the above equivalence for compatibility of the operations $\sqsubseteq, {}^{-}, \circ, {}^{T}$:

$$R \circ S \sqsubseteq T \quad \Leftrightarrow \quad \overline{T} \circ S^T \sqsubseteq \overline{R} \quad \Leftrightarrow \quad R^T \circ \overline{T} \sqsubseteq \overline{S} \tag{5}$$

The inclusion $\sqsubseteq$ is equivalent to expressions which contain the intersection and the union of the sets. Hence the compatibility condition (5) contains all of the operations $\sqcap$, $\sqcup$, ${}^{-}$, $\circ$, ^{T} on the homogeneous relations in the power set $P(M \times M)$.

Rules of calculation : The rules of calculation for homogeneous binary relations are derived from the properties of the algebraic structure of these relations. All rules of calculation for sets also hold for relations. Additional rules of calculation for the transposition and the multiplication of homogeneous binary relations are compiled in the following without proofs.

Transposition	Multiplication
$\emptyset^T = \emptyset \qquad I^T = I \qquad E^T = E$	$R \circ \emptyset = \emptyset \circ R = \emptyset$
$(R^T)^T = R$	$R \sqsubseteq S \Rightarrow Q \circ R \sqsubseteq Q \circ S$
$\overline{R^T} = \overline{R}^T$	$R \sqsubseteq S \Rightarrow R \circ Q \sqsubseteq S \circ Q$
$R \sqsubseteq S \Leftrightarrow R^T \sqsubseteq S^T$	$R \circ (S \sqcap T) \sqsubseteq R \circ S \sqcap R \circ T$
$(R \sqcap S)^T = R^T \sqcap S^T$	$(S \sqcap T) \circ R \sqsubseteq S \circ R \sqcap T \circ R$
$(R \sqcup S)^T = R^T \sqcup S^T$	$R \circ (S \sqcup T) = R \circ S \sqcup R \circ T$
$(R \circ S)^T = S^T \circ R^T$	$(S \sqcup T) \circ R = S \circ R \sqcup T \circ R$

Properties of relations : The properties of homogeneous binary relations are formulated with quantifiers in Section 2.3. With the notation and the rules of the algebra of relations, they can be represented in a compact form. The formulation of the transitivity property in the algebra of relations becomes a special case of the compatibility condition with R = S = T.

Property	Equivalent conditions
reflexive	$\bigwedge_x ((x,x) \in R) \Leftrightarrow I \sqsubseteq R$
antireflexive	$\bigwedge_x ((x,x) \notin R) \Leftrightarrow I \sqsubseteq \overline{R}$
symmetric	$\bigwedge_x \bigwedge_y ((x,y) \in R \Rightarrow (y,x) \in R) \Leftrightarrow R = R^T$
asymmetric	$\bigwedge_x \bigwedge_y ((x,y) \in R \Rightarrow (y,x) \notin R) \Leftrightarrow R \sqcap R^T = \emptyset$
antisymmetric	$\bigwedge_x \bigwedge_y ((x,y) \in R \wedge (y,x) \in R \Rightarrow x = y) \Leftrightarrow R \sqcap R^T \sqsubseteq I$
linear	$\bigwedge_x \bigwedge_y ((x,y) \in R \vee (y,x) \in R) \Leftrightarrow R \sqcup R^T = E$
connex	$\bigwedge_x \bigwedge_y (x \neq y \Rightarrow (x,y) \in R \vee (y,x) \in R) \Leftrightarrow \overline{I} \sqsubseteq R \sqcup R^T$
transitive	$\bigwedge_x \bigwedge_y \bigwedge_z ((x,y) \in R \wedge (y,z) \in R \Rightarrow (x,z) \in R)$ $\Leftrightarrow R \circ R \sqsubseteq R$

8.2.4 HETEROGENEOUS BINARY RELATIONS

Introduction : A heterogeneous binary relation is a subset of the cartesian product of two different sets. Like a homogeneous binary relation, it is a set of ordered pairs of elements. Most of the rules of calculation for homogeneous binary relations may therefore be transferred. While homogeneous relations are represented by quadratic boolean matrices, heterogeneous relations are generally represented by rectangular boolean matrices.

Definition : Let two non-empty sets A and B of elements and a binary operation R for the elements of these sets be given. The value of the binary operation xRy on the elements $x \in A$ and $y \in B$ is true or false. The corresponding heterogeneous relation R is the set of all ordered pairs (x, y) for which the binary operation is true. The relation R is a subset of the cartesian product $A \times B$.

$$R := \{(x,y) \in A \times B \mid xRy\} \subseteq A \times B$$

Matrix representation : A heterogeneous binary relation R on a set A with m elements and a set B with n elements is specified by a boolean matrix **R** with m rows and n columns. The matrix notation for homogeneous relations is transferred to heterogeneous relations. The boolean matrix **R** of a heterogeneous relation R is an $m \cdot n$-tuple of the truth values $W = \{0,1\}$, and hence an element of the $m \cdot n$-fold cartesian product $W^{m \cdot n}$.

Graphical representation : In analogy with homogeneous relations, heterogeneous relations are represented in grid diagrams or relational diagrams. The grid diagram is the graphical image of the boolean matrix.

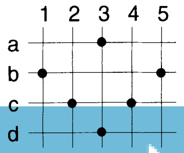
Example 1 : Representation of heterogeneous relations

Let two sets A and B be given, and let a heterogeneous relation R be given as a set of pairs of elements. The boolean matrix **R** and the two graphical representations are shown.

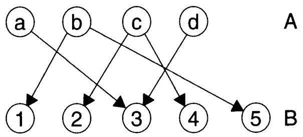
$$A = \{a, b, c, d\} \qquad B = \{1, 2, 3, 4, 5\}$$

$$R = \{(a,3), (b,1), (b,5), (c,2), (c,4), (d,3)\} \subseteq A \times B$$

grid diagram, boolean matrix and relational diagram



$$\mathbf{R} = \begin{array}{|c|c|c|c|c|c} 1 & 2 & 3 & 4 & 5 & \\ \hline 0 & 0 & 1 & 0 & 0 & a \\ \hline 1 & 0 & 0 & 0 & 1 & b \\ \hline 0 & 1 & 0 & 1 & 0 & c \\ \hline 0 & 0 & 1 & 0 & 0 & d \\ \hline \end{array}$$



Special relations : The null relation $\emptyset$ and the all relation E in the cartesian products $A \times A$, $A \times B$, $B \times A$ and $B \times B$ and the identity relation I in the cartesian products $A \times A$ and $B \times B$ are required for the theory of heterogeneous binary relations on the sets A and B. They may be distinguished by pairs of subscripts, as in $\emptyset_{AA}$, $\emptyset_{AB}$, $\emptyset_{BA}$, $\emptyset_{BB}$ or I_{AA}, I_{BB}. The explicit labeling of these special relations may be omitted, since they are implicitly determined by the rules of calculation for heterogeneous relations.

Equality and inclusion : Two heterogeneous binary relations R and S can only be checked for equality and inclusion if they are subsets of the same cartesian product $A \times B$. The rules of calculation for homogeneous relations also hold for heterogeneous relations.

Unary and binary operations : The unary and binary operations on homogeneous relations may be transferred to heterogeneous relations with the following restrictions :

(1) The operations for the complement, the intersection and the union of homogeneous relations may be transferred to heterogeneous relations of the cartesian product $A \times B$ without restrictions.

(2) The transposition of a heterogeneous relation R of the cartesian product $A \times B$ yields a heterogeneous relation R^T of the cartesian product $B \times A$. The rules for the transposition of homogeneous relations hold analogously.

(3) The multiplication $R \circ S$ of two heterogeneous relations R and S is only defined if R is a subset of $A \times B$ and S is a subset of $B \times C$. It yields a heterogeneous relation T as a subset of $A \times C$. The rules for the multiplication of homogeneous relations hold analogously.

(4) The compatibility condition (5) for homogeneous relations in Section 8.2.3 holds analogously for heterogeneous relations.

Properties of relations : The completeness and the uniqueness of a heterogeneous relation $R \subseteq A \times B$ on the sets A and B are treated in Section 2.3. A relation R may be left- or right-total as well as left- or right-unique. The corresponding conditions are formulated below, both in the notation of set theory using quantifiers and in the operational notation of the algebra of relations :

Property	Equivalent conditions
left-total	$\bigwedge_x \bigvee_y ((x,y) \in R) \Leftrightarrow R \circ E = E \Leftrightarrow I \sqsubseteq R \circ R^T \Leftrightarrow \overline{R} \sqsubseteq R \circ \overline{I}$
right-total	$\bigwedge_y \bigvee_x ((x,y) \in R) \Leftrightarrow E \circ R = E \Leftrightarrow I \sqsubseteq R^T \circ R \Leftrightarrow \overline{R} \sqsubseteq \overline{I} \circ R$
left-unique	$\bigwedge_x \bigwedge_y \bigwedge_z ((x,y) \in R \wedge (z,y) \in R \Rightarrow x = z)$ $\Leftrightarrow R \circ R^T \sqsubseteq I \Leftrightarrow \overline{I} \circ R \sqsubseteq \overline{R}$
right-unique	$\bigwedge_x \bigwedge_y \bigwedge_z ((x,y) \in R \wedge (x,z) \in R \Rightarrow y = z)$ $\Leftrightarrow R^T \circ R \sqsubseteq I \Leftrightarrow R \circ \overline{I} \sqsubseteq \overline{R}$

The equivalence of the three conditions for a left-total relation R is demonstrated in the following. The first condition $R \circ E = E$ may be read off directly from the set-theoretic expression. The second condition $I \sqsubseteq R \circ R^T$ is obtained by extending and transforming the set-theoretic expression. The third condition $\overline{R} \sqsubseteq R \circ \overline{I}$ is derived directly from the first condition using the rule $R \sqsubseteq S \Leftrightarrow \overline{R} \sqcup S = E$ for inclusion.

$$\bigwedge_x \bigvee_y ((x,y) \in R) \Leftrightarrow \bigwedge_x \bigvee_y ((x,y) \in R \wedge (y,x) \in R^T) \Leftrightarrow \bigwedge_x ((x,x) \in R \circ R^T)$$

$$\Leftrightarrow I \sqsubseteq R \circ R^T$$

$$R \circ E = E \Leftrightarrow R \circ (I \sqcup \overline{I}) = E \Leftrightarrow R \sqcup R \circ \overline{I} = E \Leftrightarrow \overline{R} \sqsubseteq R \circ \overline{I}$$

The conditions for a left-unique relation R are obtained by transforming the set-theoretic expression and the compatibility condition :

$$\bigwedge_x \bigwedge_y \bigwedge_z ((x,y) \in R \wedge (z,y) \in R \Rightarrow x = z) \Leftrightarrow$$

$$\bigwedge_x \bigwedge_y \bigwedge_z ((x,y) \in R \wedge (y,z) \in R^T \Rightarrow (x,z) \in I) \Leftrightarrow$$

$$R \circ R^T \sqsubseteq I \Leftrightarrow \overline{I} \circ R \sqsubseteq \overline{R}$$

The conditions for a right-total or a right-unique relation may be derived analogously. A left- and right-total relation is bitotal. A left- and right-unique relation is bi-unique.

Mapping : A mapping $\Phi : A \rightarrow B$ from a domain A to a target B is by definition a left-total and right-unique relation $\Phi \subseteq A \times B$. The relation Φ associates every element $x \in A$ of the domain A with exactly one element $y = \Phi(x) \in B$ of the target B. The conditions $\overline{\Phi} \sqsubseteq \Phi \circ \overline{I}$ for left-totality and $\Phi \circ \overline{I} \sqsubseteq \overline{\Phi}$ for right-uniqueness can be combined into the condition $\overline{\Phi} = \Phi \circ \overline{I}$ using the rules of the algebra of relations :

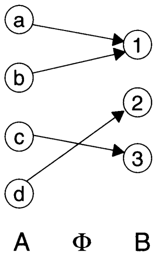
$$\Phi \text{ is a mapping} \quad \Leftrightarrow \quad \overline{\Phi} = \Phi \circ \overline{I}$$

A mapping may be injective, surjective or bijective. The properties of mappings are treated in Section 2.5. They are derived from the preceding table using the rules of the algebra of relations :

Property	Condition				
injective	left-total ∧ bi-unique	⇔	$I = \Phi \circ \Phi^T$	∧	$\Phi^T \circ \Phi \sqsubseteq I$
surjective	bitotal ∧ right-unique	⇔	$I \sqsubseteq \Phi \circ \Phi^T$	∧	$\Phi^T \circ \Phi = I$
bijective	bitotal ∧ bi-unique	⇔	$I = \Phi \circ \Phi^T$	∧	$\Phi^T \circ \Phi = I$

Example 2 : Mapping

Let a relation $\Phi \subseteq A \times B$ for a mapping $A \rightarrow B$ be given in the form of a relational diagram. Using the rules of the algebra of relations, the mapping is formally shown to be surjective. The graphical representation of the multiplication of boolean matrices is used.



$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline a & 1 & 0 & 0 \\ b & 1 & 0 & 0 \\ c & 0 & 0 & 1 \\ d & 0 & 1 & 0 \end{array} \circ \begin{array}{|cccc|} 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{array} = \begin{array}{|cccc|} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}$$

product $\Phi \circ \Phi^T \sqsupseteq I$

$$\begin{array}{c|cccc} & a & b & c & d \\ \hline 1 & 1 & 1 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 3 & 0 & 0 & 1 & 0 \end{array} \circ \begin{array}{|ccc|} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{array} = \begin{array}{|ccc|} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{array}$$

product $\Phi^T \circ \Phi = I$

Induced homogeneous binary relations : Let a homogeneous binary relation $R_A \subseteq A \times A$ on a set A and a mapping $\Phi : A \to B$ from the set A to a set B be given. The mapping Φ induces a homogeneous binary relation $R_B \subseteq B \times B$ which contains the image pair $(\Phi(a_1), \Phi(a_2))$ for all pairs (a_1, a_2) of elements from R_A :

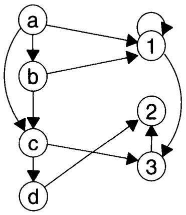
$$R_B := \{(b_1,b_2) \in B \times B \mid \bigvee_{(a_1,a_2) \in R_A} [(b_1,b_2) = (\Phi(a_1),\Phi(a_2))]\}$$

$$R_B = \{(b_1,b_2) \in B \times B \mid \bigvee_{a_1} \bigvee_{a_2} [(b_1,a_1) \in \Phi^T \wedge (a_1,a_2) \in R_A \wedge (a_2,b_2) \in \Phi]\}$$

$$R_B = \Phi^T \circ R_A \circ \Phi$$

Example 3 : Induced binary relation

Let the sets A and B and the surjective mapping $\Phi : A \to B$ from the preceding example be given. Let a binary relation R_A be defined on the set A. The induced relation R_B on the set B is to be determined. The relations R_A and R_B are represented by arrows in the following diagram.



$A, R_A \quad \Phi \quad B, R_B$

$$\begin{array}{c|cccc|cccc|ccc|c}
 & & & & & a & b & c & d & 1 & 2 & 3 & \\
 & & & & a & 0 & 1 & 1 & 0 & 1 & 0 & 0 & a \\
 & & & & b & 0 & 0 & 1 & 0 & 1 & 0 & 0 & b \\
 & & & & c & 0 & 0 & 0 & 1 & 0 & 0 & 1 & c \\
 & & & & d & 0 & 0 & 0 & 0 & 0 & 1 & 0 & d \\
\hline
 & a & b & c & d & & & & & & & & \\
1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 \\
2 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \\
3 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 3
\end{array}$$

product $\mathbf{\Phi}^T \circ \mathbf{R}_A \circ \mathbf{\Phi} = \mathbf{R}_B$

8.2.5 UNARY AND BINARY RELATIONS

Introduction : By the definition in Section 8.2.2, a unary relation is a subset of a given set M. It may also be regarded as a special case of a heterogeneous binary relation. Thus transposition and multiplication may also be applied to unary relations. This leads to a boolean vector and matrix algebra.

Unary relation : Let a non-empty set A and a one-element set $B = \{b\}$ as well as a binary operation U for the elements of these sets be given. Then the heterogeneous binary relation U contains all ordered pairs of elements (x, b) for which the binary operation $x U b$ is true.

$$U = \{(x, b) \in A \times \{b\} \mid x U b\} \subseteq A \times \{b\}$$

A unary relation u is derived from the heterogeneous binary relation U according to the following rule :

$$u = \{x \in A \mid (x, b) \in U\} \subseteq A$$

The boolean matrix **U** of the heterogeneous binary relation U contains exactly one column, since the set $B = \{b\}$ contains exactly one element. It is formally identical with the boolean vector **u** of the unary relation u if the elements $x \in A$ are arranged in the same order in **U** and **u**.

Operations on unary and binary relations : If a unary relation is regarded as a special case of a heterogeneous binary relation, unary and binary relations may occur together as operands of operations. The rules for operations on heterogeneous binary relations must be observed. The following products are important for calculations with unary and binary relations :

(1) Let the unary relations $u \subseteq A$ and $v \subseteq B$ be given. Then the product $u \circ v^T$ is a binary relation $R \subseteq A \times B$, and the product $v \circ u^T$ is the binary relation $R^T \subseteq B \times A$.

$$R = u \circ v^T = \{(x, y) \in A \times B \mid x \in u \wedge y \in v\}$$

$$R^T = v \circ u^T = \{(y, x) \in B \times A \mid x \in u \wedge y \in v\}$$

(2) Let a binary relation $R \subseteq A \times B$ and a unary relation $v \subseteq B$ be given. Then the product $R \circ v$ is a unary relation $u \subseteq A$.

$$u = R \circ v = \{x \in A \mid \bigvee_y ((x, y) \in R \wedge y \in u)\}$$

In relational expressions for unary and binary relations, the products in (1) and (2) often occur together with unions and intersections. The relevant rules of calculation are compiled in the following table without proofs.

Union $\sqcup$	Intersection $\sqcap$
$u \circ (v \sqcup w)^T = u \circ v^T \sqcup u \circ w^T$ $(u \sqcup v) \circ w^T = u \circ w^T \sqcup v \circ w^T$	$u \circ (v \sqcap w)^T \sqsubseteq u \circ v^T \sqcap u \circ w^T$ $(u \sqcap v) \circ w^T \sqsubseteq u \circ w^T \sqcap v \circ w^T$
$R \circ (u \sqcup v) = R \circ u \sqcup R \circ v$ $(R \sqcup S) \circ u = R \circ u \sqcup S \circ u$	$R \circ (u \sqcap v) \sqsubseteq R \circ u \sqcap R \circ v$ $(R \sqcap S) \circ u \sqsubseteq R \circ u \sqcap S \circ u$

The rules for inclusion and the compatibility properties from Section 8.2.3 also hold for heterogeneous binary relations. Hence they hold for unary and binary relations. The following implications and equivalences therefore hold for the products in (1) and (2) :

$$
\begin{array}{lllll}
u \sqsubseteq v & \Rightarrow & R \circ u \sqsubseteq R \circ v & & \\
R \sqsubseteq S & \Rightarrow & R \circ u \sqsubseteq S \circ u & & \\
u \circ v^T \sqsubseteq R & \Leftrightarrow & \overline{R} \circ v \sqsubseteq \overline{u} & \Leftrightarrow & \overline{R}^T \circ u \sqsubseteq \overline{v} \\
R \circ u \sqsubseteq v & \Leftrightarrow & \overline{v} \circ u^T \sqsubseteq \overline{R} & \Leftrightarrow & R^T \circ \overline{v} \sqsubseteq \overline{u}
\end{array}
$$

Unary point relations : A unary point relation x is a subset which contains exactly one element x of a given set A. It is specified by a boolean unit vector. The essential properties of unary point relations are compiled in the following :

(1) Let the unary point relations $x \subseteq A$ and $y \subseteq B$ be given. Then $x \circ x^T \sqsubseteq \mathrm{I}$ and $x \circ y^T \circ y = y \circ y^T \circ x = x$.

(2) For a non-empty binary relation $R \subseteq A \times B$, there are two unary point relations $x \subseteq A$ and $y \subseteq B$ such that $x \circ y^T \sqsubseteq R$.

(3) For two unary point relations $x \subseteq A$, $y \subseteq B$ and a non-empty relation $R \subseteq A \times B$, the following expressions are logically equivalent :

$$x \circ y^T \sqsubseteq R \quad \Leftrightarrow \quad x \sqsubseteq R \circ y \quad \Leftrightarrow \quad y \sqsubseteq R^T \circ x \quad \Leftrightarrow \quad y \circ x^T \sqsubseteq R^T$$

Properties (1) to (3) follow from the definition of unary point relations and the rules for multiplication and inclusion of unary relations. Property (3) is proved as follows. Multiplying $x \circ y^T \sqsubseteq R$ by y from the right yields :

$$x \circ y^T \sqsubseteq R \quad \Rightarrow \quad x \circ y^T \circ y \sqsubseteq R \circ y \quad \Leftrightarrow \quad x \sqsubseteq R \circ y$$

Multiplying $x \sqsubseteq R \circ y$ by y^T from the right and using $y \circ y^T \sqsubseteq \mathrm{I}$ yields :

$$x \sqsubseteq R \circ y \quad \Rightarrow \quad x \circ y^T \sqsubseteq R \circ y \circ y^T \sqsubseteq R \circ \mathrm{I} \quad \Leftrightarrow \quad x \circ y^T \sqsubseteq R$$

This proves the equivalence $x \circ y^T \sqsubseteq R \Leftrightarrow x \sqsubseteq R \circ y$. The equivalence $y \circ x^T \sqsubseteq R^T \Leftrightarrow y \sqsubseteq R^T \circ x$ is proved analogously. Transposition of $x \circ y^T \sqsubseteq R$ yields $y \circ x^T \sqsubseteq R^T$, and transposing this again yields $x \circ y^T \sqsubseteq R$. This proves the equivalence $x \circ y^T \sqsubseteq R \Leftrightarrow y \circ x^T \sqsubseteq R^T$.

Example 1 : Unary point relations

Let the unary point relations x and y be given as boolean vectors **x** and **y** and the homogeneous binary relation R as a boolean matrix **R**. Some operations are performed on these relations :

$$\mathbf{x} = \begin{array}{|c|}\hline 1\\\hline 0\\\hline 0\\\hline 0\\\hline\end{array} \qquad \mathbf{y} = \begin{array}{|c|}\hline 0\\\hline 0\\\hline 1\\\hline 0\\\hline\end{array} \qquad \mathbf{R} = \begin{array}{|c|c|c|c|}\hline 0&0&1&1\\\hline 0&1&1&0\\\hline 1&0&0&0\\\hline 1&1&1&0\\\hline\end{array} \qquad \mathbf{R}^T = \begin{array}{|c|c|c|c|}\hline 0&0&1&1\\\hline 0&1&0&1\\\hline 1&1&0&1\\\hline 1&0&0&0\\\hline\end{array}$$

$$\mathbf{x}\circ\mathbf{x}^T \sqsubseteq \mathbf{I}: \begin{array}{|c|c|c|c|}\hline 1&0&0&0\\\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 0&0&0&0\\\hline\end{array} \qquad \mathbf{y}\circ\mathbf{y}^T \sqsubseteq \mathbf{I}: \begin{array}{|c|c|c|c|}\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 0&0&1&0\\\hline 0&0&0&0\\\hline\end{array}$$

$$\mathbf{R}\circ\mathbf{y} = \begin{array}{|c|}\hline 1\\\hline 1\\\hline 0\\\hline 1\\\hline\end{array} \qquad \mathbf{R}^T\circ\mathbf{x} = \begin{array}{|c|}\hline 0\\\hline 0\\\hline 1\\\hline 1\\\hline\end{array} \qquad \mathbf{x}\circ\mathbf{y}^T = \begin{array}{|c|c|c|c|}\hline 0&0&1&0\\\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 0&0&0&0\\\hline\end{array} \qquad \mathbf{y}\circ\mathbf{x}^T = \begin{array}{|c|c|c|c|}\hline 0&0&0&0\\\hline 0&0&0&0\\\hline 1&0&0&0\\\hline 0&0&0&0\\\hline\end{array}$$

$$\mathbf{x} \sqsubseteq \mathbf{R}\circ\mathbf{y} \;\Leftrightarrow\; \mathbf{y} \sqsubseteq \mathbf{R}^T\circ\mathbf{x} \;\Leftrightarrow\; \mathbf{x}\circ\mathbf{y}^T \sqsubseteq \mathbf{R} \;\Leftrightarrow\; \mathbf{y}\circ\mathbf{x}^T \sqsubseteq \mathbf{R}^T$$

The results illustrate property (1) for unary point relations : $\mathbf{x}\circ\mathbf{x}^T$ and $\mathbf{y}\circ\mathbf{y}^T$ are contained in the identity relation $\mathbf{I}$. The results also show that the statements $\mathbf{x} \sqsubseteq \mathbf{R}\circ\mathbf{y}$, $\mathbf{y} \sqsubseteq \mathbf{R}^T\circ\mathbf{x}$, $\mathbf{x}\circ\mathbf{y}^T \sqsubseteq \mathbf{R}$ and $\mathbf{y}\circ\mathbf{x}^T \sqsubseteq \mathbf{R}^T$ hold. Thus the four statements are equivalent for the given relations.

Induced unary relations : Let a unary relation $u_A \subseteq A$ in a set A and a mapping $\Phi : A \rightarrow B$ from the set A to a set B be given. The mapping Φ induces a unary relation $u_B \subseteq B$ which contains the images $\Phi(a)$ for all elements a in u_A :

$$u_B := \{b \in B \mid \bigvee_{a \in u_A} (b = \Phi(a))\}$$

$$u_B = \{b \in B \mid \bigvee_{a} [(b,a) \in \Phi^T \;\wedge\; a \in u_A]\}$$

$$u_B = \Phi^T \circ u_A$$

8.2.6 CLOSURES

Introduction : Closures of relations are defined in Section 2.4, where their properties are formulated using quantifiers. In the following, the properties of such closures are described using the notation and rules of the algebra of relations. The closure is regarded as an extension of the relation $R \subseteq M \times M$ by further elements of the cartesian product $M \times M$. Since powers of the relation R are used in forming the transitive closures, the properties of these powers are studied before the treatment of transitive closures.

Closure : An extension of a homogeneous binary relation $R \subseteq M \times M$ is called a closure and is designated by $\langle R \rangle$ if the following conditions are satisfied :

$$
\begin{array}{lll}
\text{inclusion} & : & R \sqsubseteq \langle R \rangle \\
\text{isotonicity} & : & R \sqsubseteq S \quad \Rightarrow \quad \langle R \rangle \sqsubseteq \langle S \rangle \\
\text{idempotency} & : & \langle\langle R \rangle\rangle = \langle R \rangle
\end{array}
$$

The extension is performed such that the closure has special properties which the relation itself does not necessarily possess. Reflexive, symmetric and transitive closures are defined in the following. Closures may also have several of these properties.

Reflexive closure : The reflexive closure $\langle R \rangle_r$ of a relation $R \subseteq M \times M$ is formed by adding the elements $(x, x) \in M \times M$ to R. The closure $\langle R \rangle_r$ satisfies the condition for reflexive relations.

$$
\begin{array}{lll}
\langle R \rangle_r & := & \{(x,y) \mid (x,y) \in R \ \vee \ x = y \in M\} \\
\langle R \rangle_r & = & R \sqcup I \\
I & \sqsubseteq & \langle R \rangle_r \quad \Rightarrow \quad \langle R \rangle_r \text{ is reflexive}
\end{array}
$$

Symmetric closure : The symmetric closure $\langle R \rangle_s$ of a relation $R \subseteq M \times M$ is the union of R with its transpose R^T. If $\langle R \rangle_s$ contains the element (x,y), then (y,x) is also an element of $\langle R \rangle_s$. The closure $\langle R \rangle_s$ satisfies the condition for symmetric relations.

$$
\begin{array}{lll}
\langle R \rangle_s & := & \{(x,y) \mid (x,y) \in R \ \vee \ (y,x) \in R\} \\
\langle R \rangle_s & = & R \sqcup R^T \\
\langle R \rangle_s & = & \langle R \rangle_s{}^T \quad \Rightarrow \quad \langle R \rangle_s \text{ is symmetric}
\end{array}
$$

Powers of a relation : In Section 2.4, a connection V_R of elements $a, b \in M$ by a relation $R \subseteq M \times M$ is defined. In the algebra of relations, connections are represented by products of the relation R with itself. For example, if R contains the elements (a,b) and (b,c), then by definition the product $R \circ R$ contains the element (a,c). The element (a,c) is a connection of length 2 in R. Each of the elements of $R \circ R$ is a connection of length 2 in R. The power $R^m = R \circ \ldots \circ R$ (m-fold) contains all connections of length m between two elements of M. To determine all connections of length $m \le q$ in M by R, the union of the relations $R \sqcup R^2 \sqcup \ldots \sqcup R^q$ is formed.

Stability index : The least exponent s for which the union $R \sqcup R^2 \sqcup \ldots \sqcup R^s$ is not changed by adding terms R^m with $m > s$ is called the stability index of the relation R. The union $R \sqcup R^2 \sqcup \ldots \sqcup R^s$ contains all connections by R in M.

The stability index s of a relation R may be interpreted as follows. If there are several connections between two elements of M, then there is a shortest connection, of length q, which is contained in R^q. Among all the shortest connections between pairs of elements, there is a shortest connection of maximal length s, which is contained in the power R^s. Hence the union $R \sqcup R^2 \sqcup \ldots \sqcup R^s$ contains all connections in M by R. For a set M with n elements, the stability index s of the relation $R \subseteq M \times M$ is less than n, since the maximal length of all shortest connections in M by R cannot be greater than $n - 1$.

Transitive closure : The transitive closure $<R>_t$ of a relation $R \subseteq M \times M$ contains all elements $(x, y) \in M \times M$ which are connected in M by R. The closure $<R>_t$ satisfies the condition for transitive relations.

$$<R>_t \; := \; \{(x,y) \in M \times M \mid x \text{ and } y \text{ are connected in M by R}\}$$

$$<R>_t \; := \; R \sqcup \ldots \sqcup R^s$$

$$<R>_t \circ <R>_t \; \sqsubseteq \; <R>_t \quad \Rightarrow \quad <R>_t \text{ is transitive}$$

$$s \qquad \text{stability index of R with } <R>_t \sqcup R^{s+1} \; = \; <R>_t$$

The transitivity of the closure $<R>_t$ is proved as follows :

$$\begin{aligned} <R>_t \circ <R>_t \; \sqsubseteq \; R \sqcup (<R>_t \circ <R>_t) \; &= \; R \sqcup (R \sqcup \ldots \sqcup R^s) \circ (R \sqcup \ldots \sqcup R^s) \\ = R \sqcup R^2 \sqcup \ldots \sqcup R^{2s} \; &= \; <R>_t \end{aligned}$$

Reflexive transitive closure : The reflexive transitive closure $\langle R \rangle_{rt}$ of a relation $R \subseteq M \times M$ may alternatively be regarded as the transitive closure $\langle\langle R \rangle_r\rangle_t$ of the reflexive closure $\langle R \rangle_r$ or as the reflexive closure $\langle\langle R \rangle_t\rangle_r$ of the transitive closure $\langle R \rangle_t$. The two viewpoints lead to identical relations. The closure $\langle R \rangle_{rt} = \langle R \rangle_{tr}$ satisfies the condition for transitive relations in the special form of an equation.

$$\langle R \rangle_{rt} := \langle\langle R \rangle_r\rangle_t \qquad \langle R \rangle_{tr} := \langle\langle R \rangle_t\rangle_r$$

$$\langle R \rangle_{rt} = \langle R \rangle_{tr}$$

$$\langle R \rangle_{rt} \circ \langle R \rangle_{rt} = \langle R \rangle_{rt} \quad \Rightarrow \quad \langle R \rangle_{rt} \text{ is transitive}$$

The equality of $\langle R \rangle_{rt}$ and $\langle R \rangle_{tr}$ is proved as follows :

$$\begin{aligned}
\langle R \rangle_{tr} &= I \sqcup \langle R_t \rangle &&= I \sqcup R \sqcup \ldots \sqcup R^s \\
\langle R \rangle_{rt} &= \langle R \sqcup I \rangle_t &&= (R \sqcup I) \sqcup \ldots \sqcup (R \sqcup I)^s \\
& &&= I \sqcup R \sqcup \ldots \sqcup R^s
\end{aligned}$$

The transitivity of the closure $\langle R \rangle_{rt}$ is proved as follows :

$$\begin{aligned}
\langle R \rangle_{rt} \circ \langle R \rangle_{rt} &= (I \sqcup R \sqcup \ldots \sqcup R^s) \circ (I \sqcup R \sqcup \ldots \sqcup R^s) \\
&= I \sqcup R \sqcup \ldots \sqcup R^{2s} = \langle R \rangle_{rt}
\end{aligned}$$

Reflexive symmetric transitive closure : The reflexive symmetric transitive closure $\langle R \rangle_{rst}$ of a relation $R \subseteq M \times M$ is the transitive closure of the symmetric closure of the reflexive closure of R. It coincides with the reflexive symmetric transitive closure $\langle R \rangle_{srt}$. The closure $\langle R \rangle_{rst}$ is of special importance, as it is an equivalence relation and therefore yields a classification of the set M.

$$\langle R \rangle_{rst} := \langle\langle\langle R \rangle_r\rangle_s\rangle_t = \langle\langle R \rangle_s\rangle_{rt} = \langle R \sqcup R^T \rangle_{rt}$$

$$\langle R \rangle_{rst} = \langle R \sqcup I \sqcup R^T \rangle_t$$

Closure operations : The transitive closure $\langle R \rangle_t$ and the reflexive transitive closure $\langle R \rangle_{rt}$ are of special importance in the algebra of relations. They are designated by R^+ and R^*, respectively. According to the preceding definitions, these closures are calculated using power expressions which differ only by the 0-th power $R^0 = I$.

$$\begin{aligned}
R^+ &:= \langle R \rangle_t &&= \phantom{I \sqcup {}} R \sqcup \ldots \sqcup R^s \\
R^* &:= \langle R \rangle_{rt} &&= I \sqcup R \sqcup \ldots \sqcup R^s
\end{aligned}$$

The closures R^+ and R^* satisfy the following relationships :

$$R^+ = R \circ R^* = R^* \circ R \qquad R^+ \circ R^+ \sqsubseteq R^+$$

$$R^* = I \sqcup R^+ = (I \sqcup R)^+ \qquad R^* \circ R^* = R^*$$

The following rules of calculation hold for operations with closures :

$$
\begin{array}{llll}
\text{special relations :} & 0^* = \mathrm{I} & & 0^+ = 0 \\
 & \mathrm{I}^* = \mathrm{I}^+ & = & \mathrm{I} \\
 & \mathrm{E}^* = \mathrm{E}^+ & = & \mathrm{E} \\
\text{transposition :} & (\mathrm{R}^\mathrm{T})^* & = & (\mathrm{R}^*)^\mathrm{T} \\
 & (\mathrm{R}^\mathrm{T})^+ & = & (\mathrm{R}^+)^\mathrm{T} \\
\text{multiplication :} & (\mathrm{R} \circ \mathrm{S})^* \circ \mathrm{R} & = & \mathrm{R} \circ (\mathrm{S} \circ \mathrm{R})^* \\
\quad : & (\mathrm{R} \circ \mathrm{S})^* & = & \mathrm{I} \sqcup \mathrm{R} \circ (\mathrm{S} \circ \mathrm{R})^* \circ \mathrm{S} \\
\text{union :} & (\mathrm{R} \sqcup \mathrm{S})^* & = & (\mathrm{R}^* \circ \mathrm{S}^*)^* \\
 & (\mathrm{R} \sqcup \mathrm{S})^* & = & (\mathrm{R}^* \circ \mathrm{S})^* \circ \mathrm{R}^* \\
 & (\mathrm{R} \sqcup \mathrm{S})^* & = & \mathrm{R}^* \circ (\mathrm{S} \circ \mathrm{R}^*)^*
\end{array}
$$

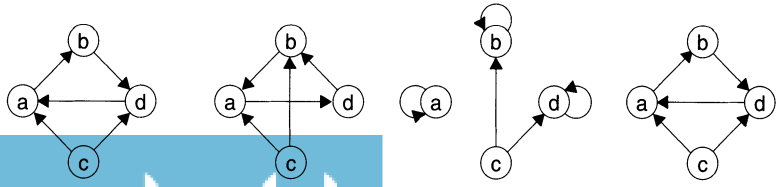
The rules of calculation for the special relations and for transposition follow directly from the definition in terms of power expressions. The rules for multiplication are proved as follows :

$$
\begin{aligned}
(\mathrm{R} \circ \mathrm{S})^* \circ \mathrm{R} &= (\mathrm{I} \sqcup \mathrm{R} \circ \mathrm{S} \sqcup \ldots \sqcup (\mathrm{R} \circ \mathrm{S})^s) \circ \mathrm{R} \\
&= \mathrm{R} \circ (\mathrm{I} \sqcup \mathrm{S} \circ \mathrm{R} \sqcup \ldots \sqcup (\mathrm{S} \circ \mathrm{R})^s) = \mathrm{R} \circ (\mathrm{S} \circ \mathrm{R})^* \\
(\mathrm{R} \circ \mathrm{S})^* &= \mathrm{I} \sqcup \mathrm{R} \circ \mathrm{S} \sqcup \ldots \sqcup (\mathrm{R} \circ \mathrm{S})^{s+1} \\
&= \mathrm{I} \sqcup \mathrm{R} \circ (\mathrm{I} \sqcup \mathrm{S} \circ \mathrm{R} \sqcup \ldots \sqcup (\mathrm{S} \circ \mathrm{R})^s) \circ \mathrm{S} \\
&= \mathrm{I} \sqcup \mathrm{R} \circ (\mathrm{S} \circ \mathrm{R})^* \circ \mathrm{S}
\end{aligned}
$$

Example : Transitive closures

Let a relation R on a set M with $n = 4$ elements be given. The relation R and the calculated powers R^2, R^3 and R^4 are shown as boolean matrices and as graph diagrams.

$$
\mathbf{R}: \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 1 & 0 & 0 \\ b & 0 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 1 & 0 & 0 & 0 \end{array}
\qquad
\mathbf{R}^2: \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 0 & 0 & 1 \\ b & 1 & 0 & 0 & 0 \\ c & 1 & 1 & 0 & 0 \\ d & 0 & 1 & 0 & 0 \end{array}
\qquad
\mathbf{R}^3: \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 1 & 0 & 0 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 0 & 1 & 0 & 1 \\ d & 0 & 0 & 0 & 1 \end{array}
\qquad
\mathbf{R}^4: \begin{array}{c|cccc} & a & b & c & d \\ \hline a & 0 & 1 & 0 & 0 \\ b & 0 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 1 & 0 & 0 & 0 \end{array}
$$

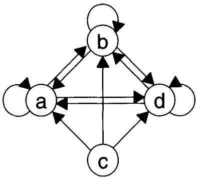


A graph diagram consist of vertices representing the elements of the set and directed edges for the pairs of elements of the relation. In the graph diagram for R, every vertex pair $(x,y) \in R$ is represented by a directed edge. A directed edge sequence is a chain of edges in which the end vertex of an edge coincides with the start vertex of the following edge. The number of edges is called the length of the edge sequence. The relation R^m contains all vertex pairs (x,y) which are connected by a directed edge sequence of length m in the graph diagram for R. For example, the graph diagram for R contains a directed edge sequence from a to d with the two edges (a,b) and (b,d). Hence the vertex pair (a,d) is contained in the relation R^2 and is represented as a directed edge in the graph diagram for R^2.

The closure R^+ contains all vertex pairs (x,y) which are connected by at least one edge sequence from x to y in the graph diagram for R. For every vertex pair (x,y), an edge sequence of minimal length may be determined. The maximum of the minimal lengths for all vertices connected by edge sequences is the stability index s for the closure R^+. Since the length of a minimal edge sequence in a set of n vertices can be at most $n - 1$, it follows that $s < n$. In the present example, the stability index is $s = 3$, since several minimal edge sequences of length $s = 3$ exist, for instance the edge sequence for the vertex pair (a,a) with the edges (a,b), (b,d), (d,a). The closures R^+ and R^* are calculated using the power expressions and are shown as graph diagrams.

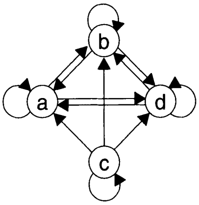
$$\mathbf{R}^+ = \mathbf{R} \sqcup \mathbf{R}^2 \sqcup \mathbf{R}^3$$

	a	b	c	d
a	1	1	0	1
b	1	1	0	1
c	1	1	0	1
d	1	1	0	1



$$\mathbf{R}^* = \mathbf{R}^+ \sqcup \mathbf{I}$$

	a	b	c	d
a	1	1	0	1
b	1	1	0	1
c	1	1	1	1
d	1	1	0	1



8.3 CLASSIFICATION OF GRAPHS

8.3.1 INTRODUCTION

A graph is a domain consisting of vertices and edges. The meaning of the vertices and edges depends on the application. In the graph diagram, the vertices are represented by labeled points. The edges connect vertices between which a relationship subsists. An edge between two vertices may be directed or undirected. In the graph diagram, directed edges are represented as arrows and undirected edges as lines. An edge which connects a vertex with itself is called a loop. The direction of a loop is irrelevant. The geometric arrangement of the vertices and the geometric shape of the edges are arbitrary. Many concepts in graph theory result from this kind of graphical representation.

The algebra of relations forms the mathematical basis of graph theory. The mathematical formulation of graphs with different properties leads to different classes of graphs. An essential aspect of the formulation is whether the edges are introduced as vertex pairs or as independent objects. In the first case, the graph represents a structured vertex set with a vertex relation. In the second case, it represents a structured vertex set and edge set with relations between vertices and edges. There are close relationships between the different types of graphs, which may be exploited in practical applications.

Directed graphs, bipartite graphs, multigraphs and hypergraphs are treated in this chapter. The mathematical formulation, the relationships and exemplary applications of these graphs are shown. The notation for expressions of the algebra of relations is simplified as usual by omitting the operator $\circ$ in products.

8.3.2 DIRECTED GRAPHS

Introduction : A directed graph is suitable for describing relationships between the elements of a set. The elements of the set are called vertices and are identified by their labels. The relationships between the vertices are called edges and are identified by an ordered vertex pair. Thus the edge set is a homogeneous binary relation on the vertex set. The properties of homogeneous binary relations and their rules of calculation may be directly transferred to directed graphs.

Definition : A domain $G = (V; R)$ is called a directed graph (digraph) if V is the vertex set and $R \subseteq V \times V$ is the edge set of the graph. An edge from the vertex $x \in V$ to the vertex $y \in V$ is designated by the ordered pair $(x,y) \in R$. The edge (x, y) is said to be directed from x to y. The vertex x is called the start vertex of the edge. The vertex y is called the end vertex of the edge.

$$G := (V; R) \qquad\qquad R \subseteq V \times V$$

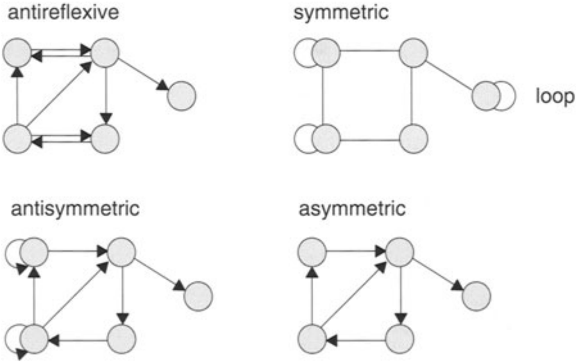
V	set of vertices
R	set of ordered vertex pairs (edge set)

The graph G is called a null graph if the vertex set is empty. It is called an empty graph if the edge set is empty. It is called a complete graph if the edge set R is the all relation $E = V \times V$.

Properties : The properties of a directed graph $(V; R)$ are determined by the properties of the homogeneous relation R. The properties of homogeneous relations described in Section 8.2.3 are therefore transferred to directed graphs. Antireflexive, symmetric, antisymmetric and asymmetric graphs are important in applications :

directed graph		$G = (V; R)$
antireflexive	:⇔	$I \sqsubseteq \overline{R}$
symmetric	:⇔	$R = R^T$
antisymmetric	:⇔	$R \sqcap R^T \sqsubseteq I$
asymmetric	:⇔	$R \sqcap R^T = \emptyset$

For an antireflexive graph, the edge set does not contain vertex pairs of the form (x, x), and the graph diagram is free of loops. Between two different vertices in the graph diagram, a symmetric graph contains either no edge or a pair of edges with opposite directions, which are combined into an undirected edge. An antisymmetric graph contains either no edges or only one directed edge between two vertices in the graph diagram. Symmetric and antisymmetric graphs may contain loops. An asymmetric graph is antisymmetric and antireflexive, and hence free of loops.



Simple graph : A graph $G = (V;\Gamma)$ is said to be simple (undirected) if the relation Γ is antireflexive and symmetric. An antireflexive and symmetric relation is called a neighborhood relation or adjacency relation. The vertices x and y of the pair $(x, y) \in \Gamma$ are neighbors. Every directed graph $G = (V;R)$ is associated with a simple graph $G_S = (V;\Gamma)$.

simple graph $\qquad G_S := (V;\Gamma)$

$$\Gamma \; = \bar{I} \sqcap (R \sqcup R^T)$$

Dual and complementary graph : For every directed graph $G = (V;R)$ there is a dual graph $G^T = (V;\, R^T)$ with the transposed relation R^T and a complementary graph $\bar{G} = (V;\, \bar{R})$ with the complementary relation $\bar{R}$.

dual graph $\qquad G^T := (V;\, R^T)$

complementary graph $\qquad \bar{G} := (V;\, \bar{R})$

Equality and inclusion : Let two directed graphs G_1 and G_2 be given. Using the algebra of relations, equality and inclusion are defined as follows for these graphs :

equality $\qquad G_1 = G_2 \; :\Leftrightarrow \; V_1 = V_2 \;\wedge\; R_1 = R_2$

partial graph $\qquad G_1 \sqsubseteq G_2 \; :\Leftrightarrow \; V_1 = V_2 \;\wedge\; R_1 \sqsubseteq R_2$

subgraph $\qquad G_1 \subseteq G_2 \; :\Leftrightarrow \; V_1 \subseteq V_2 \;\wedge\; R_1 \sqsubseteq R_2 \sqcap (V_1 \times V_1)$

A partial graph (spanning subgraph) G_1 is generated from a graph G_2 by removing edges from G_2. A subgraph G_1 is generated from a graph G_2 by first removing vertices together with the incident edges and then removing further edges from G_2.

Intersection and union : Let two directed graphs G_1 and G_2 be given. They are said to be vertex-disjoint if the intersection of the vertex sets V_1 and V_2 is empty. They are said to be edge-disjoint if the intersection of the edge sets R_1 and R_2 is empty. Vertex-disjoint graphs are edge-disjoint. The intersection and the union of two graphs are defined as follows according to the algebra of relations :

$$\begin{array}{lll} \text{intersection} & G_1 \sqcap G_2 & := (V_1 \cap V_2 \, ; R_1 \sqcap R_2) \\ \text{union} & G_1 \sqcup G_2 & := (V_1 \cup V_2 \, ; R_1 \sqcup R_2) \end{array}$$

Homomorphic mapping : Let a directed graph $G_1 = (V_1 \, ; R_1)$, a directed graph $G_2 = (V_2 \, ; R_2)$ and a mapping $\Phi : V_1 \to V_2$ from the vertex set V_1 to the vertex set V_2 be given. The mapping Φ is called a homomorphic mapping of G_1 into G_2 if for all vertex pairs $(x, y) \in R_1$ the relation R_2 contains the corresponding image pairs $(\Phi(x), \Phi(y))$.

$$\Phi \text{ is homomorphic} \quad :\Leftrightarrow \quad \bigwedge_x \bigwedge_y \; ((x,y) \in R_1 \quad \Rightarrow \quad (\Phi(x), \Phi(y)) \in R_2)$$

The defining property of a homomorphic mapping may be transformed as follows using the algebra of relations for unary point relations :

$$\begin{array}{lllll} \bigwedge_x \bigwedge_y \; ((x,y) \in R_1 & \Rightarrow & (\Phi(x), \Phi(y)) & \in R_2) & \Leftrightarrow \\ \bigwedge_x \bigwedge_y \; (x y^T \sqsubseteq R_1 & \Rightarrow & (\Phi^T x)\,(\Phi^T y)^T & \sqsubseteq R_2) & \Leftrightarrow \\ \bigwedge_x \bigwedge_y \; (x y^T \sqsubseteq R_1 & \Rightarrow & \Phi^T x \, y^T \Phi & \sqsubseteq R_2) & \Leftrightarrow \\ & & \Phi^T R_1 \, \Phi & \sqsubseteq R_2 & \end{array}$$

If the expression $\Phi^T R_1 \Phi \sqsubseteq R_2$ is multiplied by Φ from the left or by Φ^T from the right, the conditions $\mathrm{I} \sqsubseteq \Phi\Phi^T$ for left-totality and $\Phi^T\Phi \sqsubseteq \mathrm{I}$ for right-uniqueness of a mapping Φ yield the following equivalent expressions :

$$\Phi^T R_1 \Phi \sqsubseteq R_2 \quad \Leftrightarrow \quad R_1 \Phi \sqsubseteq \Phi \, R_2 \quad \Leftrightarrow \quad \Phi^T R_1 \sqsubseteq R_2 \Phi^T \quad \Leftrightarrow \quad R_1 \sqsubseteq \Phi R_2 \Phi^T$$

A graph G_1 is said to be homomorphic to the graph G_2 if there is a homomorphic mapping of G_1 into G_2. A homomorphic mapping of G_1 into G_2 preserves the structural properties of G_1.

Isomorphic mapping : Let a directed graph $G_1 = (V_1 \, ; R_1)$, a directed graph $G_2 = (V_2 \, ; R_2)$ and a bijective mapping $\Phi : V_1 \to V_2$ together with its inverse mapping $\Phi^T : V_2 \to V_1$ be given. The bijective mapping Φ is said to be isomorphic if Φ and Φ^T are homomorphic.

$$\Phi \text{ is isomorphic} \quad :\Leftrightarrow \quad (\Phi \text{ is homomorphic}) \quad \wedge \quad (\Phi^T \text{ is homomorphic})$$

The defining property of an isomorphic mapping may be transformed as follows using the defining property of homomorphic mappings :

$$\begin{array}{llll} (\Phi \text{ is homomorphic}) & \wedge & (\Phi^T \text{ is homomorphic}) & \Leftrightarrow \\ (\Phi^T R_1 \Phi \sqsubseteq R_2) & \wedge & (\Phi R_2 \Phi^T \sqsubseteq R_1) & \Leftrightarrow \\ (R_1 \sqsubseteq \Phi R_2 \Phi^T) & \wedge & (\Phi R_2 \Phi^T \sqsubseteq R_1) & \Leftrightarrow \\ R_1 = \Phi R_2 \Phi^T & & & \end{array}$$

The following equivalent expressions are obtained by multiplying the expression $R_1 = \Phi R_2 \Phi^T$ by Φ^T from the left or by Φ from the right :

$$R_1 = \Phi R_2 \Phi^T \Leftrightarrow \Phi^T R_1 = R_2 \Phi^T \Leftrightarrow R_1 \Phi = \Phi R_2 \Leftrightarrow R_2 = \Phi^T R_1 \Phi$$

Two graphs G_1 and G_2 are said to be isomorphic if there is an isomorphic mapping $\Phi : V_1 \rightarrow V_2$. Isomorphic graphs have the same structure.

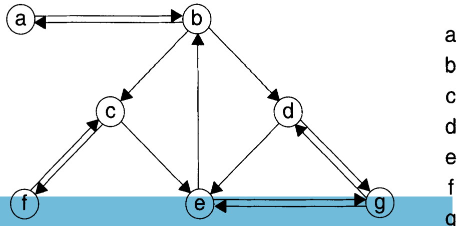
Induced directed graph : Let a directed graph $G_1 = (V_1 ; R_1)$ and a mapping $\Phi : V_1 \rightarrow V_2$ from its vertex set to a vertex set V_2 be given. Then the mapping Φ induces a directed graph $G_2 = (V_2 ; R_2)$ with the edge set R_2 which contains the image vertex pairs $(\Phi(x), \Phi(y))$ for all vertex pairs (x, y) in R_1. The edge set R_2 is determined using the rule for induced binary relations given in Section 8.2.4.

$$\begin{array}{lllll} \text{graph} & : & G_1 & = & (V_1 ; R_1) \\ \text{mapping} & : & \Phi & : & V_1 \rightarrow V_2 \\ \text{edge set} & : & R_2 & = & \Phi^T R_1 \Phi \\ \text{induced graph} & : & G_2 & = & (V_2 ; R_2) \end{array}$$

The graph G_1 is homomorphic to the induced graph G_2. If the mapping Φ is bijective, then the graphs G_1 and G_2 are isomorphic.

Example 1 : Traffic network

Let a set of junctions connected by streets be given. A street can accommodate traffic flow either in two directions or only in one direction (one-way street). The traffic network is represented by a directed graph G = (V ; R). The vertices correspond to the junctions and are collected in the vertex set V. The edges correspond to the traffic flows and are collected in the edge set R.

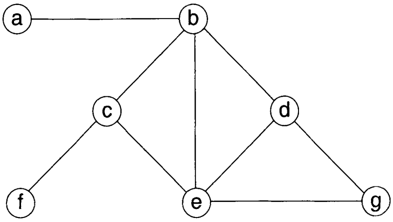


$$\begin{array}{c|ccccccc|} & a & b & c & d & e & f & g \\ \hline a & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ b & 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ c & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ d & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ e & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ f & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ g & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ \hline \end{array}$$

relation R

R

The directed graph G = (V ; R) is associated with a simple graph $G_S = (V ; \Gamma)$. The simple graph describes the neighborhood relationships of the vertices. Two vertices are neighbors if they are connected by at least one directed edge. The neighborhood relationship is represented by an undirected edge. In the graphical representation, the simple graph is obtained from the directed graph by replacing every directed edge or every pair of edges with opposite directions by an undirected edge. In the algebraic representation, the simple graph is obtained from the directed graph by determining the adjacency relation Γ from the relation R. The simple graph for the given traffic network describes the neighborhood relationships of the junctions.



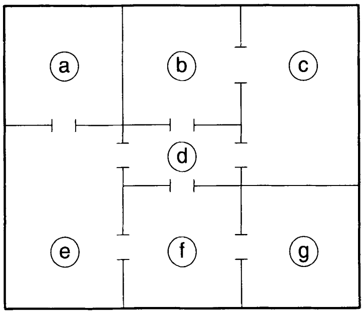
$$\begin{array}{c|ccccccc|}
 & a & b & c & d & e & f & g \\
\hline
a & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
b & 1 & 0 & 1 & 1 & 1 & 0 & 0 \\
c & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\
d & 0 & 1 & 0 & 0 & 1 & 0 & 1 \\
e & 0 & 1 & 1 & 1 & 0 & 0 & 1 \\
f & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
g & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
\hline
\end{array}$$

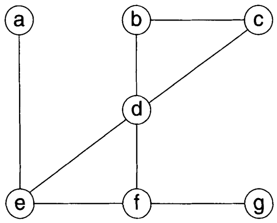
adjacency Γ

$$\mathbf{\Gamma} = \bar{\mathbf{I}} \sqcap (\mathbf{R} \sqcup \mathbf{R}^T)$$

Example 2 : Room planning

Let a room plan for a floor of a building be given. Two neighboring rooms may be connected by a door. The connections between rooms are represented in a simple graph G = (V ; Γ). The vertices correspond to the rooms and are collected in the vertex set V. The undirected edges correspond to the doors between neighboring rooms and are collected in the adjacency relation Γ.





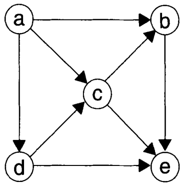
adjacency Γ

	a	b	c	d	e	f	g
a	0	0	0	0	1	0	0
b	0	0	1	1	0	0	0
c	0	1	0	1	0	0	0
d	0	1	1	0	1	1	0
e	1	0	0	1	0	1	0
f	0	0	0	1	1	0	1
g	0	0	0	0	0	1	0

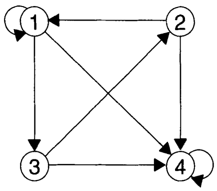
Γ

Example 3 : Homomorphic and isomorphic graphs

$G_1 = (V_1;R_1)$



$G_2 = (V_2;R_2)$

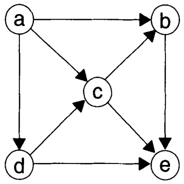


$\Phi : V_1 \rightarrow V_2$

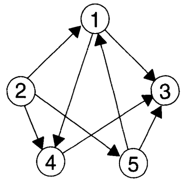
a	2
b	4
c	1
d	1
e	4

The graph G_1 is homomorphic to the graph G_2 by virtue of the homomorphic mapping Φ. For every edge from x to y in G_1 there is an edge from $i = \Phi(x)$ to $j = \Phi(y)$ in G_2. For example, for the edge from a to b in G_1 there is the edge from 2 to 4 in G_2. The graphs G_1 and G_2 are, however, not isomorphic.

$G_1 = (V_1;R_1)$



$G_2 = (V_2;R_2)$



$\Phi : V_1 \rightarrow V_2$

a	2
b	4
c	1
d	5
e	3

$V_1 \leftarrow V_2$

The graphs G_1 and G_2 are isomorphic by virtue of the isomorphic mapping Φ. The mapping Φ is bijective since every vertex of G_1 is bi-uniquely associated with exactly one vertex of G_2. For every edge from x to y in G_1 there is exactly one edge from $i = \Phi(x)$ to $j = \Phi(y)$ in G_2. Conversely, for every edge from i to j in G_2 there is exactly one edge from x to y with $i = \Phi(x)$ and $j = \Phi(y)$ in G_1.

8.3.3 BIPARTITE GRAPHS

Introduction : A bipartite graph is suitable for describing relationships between elements from two disjoint sets. The elements of the two sets are called vertices and are identified by a label in their set. The relationships between any two vertices from different vertex sets are called edges and are identified by an ordered vertex pair. Since edges from vertices of the first set to vertices of the second set as well as edges from vertices of the second set to vertices of the first set may exist, two edge sets are specified, which are heterogeneous binary relations on the two vertex sets. The properties of heterogeneous binary relations and their rules of calculation may be directly transferred to bipartite graphs.

Definition : A domain $G = (V_1,V_2 ; R_1,R_2)$ is called a bipartite graph if V_1, V_2 are disjoint vertex sets and $R_1 \subseteq V_1 \times V_2$, $R_2 \subseteq V_2 \times V_1$ are edge sets. An edge from vertex $x \in V_1$ to vertex $y \in V_2$ is designated by the ordered pair $(x,y) \in R_1$, an edge from vertex $y \in V_2$ to vertex $x \in V_1$ by the ordered pair $(y,x) \in R_2$.

$$G := (V_1,V_2 ; R_1,R_2) \qquad R_1 \subseteq V_1 \times V_2 \qquad R_2 \subseteq V_2 \times V_1$$

V_1,V_2 disjoint sets of vertices

R_1,R_2 disjoint sets of ordered vertex pairs (edge sets)

The graph G is called a null graph if one of the two vertex sets is empty. It is called an empty graph if both edge sets are empty. It is called an associating graph if only one of the two edge sets is empty : The non-empty edge set associates the vertices of one set with the vertices of the other set.

Properties : The properties of a bipartite graph are determined by the properties of the two heterogeneous relations. Left- or right-uniqueness and left- or right-totality are important properties of heterogeneous relations. They are described in Section 8.2.4. In applications, the following properties of bipartite graphs are of special importance :

bipartite graph $G = (V_1,V_2 ; R_1,R_2)$

V_1-total $:\Leftrightarrow$ R_1 left-total $\wedge$ R_2 right-total

$\Leftrightarrow$ $I \sqsubseteq R_1 R_1^T \sqcap R_2^T R_2$

V_2-total $:\Leftrightarrow$ R_1 right-total $\wedge$ R_2 left-total

$\Leftrightarrow$ $I \sqsubseteq R_1^T R_1 \sqcap R_2 R_2^T$

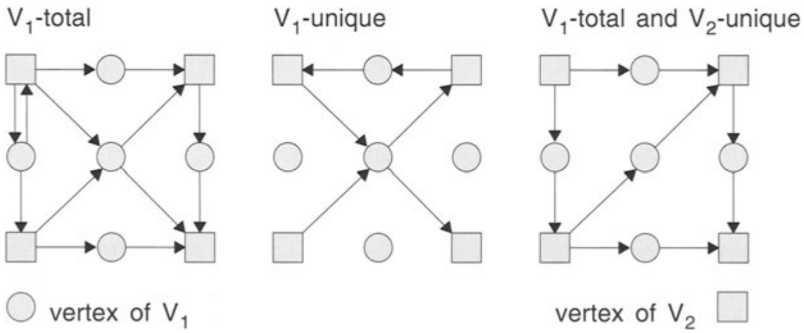
V_1-unique $:\Leftrightarrow$ R_1 left-unique $\wedge$ R_2 right-unique

$\Leftrightarrow$ $R_1 R_1^T \sqcup R_2^T R_2 \sqsubseteq I$

V_2-unique $:\Leftrightarrow$ R_1 right-unique $\wedge$ R_2 left-unique

$\Leftrightarrow$ $R_1^T R_1 \sqcup R_2 R_2^T \sqsubseteq I$

These properties of bipartite graphs are shown in the following graph diagrams. The graph diagram of a bipartite graph contains only directed edges from a vertex of one set to a vertex of the other set. A bipartite graph is V_1-total if at least one edge starts and at least one edge ends at every vertex of V_1. A bipartite graph is V_1-unique if at most one edge starts and at most one edge ends at every vertex of V_2. A bipartite graph is V_1-total and V_2-unique if exactly one edge starts and exactly one edge ends at every vertex of V_1. In this case, the relations R_1 and R_2^T are mappings from the set V_1 to the set V_2.



Associated directed graph : A bipartite graph G is transformed into a directed graph $G_D = (V; R)$ by combining the disjoint vertex sets V_1 and V_2 to the set V and the disjoint edge sets R_1 and R_2 to the set R. The homogeneous relation R for the edge set has a characteristic structure in the matrix scheme, which is a consequence of the bipartite property of the graph.

$$G_D = (V; R) \qquad V = V_1 \cup V_2 \qquad R = R_1 \sqcup R_2$$

$$\mathbf{R} = \begin{array}{c|c|c} & V_1 & V_2 \\ \hline V_1 & \mathbf{0} & \mathbf{R}_1 \\ \hline V_2 & \mathbf{R}_2 & \mathbf{0} \end{array}$$

Since every bipartite graph G may be represented as a directed graph G_S, all properties of directed graphs and their rules of calculation may be directly transferred to bipartite graphs. A bipartite graph is always antireflexive and therefore free of loops. It is symmetric or asymmetric under the following conditions :

$$\text{symmetric} \quad \Leftrightarrow \quad R = R^T \quad \Leftrightarrow \quad R_1 = R_2^T$$

$$\text{asymmetric} \quad \Leftrightarrow \quad R \sqcap R^T = \emptyset \quad \Leftrightarrow \quad R_1 \sqcap R_2^T = \emptyset$$

Directed vertex graphs : Every bipartite graph $G = (V_1,V_2 \,;\, R_1,R_2)$, is associated with two directed vertex graphs $G_1 = (V_1 \,;\, Q_1)$ and $G_2 = (V_2 \,;\, Q_2)$ for the vertex sets V_1 and V_2. The relation Q_1 of G_1 contains exactly those vertex pairs $(x, z) \in V_1 \times V_1$ for which there is a vertex $y \in V_2$ such that G contains an edge from x to y and an edge from y to z. The relation Q_2 of G_2 contains exactly those vertex pairs $(a, c) \in V_2 \times V_2$ for which there is a vertex $b \in V_1$ such that G contains an edge from a to b and an edge from b to c. The definitions of the relations Q_1 and Q_2 correspond to the products $R_1 R_2$ and $R_2 R_1$.

$$Q_1 := \{(x,z) \in V_1 \times V_1 \mid \bigvee_{y \in V_2} [(x,y) \in R_1 \ \wedge \ (y,z) \in R_2]\}$$
$$Q_2 := \{(a,c) \in V_2 \times V_2 \mid \bigvee_{b \in V_1} [(a,b) \in R_2 \ \wedge \ (b,c) \in R_1]\}$$

$$G_1 := (V_1 ; Q_1) \quad \text{with} \quad Q_1 = R_1 R_2$$
$$G_2 := (V_2 ; Q_2) \quad \text{with} \quad Q_2 = R_2 R_1$$

Simple vertex graphs : Every bipartite graph $G = (V_1,V_2 \,;\, R_1,R_2)$ is associated with two simple vertex graphs $G_{S1} = (V_2 \,;\, \Gamma_{S1})$ and $G_{S2} = (V_2 \,;\, \Gamma_{S2})$ for the vertex sets V_1 and V_2. The adjacency relation Γ_{S1} of G_{S1} contains exactly those vertex pairs $(x, z) \in V_1 \times V_1$ with $x \neq z$ for which there is a vertex $y \in V_2$ such that G contains an edge from x to y or from y to x and an edge from y to z or from z to y. The adjacency relation Γ_{S2} of G_{S2} contains exactly those vertex pairs $(a, c) \in V_2 \times V_2$ with $a \neq c$ for which there is a vertex $b \in V_1$ such that G contains an edge from a to b or from b to a and an edge from b to c or from c to b.

$$\Gamma_{S1} := \{(x,z) \in V_1 \times V_1 \mid (x \neq z) \ \wedge$$
$$\bigvee_{y \in V_2} [((x,y) \in R_1 \ \vee \ (y,x) \in R_2) \ \wedge \ ((y,z) \in R_2 \ \vee \ (z,y) \in R_1)]\}$$

$$\Gamma_{S2} := \{(a,c) \in V_2 \times V_2 \mid (a \neq c) \ \wedge$$
$$\bigvee_{b \in V_1} [((a,b) \in R_2 \ \vee \ (b,a) \in R_1) \ \wedge \ ((b,c) \in R_2 \ \vee \ (c,b) \in R_1)]\}$$

$$G_{S1} := (V_1 ; \Gamma_{S1}) \quad \text{with} \quad \Gamma_{S1} = \overline{I} \sqcap (R_1 \sqcup R_2^T)(R_2 \sqcup R_1^T)$$
$$G_{S2} := (V_2 ; \Gamma_{S2}) \quad \text{with} \quad \Gamma_{S2} = \overline{I} \sqcap (R_2 \sqcup R_1^T)(R_1 \sqcup R_2^T)$$

The directed vertex graphs $G_1 = (V_1 \,;\, Q_1)$ and $G_2 = (V_2 \,;\, Q_2)$ are associated with the simple graphs $G_{1S} = (V_1 \,;\, \Gamma_{1S})$ and $G_{2S} = (V_2 \,;\, \Gamma_{2S})$. The adjacency relation Γ_{1S} of G_{1S} contains exactly those vertex pairs $(x, z) \in V_1 \times V_1$ with $x \neq z$ for which there is an edge from x to z or from z to x in G_1. The adjacency relation Γ_{2S} of G_{2S} contains exactly those vertex pairs $(a, c) \in V_2 \times V_2$ with $a \neq c$ for which there is an edge from a to c or from c to a in G_2.

$$\begin{aligned}
\Gamma_{1S} &= \{(x,z) \in V_1 \times V_1 \mid (x \neq z) \wedge ((x,z) \in Q_1 \vee (z,x) \in Q_1)\} \\
\Gamma_{2S} &= \{(a,c) \in V_2 \times V_2 \mid (a \neq c) \wedge ((a,c) \in Q_2 \vee (c,a) \in Q_2)\}
\end{aligned}$$

$$\begin{aligned}
G_{1S} &= (V_1 ; \Gamma_{1E}) \quad \text{with} \quad \Gamma_{1E} = \overline{I} \sqcap (Q_1 \sqcup Q_1^T) = \overline{I} \sqcap (R_1 R_2 \sqcup R_2^T R_1^T) \\
G_{2S} &= (V_2 ; \Gamma_{2E}) \quad \text{with} \quad \Gamma_{2E} = \overline{I} \sqcap (Q_2 \sqcup Q_2^T) = \overline{I} \sqcap (R_2 R_1 \sqcup R_1^T R_2^T)
\end{aligned}$$

The simple graph G_{1S} associated with the directed vertex graph G_1 is a partial graph of the simple vertex graph G_{S1}. If the bipartite graph is V_1-unique, then G_{1S} and G_{S1} coincide. The same is true analogously for G_{2S} and G_{S2}.

$$\begin{aligned}
G_{1S} &\sqsubseteq G_{S1} \qquad & G_{1S} &= G_{S1} \quad \text{if G is } V_1\text{-unique} \\
G_{2S} &\sqsubseteq G_{S2} \qquad & G_{2S} &= G_{S2} \quad \text{if G is } V_2\text{-unique}
\end{aligned}$$

Proof : G_{1S} is a partial graph of G_{S1}

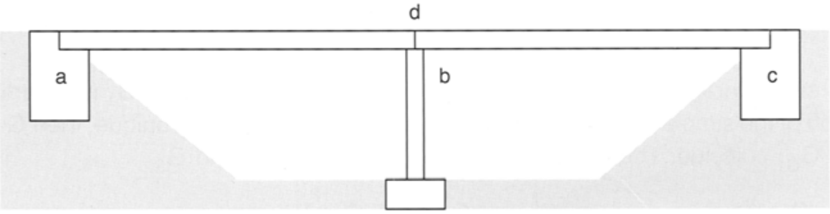
The simple graph $G_{1S} = (V_1 ; \Gamma_{1S})$ associated with the directed vertex graph G_1 is a partial graph of the simple vertex graph $G_{S1} = (V_1 ; \Gamma_{S1})$ if the adjacency relation Γ_{1S} is contained in the adjacency relation Γ_{S1}.

$$\begin{aligned}
\Gamma_{1S} &= \overline{I} \sqcap (R_1 R_2 \sqcup R_2^T R_1^T) \\
\Gamma_{S1} &= \overline{I} \sqcap (R_1 \sqcup R_2^T)(R_2 \sqcup R_1^T) \\
&= \overline{I} \sqcap (R_1 R_2 \sqcup R_1 R_1^T \sqcup R_2^T R_2 \sqcup R_2^T R_1^T) \\
&= (\overline{I} \sqcap (R_1 R_2 \sqcup R_2^T R_1^T)) \sqcup (\overline{I} \sqcap (R_1 R_1^T \sqcup R_2^T R_2)) \\
\Gamma_{S1} &= \Gamma_{1S} \sqcup (\overline{I} \sqcap (R_1 R_1^T \sqcup R_2^T R_2)) \sqsupseteq \Gamma_{1S}
\end{aligned}$$

If the bipartite graph G is V_1-unique, then by definition $R_1 R_1^T \sqcup R_2^T R_2 \sqsubseteq I$, and hence $\overline{I} \sqcap (R_1 R_1^T \sqcup R_2^T R_2) = \emptyset$. This yields the identity $\Gamma_{S1} = \Gamma_{1S}$, and hence $G_{S1} = G_{1S}$.

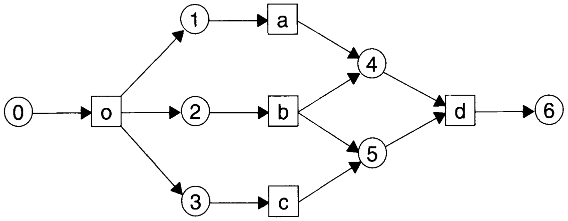
Example : Construction process

A construction process is illustrated in simplified form for the example of a bridge. The process may be represented as a bipartite graph with the construction activities and the construction states as vertices.



construction activities		construction states	
0	start	o	initial state
1	construct left support	a	left support finished
2	construct central column	b	central column finished
3	construct right support	c	right support finished
4	construct left beam		
5	construct right beam	d	beams finished
6	end	d	terminal state

The construction activities form the vertex set V_1, the construction states form the vertex set V_2 of the bipartite graph. A vertex pair $(j, x) \in V_1 \times V_2$ belongs to the relation R_1 if and only if performing the construction activity j contributes to the construction state x. A vertex pair $(y, k) \in V_2 \times V_1$ belongs to the relation R_2 if and only if the construction state y is an initial state for the construction activity k. The construction activities are represented by round vertices, the construction states by quadratic vertices.



bipartite graph

	o	a	b	c	d
0	1	0	0	0	0
1	0	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	0
4	0	0	0	0	1
5	0	0	0	0	1
6	0	0	0	0	0

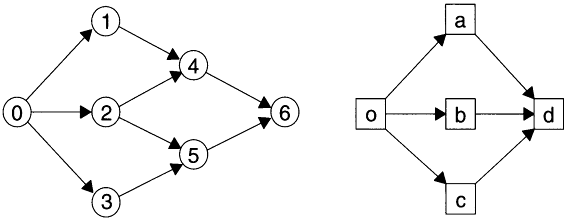
R_1

	0	1	2	3	4	5	6
o	0	1	1	1	0	0	0
a	0	0	0	0	1	0	0
b	0	0	0	0	1	1	0
c	0	0	0	0	0	1	0
d	0	0	0	0	0	0	1

R_2

relations R_1 and R_2

The bipartite graph for the construction process is asymmetric. The two directed vertex graphs of the bipartite graph describe the relationships between the construction activities and the relationships between the construction states. The directed vertex graphs are shown below.



directed vertex graphs

	0	1	2	3	4	5	6
0	0	1	1	1	0	0	0
1	0	0	0	0	1	0	0
2	0	0	0	0	1	1	0
3	0	0	0	0	0	1	0
4	0	0	0	0	0	0	1
5	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0

$\mathbf{Q}_1$

	o	a	b	c	d
o	0	1	1	1	0
a	0	0	0	0	1
b	0	0	0	0	1
c	0	0	0	0	1
d	0	0	0	0	0

$\mathbf{Q}_2$

relations Q_1 and Q_2

$$Q_1 = R_1 R_2$$

$$Q_2 = R_2 R_1$$

8.3.4 MULTIGRAPHS

Introduction : A multigraph consists of vertices and edges, which form a vertex set and an edge set. The vertices are identified by their label in the vertex set, the edges by their label in the edge set. In contrast to the case of a directed graph, the edges are not defined as a relation on the vertices, but rather as independent elements of the graph.

The edges of a multigraph are directed. In contrast to the case of a directed graph, an edge of a multigraph need not possess a start vertex or an end vertex. A vertex of the graph which is the start vertex of at least one edge is called an initial vertex. A vertex of the graph which is the end vertex of at least one edge is called a terminal vertex. Two vertices may be connected by several edges in the same direction. These edges are called parallel edges. A multigraph is a special case of a bipartite graph.

Definition : A domain $G = (V, K ; A, B)$ is called a multigraph if V is the vertex set, K is the edge set and $A, B \subseteq K \times V$ are right-unique binary relations. The relation A specifies the start vertices of the edges and is called the initial incidence. The relation B specifies the end vertices of the edges and is called the terminal incidence. Both relations are right-unique, so that $A^T A \sqsubseteq \mathrm{I}$ and $B^T B \sqsubseteq \mathrm{I}$ holds.

$G :=$	$(V, K ; A, B)$		$A, B \subseteq K \times V$
V	set of vertices		
K	set of edges		
A	initial incidence	:	$A^T A \sqsubseteq \mathrm{I}$
B	terminal incidence	:	$B^T B \sqsubseteq \mathrm{I}$

Initial vertex : A vertex of the graph G is called an initial vertex if it is the start vertex of at least one edge. If the diagonal element for a vertex in the product $\mathbf{A}^T\mathbf{A}$ is one, the vertex is an initial vertex. If it is zero, the vertex is not an initial vertex.

Terminal vertex : A vertex of the graph G is called a terminal vertex if it is the end vertex of at least one edge. If the diagonal element for a vertex in the product $\mathbf{B}^T\mathbf{B}$ is one, the vertex is a terminal vertex. If it is zero, the vertex is not a terminal vertex.

Partial edge : An edge of the graph G is called a partial edge if it lacks a start vertex or an end vertex. If the diagonal element in the product $\mathbf{A}\mathbf{A}^T$ or the diagonal element in the product $\mathbf{B}\mathbf{B}^T$ is zero for an edge, then the edge is a partial edge. If both diagonal elements are one, it is not a partial edge.

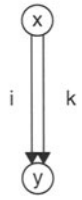
Parallel edges : Two different edges of a multigraph are called parallel if their start vertices are identical and their end vertices are identical. To find out whether two edges are parallel, one considers the product $\mathbf{AA}^T$ for the initial incidence and the product $\mathbf{BB}^T$ for the terminal incidence.

$$\mathbf{AA}^T \qquad \begin{array}{c|cccc} & i & \cdots & k & \cdots \\ \hline \vdots & 0 & & 0 & \\ x & 1 & & 1 & \\ \vdots & 0 & & 0 & \\ y & 0 & & 0 & \\ \vdots & 0 & & 0 & \end{array}$$

$$\begin{array}{c|ccccc|cccc} & \cdots & x & \cdots & y & \cdots & & & & \\ \hline i & 0 & 1 & 0 & 0 & 0 & 1 & & 1 & \\ \vdots & & & & & & & & & \\ k & 0 & 1 & 0 & 0 & 0 & 1 & & 1 & \\ \vdots & & & & & & & & & \end{array}$$

$$\mathbf{BB}^T \qquad \begin{array}{c|cccc} & i & \cdots & k & \cdots \\ \hline \vdots & 0 & & 0 & \\ x & 0 & & 0 & \\ \vdots & 0 & & 0 & \\ y & 1 & & 1 & \\ \vdots & 0 & & 0 & \end{array}$$

$$\begin{array}{c|ccccc|cccc} & \cdots & x & \cdots & y & \cdots & & & & \\ \hline i & 0 & 0 & 0 & 1 & 0 & 1 & & 1 & \\ \vdots & & & & & & & & & \\ k & 0 & 0 & 0 & 1 & 0 & 1 & & 1 & \\ \vdots & & & & & & & & & \end{array}$$



Two different edges i,k have the same start vertex x if the elements (i, i), (i, k), (k, i) and (k, k) in $\mathbf{AA}^T$ are equal to 1. The edges have the same end vertex y if the elements (i, i), (i, k), (k, i) and (k, k) in $\mathbf{BB}^T$ are equal to 1. If both conditions are satisfied, the edges i and k parallel.

Associated bipartite graph : The vertices and edges of a multigraph G are independent elements. A multigraph may therefore be regarded as a bipartite graph. According to the definition of bipartite graphs in Section 8.3.3, the transpose A^T must be used in the bipartite graph instead of the initial incidence A of the multigraph :

Bipartite graph $\qquad G = (V, K ; A^T, B)$

The bipartite graph G is V-unique, since the relations $A, B \subseteq K \times V$ are right-unique. It is K-total if the relations $A, B \subseteq K \times V$ are left-total. A K-total graph does not have partial edges. In this case, the relations A, B are mappings from the edge set K to the vertex set V.

Directed vertex and edge graph : The directed vertex graph and the directed edge graph of a multigraph are formed using the rules for bipartite graphs. The directed vertex graph $G_V = (V ; R_V)$ consists of the vertex set V and the vertex relation R_V. The directed edge graph $G_K = (K ; R_K)$ consists of the edge set K and the edge relation R_K. Two vertices are related if they are connected by an edge. Two edges are related if the end vertex of the first edge coincides with the start vertex of the second edge.

$$G_V = (V ; R_V) \qquad R_V = A^T B$$
$$G_K = (K ; R_K) \qquad R_K = BA^T$$

Simple vertex and edge graph : The simple vertex graph and the simple edge graph of a multigraph are formed using the rules for bipartite graphs. The simple vertex graph $G_{SV} = (V ; \Gamma_{SV})$ consists of the vertex set V and the vertex adjacency Γ_{SV}. Two vertices are adjacent if they are connected by an edge. The simple edge graph $G_{SK} = (K ; \Gamma_{SK})$ consists of the edge set K and the edge adjacency Γ_{SK}. Two edges are adjacent if they have a vertex in common. The vertex adjacency and the edge adjacency are calculated using the incidence M, which is defined as the union of the initial incidence A and the terminal incidence B.

$$G_{SV} = (V ; \Gamma_{SV}) \qquad \Gamma_{SV} = \overline{I} \sqcap M^T M$$

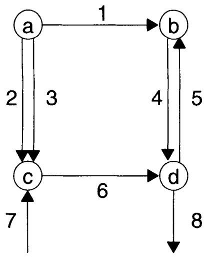
$$G_{SK} = (K ; \Gamma_{SK}) \qquad \Gamma_{SK} = \overline{I} \sqcap M M^T$$

$$\text{incidence} \qquad M = A \sqcup B$$

The simple vertex graph coincides with the simple graph associated with the directed vertex graph, since the associated bipartite graph is V-unique. The simple edge graph does not generally coincide with the simple graph associated with the directed edge graph.

Example : Multigraph

Let a multigraph with the initial incidence A and the terminal incidence B be given. The vertices are labeled by lowercase letters, the edges by numbers. The products of A and B which determine the properties of the graph are calculated.



$$\begin{array}{c|cccc|} & a & b & c & d \\ \hline 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 \\ 5 & 0 & 0 & 0 & 1 \\ 6 & 0 & 0 & 1 & 0 \\ 7 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 1 \\ \hline \end{array} \ \mathbf{A} \qquad \begin{array}{c|cccc|} & a & b & c & d \\ \hline 1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & 1 \\ 5 & 0 & 1 & 0 & 0 \\ 6 & 0 & 0 & 0 & 1 \\ 7 & 0 & 0 & 1 & 0 \\ 8 & 0 & 0 & 0 & 0 \\ \hline \end{array} \ \mathbf{B}$$

$$\begin{array}{c|cccc|} & a & b & c & d \\ \hline a & 1 & 0 & 0 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 0 \\ d & 0 & 0 & 0 & 1 \\ \hline \end{array} \ \mathbf{A^T A} \qquad \begin{array}{c|cccc|} & a & b & c & d \\ \hline a & 0 & 0 & 0 & 0 \\ b & 0 & 1 & 0 & 0 \\ c & 0 & 0 & 1 & 0 \\ d & 0 & 0 & 0 & 1 \\ \hline \end{array} \ \mathbf{B^T B} \qquad \begin{array}{c|cccc|} & a & b & c & d \\ \hline a & 0 & 1 & 1 & 0 \\ b & 0 & 0 & 0 & 1 \\ c & 0 & 0 & 0 & 1 \\ d & 0 & 1 & 0 & 0 \\ \hline \end{array} \ \mathbf{A^T B}$$

$$\begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 3 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ \hline \end{array} \quad \mathbf{AA}^{\mathrm{T}}$$

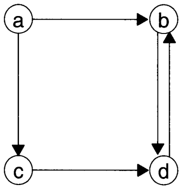
$$\begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 3 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 4 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 5 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 7 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} \quad \mathbf{BB}^{\mathrm{T}}$$

$$\begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 3 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 4 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} \quad \mathbf{BA}^{\mathrm{T}}$$

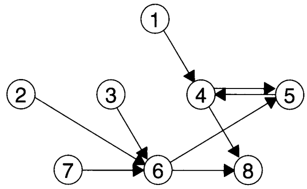
All vertices of the graph are initial vertices, since all diagonal elements in the product $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ are one. The vertex a is not a terminal vertex, since the diagonal element (a,a) in the product $\mathbf{B}^{\mathrm{T}}\mathbf{B}$ is zero. All remaining vertices of the graph are terminal vertices. Edge 7 is a partial edge, since it does not have a start vertex, so that the diagonal element (7,7) in the product $\mathbf{AA}^{\mathrm{T}}$ is zero. Edge 8 is a partial edge, since it does not have an end vertex, so that the diagonal element (8,8) in the product $\mathbf{BB}^{\mathrm{T}}$ is zero. Edges 2 and 3 are parallel, since the elements (2,2), (2,3), (3,2) and (3,3) in the products $\mathbf{AA}^{\mathrm{T}}$ and $\mathbf{BB}^{\mathrm{T}}$ are one.

The directed vertex graph with the vertex relation $R_V = A^TB$ and the directed edge graph with the edge relation $R_K = BA^T$ are represented in the following diagrams. The directed vertex graph corresponds to the multigraph without partial and parallel edges. In the directed edge graph, the edges are represented as vertices, and the edge pairs in the edge relation R_K are represented as directed edges. For example, vertex c is the end vertex of edge 2 and the start vertex of edge 6.

directed vertex graph

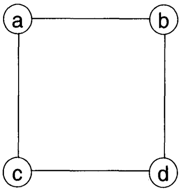


directed edge graph

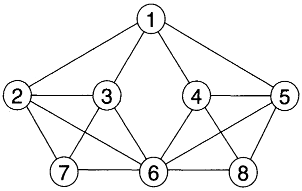


The simple vertex graph with the vertex adjacency $\Gamma_{SV} = \bar{\mathrm{I}} \sqcap \mathrm{M}^T\mathrm{M}$ and the simple edge graph with the edge adjacency $\Gamma_{SK} = \bar{\mathrm{I}} \sqcap \mathrm{M}\mathrm{M}^T$ are shown in the following diagrams.

simple vertex graph



simple edge graph



$$\begin{array}{c|cccc|} & a & b & c & d \\ \hline 1 & 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 \\ 3 & 1 & 0 & 1 & 0 \\ 4 & 0 & 1 & 0 & 1 \\ 5 & 0 & 1 & 0 & 1 \\ 6 & 0 & 0 & 1 & 1 \\ 7 & 0 & 0 & 1 & 0 \\ 8 & 0 & 0 & 0 & 1 \\ \hline \end{array}$$

$\mathbf{M} = \mathbf{A} \sqcup \mathbf{B}$

$$\begin{array}{c|cccc|} & a & b & c & d \\ \hline a & 0 & 1 & 1 & 0 \\ b & 1 & 0 & 0 & 1 \\ c & 1 & 0 & 0 & 1 \\ d & 0 & 1 & 1 & 0 \\ \hline \end{array}$$

$\Gamma_{SV} = \bar{\mathbf{I}} \sqcap \mathbf{M}^T\mathbf{M}$

$$\begin{array}{c|cccccccc|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\ \hline 1 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 1 & 0 & 0 & 1 & 1 & 0 \\ 3 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 5 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 6 & 0 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\ 7 & 0 & 1 & 1 & 0 & 0 & 1 & 0 & 0 \\ 8 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ \hline \end{array}$$

$\Gamma_{SK} = \bar{\mathbf{I}} \sqcap \mathbf{M}\mathbf{M}^T$

The simple vertex graph is obtained from the directed vertex graph by replacing the directed edges by undirected edges and omitting parallel undirected edges. The simple graph associated with the directed vertex graph therefore coincides with the simple vertex graph. The same is not true for the edge graphs.

8.3.5 HYPERGRAPHS

Introduction : A hypergraph consists of vertices and hyperedges, which form a vertex set and an edge set. The vertices are identified by their label in the vertex set, the hyperedges by their label in the edge set. A hyperedge is undirected and connects several vertices with each other. Different hyperedges which connect the same vertices are said to be parallel. For every multigraph, there is a unique associated hypergraph.

Definition : A domain $G = (V, K; M)$ is called a hypergraph if V is the vertex set, K is the edge set and $M \subseteq K \times V$ is a heterogeneous binary relation. The relation M specifies the vertices of the hyperedges and is called the incidence.

$$G := (V, K; M) \qquad\qquad M \subseteq K \times V$$

V	set of vertices
K	set of hyperedges
M	incidence

Associated bipartite graph : The vertices and hyperedges of a hypergraph G are independent objects. A hypergraph may therefore be regarded as a bipartite graph. Since the hyperedges are undirected, the bipartite graph is symmetric. According to Section 8.3.3, the following bipartite graph is associated with the hypergraph :

bipartite graph $\qquad G = (V, K; M^T, M)$

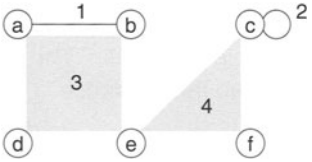
Simple vertex and edge graph : The simple vertex graph and the simple edge graph of a hypergraph are formed using the rules for bipartite graphs. The simple vertex graph $G_{SV} = (V; \Gamma_{SV})$ consists of the vertex set V and the vertex adjacency Γ_{SV}. Two vertices are adjacent if they are connected by an edge. The simple edge graph $G_{SK} = (K; \Gamma_{SK})$ consists of the edge set K and the edge adjacency Γ_{SK}. Two hyperedges are adjacent if they have a vertex in common.

$$G_{SV} = (V; \Gamma_{SV}) \qquad\qquad \Gamma_{SV} = \bar{I} \sqcap M^T M$$
$$G_{SK} = (K; \Gamma_{SK}) \qquad\qquad \Gamma_{SK} = \bar{I} \sqcap M M^T$$

Hypergraph of a multigraph : For a multigraph $G = (V, K; A, B)$ with the initial incidence A and the terminal incidence B, there is a unique associated hypergraph $G_H = (V, K; M)$ with the incidence $M = A \sqcup B$. This hypergraph corresponds to an undirected graph which may contain parallel edges and loops.

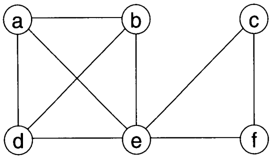
Example 1 : Team projects

The collaboration of several persons on several projects may be represented in a hypergraph. A person may be involved in different projects. A project may involve different persons. The persons form the vertex set V, the projects form the set K of hyperedges. The incidence M describes the assignment of persons to projects. In the graphical representation, the vertices are labeled by lowercase letters, the hyperedges by numbers. Hyperedges with more than two vertices are represented by polygons.

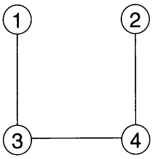


$$\begin{array}{c|cccccc|} & a & b & c & d & e & f \\ \hline 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 2 & 0 & 0 & 1 & 0 & 0 & 0 \\ 3 & 1 & 1 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 1 & 0 & 1 & 1 \\ \hline \end{array} \quad \mathbf{M}$$

The simple vertex graph $G_{SV} = (V ; \Gamma_{SV})$ with the vertex adjacency Γ_{SV} describes the relationships between different persons due to common projects. The simple edge graph $G_{SK} = (K ; \Gamma_{SK})$ with the edge adjacency Γ_{SK} describes the relationships between different projects due to common persons.



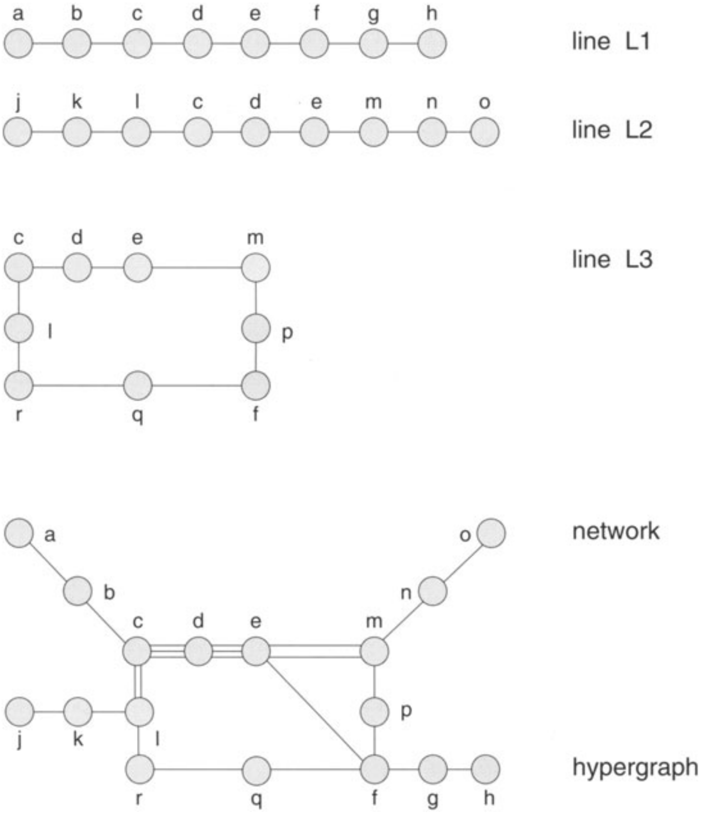
$$\begin{array}{c|cccccc|} & a & b & c & d & e & f \\ \hline a & 0 & 1 & 0 & 1 & 1 & 0 \\ b & 1 & 0 & 0 & 1 & 1 & 0 \\ c & 0 & 0 & 0 & 0 & 1 & 1 \\ d & 1 & 1 & 0 & 0 & 1 & 0 \\ e & 1 & 1 & 1 & 1 & 0 & 1 \\ f & 0 & 0 & 1 & 0 & 1 & 0 \\ \hline \end{array} \quad \boldsymbol{\Gamma}_{SV} = \bar{\mathbf{I}} \sqcap \mathbf{M}^T\mathbf{M}$$



$$\begin{array}{c|cccc|} & 1 & 2 & 3 & 4 \\ \hline 1 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 0 & 1 \\ 3 & 1 & 0 & 0 & 1 \\ 4 & 0 & 1 & 1 & 0 \\ \hline \end{array} \quad \boldsymbol{\Gamma}_{SK} = \bar{\mathbf{I}} \sqcap \mathbf{M}\mathbf{M}^T$$

Example 2 : Subway network

Let several subway lines and their stations be given. The trains travel in both directions on these lines. The totality of subway lines forms the subway network. This network is represented as a hypergraph. The stations are the vertices of the hypergraph. The legs joining two stations are the undirected edges of the hypergraph. They may be used by trains from several lines, so that parallel undirected edges occur in the hypergraph. The subway lines are subgraphs of the hypergraph.



8.4 STRUCTURE OF GRAPHS

8.4.1 INTRODUCTION

Structure : The structure of a graph is uniquely determined by the relations of the domain. For example, the structure of a directed graph (V ; R) is determined by the edge relation R. In analogy with vector algebra, topology and group theory, the question arises whether a graph may be decomposed into subgraphs which have simple structural characteristics and yield insight into the essential structural properties of the graph. Paths and cycles are examples of such subgraphs.

The foundations for the structural analysis of graphs are first treated for directed graphs with directed edges and then transferred to simple graphs with undirected edges. Multigraphs and hypergraphs may be transformed into directed and simple graphs, so that the foundations of structural analysis treated here also apply to these graphs.

Paths and cycles : Graphs consist of vertices and edges. An edge sequence in a graph is a chain of connected edges. An open edge sequence with different start and end vertices is a path. A closed edge sequence with identical start and end vertices is a cycle. Paths and cycles can be simple or elementary. Graphs without cycles are called acyclic graphs. Graphs which consist entirely of cycles are called cyclic graphs. Paths and cycles in directed and simple graphs are treated in Sections 8.4.2 and 8.4.5.

Connectedness : A graph is said to be connected if there is an edge sequence between any two vertices. Different forms of connectedness are defined for directed graphs, in particular strong and weak connectedness. Every graph which is not strongly or weakly connected has a unique decomposition into strongly or weakly connected components. The connectedness of directed and simple graphs is treated in Sections 8.4.3 and 8.4.6.

Cuts : The effects of removing edges and vertices on the connectedness of a graph are studied. Edges are cut, or vertices are excised together with the incident edges, and the connectedness of the remaining graph is studied. These considerations lead to a classification of edges and vertices and to the definition of multiple vertex-disjoint connectedness and multiple edge-disjoint connectedness of graphs. Cuts in directed and simple graphs are treated in Sections 8.4.4 and 8.4.7.

8.4.2 PATHS AND CYCLES IN DIRECTED GRAPHS

Introduction : A directed graph $G = (V\,;R)$ is a structured set. It consists of the vertex set V and a binary vertex relation R which corresponds to a set of directed edges. The vertex set V is equipped with structure by the vertex relation R. The structural properties of a directed graph are entirely determined by the properties of the relation R.

Various concepts are introduced in order to study the structural properties of directed graphs and to cast them in a mathematical form. The definition of paths and cycles in a directed graph forms the basis of the structural analysis. The existence of paths and cycles between two vertices leads to the formation of the transitive closure R^+ of the relation R. The properties of the transitive closure allow a classification into acyclic, anticyclic and cyclic graphs. The essential concepts and fundamentals for the structural analysis of directed graphs are treated in the following.

Predecessor and successor : A vertex x is called a predecessor of a vertex y if there is an edge from x to y in the graph, so that the ordered vertex pair (x,y) is contained in the relation R. If x is a predecessor of y, then y is called a successor of x.

$$\begin{array}{lcll} \text{x predecessor of y} & \Leftrightarrow & (x, y) \in R & \Leftrightarrow \\ \text{y successor of x} & \Leftrightarrow & (y, x) \in R^T & \end{array}$$

A vertex x in a vertex set V may be regarded as a unary point relation in V. In the following, this unary point relation is also designated by x. The predecessorship and the successorship of vertices $x,y \in V$ are formulated as an inclusion using such unary relations :

$$\begin{array}{lcll} \text{x predecessor of y} & \Leftrightarrow & x\,y^T \sqsubseteq R & \Leftrightarrow \\ \text{y successor of x} & \Leftrightarrow & y\,x^T \sqsubseteq R^T & \end{array}$$

The set of all predecessors of a vertex $x \in V$ is designated by $t_P(x)$, the set of all successors of x by $t_S(x)$. The sets $t_P(x)$ and $t_S(x)$ are unary relations in V and are determined as follows using the edge relation R :

$$\begin{array}{lcl} \text{predecessors of x :} & & t_P(x) = R\,x \\ \text{successors of x :} & & t_S(x) = R^T x \end{array}$$

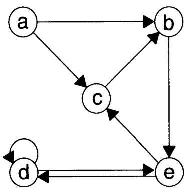
Indegree and outdegree : The number of predecessors of a vertex x is called the indegree of x and is designated by $g_P(x)$. The indegree $g_P(x)$ corresponds to the number of elements in the set $t_P(x)$, and hence to the number of directed edges which end at the vertex x. The number of successors of a vertex x is called the outdegree of x and is designated by $g_S(x)$. The outdegree $g_S(x)$ corresponds to the number of elements in the set $t_S(x)$, and hence to the number of directed edges which emanate from the vertex x.

$$\begin{array}{ll} \text{indegree} & g_P(x) = |t_P(x)| = |Rx| \\ \text{outdegree} & g_S(x) = |t_S(x)| = |R^T x| \end{array}$$

The sum of the indegrees of all vertices $x \in V$ is equal to the number of directed edges of the directed graph, and hence coincides with the number of elements of the relation R. The same is true for the outdegrees.

$$\text{sum} \qquad \sum_{x \in V} g_P(x) = \sum_{x \in V} g_S(x) = |R|$$

Example 1 : Predecessors, successors and degrees in directed graphs



$$\underset{\mathbf{R}}{\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 1 & 0 & 0 \\ b & 0 & 0 & 0 & 0 & 1 \\ c & 0 & 1 & 0 & 0 & 0 \\ d & 0 & 0 & 0 & 1 & 1 \\ e & 0 & 0 & 1 & 1 & 0 \end{array}} \qquad \underset{\mathbf{e}}{\begin{array}{c|c} a & 0 \\ b & 0 \\ c & 0 \\ d & 0 \\ e & 1 \end{array}} \qquad \underset{\mathbf{t}_P(e)}{\begin{array}{c|c} a & 0 \\ b & 1 \\ c & 0 \\ d & 1 \\ e & 0 \end{array}} \qquad \underset{\mathbf{t}_S(e)}{\begin{array}{c|c} a & 0 \\ b & 0 \\ c & 1 \\ d & 1 \\ e & 0 \end{array}}$$

x	a	b	c	d	e
$g_P(x)$	0	2	2	2	2
$g_S(x)$	2	1	1	2	2

$$\sum g_P(x) = 8$$
$$\sum g_S(x) = 8$$

The directed graph shown above consists of 5 vertices and 8 directed edges. The relation is specified by a boolean matrix **R**. The unary point relation for the vertex e is shown as a boolean unit vector **e**. The product **Re** yields the boolean vector $\mathbf{t}_P(e)$ for the set of predecessors of e. It is identical with the column of **R** which is associated with the vertex e. The product $\mathbf{R}^T\mathbf{e}$ yields the boolean vector $\mathbf{t}_S(e)$ for the set of successors of e. It is identical with the row of **R** which is associated with the vertex e. The indegrees and the outdegrees of all vertices are compiled. The sum of the indegrees and the sum of the outdegrees are both equal to the number of edges.

Edge sequence : A chain of edges is called an edge sequence if the end vertex of each edge except for the last edge is the start vertex of the following edge.

edge sequence	$< (x_0, x_1), (x_1, x_2) ,..., (x_{n-1}, x_n) >$
condition	$\bigwedge_{j=1}^{n} ((x_{j-1}, x_j) \in R)$

The start vertex x_0 of the first edge and the end vertex x_n of the last edge are called the start vertex and the end vertex of the edge sequence, respectively. The vertices x_1 to x_{n-1} are called intermediate vertices of the edge sequence. The number n of edges is called the length of the edge sequence. An edge may occur more than once in an edge sequence.

Ancestors and descendants : A vertex x is called an n-th ancestor of a vertex y if there is an edge sequence of length n from x to y in the graph. If x is an n-th ancestor of y, then y is called an n-th descendant of x. A 1-st ancestor or 1-st descendant of x is a predecessor or successor of x, respectively. The n-th ancestors and descendants of x are determined recursively from the relationships for predecessors and successors according to the following rule :

n-th ancestors of x :

$$t_P^{(k)}(x) = R\, t_P^{(k-1)}(x) \qquad \text{for} \qquad k = 1,...,n \qquad \text{with} \qquad t_P^{(0)}(x) = x$$

$$t_P^{(n)}(x) = R^n x \qquad \text{for} \qquad n > 0$$

n-th descendants of x :

$$t_S^{(k)}(x) = R^T t_S^{(k-1)}(x) \qquad \text{for} \qquad k = 1,...,n \qquad \text{with} \qquad t_S^{(0)}(x) = x$$

$$t_S^{(n)}(x) = (R^n)^T x \qquad \text{for} \qquad n > 0$$

The set of all ancestors of a vertex x is designated by $t_P^+(x)$; it is determined as the union of the sets of n-th ancestors of x. The set $t_S^+(x)$ of all descendants of x is determined analogously. The transitive closure R^+ of a relation R with stability index s, defined in Section 8.2.6, may be used to determine these sets :

ancestors of x :

$$t_P^+(x) = t_P^{(1)}(x) \sqcup ... \sqcup t_P^{(s)}(x) = Rx \sqcup ... \sqcup R^s x = R^+ x$$

descendants of x :

$$t_S^+(x) = t_S^{(1)}(x) \sqcup ... \sqcup t_S^{(s)}(x) = R^T x \sqcup ... \sqcup R^{sT} x = R^{+T} x$$

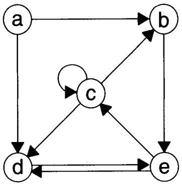
Path : A path from a start vertex x via intermediate vertices to an end vertex y is an edge sequence. In a directed graph, a path may be uniquely represented as a vertex sequence < x,...,y >. A path < x > with the same start and end vertex x contains no edges and is called an empty path. The length of an empty path is 0. There is an empty path for every vertex of a directed graph. The existence of non-empty paths in a directed graph is established as follows :

there is a path of length n from x to y $\Leftrightarrow$ $xy^T \sqsubseteq R^n$
there is a non-empty path from x to y $\Leftrightarrow$ $xy^T \sqsubseteq R^+$

Cycle : A non-empty path whose start vertex and end vertex coincide is called a cycle. A loop at a vertex is a cycle of length 1. A cycle which contains no loops is called a proper cycle. If there is a non-empty path from x to y and a non-empty path from y to x, then the concatenation of the two paths yields a cycle through x and y. The existence of cycles in a directed graph is established as follows :

there is a cycle of length $n > 0$ through x $\Leftrightarrow$ $xx^T \sqsubseteq R^n$
there is a cycle through x $\Leftrightarrow$ $xx^T \sqsubseteq R^+$
there is a cycle through x and y $\Leftrightarrow$ $xy^T \sqsubseteq R^+ \sqcap R^{+T}$

Example 2 : Ancestors and descendants, paths and cycles in directed graphs



$$\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 0 & 1 & 0 \\ b & 0 & 0 & 0 & 0 & 1 \\ c & 0 & 1 & 1 & 1 & 0 \\ d & 0 & 0 & 0 & 0 & 1 \\ e & 0 & 0 & 1 & 1 & 0 \end{array}$$

R

$$\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 0 & 0 & 0 & 1 \\ b & 0 & 0 & 1 & 1 & 0 \\ c & 0 & 1 & 1 & 1 & 1 \\ d & 0 & 0 & 1 & 1 & 0 \\ e & 0 & 1 & 1 & 1 & 1 \end{array}$$

$\mathbf{R}^2$

$$\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 1 & 1 & 1 \\ b & 0 & 1 & 1 & 1 & 1 \\ c & 0 & 1 & 1 & 1 & 1 \\ d & 0 & 1 & 1 & 1 & 1 \\ e & 0 & 1 & 1 & 1 & 1 \end{array}$$

$\mathbf{R}^+$

The relation R, the product R^2 and the transitive closure R^+ for the graph under consideration are shown as boolean matrices. The second ancestors of a vertex are read off from the column of the matrix $\mathbf{R}^2$ associated with that vertex. The second ancestors of vertex e are a, c, e. The second descendants of a vertex are read off from the row of the matrix $\mathbf{R}^2$ associated with that vertex. The second descendants of vertex b are c, d. In the same way, the ancestors and descendants are read off from the matrix $\mathbf{R}^+$. Vertex a does not have any ancestors, but it has the descendants b, c, d, e.

The existence of paths of length 2 may be read off directly from the elements of the matrix $\mathbf{R}^2$. There are paths of length 2 from vertex a to e, namely < a, b, e > and < a, d, e >. There is no path of length 2 from vertex d to e. There is a path of length 2 from vertex d to vertex d, namely the cycle < d, e, d >. There is a path of length 2 from vertex c to vertex c, namely the improper cycle < c, c, c >. The existence of non-empty paths may also be read off from the elements of the matrix $\mathbf{R}^+$. There is no path from vertex e to vertex a. There is a non-empty path from vertex e to d, namely < e, d >, < e, c, d >, < e, c, b, e, d >,... . There are non-empty paths from vertex e to e, namely the cycles < e, d, e >, < e, c, c, d, e >, < e, c, b, e >,... .

Acyclic graph : A directed graph G = (V ; R) is said to be acyclic if it does not contain any cycles. The transitive closure R^+ of an acyclic graph is asymmetric. If there is a non-empty path from x to y, then there is no non-empty path from y to x, since otherwise the concatenation of the two paths would yield a cycle.

acyclic graph $:\Leftrightarrow \quad R^+ \sqcap R^{+T} = 0$

Anticyclic graph : A directed graph G = (V ; R) is said to be anticyclic if it does not contain any proper cycles. In contrast to acyclic graphs, an anticyclic graph may contain loops at the vertices. The transitive closure R^+ of an anticyclic graph is antisymmetric.

anticyclic graph $:\Leftrightarrow \quad R^+ \sqcap R^{+T} \sqsubseteq \mathrm{I}$

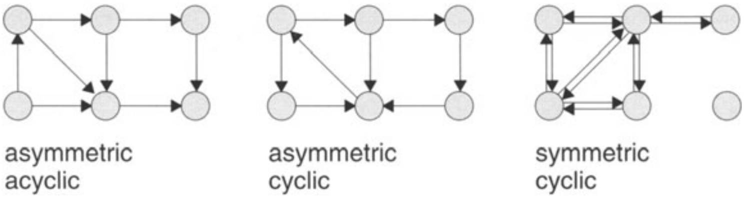
Cyclic graph : A directed graph G = (V ; R) is said to be cyclic if every non-empty path in G belongs to a cycle. The transitive closure R^+ of a cyclic graph is symmetric. If there is a non-empty path from x to y, then there is also a non-empty path from y to x, so that the concatenation of the two paths yields a cycle.

cyclic graph $:\Leftrightarrow \quad R^+ = R^{+T}$

Properties : The following relationships hold between the properties of a relation R and of its transitive closure R^+. If the transitive closure R^+ is asymmetric or antisymmetric, then the relation R is asymmetric or antisymmetric, respectively. If the relation R is symmetric, then the transitive closure R^+ is symmetric. These relationships lead to the following implications :

acyclic graph $\Rightarrow$ asymmetric graph
anticyclic graph $\Rightarrow$ antisymmetric graph
cyclic graph $\Leftarrow$ symmetric graph

Example 3 : Properties of graphs



asymmetric acyclic

asymmetric cyclic

symmetric cyclic

These examples show that while every acyclic graph is asymmetric, not every asymmetric graph is acyclic. They also show that while every symmetric graph is cyclic, not every cyclic graph is symmetric.

Simple path : A non-empty path is said to be simple if it does not contain any edge more than once. The vertices and the edges of a simple path form a subgraph of the directed graph. If the start vertex and end vertex of a simple path are different, the following relationships hold between the indegrees and the outdegrees of the vertices of the corresponding subgraph :

subgraph for a simple path $< x,...,z,...,y >$ with $x \neq y$

start vertex	$g_S(x) = g_P(x) + 1$
intermediate vertex	$g_S(z) = g_P(z)$
end vertex	$g_S(y) = g_P(y) - 1$

Simple cycle : A simple path whose start vertex and end vertex coincide is called a simple cycle. In the subgraph for a simple cycle, the indegree and the outdegree of each vertex are equal.

subgraph for a simple cycle with vertex z

vertex	$g_S(z) = g_P(z)$

Eulerian paths and cycles : A simple path with different start and end vertices is called an Eulerian path if it contains all edges of the directed graph. A simple cycle is called an Eulerian cycle if it contains all edges of the directed graph.

Elementary path : A non-empty path is said to be elementary if it does not contain any vertex more than once. The vertices and the edges of an elementary path form a subgraph. If the start vertex and end vertex of an elementary path are different, then the vertices of the corresponding subgraph have the following indegrees and outdegrees :

subgraph for an elementary path $< x,...,z,...,y >$ with $x \neq y$

start vertex	$g_S(x) = 1$	$g_P(x) = 0$
intermediate vertex	$g_S(z) = 1$	$g_P(z) = 1$
end vertex	$g_S(y) = 0$	$g_P(y) = 1$

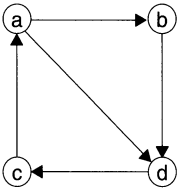
Elementary cycle : An elementary path whose start vertex and end vertex coincide is called an elementary cycle. In the subgraph of an elementary cycle, the indegree and the outdegree of every vertex are equal to 1. Note that the identical start and end vertex is counted once, not twice.

subgraph for an elementary cycle with vertex z

vertex	$g_S(z) = g_P(z) = 1$

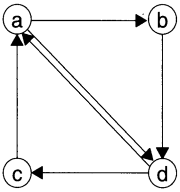
Hamiltonian paths and cycles : An elementary path with different start and end vertices is called a Hamiltonian path if it contains all vertices of the directed graph. An elementary cycle is called a Hamiltonian cycle if it contains all vertices of the directed graph.

Example 4 : Eulerian and Hamiltonian paths and cycles



x	a	b	c	d
$g_S(x)$	2	1	1	1
$g_P(x)$	1	1	1	2

Eulerian paths	< a, d, c, a, b, d >,	< a, b, d, c, a, d >
Hamiltonian paths	< a, b, d, c >,	< b, d, c, a >



x	a	b	c	d
$g_S(x)$	2	1	1	2
$g_P(x)$	2	1	1	2

Eulerian cycles	< a, d, c, a, b, d, a >,	< d, c, a, d, a, b, d >
Hamiltonian cycles	< a, b, d, c, a >,	< b, d, c, a, b >

8.4.3 CONNECTEDNESS OF DIRECTED GRAPHS

Introduction : In a directed graph $G = (V\,;R)$, a vertex may or may not be reachable from another vertex along the directed edges. The concept of reachability forms the basis for a definition of the connectedness of vertices. Different kinds of connectedness may be defined, such as strong and weak connectedness. Directed graphs which are not strongly or weakly connected may be decomposed uniquely into strongly or weakly connected subgraphs. These subgraphs are called strongly or weakly connected components, respectively. Connectedness and decompositions of directed graphs are treated in the following.

Reachability : In a directed graph $G = (V\,;R)$, a vertex $y \in V$ is said to be reachable from a vertex $x \in V$ if there is an empty or non-empty path from x to y. Vertex y is reachable from vertex x if and only if the product $x y^T$ of the associated point relations x and y is contained in the reflexive transitive closure R^*.

y is reachable from x $\quad :\Leftrightarrow \quad x y^T \sqsubseteq R^* \qquad\qquad R^* = I \sqcup R^+$

Strong connectedness : Two vertices x and y of a directed graph are said to be strongly connected if x is reachable from y and y is reachable from x. A directed graph is said to be strongly connected if all vertices are pairwise strongly connected.

x and y are strongly connected $\quad :\Leftrightarrow \quad x y^T \sqsubseteq R^* \sqcap R^{*T}$
the graph is strongly connected $\quad :\Leftrightarrow \quad R^* \sqcap R^{*T} = E \quad \Leftrightarrow \quad R^* = E$

Unilateral connectedness : Two vertices x and y of a directed graph are said to be unilaterally connected if x is reachable from y or y is reachable from x. A directed graph is said to be unilaterally connected if all vertices are pairwise unilaterally connected.

x and y are unilaterally connected $\quad :\Leftrightarrow \quad x y^T \sqsubseteq R^* \sqcup R^{*T}$
the graph is unilaterally connected $\quad :\Leftrightarrow \quad R^* \sqcup R^{*T} = E$

Weak connectedness : Two vertices x and y of a directed graph $(V\,;R)$ are said to be weakly connected if they are strongly connected in the symmetric graph $G = (V\,;R \sqcup R^T)$. A directed graph is said to be weakly connected if all vertices are pairwise weakly connected. Since the transitive closure of a symmetric relation is symmetric, this definition may be expressed as follows :

x and y are weakly connected $\quad :\Leftrightarrow \quad x y^T \sqsubseteq (R \sqcup R^T)^*$
the graph is weakly connected $\quad :\Leftrightarrow \quad (R \sqcup R^T)^* = E$

Connectedness relations : The relation R of a directed graph G = (V ; R) generally contains strong, unilateral and weak connections. A relation which contains only connections of the same type is called a connectedness relation. The connectedness relations for a directed graph G are derived from the relation R and its reflexive transitive closure R^* :

strong connectedness relation : $S = R^* \sqcap R^{*T}$
unilateral connectedness relation : $P = R^* \sqcup R^{*T}$
weak connectedness relation : $C = (R \sqcup R^T)^*$

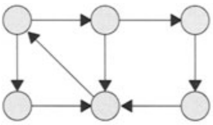
A strongly connected vertex pair is also unilaterally connected ; a unilaterally connected vertex pair is also weakly connected. Hence a strongly connected graph is also unilaterally connected, and a unilaterally connected graph is also weakly connected. For a symmetric graph, the three different kinds of connectedness coincide.

inclusion : $R^* \sqcap R^{*T} \sqsubseteq R^* \sqcup R^{*T} \sqsubseteq (R \sqcup R^T)^*$
connectedness : strong ⇒ unilateral ⇒ weak

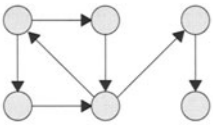
Two different vertices which are strongly connected lie on a cycle. A strongly connected graph is therefore cyclic. The converse is not true in the general case.

strongly connected graph ⇒ cyclic graph

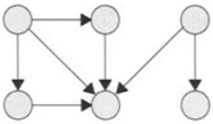
Example 1 : Connectedness of graphs



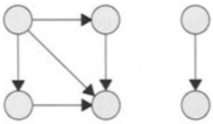
strongly connected



unilaterally connected



weakly connected



not connected

Properties of the connectedness relations : The strong connectedness relation S is reflexive, symmetric and transitive. Reflexivity and symmetry follow directly from the definition. Transitivity follows from the following consideration. If (x, y) and (y, z) are strongly connected vertex pairs, then z is reachable from x via y and x is reachable from z via y. Hence (x, z) is also a strongly connected vertex pair.

The unilateral connectedness relation P is reflexive and symmetric, but generally not transitive. This follows from the following consideration. If (x, y) and (y, z) are unilaterally connected vertex pairs, then it is possible that x is only reachable from y and z is only reachable from y. In this case, neither is x reachable from z, nor is z reachable from x. Thus (x, z) is not a unilaterally connected vertex pair.

The weak connectedness relation C is by definition the strong connectedness relation of an associated symmetric graph. This is reflexive, symmetric and transitive.

A reflexive, symmetric and transitive relation is an equivalence relation. Hence the strong and weak connectedness relations are equivalence relations. The unilateral connectedness relation is generally not an equivalence relation.

Decomposition into connected components : Let $G = (V ; R)$ be a directed graph. Its strong connectedness relation $S = R^* \sqcap R^{*T}$ and its weak connectedness relation $C = (R \sqcup R^T)^*$ are equivalence relations. Let Z stand for either of these equivalence relations. The graph (V ; R) is connected if the equivalence relation Z is the all relation E. If the graph (V ; R) is disconnected, then it may be uniquely decomposed into connected subgraphs. The subgraphs are called the connected components of the graph. The decomposition is carried out in the following steps, independent of the kind of connectedness being considered :

(1) Connected class : The vertex set V of the graph is partitioned into connected classes using the relation Z. A connected class [x] with the vertex x as a representative contains all vertices of V which are connected with x. The class [x] is a unary relation and is determined as follows :

$$[x] = Zx$$

(2) Mapping : The set K of all connected classes is the quotient set V / Z. Each vertex $x \in V$ is mapped to exactly one connected class, yielding a canonical mapping Φ :

$$\Phi : V \rightarrow K \qquad \text{with} \qquad K = V/Z$$

(3) Reduced graph : The mapping Φ from the vertex set V of the directed graph $G = (V ; R)$ to the set K of connected classes induces the reduced graph $G_K = (K ; R_K)$.

$$G_K = (K ; R_K) \qquad \text{with} \qquad R_K = \Phi^T R \Phi$$

(4) Connected component : A connected component is a connected subgraph $G_k := (V_k, R_k)$ of a directed graph $G = (V; R)$. The vertex set V_k contains all vertices of a connected class k of the graph $(V; R)$. The edge set $R_k = R \sqcap (V_k \times V_k)$ contains the edges from R whose vertices belong to V_k. The union of all connected components G_k is generally a partial graph of G, since the union of all vertex sets V_k is the vertex set V and the union of all edge sets R_k is only a subset of the edge set R.

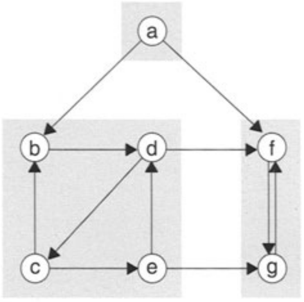
$$\bigsqcup_{k \in K} G_k \sqsubseteq G$$

Decomposition into strongly connected components : The vertex set V of a directed graph $G = (V; R)$ may be decomposed into strongly connected classes using its strong connectedness relation $Z = S = R^* \sqcap R^{*T}$. Two different classes cannot be strongly connected in the reduced graph $G_K = (K; R_K)$, since strongly connected vertices belong to the same class. Each connected component $G_k = (V_k; R_k)$ has a symmetric transitive closure R_k^+ and is therefore a cyclic graph. The reduced graph $G_K = (K; R_K)$ has an antisymmetric transitive closure R_K^+ and is therefore an anticyclic graph.

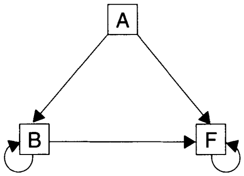
Decomposition into weakly connected components : The vertex set V of a directed graph $G = (V; R)$ may be decomposed into weakly connected classes using its weak connectedness relation $Z = C = (R \sqcup R^T)^*$. Two different classes cannot be weakly connected in the reduced graph $G_K = (K; R_K)$, since weakly connected vertices belong to the same class and the two vertices of an edge are at least weakly connected. Hence every directed graph is the union of its weakly connected components.

$$G = \bigsqcup_{k \in K} G_k$$

Example 2 : Decomposition into strongly connected components



directed graph



reduced graph

relation **R**

	a	b	c	d	e	f	g
a	0	1	0	0	0	1	0
b	0	0	0	1	0	0	0
c	0	1	0	0	1	0	0
d	0	0	1	0	0	1	0
e	0	0	0	1	0	0	1
f	0	0	0	0	0	0	1
g	0	0	0	0	0	1	0

closure $\mathbf{R}^*$

	a	b	c	d	e	f	g
a	1	1	1	1	1	1	1
b	0	1	1	1	1	1	1
c	0	1	1	1	1	1	1
d	0	1	1	1	1	1	1
e	0	1	1	1	1	1	1
f	0	0	0	0	0	1	1
g	0	0	0	0	0	1	1

$\mathbf{S} = \mathbf{R}^* \sqcap \mathbf{R}^{*\mathrm{T}}$

	a	b	c	d	e	f	g
a	1	0	0	0	0	0	0
b	0	1	1	1	1	0	0
c	0	1	1	1	1	0	0
d	0	1	1	1	1	0	0
e	0	1	1	1	1	0	0
f	0	0	0	0	0	1	1
g	0	0	0	0	0	1	1

Let the directed graph $G = (V\,;R)$ shown in the diagram be given. The relation R, the reflexive transitive closure R^* and the strong connectedness relation S are shown as boolean matrices. The graph is not strongly connected, since the reflexive transitive closure R^* is not equal to the all relation E. It is decomposed into its strongly connected components.

The strongly connected classes [a], [b] and [f] are determined using the connectedness relation S. The class [a] contains the vertex a, the class [b] contains the vertices b, c, d, e, and the class [f] contains the vertices f, g. Each vertex of the graph G is mapped to exactly one strongly connected class. The vertex set is thus partitioned into three strongly connected classes, which are designated by the uppercase letters A, B, F of their representatives a, b, f in order to simplify the diagram.

The boolean matrix for the mapping $\Phi: V \rightarrow \{A, B, F\}$ is formed columnwise from the boolean vectors for the unary relations [a], [b], [f]. The edge relation $R_K = \Phi^T R\, \Phi$ of the reduced graph is calculated as a product of boolean matrices.

$$\begin{array}{c|ccc|}
 & A & B & F \\
\hline
a & 1 & 0 & 0 \\
b & 0 & 1 & 0 \\
c & 0 & 1 & 0 \\
d & 0 & 1 & 0 \\
e & 0 & 1 & 0 \\
f & 0 & 0 & 1 \\
g & 0 & 0 & 1 \\
\hline
\end{array}$$

Φ

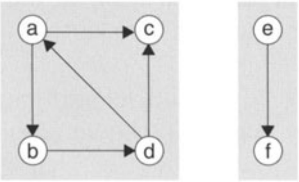
mapping

$$R_K = \Phi^T R \Phi$$

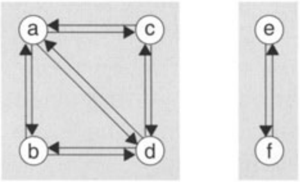
$$\begin{array}{c|ccccccc|c|ccccccc|ccc|}
 & & & & & & & & & a & b & c & d & e & f & g & A & B & F \\
\hline
 & & & & & & & & a & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\
 & & & & & & & & b & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\
 & & & & & & & & c & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\
 & & & & & & & & d & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 1 & 0 \\
 & & & & & & & & e & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\
 & & & & & & & & f & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
 & & & & & & & & g & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
 & a & b & c & d & e & f & g & & & & & & & & & & & \\
\hline
A & 1 & 0 & 0 & 0 & 0 & 0 & 0 & & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 & 1 \\
B & 0 & 1 & 1 & 1 & 1 & 0 & 0 & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 1 \\
F & 0 & 0 & 0 & 0 & 0 & 1 & 1 & & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\
\hline
\end{array}$$

The decomposition of the directed graph yields three strongly connected components, which are cyclic graphs. Like the strongly connected classes, they are designated by A, B, F. They form a reduced graph which is anticyclic. In the diagram of the directed graph at the beginning of the example, the strongly connected components are shaded. The reduced graph is shown alongside, with the vertices A, B, F and the directed edges corresponding to the relation R_K calculated above.

Example 3 : Decomposition into weakly connected components



directed graph with weakly connected components



symmetric graph with strongly connected components

The figure shows a directed graph with the vertex set $V = \{a,...,f\}$, as well as the associated symmetric graph. The weakly connected components of the directed graph and the strongly connected components of the symmetric graph are shaded.

The decomposition of a directed graph into its weakly connected classes is reduced to the decomposition of the symmetric graph into its strongly connected classes. The strongly connected components of the symmetric graph are not connected by edges, and hence neither are the weakly connected components of the directed graph. The directed graph is the union of its weakly connected components.

8.4.4 CUTS IN DIRECTED GRAPHS

Introduction : The reachability and connectedness of vertices in a directed graph are treated in the preceding section. In this section, the effects of removing edges or vertices on reachability and connectedness in the remaining graph are studied. For this purpose, the concept of cuts is introduced.

In a directed graph, edges may be cut or vertices may be excised. Cutting an edge means removing the edge from the graph; this leads to a partial graph. Excising a vertex means removing the vertex as well as the incident edges from the graph; this leads to a subgraph.

The vertices and edges of a directed graph are classified according to how their removal affects reachability and connectedness in the graph. This leads to concepts such as vertex cuts and edge cuts or vertex-disjoint and edge-disjoint paths. These are used to define further structural properties of graphs such as edge-disjoint connectedness and vertex-disjoint connectedness. These concepts and the corresponding structural properties of directed graphs are treated in the following.

Basic edge : An edge (x, y) in a directed graph G = (V ; R) is called a basic edge (separating edge) if the vertex y is reachable from the vertex x only via this edge. If the basic edge is removed, then y is no longer reachable from x.

Chord : An edge (x, y) in a directed graph G = (V ; R) is called a chord if the vertex y is also reachable from the vertex x via other edges. The chord (x, y) is the shortest path from x to y.

Basic graph : A partial graph B = (V ; Q) of a directed graph G = (V ; R) is called a basic graph for G if the following conditions are satisfied :

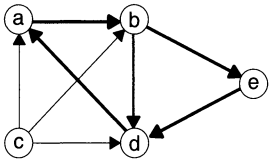
1. If a vertex y is reachable from a vertex x in the directed graph G, then y is also reachable from x in the partial graph B.
2. If an edge from x to y is removed from the partial graph B, then y is no longer reachable from x in the partial graph.

Construction of basic graphs : A basic graph B = (V ; Q) for a directed graph G = (V ; R) is generally not unique. A basic graph contains all basic edges of the graph. It may be iteratively constructed from a directed graph by removing a chord from the current graph in every step. The transitive closure R^+ of the directed graph is identical with the transitive closure Q^+ of a basic graph.

Example 1 : Basic edges and basic graph

The directed graph G = (V ; R) shown below has four basic edges and four chords. The basic edges are highlighted in the diagram. A basic graph B = (V ; Q) of G is shown. It contains the four basic edges and also a chord of the directed graph. The transitive closure R^+ of the directed graph is identical with the transitive closure Q^+ of the basic graph.

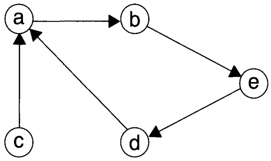
directed graph G = (V ; R)



transitive closure R^+

	a	b	c	d	e
a	1	1	0	1	1
b	1	1	0	1	1
c	1	1	0	1	1
d	1	1	0	1	1
e	1	1	0	1	1

basic graph B = (V ; Q)



transitive closure Q^+

	a	b	c	d	e
a	1	1	0	1	1
b	1	1	0	1	1
c	1	1	0	1	1
d	1	1	0	1	1
e	1	1	0	1	1

Edge cut : Let two different vertices x and y of a directed graph G = (V ; R) be given. An edge set $T(x,y) \sqsubseteq R$ is called an edge cut if removing its edges from the graph G has the effect that y is no longer reachable from x. An edge cut T(x,y) is called a minimal edge cut if no other edge cut S(x,y) contains fewer edges than T(x,y). The number of edges in a minimal edge cut is called the minimal edge cut size and is designated by min t(x,y). If a minimal edge cut contains exactly one edge, then this edge is a basic edge of the directed graph.

Both the set of all edges emanating from the vertex x and the set of all edges ending at the vertex y are edge cuts. Hence the minimal edge cut size is bounded from above by the outdegree of x and the indegree of y.

$$\min t(x,y) \quad \leq \quad \min \{g_S(x),\ g_P(y)\}$$

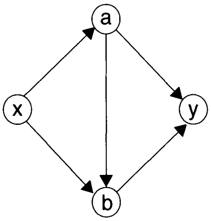
Edge-disjoint paths : Let x and y be two different vertices of a directed graph G = (V ; R). Two simple paths from x to y are said to be edge-disjoint if they have no edge in common. An edge-disjoint path set W(x,y) contains paths from x to y which are pairwise edge-disjoint. It is called a maximal edge-disjoint path set if no other edge-disjoint path set U(x,y) contains more paths than W(x,y). The number of paths in a maximal edge-disjoint path set is called the maximal number of edge-disjoint paths and is designated by max w(x,y).

Every edge emanating from the vertex x and every edge ending at the vertex y can occur in at most one edge-disjoint path from x to y. The maximal number of edge-disjoint paths is therefore bounded from above by the outdegree of x and the indegree of y.

$$\max w(x, y) \;\leq\; \min \{g_S(x),\ g_P(y)\}$$

Example 2 : Construction procedures

The following example illustrates two different procedures for constructing an edge-disjoint path set and a corresponding edge cut. Let the illustrated directed graph with the vertices x and y be given. It possesses a maximal edge-disjoint path set W(x,y) with two paths and three minimal edge cuts T(x,y) with two edges each.



maximal edge-disjoint path set :

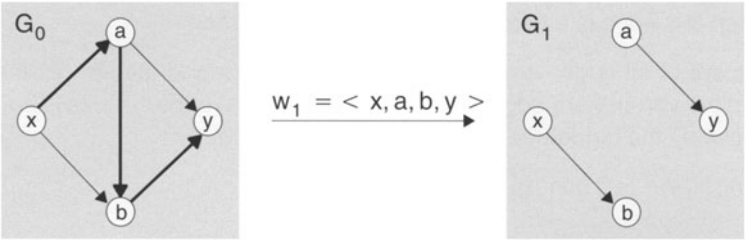
$W(x, y) = \{ < x, a, y >,\ < x, b, y > \}$

minimal edge cuts :

$T(x, y)\ :\ \{(x, a), (x, b)\}, \{(x, a), (b, y)\}, \{(a, y), (b, y)\}$

In the first procedure, edge-disjoint paths of the graph G are constructed in the following steps :

1. Look for a simple path from x to y in the graph $G_0 = G$. Let this path be $w_1 = < x, a, b, y >$. A partial graph G_1 is formed from the graph G_0 by removing all edges of the path w_1 from G_0. This prevents the edges of w_1 from being used again in a later step.

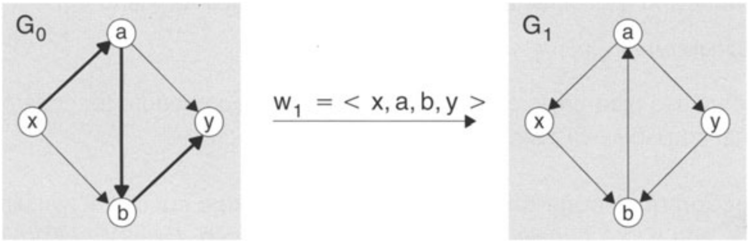


2. Look for a simple path from x to y in the graph G_1. Since there is no such path, the construction procedure terminates.

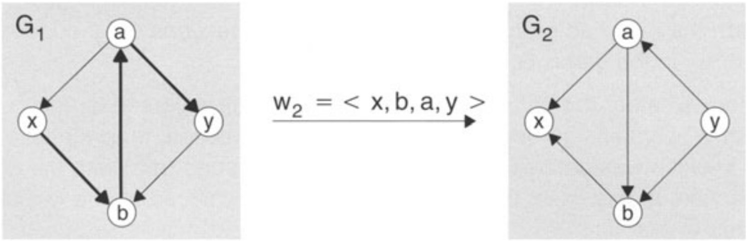
After this construction procedure, the edge-disjoint path set $W(x,y) = \{w_1\}$ consists only of the path $w_1 = < x, a, b, y >$. It is not maximal. A corresponding edge cut $T(x,y) = \{(x,a), (a,b), (b,y)\}$ consists of all edges of the path w_1. It is not minimal and contains a minimal edge cut $\{(x,a), (b,y)\}$ as a subset.

In the second procedure, edge-disjoint paths of the graph G are constructed in the following steps :

1. Look for a simple path from x to y in the graph $G_0 = G$. Let this path be $w_1 = < x, a, b, y >$. A modified graph G_1 is formed from the graph G_0 by reversing all edges of the path w_1 in G_0. This prevents the edges of w_1 from being used again with the same direction in a later step.



2. Look for a simple path from x to y in the graph G_1. Let this path be $w_2 = < x, b, a, y >$. A modified graph G_2 is formed from the graph G_1 by reversing all edges of the path w_2 in G_1. This prevents the edges of w_2 from being used again with the same direction in a later step.



Check whether the path w_2 contains a reversed edge of the path w_1 determined earlier. The path w_2 contains the edge (b,a), which is the reverse of the edge (a,b) in the path w_1. Two shorter paths $\overline{w}_1$ and $\overline{w}_2$ are constructed from the two paths w_1 and w_2. The path $\overline{w}_1 = < x, a, y >$ is the concatenation of the first subpath $< x, a >$ of w_1 and the last subpath $< a, y>$ of w_2. The path $\overline{w}_2 = < x, b, y >$ is the concatenation of the first subpath $<x, b>$ of w_2 and the last subpath $< b, y >$ of w_1. The path $\overline{w}_1$ does not contain the edge (a,b), and the path $\overline{w}_2$ does not contain the reversed edge (b,a). The graph G_2 contains the edge (a,b) in its original direction, so that this edge is available for the construction of further paths. The paths w_1 and w_2 are replaced by $\overline{w}_1$ and $\overline{w}_2$.

3. Look for a simple path from x to y in the graph G_2. Since there is no such path, the construction procedure terminates.

After this construction procedure, the edge-disjoint path set $W(x,y) = \{\bar{w}_1, \bar{w}_2\}$ consists of the paths $\bar{w}_1 = < x, a, y >$ and $\bar{w}_2 = < x, b, y >$. It is maximal. A corresponding edge cut $T(x,y)$ consists of one edge from each edge-disjoint path, for example $T(x,y) = \{(x, a), (b, y)\}$. It is minimal.

Edge-disjoint paths and edge cuts : Let two different vertices x and y of a directed graph $G = (V ; R)$ be given. The maximal number of edge-disjoint paths leading from x to y is equal to the minimal edge cut size for x and y.

$$\max w(x,y) \quad = \quad \min t(x,y)$$

The proof of this theorem is contained in the following procedure for constructing a maximal edge-disjoint path set and a minimal edge cut.

Construction of an edge-disjoint path set and an edge cut : Let two different vertices x and y in a directed graph $G = (V ; R)$ be given. A set $W(x,y)$ of edge-disjoint paths from x to y is constructed iteratively by modifying the original graph. At the beginning the path set $W(x,y)$ is empty, and the graph G_0 is equal to the original graph G. In each step $k = 1, 2, \ldots$ a path w_k of the path set $W(x,y)$ and a modified graph G_k are determined in the following steps :

1. Look for a simple path w_k from x to y in the graph G_{k-1}. If there is no such path, the path set $W(x,y)$ is complete, and the construction terminates.
2. Form the modified graph G_k by reversing the directions of all edges of the path w_k in the graph G_{k-1}.
3. Check whether the path w_k contains an edge with reversed direction. If the path w_k contains an edge (b,a) with reversed direction, then a path w_i with $i < k$ which was determined earlier contains the edge (a,b) with the original direction. In this case, the paths w_i and w_k are replaced by the two shorter paths $\bar{w}_i$ and $\bar{w}_k$.

   $w_i = < x,\ldots,r,a,b,s,\ldots,y>$ $\qquad$ $w_k = < x,\ldots,p,b,a,q,\ldots,y>$

   $\bar{w}_i = < x,\ldots,r,a,q,\ldots,y>$ $\qquad$ $\bar{w}_k = < x,\ldots,p,b,s,\ldots,y>$

   The new path $\bar{w}_i$ does not contain the edge (a,b), and the new path $\bar{w}_k$ does not contain the reversed edge (b,a). Thus the original edge (a,b) is available again for the construction of further paths. Due to the double reversal in steps i and k, this edge is contained in the modified graph G_k. Step 3 is repeated until the path w_k does not contain any edges with reversed direction.

The graph modified by the construction of the edge-disjoint path set W(x,y) is designated by G'. The directions of all edges of the edge-disjoint paths in the modified graph G' are reversed with respect to their directions in the original graph G. The modified graph G' is used as follows to construct an edge cut T(x,y) associated with W(x,y) :

1. In the modified graph G', all vertices reachable from the vertex x are determined and collected in the vertex set X. The vertex x belongs to the vertex set X, since it is reachable from itself.

2. All vertices which do not belong to the vertex set X are collected in the vertex set Y. The vertex y belongs to the vertex set Y, since there is no path from x to y in the graph G'.

3. The edge cut T(x,y) contains all edges (a,b) of the original graph G which lead from a vertex $a \in X$ to a vertex $b \in Y$.

The following relationships hold between the edge-disjoint path set W(x,y) and the corresponding edge cut T(x,y) :

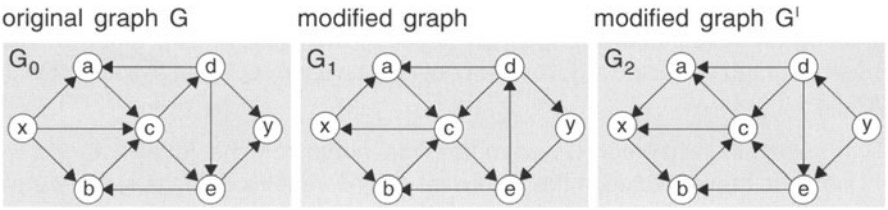
1. The edge cut T(x,y) contains only edges which occur in edge-disjoint paths. If there were an edge (a,b) with $a \in X$ and $b \in Y$ which occured in none of the edge-disjoint paths, then its direction would not be reversed in the modified graph G'. Hence b would have to be reachable from a in G', and would therefore belong to the vertex set X. This contradicts the definition of X.

2. A path $w_k = <x,...,a,b,...,y> \in W(x,y)$ consists of a subpath $w_x = <x,...,a>$ with vertices from X and a subpath $w_y = <b,...,y>$ with vertices from Y. The subpath w_x can only contain vertices from X, since in G' every vertex of w_x is reachable from a. If a vertex c from w_y would belong to X, b would also have to belong to X, since b is reachable from c in G'. This contradicts the hypothesis $b \in Y$. Thus from each path $w_k \in W(x,y)$ the edge cut T(x,y) contains exactly one edge (a,b) with $a \in X$ and $b \in Y$.

3. Since the edge cut T(x,y) contains exactly one edge from every path in W(x,y) and all edges in T(x,y) occur in paths from W(x,y), the size t(x,y) of the edge cut T(x,y) is equal to the number w(x,y) of edge-disjoint paths in W(x,y), that is $t(x,y) = w(x,y)$.

4. Each path of an edge-disjoint path set must contain at least one edge of an edge cut. Hence the edge cut size t(x,y) is an upper bound for the maximal number of edge-disjoint paths. Conversely, the number w(x,y) of edge-disjoint paths is a lower bound for the minimal edge cut size. Since $t(x,y) = w(x,y)$, it follows that t(x,y) is minimal and w(x,y) is maximal.

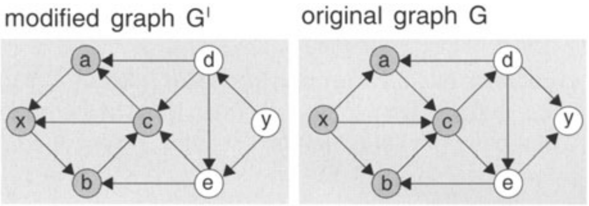
Example 3 : Edge-disjoint paths and edge cuts


original graph G
modified graph
modified graph G'

$w_1 = < x, c, d, e, y >$	$w_2 = < x, a, c, e, d, y >$	no further path
$\bar{w}_1 = < x, c, d, y >$	$\bar{w}_2 = < x, a, c, e, y >$	$W(x, y) = \{\bar{w}_1, \bar{w}_2\}$

The set W(x,y) of edge-disjoint paths from x to y in the graph G shown above is constructed iteratively. In the first step, the simple path w_1 is chosen in the original graph $G_0 = G$, and the modified graph G_1 is formed by reversing all edges of w_1 in G_0. In the second step, the simple path w_2 is chosen in the modified graph G_1, and the modified graph G_2 is formed by reversing all edges of w_2 in G_1. Since the edge (d,e) occurs in w_1 and the reversed edge (e,d) occurs in w_2, the paths w_1 and w_2 are replaced by the paths $\bar{w}_1$ and $\bar{w}_2$. In the modified graph G_2, there is no further path from x to y, and hence $W(x,y) = \{ \bar{w}_1, \bar{w}_2 \}$ is a maximal edge-disjoint path set.

In the modified graph $G' = G_2$, the edges of the paths $\bar{w}_1$ and $\bar{w}_2$ have been reversed. The vertex set X = {x,a,b,c} contains the vertices reachable from x in G', the vertex set Y = {d,e,y} contains the remaining vertices of G'. The minimal edge cut T(x,y) = {(c,d), (c,e)} associated with the maximal edge-disjoint path set W(x,y) contains the edges of the original graph G which lead from a vertex in X to a vertex in Y. The set T(x,y) contains one edge from each of the paths $\bar{w}_1$ and $\bar{w}_2$.


modified graph G'
original graph G

$X = \{x, a, b, c\}$ $\quad$ $T(x, y) = \{(c, d), (c, e)\}$

$Y = \{d, e, y\}$

Vertex cut : Let two different vertices x and y of a directed graph G = (V ; R) be given. A vertex set $S(x, y) \subseteq V - \{x, y\}$ is called a vertex cut if removing its vertices and the incident edges from the graph G has the effect that y is no longer reachable from x. If there is an edge from x to y, then there is no vertex cut. A vertex cut S(x, y) is called a minimal vertex cut if no other vertex cut Q(x, y) contains fewer vertices than S(x,y). The number of vertices of a minimal vertex cut is called the minimal vertex cut size and is designated by min s(x,y). If a minimal vertex cut contains exactly one vertex, then this vertex is called a separating vertex of the directed graph.

If there is no edge from x to y, then both the set of all successors of x and the set of all predecessors of y are vertex cuts. The minimal vertex cut size is therefore bounded from above by the outdegree of x and the indegree of y.

$$\min s(x,y) \quad \leq \quad \min\{g_S(x),\ g_P(y)\} \qquad\qquad (x,y) \notin R$$

Vertex-disjoint paths : Let two different vertices x and y of a directed graph G = (V ; R) be given. Two elementary paths from x to y are said to be vertex-disjoint if they have no vertex in common other than x and y. A vertex-disjoint path set U(x,y) contains paths from x to y which are pairwise vertex-disjoint. A vertex-disjoint path set U(x,y) is called a maximal vertex-disjoint path set if no other vertex-disjoint path set P(x,y) contains more paths than U(x,y). The number of paths in a maximal vertex-disjoint path set is called the maximal number of vertex-disjoint paths and is designated by max u(x,y).

Every successor of x and every predecessor of y can occur in at most one vertex-disjoint path from x to y. The maximal number of vertex-disjoint paths is therefore bounded from above by the outdegree of x and the indegree of y.

$$\max u(x,y) \quad \leq \quad \min\{g_S(x),\ g_P(y)\}$$

Vertex-disjoint paths and vertex cuts : Let two different vertices x and y of a directed graph G = (V ; R) be given which are not connected by an edge from x to y. The maximal number of vertex-disjoint paths leading from x to y is equal to the minimal vertex cut size for x and y.

$$\max u(x,y) \quad = \quad \min s(x,y)$$

The proof of this theorem is contained in the following procedure for constructing a maximal vertex-disjoint path set and a minimal vertex cut.

Construction of a vertex-disjoint path set and a vertex cut : Let two different vertices x and y in a directed graph G = (V ; R) be given which are not connected by an edge from x to y. The determination of a maximal set U(x,y) of vertex-disjoint paths from x to y and a minimal vertex cut S(x,y) in the graph G is reduced to the determination of a maximal set of edge-disjoint paths and a minimal edge cut in a substitute graph G_E. The substitute graph G_E is constructed from the graph G as follows :

- Every vertex a of G is replaced by two vertices a' and a'' and an edge (a',a'').
- Every edge (a,b) of G is replaced by an edge (a'',b').

In the substitute graph G_E, a maximal set $W(x'',y')$ of edge-disjoint paths from x'' to y' and a minimal edge cut $T(x'',y')$ are determined. The following relationships hold between the maximal edge-disjoint path set $W(x'',y')$ in G_E and the maximal vertex-disjoint path set $U(x,y)$ in G, and between the minimal edge cut $T(x'',y')$ in G_E and the minimal vertex cut $S(x,y)$ in G :

(1) There is a one-to-one correspondence of paths $w = < x'',...,z', z'',...,y' >$ from x'' to y' in G_E with paths $u = <x,...,z,...,y>$ from x to y in G. The vertices z', z'' and the edge (z', z'') in w correspond to the intermediate vertex z in u. Two paths from x to y in G are vertex-disjoint if and only if the corresponding paths from x'' to y' in G_E are edge-disjoint. Thus there is also a one-to-one correspondence between maximal edge-disjoint path sets $W(x'',y')$ in G_E and maximal vertex-disjoint path sets $U(x,y)$ in G. Hence the maximal number max $w(x'',y')$ of edge-disjoint paths in G_E is equal to the maximal number max $u(x,y)$ of vertex-disjoint paths in G.

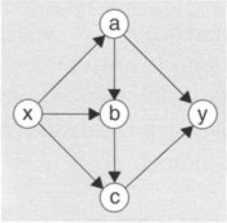
(2) For every maximal edge-disjoint path set $W(x'',y')$ in G_E there is a minimal edge cut $T(x'',y')$ which contains exactly one edge from every edge-disjoint path. If $T(x'',y')$ contains an edge (a'',z') with $z' \neq y'$ or an edge (z'',a') with $z'' \neq x''$, then by virtue of the construction of G this edge may be replaced by the edge (z',z''). Then $T(x'',y')$ contains only edges of type (z',z''). An edge (z',z'') in G_E corresponds to the vertex z. Thus there is a one-to-one correspondence between minimal edge cuts $T(x'',y')$ with edges of type (z',z'') in G_E and minimal vertex cuts $S(x,y)$ with intermediate vertices $z \neq x,y$ in G. Hence the minimal edge cut size min $t(x'',y')$ in G_E is equal to the minimal vertex cut size min $s(x,y)$ in G.

(3) Since in the substitute graph G_E the maximal number max $w(x'',y')$ of edge-disjoint paths is equal to the minimal edge cut size min $t(x'',y')$, it follows by (1) and (2) that in the graph G the maximal number max $u(x,y)$ of vertex-disjoint paths is equal to the minimal vertex cut size min $s(x, y)$.

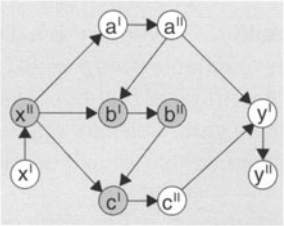
Example 4 : Vertex-disjoint paths and vertex cuts

In the directed graph G shown below, a vertex-disjoint path set and a vertex cut are constructed as follows using the substitute graph G_E and the modified substitute graph G'_E :

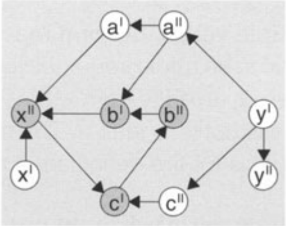
graph G



substitute graph G_E



modified substitute graph G'_E



$u_1 = < x, a, y >$ $\quad$ $w_1 = < x'', a', a'', y' >$ $\quad$ $X'' = \{x'', b', b'', c'\}$

$u_2 = < x, b, c, y >$ $\quad$ $w_2 = < x'', b', b'', c', c'', y'>$ $\quad$ $Y' = \{x', a', a'', c'', y', y''\}$

$U(x,y) = \{u_1, u_2\}$ $\quad$ $W(x'', y') = \{w_1, w_2\}$

$S(x,y) = \{a, c\}$ $\quad$ $T(x'', y') = \{(x'', a'), (c', c'')\} \;\rightarrow\; \{(a', a''), (c', c'')\}$

The substitute graph G_E for the graph G is constructed. In the substitute graph G_E, the edge-disjoint paths w_1 and w_2 from x'' to y' are determined. The substitute graph G_E is transformed into the modified substitute graph G'_E by reversing the direction of every edge in w_1 and w_2. In the modified substitute graph G'_E, all vertices reachable from x'' are collected in the vertex set X'', and all remaining vertices are collected in the vertex set Y'. The minimal edge cut $T(x'', y')$ contains all edges of the substitute graph G_E which lead from a vertex in X'' to a vertex in Y'. The edge (x'', a') is replaced by the edge (a', a''). The edges (a', a'') and (c', c'') in the substitute graph G_E correspond to the vertices a and c in the graph G. The edge-disjoint paths in the substitute graph G_E correspond to the vertex-disjoint paths in the graph G.

Multiple edge-disjoint reachability : A vertex y in a directed graph $G = (V; R)$ is said to be n-fold edge-disjointly reachable from a vertex x if x and y are identical or the maximal number of edge-disjoint paths from x to y is not less than n. The multiple edge-disjoint reachability of vertices forms the basis for the definition of multiple edge-disjoint connectedness.

Multiple edge-disjoint connectedness : Two vertices x and y are said to be n-fold edge-disjointly connected if x is at least n-fold edge-disjointly reachable from y and vice versa. A directed graph is said to be n-fold edge-disjointly connected (n-edge connected) if all vertices are pairwise n-fold edge-disjointly connected. The strong connectedness of a directed graph corresponds to simple (1-fold) edge-disjoint connectedness.

The maximal multiplicity max m of the edge-disjoint connectedness of a directed graph is equal to the minimum of the maximal number of edge-disjoint paths for all vertex pairs (x, y) with x ≠ y. The upper bound for the maximal number of edge-disjoint paths yields an upper bound for the maximal multiplicity.

$$\max m \;=\; \min_{x \neq y} \{\max w(x,y)\} \;\leq\; \min_{x} \{g_S(x), g_P(x)\}$$

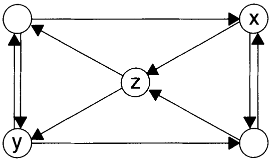
Multiple vertex-disjoint reachability : A vertex y in a directed graph G = (V ; R) is said to be n-fold vertex-disjointly reachable from a vertex x if x and y are identical, if there is an edge from x to y or if the maximal number of vertex-disjoint paths from x to y is not less than n. The multiple vertex-disjoint reachability of a vertex forms the basis for the definition of multiple vertex-disjoint connectedness.

Multiple vertex-disjoint connectedness : Two vertices x and y are said to be n-fold vertex-disjointly connected if x is at least n-fold vertex-disjointly reachable from y and vice versa. A directed graph is said to be n-fold vertex-disjointly connected (n-vertex connected, n-connected) if all vertices are pairwise n-fold vertex-disjointly connected. The strong connectedness of a directed graph corresponds to simple (1-fold) vertex-disjoint connectedness.

The maximal multiplicity max n of the vertex-disjoint connectedness of a directed graph which is not complete is equal to the minimum of the maximal number of vertex-disjoint paths for all vertex pairs (x, y) with x ≠ y which are not connected by an edge from x to y.

$$\max n \;=\; \min_{x \neq y} \{\max u(x,y) \;\mid\; (x,y) \notin R\}$$

Example 5 : Multiple edge- and vertex-disjoint connectedness



directed graph

The directed graph shown above is strongly connected, and hence simply edge-disjointly connected. It is also doubly (2-fold) edge-disjointly connected, since from each vertex each other vertex is reachable via exactly two edge-disjoint paths. The graph has no higher edge-disjoint connectedness, since every vertex has indegree 2 and outdegree 2 and the degree of connectedness is bounded from above by the minimal indegree and outdegree of the vertices.

The directed graph is strongly connected, and therefore simply vertex-disjointly connected. It is not doubly vertex-disjointly connected, since for instance there is only one vertex-disjoint path from the vertex x to the vertex y with the intermediate vertex z as a separating vertex.

8.4.5 PATHS AND CYCLES IN SIMPLE GRAPHS

Introduction : A simple graph $G = (V\,;\Gamma)$ consists of a vertex set V and an adjacency relation Γ for the neighborhood of vertices. The adjacency of two vertices is represented by an undirected edge which corresponds to a pair of edges with opposite directions. The graph is free of loops. The adjacency relation Γ is symmetric and antireflexive.

A simple graph is treated as a symmetric and antireflexive special case of a directed graph. The fundamentals of paths and cycles for directed graphs can largely be transferred to simple graphs. The fundamentals of the structural analysis of simple graphs are treated in the following.

Neighbors : Two vertices x and y are called neighbors if there is an undirected edge between x and y in the simple graph, so that the vertex pairs (x, y) and (y, x) are contained in the adjacency relation Γ.

$$\text{x and y are neighbors} \quad \Leftrightarrow \quad (x, y) \in \Gamma \quad \Leftrightarrow \quad (y, x) \in \Gamma$$

If the vertices x and y are represented as unary point relations in V, which are also designated by x and y, then their neighborhood is determined as follows using the algebra of relations :

$$\text{x and y are neighbors} \quad \Leftrightarrow \quad x y^T \sqsubseteq \Gamma \quad \Leftrightarrow \quad y x^T \sqsubseteq \Gamma$$

The set t (x) of all neighbors of a vertex x is calculated as a unary relation in the vertex set V as follows :

$$\text{neighbors of x} \quad : \quad t(x) = \Gamma x$$

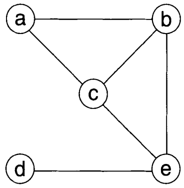
Degree : The number of neighbors of a vertex x is called the degree of the vertex and is designated by g (x). The degree g (x) corresponds to the number of elements in the set t(x) and hence to the number of undirected edges at the vertex x.

$$\text{degree of a vertex} \quad : \quad g(x) = |t(x)| = |\Gamma x|$$

If a simple graph contains k undirected edges, then the sum of the degrees of all vertices $x \in V$ is 2k, and hence equal to the number of elements in Γ.

$$\text{sum of degrees} \quad : \quad \sum_{x \in V} g(x) = \sum_{x \in V} |\Gamma x| = |\Gamma| = 2k$$

Example 1 : Neighbors and degrees



$$\begin{array}{c|ccccc} & a & b & c & d & e \\ \hline a & 0 & 1 & 1 & 0 & 0 \\ b & 1 & 0 & 1 & 0 & 1 \\ c & 1 & 1 & 0 & 0 & 1 \\ d & 0 & 0 & 0 & 0 & 1 \\ e & 0 & 1 & 1 & 1 & 0 \end{array} \qquad \begin{array}{c|c} a & 0 \\ b & 0 \\ c & 0 \\ d & 0 \\ e & 1 \end{array} \qquad \begin{array}{c|c} a & 0 \\ b & 1 \\ c & 1 \\ d & 1 \\ e & 0 \end{array}$$

$\mathbf{\Gamma}$ $\qquad$ **e** $\qquad$ **t**(e)

x	a	b	c	d	e
g(x)	2	3	3	1	3

$$\sum g(x) = |\Gamma| = 12$$

The simple graph shown above consists of 5 vertices and 6 undirected edges. The symmetric adjacency relation is specified by a boolean matrix $\mathbf{\Gamma}$. The unary point relation for the vertex e is shown as a boolean unit vector **e**. The product $\mathbf{\Gamma e}$ yields the boolean vector **t**(e) for all neighbors of e. It coincides with the column of $\mathbf{\Gamma}$ which is associated with the vertex e. The degrees of all vertices are compiled. The sum of the degrees of all vertices is equal to twice the number of undirected edges.

Edge sequence : A chain of edges is called an edge sequence if the end vertex of each edge except for the last edge is the start vertex of the following edge.

edge sequence $\quad < (x_0, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n) >$

condition $\quad \bigwedge_{j=1}^{n} ((x_{j-1}, x_j) \in \Gamma)$

The start vertex x_0 of the first edge and the end vertex x_n of the last edge are called the start vertex and end vertex of the edge sequence, respectively. The vertices x_1 to x_{n-1} are called intermediate vertices of the edge sequence. The number n of edges is called the length of the edge sequence. If there is an edge sequence from x_0 to x_n, then by virtue of the symmetry of simple graphs there is also an edge sequence in the reverse direction from x_n to x_0.

Descendants : A vertex y is called an n-th descendant of a vertex x if there is an edge sequence of length n from x to y in the graph. If y is an n-th descendant of x, then x is also an n-th descendant of y, since for an edge sequence from x to y there is also a reverse edge sequence from y to x. The descendants are calculated as for directed graphs using the n-th power Γ^n and the transitive closure Γ^+ of the adjacency relation Γ.

n-th descendants of x $\quad : \quad t^{(n)}(x) = \Gamma^n x$

all descendants of x $\quad : \quad t^{+}(x) = \Gamma^{+} x$

Path : A path from a start vertex x via intermediate vertices to an end vertex y is an edge sequence. In a simple graph, it can be uniquely represented as a vertex sequence < x,...,y >. A path < x > with the same start and end vertex x contains no edges and is called an empty path. The length of an empty path is 0. There is an empty path for every vertex of a simple graph. The existence of non-empty paths in a simple graph is established as follows :

there is a path of length n from x to y $\Leftrightarrow$ $xy^T \sqsubseteq \Gamma^n$

there is a non-empty path from x to y $\Leftrightarrow$ $xy^T \sqsubseteq \Gamma^+$

Cycle : A non-empty path whose start and end vertex coincide is called a cycle. Due to the symmetry of the adjacency relation Γ, a simple graph contains a large number of trivial cycles. If a path is first traversed in one direction and then retraced in the other direction, a trivial cycle is obtained. However, trivial cycles are not significant for the structure of simple graphs. Since the conditions for the existence of cycles in directed graphs also hold for trivial cycles when applied to simple graphs, they cannot be used to study simple graphs.

Simple path : A non-empty path is said to be simple if it does not contain any undirected edge more than once. The vertices and the undirected edges of a simple path form a subgraph of the simple graph. If the start vertex and end vertex of a simple path are different, then the degrees of the vertices in the subgraph have the following properties :

subgraph for a simple path < x,...,z,...,y > with $x \neq y$

start vertex	:	$g(x)$	odd
intermediate vertex	:	$g(z)$	even
end vertex	:	$g(y)$	odd

Simple cycle : A simple path whose start vertex and end vertex coincide is called a simple cycle. In the subgraph for a simple cycle, the degree of every vertex is even.

subgraph for a simple cycle with vertex z

vertex	:	$g(z)$	even

Eulerian paths and cycles : A simple path with different start and end vertices is called an Eulerian path if it contains all undirected edges of the simple graph. A simple cycle is called an Eulerian cycle if it contains all undirected edges of the simple graph.

Elementary path : A non-empty path is said to be elementary if it does not contain any vertex more than once. The vertices and the undirected edges of an elementary path form a subgraph of the simple graph. If the start vertex and end vertex of a simple path are different, then the degrees of the vertices in the subgraph are :

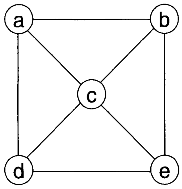
subgraph for a simple path $< x,..., z,...,y >$ with $x \neq y$
start vertex : $g(x) = 1$
intermediate vertex : $g(z) = 2$
end vertex : $g(y) = 1$

Elementary cycle : An elementary path whose start vertex and end vertex coincide is called an elementary cycle. In the subgraph for an elementary cycle, the degree of every vertex is 2. Note that the identical start and end vertex of the cycle is counted once, not twice.

subgraph for an elementary cycle with vertex z
vertex z : $g(z) = 2$

Hamiltonian paths and cycles : An elementary path with different start and end vertices is called a Hamiltonian path if it contains all vertices of the simple graph. An elementary cycle is called a Hamiltonian cycle if it contains all vertices of the simple graph.

Example 2 : Paths and cycles



x	a	b	c	d	e
g(x)	3	3	4	3	3

simple paths	< a, c, e, d, c, b >	< a, b, e, c, b >
elementary paths	< a, b, c >	< b, c, d >
simple cycles	< a, b, c, d, e, c, a >	< e, b, c, d, a, c, e >
elementary cycles	< a, b, c, a >	< a, b, e, c, a >

The simple graph shown above does not contain any Eulerian cycles, since the degrees of vertices a,b,d,e are odd and hence the necessary condition for a simple cycle containing all edges of the graph is not satisfied. However, the graph contains several Hamiltonian cycles. For example, the cycle $< a, b, c, e, d, a >$ is a Hamiltonian cycle.

8.4.6 CONNECTEDNESS OF SIMPLE GRAPHS

Introduction : The connectedness properties of directed graphs may be transferred directly to simple graphs. The symmetry property of simple graphs leads to essential simplifications. The different forms of connectedness of directed graphs coincide for simple graphs and are all referred to as simple connectedness. The fundamentals for the connectedness and the decomposition of simple graphs are treated in the following.

Connectability : In a simple graph $G = (V ; \Gamma)$ two vertices $x, y \in V$ are said to be connectable if there is an empty or non-empty path between x and y. The vertices x and y are connectable if and only if the product xy^T of the associated point relations is contained in the reflexive transitive closure Γ^*.

$$x \text{ and } y \text{ are connectable} \quad :\Leftrightarrow \quad x\,y^T \sqsubseteq \Gamma^* \qquad \Gamma^* = I \sqcup \Gamma^+$$

Simple connectedness : Two vertices x and y in a simple graph are simply connected if x and y are connectable. A simple graph is simply connected if all vertices in V are pairwise simply connected. A distinction between strong, unilateral and weak connectedness is not possible for simple graphs, since the adjacency relation Γ is symmetric, which implies $\Gamma^* \sqcap \Gamma^{*T} = \Gamma^* \sqcup \Gamma^{*T} = (\Gamma \sqcup \Gamma^T)^* = \Gamma^*$.

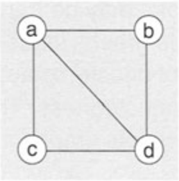
$$\begin{array}{lll} x \text{ and } y \text{ are simply connected} & :\Leftrightarrow & xy^T \sqsubseteq \Gamma^* \\ \text{the graph is simply connected} & :\Leftrightarrow & \Gamma^* = E \end{array}$$

Connectedness relation : Like the strong and the weak connectedness relation for directed graphs, the simple connectedness relation $Z = \Gamma^*$ for simple graphs is an equivalence relation. It forms the basis for a decomposition of simple graphs into their simply connected components.

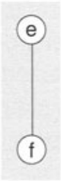
Decomposition into simply connected components : A simple graph $G = (V ; \Gamma)$ may be decomposed into simply connected components using the simple connectedness relation $Z = \Gamma^*$. The vertex set V is mapped to the quotient set $K = V/Z$. The vertex set V_k of a connected component $G_k := (V_k ; \Gamma_k)$ contains all vertices of a connected class $k \in K$. The edge set $\Gamma_k := \Gamma \sqcap (V_k \times V_k)$ contains the edges from Γ whose vertices belong to V_k. There are no edges between the elements of the reduced graph. The simple graph is the union of its simply connected components G_k.

$$G = \bigsqcup_{k \in K} G_k$$

Example : Simple connectedness of a graph



G_1



G_2

	a	b	c	d	e	f
a	0	1	1	1	0	0
b	1	0	0	1	0	0
c	1	0	0	1	0	0
d	1	1	1	0	0	0
e	0	0	0	0	0	1
f	0	0	0	0	1	0

Γ

	a	b	c	d	e	f
a	1	1	1	1	0	0
b	1	1	1	1	0	0
c	1	1	1	1	0	0
d	1	1	1	1	0	0
e	0	0	0	0	1	1
f	0	0	0	0	1	1

closure Γ^*

The reflexive transitive closure of the graph shown above is represented as a boolean matrix Γ^*, from which the simply connected classes may be read off directly. The class [a] corresponds to the column for a in the matrix Γ^*. It contains the vertices a,b,c,d and is the vertex set for the simply connected component G_1. The class [e] corresponds to the column e in the matrix Γ^*. It contains the vertices e,f and is the vertex set for the simply connected component G_2.

8.4.7 CUTS IN SIMPLE GRAPHS

Introduction : The connectability and connectedness of vertices in a simple graph are treated in the preceding section. In this section, the effects of removing edges or vertices on the connectability and connectedness in the remaining graph are studied. For this purpose, the concept of cuts is introduced as in the case of directed graphs.

Edges are classified into bridges and cycle edges according to how their removal affects the connectedness of the graph. If a bridge is cut, the connectedness of the graph is partially lost; if a cycle edge is cut, connectedness is preserved. Acyclic graphs contain only bridges, cyclic graphs contain only cycle edges. Simple cyclic connectedness is defined for simple graphs. Graphs which are not simply cyclically connected may be uniquely decomposed into simply cyclically connected components.

If the connectedness of a graph is partially lost by the excision of a vertex, this vertex is called an articulation vertex. A graph without articulation vertices is elementarily cyclically connected. A graph with articulation vertices may be uniquely decomposed into elementarily cyclically connected blocks.

The definitions of simple and elementary cyclic connectedness allow a deeper analysis of the connectedness properties of simple graphs. They are special cases of multiple edge- and vertex-disjoint connectedness, which is described for directed graphs in Section 8.4.4. The concepts and fundamentals for cyclic connectedness of graphs are treated in the following.

Bridge : An undirected edge between two vertices x and y in a simple graph $G = (V\,;\Gamma)$ is called a bridge if x and y are connectable only via this edge. If a bridge from x to y is removed from the simple graph, then x and y are no longer connectable. If a bridge is removed from a simply connected graph, the graph is divided into two simply connected components.

Cycle edge : An undirected edge between two vertices x and y in a simple graph $G = (V\,;\Gamma)$ is called a cycle edge if it is contained in a simple cycle. If a cycle edge between x and y is removed from the simple graph, x and y remain connectable. If a cycle edge is removed from a simply connected graph, simple connectedness is preserved.

Decomposition of the adjacency relation : Every undirected edge of a simple graph $G = (V ; \Gamma)$ is either a bridge or a cycle edge. The adjacency relation Γ may therefore be uniquely decomposed into a part Γ_B for the bridges and a part Γ_Z for the cycle edges.

adjacency relation for G	:	$\Gamma = \Gamma_B \sqcup \Gamma_Z$ with $\Gamma_B \sqcap \Gamma_Z = \emptyset$
adjacency relation for bridges	:	Γ_B
adjacency relation for cycle edges	:	Γ_Z

Simple acyclic and cyclic graphs : A simple acyclic graph contains only bridges and no cycle edges. A simple cyclic graph contains only cycle edges and no bridges.

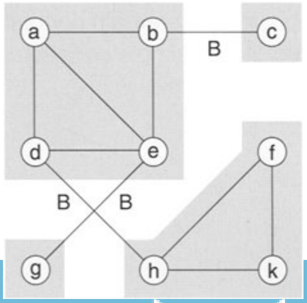
simple acyclic graph :⇔ $\Gamma_Z = \emptyset$
simple cyclic graph :⇔ $\Gamma_B = \emptyset$

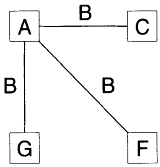
Simple cyclic connectedness : Two vertices x and y are said to be simply cyclically connected if x and y are identical or there is a simple cycle in which they both occur. A simple graph is said to be simply cyclically connected if all vertices are pairwise simply cyclically connected.

x and y are simply cyclically connected :⇔ $xy^T \sqsubseteq \Gamma_Z^*$
the graph is simply cyclically connected :⇔ $\Gamma_Z^* = E$

Decomposition into simply cyclically connected components : The simple cyclic connectedness relation $Z = \Gamma_Z^*$ is an equivalence relation. A simple graph $G = (V ; \Gamma)$ may therefore be uniquely decomposed into simply cyclically connected components. The decomposition is carried out as in the case of directed graphs. Every simply cyclically connected component is a simple cyclic subgraph. The reduced simple graph is a simple acyclic graph if the loops at the vertices are disregarded.

Example 1 : Decomposition into simply cyclically connected components





reduced graph without loops
B bridges

bridge part Γ_B

	a	b	c	d	e	f	g	h	k
a	0	0	0	0	0	0	0	0	0
b	0	0	1	0	0	0	0	0	0
c	0	1	0	0	0	0	0	0	0
d	0	0	0	0	0	0	0	1	0
e	0	0	0	0	0	0	1	0	0
f	0	0	0	0	0	0	0	0	0
g	0	0	0	0	1	0	0	0	0
h	0	0	0	1	0	0	0	0	0
k	0	0	0	0	0	0	0	0	0

cycle part Γ_Z

	a	b	c	d	e	f	g	h	k
a	0	1	0	1	1	0	0	0	0
b	1	0	0	0	1	0	0	0	0
c	0	0	0	0	0	0	0	0	0
d	1	0	0	0	1	0	0	0	0
e	1	1	0	1	0	0	0	0	0
f	0	0	0	0	0	0	0	1	1
g	0	0	0	0	0	0	0	0	0
h	0	0	0	0	0	1	0	0	1
k	0	0	0	0	0	1	0	1	0

closure Γ_Z^*

	a	b	c	d	e	f	g	h	k
a	1	1	0	1	1	0	0	0	0
b	1	1	0	1	1	0	0	0	0
c	0	0	1	0	0	0	0	0	0
d	1	1	0	1	1	0	0	0	0
e	1	1	0	1	1	0	0	0	0
f	0	0	0	0	0	1	0	1	1
g	0	0	0	0	0	0	1	0	0
h	0	0	0	0	0	1	0	1	1
k	0	0	0	0	0	1	0	1	1

The simple graph shown above is simply connected, but not simply cyclically connected. The adjacency relations for the bridge part Γ_B and for the cycle part Γ_Z as well as the reflexive transitive closure Γ_Z^* for the cycle part are shown as boolean matrices. The undirected edges (b, c), (d, h), (e, g) are bridges. All remaining edges are cycle edges. The simply cyclically connected classes can be read off directly from the boolean matrix for the reflexive transitive closure Γ_Z^*. The simple graph possesses the simply cyclically connected classes A, C, G, F with the vertex sets {a, b, d, e}, {c}, {g}, {f, h, k}. These classes form the vertex set of the reduced graph, which is a simple acyclic graph except for the loops. The bridges of the simple graph induce edges between the connected classes of the reduced graph.

Articulation vertex : A vertex a of a simple graph G = (V ; Γ) is called an articulation vertex if two different vertices x ≠ a and y ≠ a are connectable only via a. If the articulation vertex a is removed from the simple graph together with its edges, then x and y are no longer connectable. If an articulation vertex is excised from a simply connected graph, the graph is divided into several simply connected components.

Elementary cyclic connectedness : Two vertices x and y of a simple graph are said to be elementarily cyclically connected if x and y are identical, they are neighbors or there is an elementary cycle in which they both occur. A simple graph is said to be elementarily cyclically connected if all vertices are pairwise elementarily cyclically connected. An elementarily cyclically connected graph does not contain any articulation vertices.

Block : A subgraph of a simple graph $G = (V ; \Gamma)$ is called a block if it is elementarily cyclically connected. A block is said to be proper if it is not contained in another block as a subgraph. If a simple graph contains an elementary cycle, then all vertices and all undirected edges of this cycle belong to a proper block.

Relationships between blocks : A simple graph $G = (V ; \Gamma)$ may possess several proper blocks. Two different proper blocks have the following properties :

(1) Two different proper blocks have either one vertex or no vertices in common.

(2) If two different proper blocks have a vertex in common, this vertex is an articulation vertex of the simple graph.

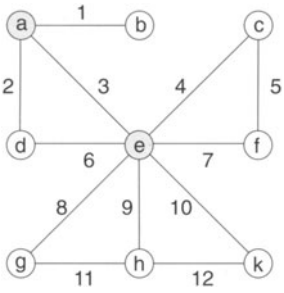
(3) Two different proper blocks are not connected by edges.

Block decomposition : A simple graph $G = (V ; \Gamma)$ may be uniquely decomposed into proper blocks $B_e = (V_e ; \Gamma_e)$. Since two different proper blocks have at most one vertex in common and are not connected by edges, every edge of the simple graph is associated with a unique proper block. The edge sets Γ_e of all blocks are therefore disjoint subsets of the edge set Γ of the simple graph. The vertex sets V_e of the blocks are generally not disjoint subsets of the vertex set V of the simple graph, since articulation vertices are contained in different vertex sets V_e.

$$G = \bigsqcup_e B_e$$

The block structure of a simple graph is represented in a block graph. The vertices of the block graph correspond to the proper blocks. The edges of the block graph indicate that the two proper blocks share a common vertex, which is an articulation vertex of the simple graph. A block graph may also be represented as a hypergraph in which every hyperedge corresponds to an articulation vertex.

Example 2 : Articulation vertices and block decomposition



proper blocks

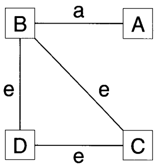
A = ({a,b} ; {1})

B = ({a,d,e} ; {2,3,6})

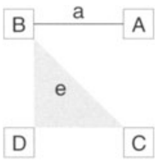
C = ({c,e,f} ; {4,5,7})

D = ({e,g,h,k} ; {8,9,10,11,12})

In the simple graph shown above, lowercase letters identify vertices and numbers identify edges. The simple graph is simply connected, but not elementarily cyclically connected. For example, the vertices a and k do not lie on an elementary cycle. The graph has the articulation vertices a and e. For example, the vertices b and d are connectable only via a, the vertices c and g only via e. The graph possesses four proper blocks A, B, C and D, which are elementarily cyclically connected. The vertices and edges of the blocks are specified above. Every undirected edge is contained in exactly one block. The corresponding block graph with the blocks as vertices and the articulation vertices as edges is shown as a simple graph and as a hypergraph.



simple graph



hypergraph

Multiple edge- and vertex-disjoint connectedness : The fundamentals for multiple edge- and vertex-disjoint connectedness of directed graphs are described in Section 8.4.4. They may be directly transferred to simple graphs, taking into account the symmetry of these graphs. Undirected edges take the place of directed edges, and connectability takes the place of reachability of vertices. The forms of connectedness of simple graphs treated here are special cases of multiple edge- or vertex-disjoint connectedness which are particularly important in applications to practical problems. Simple connectedness corresponds to simple edge-disjoint connectedness and simple vertex-disjoint connectedness. Simple cyclic connectedness corresponds to two-fold edge-disjoint connectedness, and elementary cyclic connectedness corresponds to two-fold vertex-disjoint connectedness.

8.4.8 ACYCLIC GRAPHS

Introduction : The acyclicity of a graph leads to special structural properties of the graph. In studying these properties, a distinction is made between directed acyclic graphs with directed edges and simple acyclic graphs with undirected edges.

Directed acyclic graphs possess an order structure. The vertex set is an ordered set. The directed edges describe the order relation in the vertex set. Due to the order structure, the vertices can be sorted. Edges can be removed from the graph in such a manner that the order structure is preserved. The minimal structure-preserving edge set is unique. The vertex set and the minimal edge set form the basic graph.

Simple acyclic graphs do not have an order structure, since their edges are undirected. They form undirected trees or forests.

Directed acyclic graph : A directed acyclic graph $G = (V ; R)$ is asymmetric and does not contain cycles. Every path from a vertex x to a vertex y is elementary. The closure R^+ is asymmetric and transitive. Hence it is a strict order relation. The theoretical foundations of strict order relations may therefore be applied to directed acyclic graphs.

Rank : Every vertex x of a directed acyclic graph $G = (V ; R)$ is assigned a rank r(x), which is a natural number with the following properties :

(1) A vertex x has the rank $r(x) = 0$ if it does not have any ancestors.

(2) A vertex x has the rank $r(x) = k > 0$ if it has a k-th ancestor and no $(k+1)$-th ancestors.

It is only possible to assign ranks if the directed graph G is acyclic. If there is a cycle through the vertex x, then for every k-th ancestor of x in the cycle there is a predecessor in the cycle, and hence also a $(k+1)$-th ancestor of x. The directed graph must therefore be free of cycles.

If the rank r(x) of a vertex x is k, then by definition the vertex x has a k-th ancestor but no $(k+1)$-th ancestor. Thus there must be a path of length k but no path of length $k+1$ from a vertex without predecessor in G to x. Hence the rank r(x) is the length k of a longest path from a vertex without predecessor in G to x.

Topological sorting : The determination of the ranks of the vertices of a directed graph $G = (V ; R)$ is called topological sorting. The vertex set $V = V_0$ is topologically sorted by iteratively reducing it to the empty vertex set $\emptyset$. In step k, the vertex set V_k is determined whose vertices $x \in V_k$ have a k-th ancestor in G and are therefore of rank $r(x) \geq k$. The vertex set V_k contains all predecessors of the vertices in the vertex set V_{k-1}. This iterative reduction is formulated as follows using unary relations :

initial values	:	$v_0 = e$	all relation
reduction	:	$v_k = R^T v_{k-1}$	$k = 1,...,n$
termination	:	$v_n = \emptyset$	null relation

A vertex x of the vertex set V_k is of degree $r(x) = k$ if it does not belong to the vertex set V_{k+1}. The set W_k of all vertices of rank k is therefore the difference $V_k - V_{k+1}$, which is calculated as the intersection of V_k and the complement of V_{k+1}. It is called the k-th vertex class and is determined as a unary relation as follows :

k-th vertex class : $w_k = v_k \sqcap \overline{v}_{k+1}$ $\qquad k = 0,...,n-1$

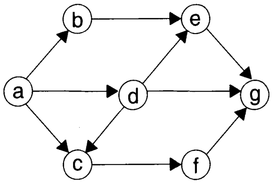
Order structure : Topologically sorting a directed acyclic graph $G = (V ; R)$ yields a partition of the vertex set into disjoint vertex classes W_k with $k = 0,...,n-1$. The partition has the following ordinal properties :

- The vertex class W_0 contains all vertices of the lowest rank 0. These vertices have no ancestors in G, and hence no predecessors. They are therefore minimal. Since there are no other vertices without predecessors, W_0 contains all minimal vertices.
- The vertex class W_{n-1} contains all vertices of the highest rank $n-1$. These vertices have no descendants in G, and hence no successors. They are therefore maximal. Since there may generally also be other vertices without successor, W_{n-1} generally does not contain all maximal vertices.
- Every vertex x in the vertex class W_k with $k > 0$ has at least one predecessor y in the vertex class W_{k-1}. If $x \in W_k$ did not have a predecessor $y \in W_{k-1}$, then x would not have any k-th ancestors, and would therefore not belong to W_k.
- A vertex has neither a predecessor nor a successor in its own vertex class. If y were a predecessor of x and hence x a successor of y, then the rank of y would have to be less than the rank of x and x, y could not belong to the same vertex class.

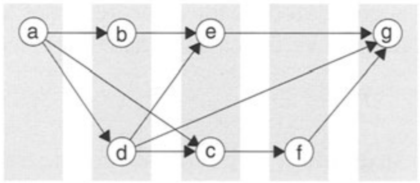
Example 1 : Topological sorting

Let a directed acyclic graph be given. The calculation steps for sorting this graph topologically are shown. The sorting leads to the formation of classes in the vertex set of the graph. The sorted graph and its classes are represented graphically.

acyclic graph



vertex classes of the graph



topological sorting $\mathbf{v}_k = \mathbf{R}^T \mathbf{v}_{k-1}$

$$\begin{array}{c|ccccccc|}
 & a & b & c & d & e & f & g \\
\hline
a & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
b & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
c & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\
d & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
e & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
f & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\
g & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\
\hline
\end{array}$$

$\mathbf{R}^T$

$$\begin{array}{cccccc}
\begin{bmatrix}1\\1\\1\\1\\1\\1\\1\end{bmatrix} &
\begin{bmatrix}0\\1\\1\\1\\1\\1\\1\end{bmatrix} &
\begin{bmatrix}0\\0\\1\\0\\1\\1\\1\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\1\\1\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\0\\1\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\0\\0\end{bmatrix} \\
\mathbf{v}_0 & \mathbf{v}_1 & \mathbf{v}_2 & \mathbf{v}_3 & \mathbf{v}_4 & \mathbf{v}_5
\end{array}$$

vertex classes

$\mathbf{w}_k = \mathbf{v}_k \sqcap \overline{\mathbf{v}}_{k+1}$

$$\begin{array}{ccccc}
\begin{bmatrix}1\\0\\0\\0\\0\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\1\\0\\1\\0\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\1\\0\\1\\0\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\1\\0\end{bmatrix} &
\begin{bmatrix}0\\0\\0\\0\\0\\0\\1\end{bmatrix} \\
\mathbf{w}_0 & \mathbf{w}_1 & \mathbf{w}_2 & \mathbf{w}_3 & \mathbf{w}_4
\end{array}$$

Basic edges and chords : A directed acyclic graph G = (V ; R) has basic edges and chords. An edge from x to y is called a basic edge if y is reachable from x only via this edge. Otherwise it is called a chord. Since a directed acyclic graph does not contain cycles, an edge from x to y is a chord if and only if there is a path of length n > 1 from x to y.

path from x to y with n > 1 $\Leftrightarrow$ $xy^T \sqsubseteq \bigsqcup_{n>1} R^n = R \bigsqcup_{n>0} R^n = RR^+$

chord (x,y) $\Leftrightarrow$ $xy^T \sqsubseteq R \sqcap RR^+$

basic edge (x,y) $\Leftrightarrow$ $xy^T \sqsubseteq R \sqcap \overline{RR^+}$

Basic path : A directed acyclic graph G = (V ; R) does not contain cycles. If there are one or more paths from x to y, then there is at least one path of maximal length. A path of maximal length is called a basic path. A basic path contains only basic edges.

Proof : A basic path contains only basic edges.

Consider a path from x to y of maximal length m which contains an edge from a to b. If the edge from a to b were a chord, there would have to be a path from a to b of length greater than 1, and hence also a path from x to y of length greater than m. But this contradicts the hypothesis. It follows that all edges of a path from x to y of maximal length are basic edges.

Basic graph : The graph B = (V ; Q) is a basic graph of a directed acyclic graph G = (V ; R) if Q contains only the basic edges in R. The basic graph B is constructed by removing all chords from R. The basic graph B is unique. The transitive closures R^+ and Q^+ coincide.

$$\text{basic graph } B = (V\,;\,Q) \qquad \text{with} \qquad Q = R \sqcap \overline{RR^+}$$

Proof : The transitive closures R^+ and Q^+ coincide.

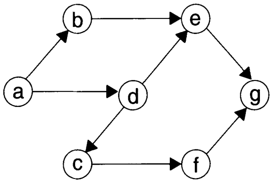
For every chord $(x, y) \in R$ there is by definition a path from x to y of length $n > 1$. Thus there is also a path of maximal length from x to y which is a basic path and consists only of basic edges. Hence y is still reachable from x if the chord (x, y) is removed from R, so that the chord (x, y) yields no additional contribution to the closure R^+. Hence the closures R^+ and Q^+ coincide.

Order diagram : In the topological sorting of a directed acyclic graph G = (V ; R), the rank r(x) of a vertex $x \in V$ is equal to the length of a longest path from a vertex without predecessor to x. This path is a basic path consisting only of basic edges. Hence removing chords from R does not change the rank r(x) of a vertex x, so that topologically sorting the graph G = (V ; R) and its basic graph B = (V ; Q) leads to the same result. The representation of the order structure of the basic graph with its vertex classes is an order diagram according to Section 4.2.

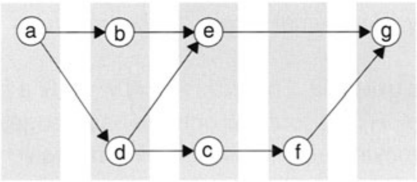
Example 2 : Basic graph and order diagram

Let the directed acyclic graph G = (V ; R) from Example 1 be given. The edges (a, c) and (d, g) are chords, since there are basic paths < a, d, c > and < d, e, g >. The basic graph B = (V ; Q) is constructed from the graph G by removing these chords from G. The edge set Q of the basic graph is calculated using the formula specified above. The basic graph B and the order diagram are represented graphically. The directed acyclic graph G and the basic graph B possess the same vertex classes.

basic graph



vertex classes of the basic graph



Simple acyclic graph : A simple acyclic graph G = (V ; Γ) does not contain any simple cycles. All undirected edges of the graph G are bridges. Removing an edge destroys the original connectedness of the graph G.

Tree : A simple acyclic graph which is simply connected is called a tree. A tree with n vertices has exactly n – 1 undirected edges.

tree	:	$n - k = 1$
number of vertices	:	n
number of edges	:	k

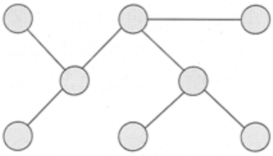
A tree is constructed as follows : A simple graph with only one vertex and no undirected edges is simply connected, does not contain simple cycles and is therefore a tree. Simple connectedness and absence of simple cycles are preserved if the tree is iteratively extended by adding a new vertex with a new undirected edge to an existing vertex in each step. For n vertices, this construction leads to n – 1 edges.

In a tree, the path between two different vertices x and y is unique. If there were several different paths between x and y, there would be cycles, but this is ruled out by the definition of a tree.

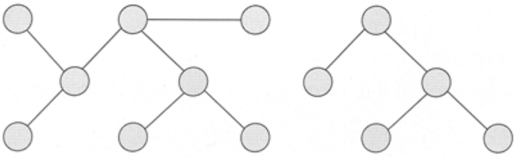
Forest : A simple acyclic graph with several simply connected components is called a forest. Every simply connected component is a tree. By the definition of trees, a forest with n vertices and k undirected edges contains exactly n – k trees.

forest	:	$n - k = c$
number of vertices	:	n
number of edges	:	k
number of components	:	c

Example 3 : Trees and forests



tree : n=8 k=7 c=1



forest : n=13 k=11 c=2

8.4.9 ROOTED GRAPHS AND ROOTED TREES

Introduction : A vertex of a graph from which all remaining vertices are reachable is called a root of the graph. Rooted graphs and rooted trees are of fundamental importance in computer science. For example, finite automata, syntax diagrams and flow diagrams are treated as rooted graphs. All hierarchical structures are regarded as rooted trees. Searching for all vertices of a graph which are reachable from a given vertex leads to a search tree which corresponds to a rooted tree and forms a skeleton of the graph. The fundamentals for rooted graphs, rooted trees and search trees are treated in the following.

Root : A vertex w is called a root (root vertex) of a directed graph $G = (V;R)$ if all vertices of the graph are reachable from the vertex w. If a directed graph is not weakly connected, then it has no root. If it is strongly connected, then every vertex of the graph is a root.

$$w \text{ is a root} \quad :\Leftrightarrow \quad w\, e^T \sqsubseteq R^*$$

Rooted graph : A directed graph $G = (V;R)$ is called a rooted graph if it contains at least one root. In a rooted graph, there is a special form of connectedness between pairs of vertices, called quasi-strong connectedness. Two vertices x and y are quasi-strongly connected if there is a vertex z from which the vertices x and y are both reachable. In this case, there is a path from x to z in the dual graph G^T and a path from z to y in the graph G, so that $(x,z) \in R^{*T}$ and $(z,y) \in R^*$, and hence $(x,y) \in R^{*T}R^*$. In a rooted graph, all vertices are pairwise quasi-strongly connected via a root, so that $R^{*T}R = E$ holds.

$$\begin{array}{lll} \text{x and y are quasi-strongly connected} & :\Leftrightarrow & x\,y^T \sqsubseteq R^{*T}R^* \\ G = (V;R) \text{ is a rooted graph} & :\Leftrightarrow & R^{*T}R^* = E \end{array}$$

Acyclic rooted graph : A directed graph $G = (V;R)$ is acyclic if $R^+ \sqcap R^{+T} = \emptyset$ holds. It is a rooted graph if $R^{*T}R^* = E$ holds. An acyclic rooted graph has exactly one root. The existence of several roots would contradict the absence of cycles.

$$G = (V;R) \text{ is an acyclic rooted graph} \;\Leftrightarrow\; R^+ \sqcap R^{+T} = \emptyset \;\wedge\; R^{*T}R^* = E$$

Rooted tree : An acyclic rooted graph $G = (V;R)$ is called a rooted tree if R is left-unique, so that $RR^T \sqsubseteq I$ holds.

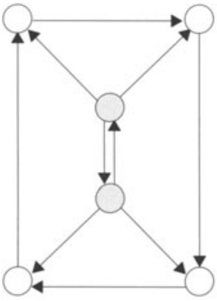
$$G = (V;R) \text{ is a rooted tree} \;:\Leftrightarrow\; RR^T \sqsubseteq I \;\wedge\; R^+ \sqcap R^{+T} = \emptyset \;\wedge\; R^{*T}R^* = E$$

A rooted tree with the root w has the following properties :

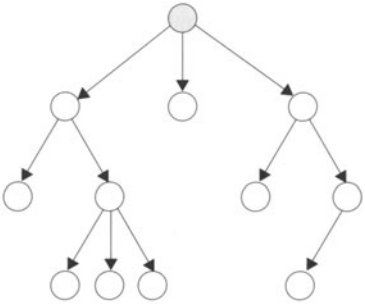
- The root w has no predecessor.
- Every vertex $x \neq w$ has exactly one predecessor.
- Every vertex $x \neq w$ is reachable along exactly one path from w to x.
- A rooted tree with n vertices has exactly n – 1 edges.

Forest of rooted trees : A directed graph is called a forest of rooted trees if every weakly connected component is a rooted tree.

Example 1 : Rooted graphs and rooted trees



rooted graph
with 2 roots



rooted tree
with 12 vertices and 11 edges

Search tree : Let a vertex a in a directed graph G be given. A rooted tree with root a which contains all descendants of a in G is called a search tree at the vertex a. A search tree is constructed by an iterative search, starting from the vertex a. Breadth-first search and depth-first search are distinguished.

Breadth-first search : In a breadth-first search, a vertex sequence F is maintained, which at first contains only the root a. As long as the vertex sequence F is not empty, the following steps are carried out in a loop :

- If the vertex at the beginning of F has a successor which has not been visited yet, such a successor is appended to the end of the sequence F.
- If the vertex at the beginning of F has no successor which has not been visited yet, it is removed from the sequence F.

The vertices visited and the edges used in the course of the breadth-first search form the breadth-first search tree. For every visited vertex x, the search tree contains a path of minimal length from a to x. This property is of fundamental importance for determining paths of minimal length between the vertices of a directed graph.

Depth-first search : In a depth-first search, a vertex sequence F is maintained, which at first contains only the root a. As long as the vertex sequence F is not empty, the following steps are carried out in a loop :

- If the vertex at the end of F has a successor which has not been visited yet, such a successor is appended to the end of the sequence F.
- If the vertex at the end of F has no successor which has not been visited yet, it is removed from the sequence F.

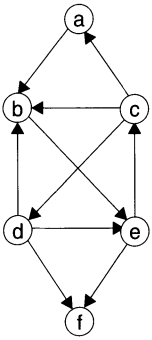
The vertices visited and the edges used in the course of the depth-first search form the depth-first search tree.

Properties : Breadth-first search and depth-first search lead to different search trees. The depth of a search tree is the length of a longest path from the root a to a visited vertex without successor. The breadth of a search tree is the maximal number of visited vertices without successor. Among all search trees, a breadth-first search tree has maximal breadth and minimal depth. A depth-first search tree generally has small breadth and great depth.

Example 2 : Breadth-first search and depth-first search

Let the directed graph shown below be given. The descendants of the vertex c are to be determined by breadth-first search and by depth-first search. The iterative construction of the vertex sequence F for the breadth-first search and the depth-first search is shown.

directed graph



breadth-first search sequence F

c					
c	a				
c	a	b			
c	a	b	d		
	a	b	d		
		b	d		
		b	d	e	
			d	e	
			d	e	f
				e	f
					f

depth-first search sequence F

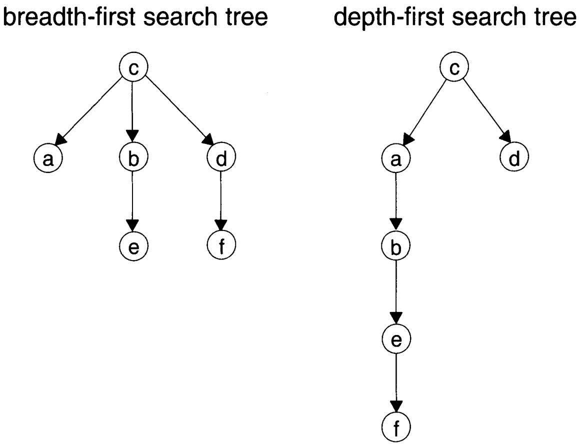
c				
c	a			
c	a	b		
c	a	b	e	
c	a	b	e	f
c	a	b	e	
c	a	b		
c	a			
c				
c	d			
c				

The vertex c is the root of the breadth-first search tree and the depth-first search tree. The breadth-first search tree is constructed according to the following rule, starting from the root c :

- If a new vertex y is appended to the end of the sequence F with start vertex x, the new vertex y and the edge from x to y are added to the breadth-first search tree.

The depth-first search tree is constructed according to the following rule, starting from the root c :

- If a new vertex y is appended to the end of the sequence F with end vertex x, the new vertex y and the edge from x to y are added to the depth-first search tree.


breadth-first search tree depth-first search tree

8.5 PATHS IN NETWORKS

8.5.1 INTRODUCTION

Network : Let a directed graph be given. Let a weight be associated with each edge of the directed graph. A directed graph with edge weights is called a weighted graph or a network. The form and meaning of the edge weights depends on the application. For example, every edge in a network may be weighted by a real number which represents a length or by a character serving as a label.

Path problem : The determination of paths with specific properties in networks is called a path problem. Different path problems can be formulated for different applications. A general distinction is made between structure problems and extreme value problems.

In structure problems, specific structural properties of paths between two vertices in a network are determined. Examples include determining the existence of paths, determining simple or elementary paths and determining common edges or intermediate vertices of all paths between two given vertices.

In extreme value problems, minimal or maximal properties of paths between two vertices in a network are determined. Examples of path problems with minimal properties include determining a shortest path between two locations in a traffic network and determining a cost-effective path between two locations in a transport network. Examples of path problems with maximal properties include determining a most reliable path between two vertices in a communication network and determining a critical path in a network schedule for a construction project.

Path : A path from i to k is an edge sequence with start vertex i and end vertex k. The path is said to be weighted if it is associated with a weight determined from the weights of its edges according to a given rule. Different path problems involve different rules for assigning weights to paths. For example, the length of a path is determined as the sum of the lengths of its edges, while the label of a path is determined as the concatenation of the labels of its edges.

Path set : A set of paths with common start vertex i and common end vertex k is called a path set. The path set is said to be weighted if it is associated with a weight determined from the weights of its paths according to a given rule. Different path problems involve different rules for assigning weights to path sets. For example, the length of a shortest path in a path set is determined as the minimum of the lengths of paths contained in the path set.

Path algebra : The union $\sqcup$ and the concatenation $\circ$ are defined as binary operations for path sets and their weights. The rules for the union and concatenation of weighted path sets are formulated such that the weights are determined directly without explicitly constructing the path sets. This leads to path algebras for networks. A path algebra is said to be either boolean, real or literal if the weights of the path sets are respectively boolean, real or literal.

The path algebras for the different path problems may be generalized by abstraction. They are conveniently formulated in matrix and vector notation. Using path algebras reduces the solution of path problems to the solution of systems of equations.

8.5.2 PATH ALGEBRA

Introduction : A directed graph consists of vertices and edges. The path algebra for directed graphs is conveniently formulated if the vertices are labeld by natural numbers and the edges by characters from an alphabet. The labels serve to identify vertices and edges uniquely.

A path in a graph is an edge sequence. Since every edge is labeled by a character, a path is labeled by a character string. A character string with characters from an alphabet is called a word over the alphabet. Every path in the graph is uniquely identified by a word. The path algebra is thereby reduced to an algebra of characters and words, called literal algebra.

A set of paths with common start vertex and common end vertex is called a path set. On the basis of the algebra of sets and the literal algebra, the binary operations of union (symbol $\sqcup$) and concatenation (symbol $\circ$) are defined for path sets.

A weight is assigned to every path set. The binary operations of union (symbol $\sqcup$) and concatenation (symbol $\circ$) are defined for the set of weights. The operations for the weights are generally different from the operations for the path sets and depend on the path problem considered.

Assigning a weight z to a path set a defines a mapping $z = f(a)$. This mapping is homomorphic if the following statements hold for the union of two path sets a, b and the concatenation of two path sets c, d :

$$\begin{aligned} f(a \sqcup b) &= f(a) \sqcup f(b) \\ f(c \circ d) &= f(c) \circ f(d) \end{aligned}$$

If the homomorphism condition is satisfied, operations on the weights of path sets may be performed directly without performing operations on the path sets themselves. The homomorphism condition is therefore of fundamental importance for a path algebra.

Alphabet and words : A finite character set is called an alphabet and is designated by $\mathbb{A}$. A finite character string with zero, one or several characters is called a word. The character string without characters is called the empty word and is designated by λ. The set of all words including the empty word λ is designated by $\mathbb{A}^*$.

Two words $a, b \in \mathbb{A}^*$ are concatenated to form a single word by appending the character string of the second word to the character string of the first word. The concatenation $\circ$ is an associative operation in the set $\mathbb{A}^*$ with the empty word λ acting as the unit element.

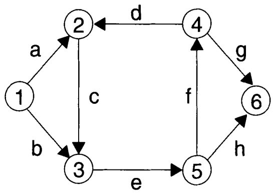
$$\begin{aligned} &\text{associative} &:&\quad a \circ (b \circ c) = (a \circ b) \circ c \\ &\text{unit element} &:&\quad a \circ \lambda = a = \lambda \circ a \end{aligned}$$

Edge and path labels : Every edge of a graph is labeled by a character from an alphabet $\mathbb{A}$. The literal labeling of the edges is said to be unique if any two different edges are labeled by different characters. If the edge labels are unique, the character for an edge also serves as an edge identifier. Edge labels are assumed to be unique in the formulation of path algebras.

Every path in a graph is an edge sequence and is labeled by a character string, which is a word in the set $\mathbb{A}^*$ of words. If the edges are labeled uniquely, then the paths are also labeled uniquely. A path without edges from a vertex k to the same vertex k is labeled by the empty word λ.

Example 1 : Literal labeling

Let the following graph with the vertices 1,...,6 and the unique edge labels a,...,h be given :



The path from vertex 1 to vertex 1 is an empty path without edges and is labeled by the empty word λ. The word a with only one character is the label of a path from 1 to 2, which consists only of the edge a. The word ace is the label of a path from 1 to 5. The word fg is the label of a path from 5 to 6. Concatenating the word ace with the word fg yields the word acefg as the label for a path from 1 via 5 to 6.

$$u = ace \qquad v = fg \qquad u \circ v = ace \circ fg = acefg$$

Path set : Let a directed graph with unique edge labels be given. A set of paths for a vertex pair (i, k) in the graph is called a complete path set and is designated by W_{ik} if it contains all paths from vertex i to vertex k. A subset a_{ik} of the complete path set W_{ik} is called a path set. The set of all possible subsets a_{ik} of W_{ik} is the power set of the complete path set and is designated by $P(W_{ik})$. Every path set a_{ik} is an element of the power set $P(W_{ik})$, that is $a_{ik} \in P(W_{ik})$.

The zero set, the unit set and the elementary path set are special path sets. The path set which contains no path is called the zero set and is designated by $0_W = \{\,\}$. The path set which contains only the empty path without edges from a vertex i to the same vertex i is called the unit set and is designated by $1_W = \{\lambda\}$. A path set a_{ik} is said to be elementary if it contains exactly one path which consists only of the edge from vertex i to vertex k.

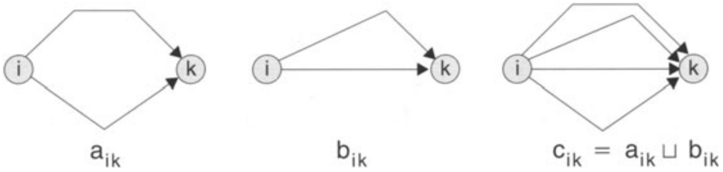
zero set	0_W	=	$\{\,\}$
unit set	1_W	=	$\{\lambda\}$
elementary path set	a_{ik}	=	$\{<i,k>\}$

Path set matrix : Let a directed graph with n vertices be given. The path sets for all vertex pairs of the graph are arranged in an n × n matrix. An n × n matrix is called a complete path set matrix and is designated by **W** if it contains the complete path set W_{ik} for every vertex pair (i, k) of the graph. An n × n matrix is called a path set matrix **A** with $\mathbf{A} \subseteq \mathbf{W}$ if it contains a path set $a_{ik} \subseteq W_{ik}$ for every vertex pair (i, k) in the graph. The set of all possible path set matrices $\mathbf{A} \subseteq \mathbf{W}$ is called the power set of the complete path set matrix and is designated by $P(\mathbf{W})$. A path set matrix **A** is an element of the power set $P(\mathbf{W})$, that is $\mathbf{A} \in P(\mathbf{W})$.

The zero matrix, the identity matrix and the elementary path set matrix are special path set matrices. A path set matrix is called a zero matrix and is designated by $\mathbf{0}_W$ if it contains the zero set 0_W for every vertex pair (i, k). A path set matrix is called an identity matrix and is designated by $\mathbf{I}_W$ if it contains the unit set 1_W for every vertex pair (k, k) and the zero set 0_W for all remaining vertex pairs. A path set matrix is said to be elementary if it contains the elementary path set a_{ik} for every vertex pair (i, k) with an edge from vertex i to vertex k and the zero set 0_W for all remaining vertex pairs. A directed graph with unique edge labels is uniquely described by the elementary path set matrix.

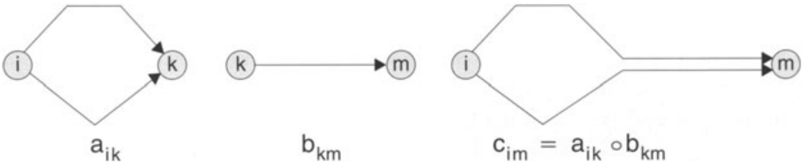
Operations on path sets : Let the path sets $a_{ik} \in P(W_{ik})$ and $b_{ik} \in P(W_{ik})$ be given. The path set $c_{ik} \in P(W_{ik})$ which contains all paths which are contained in a_{ik} or in b_{ik} is called the union of a_{ik} and b_{ik}.

union : $c_{ik} = a_{ik} \sqcup b_{ik} := \{x \mid x \in a_{ik} \ \vee \ x \in b_{ik}\}$



Let the path sets $a_{ik} \in P(W_{ik})$ and $b_{km} \in P(W_{km})$ be given. The path set $c_{im} \in P(W_{im})$ which contains the paths which are formed by concatenating a path $x \in a_{ik}$ and a path $y \in b_{km}$ is called the concatenation of a_{ik} and b_{km}.

concatenation : $c_{im} = a_{ik} \circ b_{km} := \{x \circ y \mid x \in a_{ik} \ \wedge \ y \in b_{km}\}$



Algebraic structure of path sets : The operations on path sets have the following properties :

(1) The union $\sqcup$ is idempotent, since by the definition of the union the statement $a_{ik} \sqcup a_{ik} = a_{ik}$ holds.

(2) The union $\sqcup$ is associative, since by the definition of the union the statement $a_{ik} \sqcup (b_{ik} \sqcup c_{ik}) = (a_{ik} \sqcup b_{ik}) \sqcup c_{ik}$ holds.

(3) The union $\sqcup$ is commutative, since by the definition of the union the statement $a_{ik} \sqcup b_{ik} = b_{ik} \sqcup a_{ik}$ holds.

(4) The union $\sqcup$ has the zero set $0_W = \{\ \}$ as an identity element, since by the definition of the union $a_{ik} \sqcup 0_W = \{x \mid x \in a_{ik} \vee x \in \{\ \}\} = \{x \mid x \in a_{ik}\} = a_{ik}$ and analogously $0_W \sqcup a_{ik} = a_{ik}$.

(5) The concatenation $\circ$ is associative, since by the definition of the concatenation the statement $a_{ik} \circ (b_{km} \circ c_{mj}) = (a_{ik} \circ b_{km}) \circ c_{mj}$ holds.

(6) The concatenation $\circ$ has the unit set $1_W = \{\lambda\}$ as an identity element, since by the definition of the concatenation $a_{ik} \circ 1_W = \{x \circ y \mid x \in a_{ik} \wedge y \in \{\lambda\}\} = \{x \circ \lambda \mid x \in a_{ik}\} = \{x \mid x \in a_{ik}\} = a_{ik}$ and analogously $1_W \circ a_{ik} = a_{ik}$.

(7) The concatenation $\circ$ has the zero set $0_W = \{\ \}$ as an invariant element, since by the definition of the concatenation $a_{ik} \circ 0_W = \{x \circ y \mid x \in a_{ik} \wedge y \in \{\ \}\} = \{x \circ y \mid x \in a_{ik} \wedge \text{false}\} = \{x \circ y \mid \text{false}\} = \{\} = 0_W$ and likewise $0_W \circ a_{ik} = 0_W$.

(8) The concatenation $\circ$ is distributive with respect to the union, since by their definitions the statements $a_{ik} \circ (b_{km} \sqcup c_{km}) = (a_{ik} \circ b_{km}) \sqcup (a_{ik} \circ c_{km})$ and $(a_{ik} \sqcup b_{ik}) \circ c_{km} = (a_{ik} \circ c_{km}) \sqcup (b_{ik} \circ c_{km})$ hold.

Operations on path set matrices : Let the path set matrices $\mathbf{A}, \mathbf{B} \in P(\mathbf{W})$ for a directed graph with n vertices be given. In analogy with the algebra of relations, the binary operations of union (symbol $\sqcup$) and concatenation (symbol $\circ$) are defined. The operations $\sqcup$ and $\circ$ already defined for path sets are used for the matrix elements.

union : $\mathbf{C} = \mathbf{A} \sqcup \mathbf{B} := [a_{ik} \sqcup b_{ik}]$

concatenation : $\mathbf{C} = \mathbf{A} \circ \mathbf{B} := [\bigsqcup_{m=1}^{n} (a_{im} \circ b_{mk})]$

For every vertex pair (i, k), the path set matrix $\mathbf{A} \sqcup \mathbf{B}$ contains the paths which are contained in the path set a_{ik} or in the path set b_{ik}. For every vertex pair (i, k), the path set matrix $\mathbf{A} \circ \mathbf{B}$ contains the paths formed by concatenating all paths in a_{im} with all paths in b_{mk} for all vertices m. The concatenation of two path set matrices is also called their product.

Algebraic structure of path set matrices : The algebraic structure of path sets is directly transferred to path set matrices. The domain ($P(\mathbf{W})\,;\sqcup,\circ$) with the power set $P(\mathbf{W})$ of the complete path set matrix and the binary operations $\sqcup$ and $\circ$ is a special semiring as described in Example 3 in Section 3.4.2; it is called a path algebra. It has the following properties for the path set matrices $\mathbf{A}, \mathbf{B}, \mathbf{C} \in P(\mathbf{W})$:

Property	Union $\sqcup$	Concatenation $\circ$
idempotent	$\mathbf{A} \sqcup \mathbf{A} = \mathbf{A}$	
associative	$\mathbf{A} \sqcup (\mathbf{B} \sqcup \mathbf{C}) = (\mathbf{A} \sqcup \mathbf{B}) \sqcup \mathbf{C}$	$\mathbf{A} \circ (\mathbf{B} \circ \mathbf{C}) = (\mathbf{A} \circ \mathbf{B}) \circ \mathbf{C}$
distributive	$\mathbf{A} \circ (\mathbf{B} \sqcup \mathbf{C}) = (\mathbf{A} \circ \mathbf{B}) \sqcup (\mathbf{A} \circ \mathbf{C})$	$(\mathbf{A} \sqcup \mathbf{B}) \circ \mathbf{C} = (\mathbf{A} \circ \mathbf{C}) \sqcup (\mathbf{B} \circ \mathbf{C})$
commutative	$\mathbf{A} \sqcup \mathbf{B} = \mathbf{B} \sqcup \mathbf{A}$	
zero element	$\mathbf{0}_W \sqcup \mathbf{A} = \mathbf{A} = \mathbf{A} \sqcup \mathbf{0}_W$	$\mathbf{0}_W \circ \mathbf{A} = \mathbf{0}_W = \mathbf{A} \circ \mathbf{0}_W$
unit element		$\mathbf{1}_W \circ \mathbf{A} = \mathbf{A} = \mathbf{A} \circ \mathbf{1}_W$

Closure of the elementary path set matrix : Let an elementary path set matrix $\mathbf{A}$ for a directed graph with n vertices be given. In analogy with the algebra of relations, the closure $\mathbf{A}^*$ is defined as the union of the powers $\mathbf{A}^m$ with $m \geq 0$. For every vertex pair (i, k), the power $\mathbf{A}^m$ contains all paths which lead from vertex i to vertex k and consist of exactly m edges. The power $\mathbf{A}^0$ is the identity matrix $\mathbf{I}_W$. For every vertex pair (i, k), the closure $\mathbf{A}^*$ contains all paths which lead from vertex i to vertex k. It therefore coincides with the complete path set matrix $\mathbf{W}$.

$$\mathbf{A}^* := \mathbf{I}_W \sqcup \mathbf{A} \sqcup \mathbf{A}^2 \sqcup \mathbf{A}^3 \sqcup \ldots = \mathbf{W}$$

If the power expression for the closure $\mathbf{A}^*$ does not change beyond a certain finite exponent q, the path set matrix $\mathbf{A}$ is said to be stable and the exponent q is called its stability index. For every vertex pair (i, k), the closure $\mathbf{A}^*$ of a stable path set matrix $\mathbf{A}$ with stability index q contains all paths which lead from vertex i to vertex k and consist of at most q edges. The elementary path set matrix for an acyclic graph with n vertices is stable with a stability index $q < n$, since a path in this graph consists of at most $n - 1$ edges. The elementary path set matrix of a graph containing a cycle is not stable, since a path in this graph can traverse the cycle an arbitrary number of times and may hence consist of an arbitrary number of edges.

Systems of equations for path sets : For a given vertex k, the path sets whose paths lead from each of the vertices i = 1,...,n to k may be read off in column k of the closure $\mathbf{A}^*$. The k-th column of the closure $\mathbf{A}^*$ is designated by **x**, the unit vector with the unit set 1_W in row k by $\mathbf{e}_k$. If the closure $\mathbf{A}^*$ is known, then **x** is calculated as follows :

$$\mathbf{x} \;=\; \mathbf{A}^* \circ \mathbf{e}_k$$

By substituting the calculational rule for the closure $\mathbf{A}^*$, the following relationship between the elementary path set matrix **A** and the vector **x** is obtained :

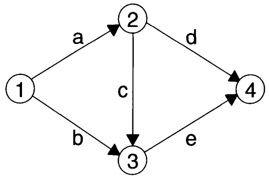
$$\begin{aligned} \mathbf{A}^* &= \mathbf{I}_W \sqcup \mathbf{A} \sqcup \mathbf{A}^2 \sqcup \ldots \;\Rightarrow\; \mathbf{A}^* = \mathbf{A} \circ \mathbf{A}^* \sqcup \mathbf{I}_W \\ \mathbf{x} &= (\mathbf{A} \circ \mathbf{A}^* \sqcup \mathbf{I}_W) \circ \mathbf{e}_k = \mathbf{A} \circ (\mathbf{A}^* \circ \mathbf{e}_k) \sqcup (\mathbf{I}_W \circ \mathbf{e}_k) \\ \mathbf{x} &= \mathbf{A} \circ \mathbf{x} \sqcup \mathbf{e}_k \end{aligned}$$

For a given vertex i, the path sets whose paths lead from i to each of the vertices k = 1,...,n may be read off in row i of the closure $\mathbf{A}^*$. The transpose of row i of the closure $\mathbf{A}^*$ is designated by **y**, the unit vector with the unit set 1_W in row i by $\mathbf{e}_i$. In analogy with the result for column k of $\mathbf{A}^*$, row i satisfies the following equation :

$$\mathbf{y} \;=\; \mathbf{A}^T \circ \mathbf{y} \sqcup \mathbf{e}_i$$

Example 2 : Path set matrices and operations

Let a directed acyclic graph with the vertices 1,...,4 and the edges a,...,e be given. To simplify the notation for path sets, in this example the zero set $0_W = \{\,\}$ is designated by 0 and the unit set $1_W = \{\lambda\}$ by 1.



The elementary path set matrix **A** for the directed graph is constructed as follows. If there is no edge from vertex i to vertex k, then the path set a_{ik} is the zero set. If there is an edge x from vertex i to vertex k, then the path set $a_{ik} = \{x\}$ is elementary.

A =

0	{a}	{b}	0
0	0	{c}	{d}
0	0	0	{e}
0	0	0	0

For every vertex pair (i, k), the power $\mathbf{A}^m$ with $m > 0$ contains all paths which lead from vertex i to vertex k and consist of exactly m edges. It is calculated as the product $\mathbf{A}^{m-1} \circ \mathbf{A}$ according to the rules for the concatenation of path set matrices. The calculation of the powers $\mathbf{A}^2$ and $\mathbf{A}^3$ is shown.

$$\begin{array}{ccc}
 & \mathbf{A}\ \begin{array}{|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 \\ \hline 0 & 0 & \{c\} & \{d\} \\ \hline 0 & 0 & 0 & \{e\} \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} & \begin{array}{|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 \\ \hline 0 & 0 & \{c\} & \{d\} \\ \hline 0 & 0 & 0 & \{e\} \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} \\
\begin{array}{|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 \\ \hline 0 & 0 & \{c\} & \{d\} \\ \hline 0 & 0 & 0 & \{e\} \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} & \begin{array}{|c|c|c|c|}\hline 0 & 0 & \{ac\} & \{ad,be\} \\ \hline 0 & 0 & 0 & \{ce\} \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} & \begin{array}{|c|c|c|c|}\hline 0 & 0 & 0 & \{ace\} \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} \\
\mathbf{A} & \mathbf{A}^2 & \mathbf{A}^3
\end{array}$$

The power $\mathbf{A}^4$ is the zero matrix $\mathbf{0}_W$. Hence the path set matrix $\mathbf{A}$ is stable, and its stability index is $q = 3$. The closure $\mathbf{A}^*$ is the union of the powers $\mathbf{A}^0 = \mathbf{I}_W$, $\mathbf{A}^1 = \mathbf{A}$, $\mathbf{A}^2$, $\mathbf{A}^3$. For every vertex pair (i, k), it contains the set of all paths from vertex i to vertex k; hence it coincides with the complete path set matrix $\mathbf{W}$.

$$\mathbf{A}^* \quad = \quad \begin{array}{|c|c|c|c|}\hline 1 & \{a\} & \{b,ac\} & \{ad,be,ace\} \\ \hline 0 & 1 & \{c\} & \{d,ce\} \\ \hline 0 & 0 & 1 & \{e\} \\ \hline 0 & 0 & 0 & 1 \\ \hline \end{array} \quad = \quad \mathbf{I} \sqcup \mathbf{A} \sqcup \mathbf{A}^2 \sqcup \mathbf{A}^3 \quad = \quad \mathbf{W}$$

The paths from each of the vertices $k = 1,...,4$ to the vertex 4 are read off directly from the elements a_{k4}^* in column 4 of the closure $\mathbf{A}^*$. They may also be determined as solutions of the system of equations $\mathbf{x} = \mathbf{A} \circ \mathbf{x} \sqcup \mathbf{e}_4$. The variables x_k in the vector $\mathbf{x}$ correspond to the path sets a_{k4}^* of the closure $\mathbf{A}^*$. The vector $\mathbf{e}_4$ is the unit vector with the unit set 1 in row 4.

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline \end{array} \quad = \quad \begin{array}{|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 \\ \hline 0 & 0 & \{c\} & \{d\} \\ \hline 0 & 0 & 0 & \{e\} \\ \hline 0 & 0 & 0 & 0 \\ \hline \end{array} \quad \circ \quad \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline \end{array} \quad \sqcup \quad \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline \end{array}$$

In this example, the path set matrix **A** is an upper triangular matrix whose non-zero elements lie above the diagonal. The system of equations may therefore be solved by back substitution. The back substitution begins in row 4 and ends in row 1 of the system of equations. Since $0 \circ a = 0$ and $0 \sqcup a = a$, the zero operations are not explicitly shown. The concatenation $\circ$ takes precedence over the union $\sqcup$.

$$\begin{array}{lllll}
x_4 & = 1 & & & = 1 \\
x_3 & = \{e\} \circ x_4 & = \{e\} \circ 1 & & = \{e\} \\
x_2 & = \{d\} \circ x_4 \sqcup \{c\} \circ x_3 & = \{d\} \circ 1 & \sqcup\ \{c\} \circ \{e\} & = \{d, ce\} \\
x_1 & = \{b\} \circ x_3 \sqcup \{a\} \circ x_2 & = \{b\} \circ \{e\} & \sqcup\ \{a\} \circ \{d, ce\} & = \{be, ad, ace\}
\end{array}$$

Weighted path set : Every path set $a_{ik} \in P(W_{ik})$ is assigned a unique weight $z_{ik} \in Z$ from a weight set Z. The zero set 0_W is assigned the zero element 0_Z, and the unit set 1_W is assigned the unit element 1_Z. Associating the path set a_{ik} with the weight z_{ik} defines a mapping.

$$\begin{array}{lrcl}
\text{mapping} & f(a_{ik}) & = & z_{ik} \in Z \\
\text{zero element} & f(0_W) & = & 0_Z \in Z \\
\text{unit element} & f(1_W) & = & 1_Z \in Z
\end{array}$$

As for path sets, the binary operations $\sqcup$ (union) and $\circ$ (concatenation) are defined for the weights. The operations for weights and the operations for path sets are generally different. The mapping f is said to be homomorphic if the following statements hold :

$$\begin{array}{rcl}
f(a_{ik} \sqcup b_{ik}) & = & f(a_{ik}) \sqcup f(b_{ik}) \\
f(a_{ik} \circ b_{km}) & = & f(a_{ik}) \circ f(b_{km})
\end{array}$$

If the mapping f is homomorphic, the weights of path sets may be determined without explicitly determining the path sets, since the following implications hold for $x_{ij} = f(a_{ij})$, $y_{ij} = f(b_{ij})$ and $z_{ij} = f(c_{ij})$:

$$\begin{array}{lcl}
c_{ik} = a_{ik} \sqcup b_{ik} & \Rightarrow & z_{ik} = x_{ik} \sqcup y_{ik} \\
c_{im} = a_{ik} \circ b_{km} & \Rightarrow & z_{im} = x_{ik} \circ y_{km}
\end{array}$$

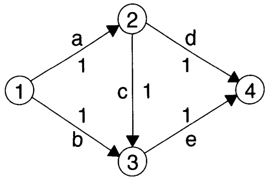
Using these implications, the properties of path sets with the operations $\sqcup$ and $\circ$ may be transferred to the properties of weights with the operations $\sqcup$ and $\circ$.

Algebraic structure of weighted path sets : Let the path sets of a directed graph be homomorphically mapped to weights. Then the domain $(Z\,;\sqcup,\circ)$ with the weight set Z and the binary operations $\sqcup$ and $\circ$ is a path algebra. It has the following properties for the elements $x, y, z \in Z$:

Property	Union $\sqcup$	Concatenation $\circ$
idempotent	$x \sqcup x \;=\; x$	
associative	$(x \sqcup y) \sqcup z \;=\; x \sqcup (y \sqcup z)$	$(x \circ y) \circ z \;=\; x \circ (y \circ z)$
distributive	$x \circ (y \sqcup z) \;=\; (x \circ y) \sqcup (x \circ z)$	$(x \sqcup y) \circ z \;=\; (x \circ z) \sqcup (y \circ z)$
commutative	$x \sqcup y \;=\; y \sqcup x$	
zero element	$0_Z \sqcup x \;=\; x \;=\; x \sqcup 0_Z$	$0_Z \circ x \;=\; 0_Z \;=\; x \circ 0_Z$
unit element		$1_Z \circ x \;=\; x \;=\; x \circ 1_Z$

Example 3 : Path sets and weights

Let the directed acyclic graph from Example 2 be given. Let the length of a path from vertex i to vertex k be the number of its edges. For every vertex pair (i, k) of the graph, the minimal length of a path from i to k is to be determined. A path algebra is to be formulated for this path problem.



A path is designated by a word which contains the characters for the edges. The length of the path is the number of edges, and hence the number of characters in the word. A path set a_{ik} contains paths from i to k. The minimum of the lengths of all paths contained in a_{ik} is taken as the weight of the path set a_{ik}.

If a_{ik} is the zero set 0_W, then it contains no path from i to k. The weight of the zero set 0_W is taken to be the zero element $0_Z = \infty$. This corresponds to the fact that the zero set 0_W contains no path with a finite number edges. If a_{ik} for $i = k$ is the unit set 1_W, then the path set contains only the empty path λ without edges. The weight of the unit set 1_W is the unit element $1_Z = 0$.

If a_{ik} is an elementary path set, then it contains exactly one path with the edge from i to k. Its weight is therefore the natural number 1. This weight for an elementary path set a_{ik} corresponds to the weight of the edge from i to k. In the graph shown above, the weight of every edge is marked as 1.

A general path set with at least one non-empty path is assigned a natural number as a weight. With these assignments, the weight mapping f is defined as follows :

$$\begin{array}{lclll} f(0_W) & = & 0_Z & := \infty & \text{zero element} \\ f(1_W) & = & 1_Z & := 0 & \text{unit element} \\ f(a_{ik}) & = & 1 & \in \mathbb{N}' & a_{ik} \text{ is elementary} \\ f(a_{ik}) & = & z_{ik} & \in \mathbb{N}' & a_{ik} \notin \{0_W, 1_W\} \end{array}$$

Let the path sets a_{ik} and b_{ik} with the weights $x_{ik} = f(a_{ik})$ and $y_{ik} = f(b_{ik})$ for the minimal path lengths be given. By the homomorphism condition the union $a_{ik} \sqcup b_{ik}$ of the path sets leads to the union $x_{ik} \sqcup y_{ik}$ of the weights. The minimal path length of the union $a_{ik} \sqcup b_{ik}$ is the minimum $\min\{x_{ik}, y_{ik}\}$ of the minimal path lengths x_{ik} and y_{ik}. Thus the union of the weights is defined as follows :

$$x_{ik} \sqcup y_{ik} := \min\{x_{ik}, y_{ik}\}$$

Let the path sets a_{ik} and b_{km} with the weights $x_{ik} = f(a_{ik})$ and $y_{km} = f(b_{km})$ for the minimal path lengths be given. By the homomorphism condition the concatenation $a_{ik} \circ b_{km}$ of the path sets leads to the concatenation $x_{ik} \circ y_{km}$ of the weights. The minimal path length of the concatenation $a_{ik} \circ b_{km}$ is the sum $x_{ik} + y_{km}$ of the minimal path lengths x_{ik} and y_{km}. Thus the concatenation of the weights is defined as follows :

$$x_{ik} \circ y_{km} := x_{ik} + y_{km}$$

The weight set Z for the minimal path lengths is the set of natural numbers $\mathbb{N}'$ augmented by the zero element $0_Z = \infty$ and the unit element $1_Z = 0$. Thus $Z = \mathbb{N}' \cup \{0\} \cup \{\infty\}$. For the weight set Z, the union $\sqcup$ is defined as the minimum operation and the concatenation $\circ$ is defined as the sum operation. The domain $(Z ; \sqcup, \circ)$ is a path algebra, since for $x, y, z \in Z$ it has the required properties :

(1) The union $\sqcup$ is idempotent, associative and commutative :

$$\begin{array}{l} x \sqcup x = \min\{x, x\} = x \\ (x \sqcup y) \sqcup z = \min\{\min\{x, y\}, z\} = \min\{x, \min\{y, z\}\} = x \sqcup (y \sqcup z) \\ x \sqcup y = \min\{x, y\} = \min\{y, x\} = y \sqcup x \end{array}$$

(2) The concatenation $\circ$ is associative :

$$(x \circ y) \circ z = (x + y) + z = x + (y + z) = x \circ (y \circ z)$$

(3) The union $\sqcup$ and the concatenation $\circ$ are distributive :

$$\begin{array}{l} x \circ (y \sqcup z) = x + \min\{y, z\} = \min\{x + y, x + z\} = (x \circ y) \sqcup (x \circ z) \\ (x \sqcup y) \circ z = \min\{x, y\} + z = \min\{x + z, y + z\} = (x \circ z) \sqcup (y \circ z) \end{array}$$

(4) The zero element $0_Z = \infty$ is the identity element for the union $\sqcup$ and the invariant element for the concatenation $\circ$:

$$\begin{aligned} \infty \sqcup x &= \min\{\infty, x\} = x = \min\{x, \infty\} = x \sqcup \infty \\ \infty \circ x &= \infty + x = \infty = x + \infty = x \circ \infty \end{aligned}$$

(5) The unit element $1_Z = 0$ is the identity element for the concatenation $\circ$:

$$0 \circ x = 0 + x = x = x + 0 = x \circ 0$$

Weight matrix : Let a directed graph with n vertices, a weight set Z and a homomorphic mapping f be given. Then every path set matrix **A** may be mapped homomorphically to a weight matrix **Z**. Every path set a_{ik} of **A** is mapped to the weight $z_{ik} = f(a_{ik}) \in Z$ of **Z**. As in the case of path set matrices, the zero matrix $\mathbf{0}_Z$, the identity matrix $\mathbf{I}_Z$ and the elementary weight matrix are special weight matrices.

Operations on weight matrices : Let weight matrices **X**, **Y** for the path set matrices **A**, **B** of a directed graph be given. As in the case of path set matrices, the binary operations of union (symbol $\sqcup$) and concatenation (symbol $\circ$) are defined by applying the operations $\sqcup$ and $\circ$ defined for weights to the matrix elements.

$$\begin{aligned} &\text{union} &&: \quad \mathbf{Z} = \mathbf{X} \sqcup \mathbf{Y} := [x_{ik} \sqcup y_{ik}] \\ &\text{concatenation} &&: \quad \mathbf{Z} = \mathbf{X} \circ \mathbf{Y} := [\bigsqcup_{m=1}^{n} (x_{im} \circ y_{mk})] \end{aligned}$$

Algebraic structure of weight matrices : Since path set matrices are homomorphically mapped to weight matrices, the algebraic structures of path set matrices and weight matrices and their operations $\sqcup$ and $\circ$ are compatible. In the set of all possible weight matrices for a directed graph, the zero matrix $\mathbf{0}_Z$ is the identity element for the union and the identity matrix $\mathbf{I}_Z$ is the identity element for the concatenation.

Closure of the elementary weight matrix : Let an elementary path set matrix **A** for a directed graph with n vertices be given. The elementary path set matrix **A** is assigned the elementary weight matrix **Z**, which contains the weights of the edges of the graph. Since the weighting is a homomorphic mapping, the closure $\mathbf{Z}^*$ of the weight matrix may be determined directly from the union of the powers of **Z** without explicitly calculating $\mathbf{A}^*$. For every vertex pair (i, k), the closure $\mathbf{Z}^*$ contains the weight for the set of all paths which lead from vertex i to vertex k.

$$\mathbf{Z}^* = \mathbf{I}_Z \sqcup \mathbf{Z} \sqcup \mathbf{Z}^2 \sqcup \mathbf{Z}^3 \sqcup \ldots$$

If the power expression for the closure $\mathbf{Z}^*$ does not change beyond a certain exponent s, then the weight matrix **Z** is said to be stable and the exponent s is called its stability index. The weight matrix **Z** may be stable even if the path set matrix **A** is not stable.

Systems of equations for weights : For a given vertex k, the weights of the path sets whose paths lead from each of the vertices i = 1,...,n to k may be read off in column k of the closure $\mathbf{Z}^*$. Column k of the closure $\mathbf{Z}^*$ is designated by **x**, the unit vector with the unit element 1_Z in row k by $\mathbf{e}_k$. If the closure $\mathbf{Z}^*$ is known, then **x** is calculated as follows :

$$\mathbf{x} = \mathbf{Z}^* \circ \mathbf{e}_k$$

By analogy with the equations for path sets, the vector **x** is the solution of the following system of equations :

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_k$$

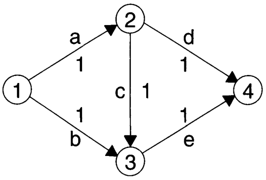
For a given vertex i, the path sets whose paths lead from i to each of the vertices k = 1,...,n may be read off in row i of the closure $\mathbf{Z}^*$. The transpose of row i of the closure $\mathbf{Z}^*$ is designated by **y**, the unit vector with the unit element 1_Z in row i by $\mathbf{e}_i$. By analogy with the result for column k of $\mathbf{Z}^*$, row i satisfies :

$$\mathbf{y} = \mathbf{Z}^T \circ \mathbf{y} \sqcup \mathbf{e}_i$$

General methods for the solution of systems of equations in a path algebra are treated in Section 8.5.7.

Example 4 : Weight matrices and operations

Let the directed acyclic graph from Example 2 with the weights for the minimal path lengths from Example 3 be given. To simplify the notation for the weights of path sets, the zero element $0_Z = \infty$ is designated by n and the unit element $1_Z = 0$ by e in this example.



The elementary weight matrix **Z** for the directed graph is constructed as follows. If there is no edge from vertex i to vertex k, then the path set a_{ik} is the zero set and is weighted with the zero element $z_{ik} = n$. If there is an edge from vertex i to vertex k, then the path set $a_{ik} = \{x\}$ is elementary and is weighted with $z_{ik} = 1$.

$$\mathbf{Z} = \begin{array}{|c|c|c|c|} \hline n & 1 & 1 & n \\ \hline n & n & 1 & 1 \\ \hline n & n & n & 1 \\ \hline n & n & n & n \\ \hline \end{array}$$

For every vertex pair (i, k), the power $\mathbf{Z}^m$ with $m > 0$ contains the path length m if a path of length m exists, and the zero element n otherwise. It is calculated as the product $\mathbf{Z}^{m-1} \circ \mathbf{Z}$ according to the rules for the concatenation of weight matrices, using the definitions for the operations of union, $x_{ik} \sqcup y_{ik} := \min\{x_{ik}, y_{ik}\}$, and of concatenation, $x_{ik} \circ y_{km} := x_{ik} + y_{km}$. The calculation of the powers $\mathbf{Z}^2$ and $\mathbf{Z}^3$ is shown.

$$\begin{array}{ccc}
 & \mathbf{Z}\ \begin{array}{|c|c|c|c|}\hline n & 1 & 1 & n \\ \hline n & n & 1 & 1 \\ \hline n & n & n & 1 \\ \hline n & n & n & n \\ \hline\end{array} & \begin{array}{|c|c|c|c|}\hline n & 1 & 1 & n \\ \hline n & n & 1 & 1 \\ \hline n & n & n & 1 \\ \hline n & n & n & n \\ \hline\end{array} \\
\begin{array}{|c|c|c|c|}\hline n & 1 & 1 & n \\ \hline n & n & 1 & 1 \\ \hline n & n & n & 1 \\ \hline n & n & n & n \\ \hline\end{array} & \begin{array}{|c|c|c|c|}\hline n & n & 2 & 2 \\ \hline n & n & n & 2 \\ \hline n & n & n & n \\ \hline n & n & n & n \\ \hline\end{array} & \begin{array}{|c|c|c|c|}\hline n & n & n & 3 \\ \hline n & n & n & n \\ \hline n & n & n & n \\ \hline n & n & n & n \\ \hline\end{array} \\
\mathbf{Z} & \mathbf{Z}^2 & \mathbf{Z}^3
\end{array}$$

The power $\mathbf{Z}^4$ is the zero matrix $\mathbf{0}_Z$. The weight matrix $\mathbf{Z}$ is therefore stable, and its stability index is $q = 3$. The closure $\mathbf{Z}^*$ is the union of the powers $\mathbf{Z}^0 = \mathbf{I}_Z$, $\mathbf{Z}^1 = \mathbf{Z}$, $\mathbf{Z}^2$, $\mathbf{Z}^3$. The union corresponds to a minimum operation. Hence for every vertex pair (i, k) the closure $\mathbf{Z}^*$ contains the minimal path length z^*_{ik} of a path from vertex i to vertex k. If there is no path from vertex i to vertex k, then z^*_{ik} is the zero element n.

$$\mathbf{Z}^* = \begin{array}{|c|c|c|c|}\hline e & 1 & 1 & 2 \\ \hline n & e & 1 & 1 \\ \hline n & n & e & 1 \\ \hline n & n & n & e \\ \hline\end{array} = \mathbf{I}_Z \sqcup \mathbf{Z} \sqcup \mathbf{Z}^2 \sqcup \mathbf{Z}^3$$

The minimal path lengths from each of the vertices $k = 1,\dots,4$ to the vertex 4 are read off directly from the elements z^*_{k4} in column 4 of the closure $\mathbf{Z}^*$. They may also be determined directly as solutions of the system of equations $\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_4$. The variables x_k of the vector $\mathbf{x}$ correspond to the minimal path lengths z^*_{k4} of the closure $\mathbf{Z}^*$. The vector $\mathbf{e}_4$ is the unit vector with the unit element e in row 4.

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline\end{array} = \begin{array}{|c|c|c|c|}\hline n & 1 & 1 & n \\ \hline n & n & 1 & 1 \\ \hline n & n & n & 1 \\ \hline n & n & n & n \\ \hline\end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline\end{array} \sqcup \begin{array}{|c|}\hline n \\ \hline n \\ \hline n \\ \hline e \\ \hline\end{array} \qquad \begin{array}{l} n := 0_Z = \infty \\ e := 1_Z = 0 \end{array}$$

In this example, the weight matrix **Z** is an upper triangular matrix whose non-zero elements all lie above the diagonal. The system of equations may therefore be solved by back substitution. The back substitution begins in row 4 and ends in row 1 of the system of equations. Since $n \circ z = n$ and $n \sqcup z = z$, the zero operations are not explicitly shown. The concatenation $\circ$ takes precedence over the union $\sqcup$.

$$\begin{array}{llllllll}
x_4 & = & e & & & & = & 0 \\
x_3 & = & 1 \circ x_4 & = & 1 + x_4 & = 1 + 0 & = & 1 \\
x_2 & = & 1 \circ x_4 \sqcup 1 \circ x_3 & = & \min\{1 + x_4, 1 + x_3\} & = \min\{1,2\} & = & 1 \\
x_1 & = & 1 \circ x_3 \sqcup 1 \circ x_2 & = & \min\{1 + x_3, 1 + x_2\} & = \min\{2,2\} & = & 2
\end{array}$$

The minimal path length from vertex 1 to vertex 4 is $x_1 = z^*_{14} = 2$. The graph contains two paths, ad and be, with the minimal path length 2. If the labels of the paths of minimal length are to be determined, a suitable literal path algebra needs to be formulated (see Section 8.5.5).

Path algebras for graphs : Different path algebras may be schematically constructed for different path problems in networks. The construction is carried out in the following steps :

1. specify the path problem
2. define the weight set Z for path sets
3. define the operations $\sqcup$ and $\circ$ on the weight set Z
4. prove the properties for the domain $(Z ; \sqcup , \circ)$ of the weights
5. study the stability of the elementary weight matrix **Z**

The weights may be boolean values, real numbers or literals. Accordingly, boolean, real and literal path algebras are distinguished. In the following sections they are treated for different path problems according to the construction scheme described above.

8.5.3 BOOLEAN PATH ALGEBRA

Problem : Let a directed graph be given. For every vertex pair (i, k) of the graph, it is to be determined whether there is at least one path from i to k. This problem is solved in Section 8.4.2 using the algebra of relations by iteratively determining all descendants of the vertex i. The same problem is now solved by weighting the path sets with the boolean path existence values {0, 1}. The path algebra for determining the existence of paths is called a boolean path algebra.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. A boolean value of 0 or 1 describing path existence is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = 0$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = 1$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then $z_{ik} = 1$. Thus the weight mapping is defined as follows :

$$\begin{aligned} f(0_W) &= 0_Z &&= 0 \\ f(1_W) &= 1_Z &&= 1 \\ f(a_{ik}) &= z_{ik} &&= 1 \qquad \text{for} \qquad a_{ik} \notin \{0_W, 1_W\} \end{aligned}$$

Operations : The operations $\sqcup$ and $\circ$ are defined for the weight set $Z = \{0, 1\}$. Let the path sets a_{ik}, b_{ik} be weighted with the path existences x_{ik}, y_{ik}. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the logical disjunction $x_{ik} \vee y_{ik}$. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the logical conjunction $x_{ik} \wedge y_{km}$.

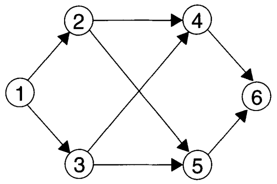
$$\begin{aligned} &\text{union} && : \quad x_{ik} \sqcup y_{ik} := x_{ik} \vee y_{ik} \\ &\text{concatenation} && : \quad x_{ik} \circ y_{km} := x_{ik} \wedge y_{km} \end{aligned}$$

The domain $(\{0, 1\} ; \sqcup, \circ)$ is a boolean path algebra with the zero element 0 and the unit element 1. The operations $\sqcup$ and $\circ$ possess the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix **Z** is 1. Otherwise $z_{ik} = 0$. The matrix **Z** is stable. If the graph contains a path from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is 1. Otherwise $z^*_{ik} = 0$. The closure $\mathbf{Z}^*$ contains the diagonal elements $z^*_{kk} = 1$. The elementary weight matrix **Z** is identical with a boolean matrix **R**, and the closure $\mathbf{Z}^*$ is identical with the closure $\mathbf{R}^*$ of the algebra of relations.

Example : Existence of paths

Let the directed graph illustrated below with the elementary weight matrix **Z** and the corresponding closure $\mathbf{Z}^*$ be given. All vertices of the graph which are the start vertex of a path to the vertex 5 are to be determined.



$$\mathbf{Z} = \begin{array}{c|cccccc} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 0 & 1 & 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 & 1 & 1 & 0 \\ 3 & 0 & 0 & 0 & 1 & 1 & 0 \\ 4 & 0 & 0 & 0 & 0 & 0 & 1 \\ 5 & 0 & 0 & 0 & 0 & 0 & 1 \\ 6 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \qquad \mathbf{Z}^* = \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

If the closure $\mathbf{Z}^*$ is known, the solution may be read off from column 5 of $\mathbf{Z}^*$. The vertices $i \in \{1, 2, 3, 5\}$ for which the element z^*_{i5} has the value 1 are start vertices of a path to the vertex 5. Formally, the solution vector $\mathbf{x}$ is determined by multiplying $\mathbf{Z}^*$ by the unit vector $\mathbf{e}_5$:

$$\mathbf{x} = \mathbf{Z}^* \circ \mathbf{e}_5$$

$$\mathbf{x} = \begin{array}{|c|c|c|c|c|c|} \hline 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 0 & 1 & 0 & 1 & 1 & 1 \\ \hline 0 & 0 & 1 & 1 & 1 & 1 \\ \hline 0 & 0 & 0 & 1 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 1 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ \hline \end{array} \circ \begin{array}{|c|} \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline 0 \\ \hline \end{array} = \begin{array}{|c|} \hline 1 \\ \hline 1 \\ \hline 1 \\ \hline 0 \\ \hline 1 \\ \hline 0 \\ \hline \end{array}$$

If the closure $\mathbf{Z}^*$ is not known, the column vector $\mathbf{x}$ is determined by solving the system of equations constructed according to Section 8.5.2. In this example $\mathbf{x}$ may be determined directly by back substitution, since the matrix $\mathbf{Z}$ contains only zero elements on and below the diagonal.

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_5$$

$$\begin{array}{|c|} \hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|c|} \hline 0 & 1 & 1 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 0 & 1 & 1 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} \circ \begin{array}{|c|} \hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} \sqcup \begin{array}{|c|} \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline 0 \\ \hline \end{array}$$

The back substitution begins with the last equation and proceeds as follows :

$$\begin{array}{lclcllcl} x_6 & = & 0 \sqcup 0 & & & & = & 0 \\ x_5 & = & 1 \circ x_6 \sqcup 1 & = & x_6 \sqcup 1 & = 0 \vee 1 & = & 1 \\ x_4 & = & 1 \circ x_6 \sqcup 0 & = & x_6 \sqcup 0 & = 0 \vee 0 & = & 0 \\ x_3 & = & 1 \circ x_4 \sqcup 1 \circ x_5 \sqcup 0 & = & x_4 \sqcup x_5 & = 0 \vee 1 & = & 1 \\ x_2 & = & 1 \circ x_4 \sqcup 1 \circ x_5 \sqcup 0 & = & x_4 \sqcup x_5 & = 0 \vee 1 & = & 1 \\ x_1 & = & 1 \circ x_2 \sqcup 1 \circ x_3 \sqcup 0 & = & x_2 \sqcup x_3 & = 1 \vee 1 & = & 1 \end{array}$$

8.5.4 REAL PATH ALGEBRA

Introduction : A path algebra is said to be real if the path sets are weighted by real numbers. Different real path problems lead to path algebras which differ in the definition of the binary operations $\sqcup$ (union) and $\circ$ (concatenation) for the weight set. In applications of real path algebras, the following extreme value problems for a vertex pair (i, k) are important :

- determination of the minimal path length
- determination of the maximal path length
- determination of the maximal path reliability
- determination of the maximal path capacity

8.5.4.1 Minimal path length

Problem : Let every edge of a directed graph be assigned a non-negative real number as an edge length. Let the length of a path from vertex i to vertex k be the sum of the lengths of all edges contained in the path. There may be different paths with different path lengths from vertex i to vertex k. The minimal path length from i to k is to be determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The minimum of the lengths of the paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = \infty$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = 0$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then z_{ik} is a non-negative real number. Thus the weight mapping is defined as follows :

$$
\begin{aligned}
f(0_W) &= 0_Z = \infty \\
f(1_W) &= 1_Z = 0 \\
f(a_{ik}) &= z_{ik} \in \mathbb{R}_0^+ \qquad \text{for} \qquad a_{ik} \notin \{0_W, 1_W\}
\end{aligned}
$$

Operations : For the weight set $Z = \mathbb{R}_0^+ \cup \{\infty\}$, the operations $\sqcup$ and $\circ$ are defined as follows. Let the path sets a_{ik}, b_{ik} be weighted by the minimal path lengths x_{ik}, y_{ik}. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the minimum $\min\{x_{ik}, y_{ik}\}$ of the two minimal path lengths. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the sum $x_{ik} + y_{km}$ of the two minimal path lengths.

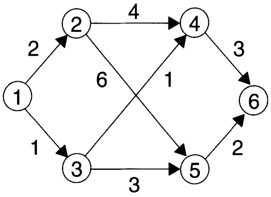
$$
\begin{array}{lll}
\text{union} & : & x_{ik} \sqcup y_{ik} := \min\{x_{ik}, y_{ik}\} \\
\text{concatenation} & : & x_{ik} \circ y_{km} := x_{ik} + y_{km}
\end{array}
$$

The domain $(\mathbb{R}_0^+ \cup \{\infty\} ; \sqcup, \circ)$ is a real path algebra with the zero element ∞ and the unit element 0. The operations $\sqcup$ and $\circ$ possess the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix **Z** is equal to the edge length. Otherwise z_{ik} is the zero element. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the minimal path length from i to k. Otherwise z^*_{ik} is the zero element. Every diagonal element z^*_{kk} of the closure $\mathbf{Z}^*$ is the unit element.

Example : Minimal path length

Let the illustrated acyclic directed graph with the elementary weight matrix **Z** for the edge lengths be given. The minimal path lengths from vertex 1 to all other vertices of the graph are to be determined.



$$\begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & n & 2 & 1 & n & n & n \\ 2 & n & n & n & 4 & 6 & n \\ 3 & n & n & n & 1 & 3 & n \\ 4 & n & n & n & n & n & 3 \\ 5 & n & n & n & n & n & 2 \\ 6 & n & n & n & n & n & n \\ \hline \end{array} \quad \mathbf{Z} \qquad n := \infty$$

Since the closure $\mathbf{Z}^*$ is not known, its first row is determined by solving the system of equations constructed according to Section 8.5.2. In this example, the solution vector **x** may be determined directly by forward substitution, since the matrix $\mathbf{Z}^T$ contains only zero elements on and above the diagonal.

$$\mathbf{x} = \mathbf{Z}^T \circ \mathbf{x} \sqcup \mathbf{e}_1$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} n & n & n & n & n & n \\ 2 & n & n & n & n & n \\ 1 & n & n & n & n & n \\ n & 4 & 1 & n & n & n \\ n & 6 & 3 & n & n & n \\ n & n & n & 3 & 2 & n \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \sqcup \begin{bmatrix} e \\ n \\ n \\ n \\ n \\ n \end{bmatrix} \qquad \begin{array}{l} n := \infty \\ e := 0 \end{array}$$

The forward substitution begins with the first row and proceeds as follows :

$$\begin{array}{lllll} x_1 & = \min\{n, e\} & = \min\{\infty, 0\} & = & 0 \\ x_2 & = \min\{2 + x_1, n\} & = \min\{2, \infty\} & = & 2 \\ x_3 & = \min\{1 + x_1, n\} & = \min\{1, \infty\} & = & 1 \\ x_4 & = \min\{4 + x_2, 1 + x_3, n\} & = \min\{6, 2, \infty\} & = & 2 \\ x_5 & = \min\{6 + x_2, 3 + x_3, n\} & = \min\{8, 4, \infty\} & = & 4 \\ x_6 & = \min\{3 + x_4, 2 + x_5, n\} & = \min\{5, 6, \infty\} & = & 5 \end{array}$$

The solution shows that there are paths from vertex 1 to all other vertices. There is a path of minimal path length 5 from vertex 1 to vertex 6.

8.5.4.2 Maximal path length

Problem : Let every edge of a directed graph be assigned a non-negative real number as an edge length. For a given vertex pair (i, k), the maximal path length from i to k is to be determined. If the vertices i and k lie on a cycle of positive length, then no maximal path length exists. Paths of maximal length are therefore determined for acyclic graphs only.

Weights : Let a path set a_{ik} containing paths from vertex i to vertex k be given. The maximum of the lengths of the paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = -\infty$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = 0$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then z_{ik} is a non-negative real number. Thus the weight mapping is defined as follows :

$$\begin{aligned} f(0_W) &= 0_Z = -\infty \\ f(1_W) &= 1_Z = 0 \\ f(a_{ik}) &= z_{ik} \in \mathbb{R}_0^+ \qquad \text{for} \qquad a_{ik} \notin \{0_W, 1_W\} \end{aligned}$$

Operations : For the weight set $Z = \mathbb{R}_0^+ \cup \{-\infty\}$ the operations $\sqcup$ and $\circ$ are defined as follows. Let the path sets a_{ik}, b_{ik} be weighted by the maximal path lengths x_{ik}, y_{ik}. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the maximum $\max\{x_{ik}, y_{ik}\}$ of the two maximal path lengths. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the sum $x_{ik} + y_{km}$ of the two maximal path lengths.

$$\begin{array}{lll} \text{union} & : & x_{ik} \sqcup y_{ik} := \max\{x_{ik}, y_{ik}\} \\ \text{concatenation} & : & x_{ik} \circ y_{km} := x_{ik} + y_{km} \end{array}$$

The domain $(\mathbb{R}_0^+ \cup \{-\infty\}; \sqcup, \circ)$ is a real path algebra with the zero element $-\infty$ and the unit element 0. The operations $\sqcup$ and $\circ$ possess the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix Z is equal to the edge length. Otherwise z_{ik} is the zero element. The matrix **Z** is only stable if the graph is acyclic. If the graph contains paths from vertex i to vertex k, then the element z_{ik}^* of the closure $\mathbf{Z}^*$ is equal to the maximal path length from i to k. Otherwise z_{ik}^* is the zero element. Every diagonal element z_{kk}^* of the closure $\mathbf{Z}^*$ is the unit element.

Example : Maximal path length

Let the acyclic directed graph from the example in the preceding section be given, for which the minimal path lengths from vertex 1 to all other vertices of the graph were determined. In this example, the maximal path lengths from vertex 1 to all other vertices of the graph are to be determined. Like the minimal path lengths, the maximal path lengths are determined by solving the following system of equations using forward substitution :

$$\mathbf{x} = \mathbf{Z}^{\mathrm{T}} \circ \mathbf{x} \sqcup \mathbf{e}_1$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} n & n & n & n & n & n \\ 2 & n & n & n & n & n \\ 1 & n & n & n & n & n \\ n & 4 & 1 & n & n & n \\ n & 6 & 3 & n & n & n \\ n & n & n & 3 & 2 & n \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \sqcup \begin{bmatrix} e \\ n \\ n \\ n \\ n \\ n \end{bmatrix} \qquad \begin{aligned} n &:= -\infty \\ e &:= 0 \end{aligned}$$

The solution of this system of equations differs from that in the preceding section, since the operations $\circ$ and $\sqcup$ are defined differently. The forward substitution begins with the first row and proceeds as follows :

$$\begin{aligned}
x_1 &= \max\{n, e\} & & \max\{-\infty, 0\} &&= 0 \\
x_2 &= \max\{2 + x_1, n\} &&= \max\{2, -\infty\} &&= 2 \\
x_3 &= \max\{1 + x_1, n\} &&= \max\{1, -\infty\} &&= 1 \\
x_4 &= \max\{4 + x_2, 1 + x_3, n\} &&= \max\{6, 2, -\infty\} &&= 6 \\
x_5 &= \max\{6 + x_2, 3 + x_3, n\} &&= \max\{8, 4, -\infty\} &&= 8 \\
x_6 &= \max\{3 + x_4, 2 + x_5, n\} &&= \max\{9,10, -\infty\} &&= 10
\end{aligned}$$

The solution shows that there are paths from vertex 1 to all other vertices of the graph. There is a path of maximal path length 10 from vertex 1 to vertex 6.

8.5.4.3 Maximal path reliability

Problem : Let every edge of a directed graph be assigned a real number in the interval]0.0, 1.0] as an edge reliability. Let the reliability of a path from vertex i to vertex k be the product of the reliabilities of all edges contained in the path. There may be different paths with different reliabilities from vertex i to vertex k. The maximal reliability of a path from vertex i to vertex k is to be determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The maximum of the reliabilities of the paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = 0.0$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = 1.0$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then $0.0 < z_{ik} \le 1.0$. Thus the weight mapping is defined as follows :

$$\begin{aligned} f(0_W) &= 0_Z = 0.0 \\ f(1_W) &= 1_Z = 1.0 \\ f(a_{ik}) &= z_{ik} \in\ ]0.0,1.0] \quad \text{for} \quad a_{ik} \notin \{0_W, 1_W\} \end{aligned}$$

Operations : The operations $\sqcup$ and $\circ$ are defined for the weight set $Z = [0.0,1.0]$. Let the path sets a_{ik}, b_{ik} be weighted by the maximal path reliabilities x_{ik}, y_{ik}. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the maximum $\max\{x_{ik}, y_{ik}\}$ of the two maximal path reliabilities. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the product $x_{ik} \cdot y_{km}$ of the two maximal path reliabilities.

$$\begin{aligned} &\text{union} && : && x_{ik} \sqcup y_{ik} && := && \max\{x_{ik}, y_{ik}\} \\ &\text{concatenation} && : && x_{ik} \circ y_{km} && := && x_{ik} \cdot y_{km} \end{aligned}$$

The domain ([0.0,1.0] ; $\sqcup$, $\circ$) is a real path algebra with the zero element 0.0 and the unit element 1.0. The operations $\sqcup$ and $\circ$ possess the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix $\mathbf{Z}$ is equal to the edge reliability. Otherwise z_{ik} is the zero element. The matrix $\mathbf{Z}$ is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is the maximal path reliability from i to k. Otherwise z^*_{ik} is the zero element. Every diagonal element z^*_{kk} of the closure $\mathbf{Z}^*$ is the unit element.

8.5.4.4 Maximal path capacity

Problem : Let every edge of a directed graph be assigned a positive real number as an edge capacity. Let the capacity of a path from vertex i to vertex k be the minimum of the capacities of all edges contained in the path. There may be different paths with different capacities from vertex i to vertex k. The maximal capacity of a path from i to k is to be determined.

Weights : Let a path set a_{ik} containing paths from vertex i to vertex k be given. The maximum of the capacities of the paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = 0$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = \infty$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then z_{ik} is a positive real number or ∞. Thus the weight mapping is defined as follows :

$$\begin{aligned} f(0_W) &= 0_Z = 0 \\ f(1_W) &= 1_Z = \infty \\ f(a_{ik}) &= z_{ik} \in \mathbb{R}^+ \cup \{\infty\} \quad \text{for} \quad a_{ik} \notin \{0_W, 1_W\} \end{aligned}$$

Operations : For the weight set $Z = \mathbb{R}_0^+ \cup \{\infty\}$, the operations $\sqcup$ and $\circ$ are defined as follows. Let the path sets a_{ik}, b_{ik} be weighted by the maximal path capacities x_{ik}, y_{ik}. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the maximum $\max\{x_{ik}, y_{ik}\}$ of the two maximal path capacities. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the minimum $\min\{x_{ik}, y_{km}\}$ of the two maximal path capacities.

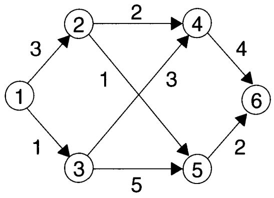
$$\begin{aligned} &\text{union} && : \quad x_{ik} \sqcup y_{ik} := \max\{x_{ik}, y_{ik}\} \\ &\text{concatenation} && : \quad x_{ik} \circ y_{km} := \min\{x_{ik}, y_{km}\} \end{aligned}$$

The domain $(\mathbb{R}_0^+ \cup \{\infty\}; \sqcup, \circ)$ is a real path algebra with the zero element 0 and the unit element ∞. The operations $\sqcup$ and $\circ$ possess the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix **Z** is equal to the edge capacity. Otherwise z_{ik} is the zero element. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the maximal path capacity from i to k. Otherwise z^*_{ik} is the zero element. Every diagonal element z^*_{kk} of the closure $\mathbf{Z}^*$ is the unit element.

Example : Maximal path capacity

Let the acyclic directed graph shown below with the elementary weight matrix **Z** for the edge capacities be given. The maximal path capacities from vertex 2 to all other vertices of the graph are to be determined.



$$\begin{array}{c|cccccc|} & 1 & 2 & 3 & 4 & 5 & 6 \\ \hline 1 & n & 3 & 1 & n & n & n \\ 2 & n & n & n & 2 & 1 & n \\ 3 & n & n & n & 3 & 5 & n \\ 4 & n & n & n & n & n & 4 \\ 5 & n & n & n & n & n & 2 \\ 6 & n & n & n & n & n & n \\ \hline \end{array} \; \mathbf{Z} \qquad n := 0$$

Since the closure $\mathbf{Z}^*$ is not known, its second row is determined by solving the system of equations constructed according to Section 8.5.2.

$$\mathbf{x} = \mathbf{Z}^T \circ \mathbf{x} \sqcup \mathbf{e}_2$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} n & n & n & n & n & n \\ 3 & n & n & n & n & n \\ 1 & n & n & n & n & n \\ n & 2 & 3 & n & n & n \\ n & 1 & 5 & n & n & n \\ n & n & n & 4 & 2 & n \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \sqcup \begin{bmatrix} n \\ e \\ n \\ n \\ n \\ n \end{bmatrix} \qquad \begin{array}{l} n := 0 \\ e := \infty \end{array}$$

The maximal path capacities are determined by forward substitution :

$$\begin{array}{lclcllcl} x_1 & = & \max\{n,n\} & & = & \max\{0,0\} & = & 0 \\ x_2 & = & \max\{\min\{3,x_1\}, e\} & & = & \max\{0,\infty\} & = & \infty \\ x_3 & = & \max\{\min\{1,x_1\}, n\} & & = & \max\{0,0\} & = & 0 \\ x_4 & = & \max\{\min\{2,x_2\}, \min\{3,x_3\}, n\} & & = & \max\{2,0,0\} & = & 2 \\ x_5 & = & \max\{\min\{1,x_2\}, \min\{5,x_3\}, n\} & & = & \max\{1,0,0\} & = & 1 \\ x_6 & = & \max\{\min\{4,x_4\}, \min\{2,x_5\}, n\} & & = & \max\{2,1,0\} & = & 2 \end{array}$$

The solution shows that there are paths from vertex 2 to the vertices $k \in \{2, 4, 5, 6\}$. There is a path of maximal path capacity 2 from vertex 2 to vertex 6.

8.5.5 LITERAL PATH ALGEBRA

Introduction : The literal labeling of graphs is treated in Section 8.5.2. It forms the basis for literal path algebras. Literal path algebras for different path problems differ in the definition of the literal weight set and the definitions of the operations. The literal path algebras are particularly important for structure problems in graph theory :

- determination of the simple paths and cycles
- determination of the elementary paths and cycles
- determination of the separating edges and vertices
- determination of the shortest or the longest paths and cycles

8.5.5.1 Path edges

Problem : Let the edges of a directed graph be uniquely labeled by the characters of an alphabet $\mathbb{A}$. Let every path from vertex i to vertex k of the graph be mapped to the set of characters for its edges. The set of characters for all edges which are contained in at least one of the paths from vertex i to vertex k is to be determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of characters for the edges contained in the paths of a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = 0_Z = \{\infty\}$ with the character ∞ which does not belong to the alphabet $\mathbb{A}$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = \emptyset = \{\}$. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then z_{ik} is a subset of the alphabet $\mathbb{A}$, and hence an element of the power set $P(\mathbb{A})$. Thus the weight mapping is defined as follows :

$$f(0_W) = 0_Z = \{\infty\}$$

$$f(1_W) = 1_Z = \{\}$$

$$f(a_{ik}) = z_{ik} \in P(\mathbb{A}) \quad \text{for} \quad a_{ik} \notin \{0_W, 1_W\}$$

Operations : For the weight set $Z = P(\mathbb{A}) \cup \{0_Z\}$, the operations $\sqcup$ and $\circ$ are defined as follows. Let the path sets a_{ik}, b_{ik} be weighted by the character sets x_{ik}, y_{ik} which are subsets of the alphabet $\mathbb{A}$. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the union $x_{ik} \cup y_{ik}$ of the two character sets. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is also the union $x_{ik} \cup y_{km}$ of the two character sets.

union : $x_{ik} \sqcup y_{ik} := x_{ik} \cup y_{ik}$ $\qquad x_{ik}, y_{ik} \in P(\mathbb{A})$

concatenation : $x_{ik} \circ y_{km} := x_{ik} \cup y_{km}$ $\qquad x_{ik}, y_{km} \in P(\mathbb{A})$

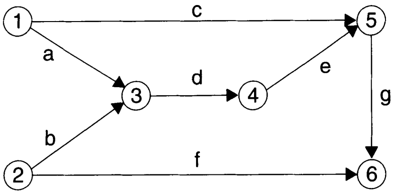
The domain $(P(\mathbb{A}) \cup \{0_Z\}; \sqcup, \circ)$ is a literal path algebra with the zero element $0_Z = \{\infty\}$ and the unit element $1_Z = \{\,\}$. The operations $\sqcup$ and $\circ$ have the properties required in Section 8.5.2 if the union $\sqcup$ and the concatenation $\circ$ for the zero element are defined as follows :

$$\begin{array}{llll} \text{union} & : \quad x_{ik} \sqcup 0_Z := x_{ik} & \qquad & 0_Z \sqcup y_{ik} := y_{ik} \\ \text{concatenation} & : \quad x_{ik} \circ 0_Z := 0_Z & \qquad & 0_Z \circ y_{km} := 0_Z \end{array}$$

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix **Z** is a one-element set containing only the character for the edge from i to k. Otherwise z_{ik} is the zero element. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of characters for the edges contained in at least one of the paths from i to k. Otherwise z^*_{ik} is the zero element. If the graph contains cycles through the vertex k, then the element z^*_{kk} of the closure $\mathbf{Z}^*$ is the set of characters for the edges contained in at least one cycle through k. Otherwise z^*_{kk} is the unit element.

Example : Path edges

Let the illustrated acyclic graph with literal edge labels be given. The path edges for the paths from each vertex of the graph to vertex 6 are to be determined.



Since the closure $\mathbf{Z}^*$ is not known, its sixth column is determined according to Section 8.5.2 as the solution of the following system of equations :

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_6$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \{a\} & 0 & \{c\} & 0 \\ 0 & 0 & \{b\} & 0 & 0 & \{f\} \\ 0 & 0 & 0 & \{d\} & 0 & 0 \\ 0 & 0 & 0 & 0 & \{e\} & 0 \\ 0 & 0 & 0 & 0 & 0 & \{g\} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} \sqcup \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \qquad \begin{array}{l} 0 := 0_Z = \{\infty\} \\ 1 := \emptyset \;= \{\,\} \end{array}$$

The path edges are determined by back substitution :

$$\begin{array}{lllllll}
x_6 & = & 0 \circ x_6 & \sqcup\ 1 & = & 0 \sqcup 1 & = 1 \\
x_5 & = & \{g\} \circ x_6 & \sqcup\ 0 & = & \{g\} \circ 1 & = \{g\} \\
x_4 & = & \{e\} \circ x_5 & \sqcup\ 0 & = & \{e\} \circ \{g\} & = \{e, g\} \\
x_3 & = & \{d\} \circ x_4 & \sqcup\ 0 & = & \{d\} \circ \{e, g\} & = \{d, e, g\} \\
x_2 & = & \{b\} \circ x_3 \sqcup \{f\} \circ x_6 & \sqcup\ 0 & = & \{b\} \circ \{d, e, g\} \sqcup \{f\} \circ 1 & = \{b, d, e, f, g\} \\
x_1 & = & \{a\} \circ x_3 \sqcup \{c\} \circ x_5 & \sqcup\ 0 & = & \{a\} \circ \{d, e, g\} \sqcup \{c\} \circ \{g\} & = \{a, c, d, e, g\}
\end{array}$$

The solution shows that there is at least one path from each vertex of the graph to the vertex 6. The set of edges which are contained in at least one of the paths from vertex 1 to vertex 6 is $x_1 = \{a, c, d, e, g\}$.

8.5.5.2 Common path edges

Problem : Let the edges of a directed graph be uniquely labeled by the characters of an alphabet $\mathbb{A}$. Let every path in the graph from vertex i to vertex k be mapped to the set of characters for its edges. The set of characters of the edges which are contained in each of the paths from vertex i to vertex k is to be determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of characters of the edges which are contained in each of the paths of a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. The weight mapping for common path edges has the same form as the weight mapping for path edges defined in Section 8.5.5.1.

Operations : For the weight set $Z = P(\mathbb{A}) \cup \{0_Z\}$, the operations $\sqcup$ and $\circ$ are defined as follows. Let the path sets a_{ik}, b_{ik} be weighted by the character sets x_{ik}, y_{ik} which are subsets of the alphabet $\mathbb{A}$. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the intersection $x_{ik} \cap y_{ik}$ of the two character sets. The concatenation is defined in the same way as for path edges in Section 8.5.5.1.

$$\begin{array}{llll}
\text{union} & : & x_{ik} \sqcup y_{ik} := x_{ik} \cap y_{ik} & x_{ik}, y_{ik} \in P(\mathbb{A}) \\
\text{concatenation} & : & x_{ik} \circ y_{km} := x_{ik} \cup y_{ik} & x_{ik}, y_{km} \in P(\mathbb{A})
\end{array}$$

The domain $(P(\mathbb{A}) \cup \{0_Z\}; \sqcup, \circ)$ is a literal path algebra with the zero element $0_Z = \{\infty\}$ and the unit element $1_Z = \{\,\}$. The operations $\sqcup$ and $\circ$ have the properties required in Section 8.5.2 if the operations $\sqcup$ and $\circ$ for the zero element are defined as follows in analogy with the algebra for path edges :

$$\begin{array}{llll}
\text{union} & : & x_{ik} \sqcup 0_Z := x_{ik} & 0_Z \sqcup y_{ik} := y_{ik} \\
\text{concatenation} & : & x_{ik} \circ 0_Z := 0_Z & 0_Z \circ y_{km} := 0_Z
\end{array}$$

Weight matrices : The elementary weight matrices **Z** of the literal path algebras for path edges and common path edges coincide. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of characters for the common edges of the paths from i to k. Otherwise z^*_{ik} is the zero element. Each diagonal element z^*_{kk} of the closure $\mathbf{Z}^*$ is the unit element.

Example : Common path edges

Let the acyclic graph with literal edge labels from the example in the preceding section be given, for which the path edges from each vertex of the graph to the vertex 6 were determined. In this example, the common path edges for the paths from each vertex of the graph to the vertex 6 are to be determined. The system of equations for common path edges is formally identical with the system of equations for path edges. It is solved by back substitution using the operations defined for common path edges :

$$0 := 0_Z = \{\infty\} \qquad 1 := 1_Z = \{\,\}$$

$$\begin{array}{lllllll}
x_6 & = & 0 \circ x_6 & \sqcup\ 1 & = & 0 \sqcup 1 & = 1 \\
x_5 & = & \{g\} \circ x_6 & \sqcup\ 0 & = & \{g\} \circ 1 & = \{g\} \\
x_4 & = & \{e\} \circ x_5 & \sqcup\ 0 & = & \{e\} \circ \{g\} & = \{e,g\} \\
x_3 & = & \{d\} \circ x_4 & \sqcup\ 0 & = & \{d\} \circ \{e,g\} & = \{d,e,g\} \\
x_2 & = & \{b\} \circ x_3 \sqcup \{f\} \circ x_6 & \sqcup\ 0 & = & \{b\} \circ \{d,e,g\} \sqcup \{f\} \circ 1 & = 1 \\
x_1 & = & \{a\} \circ x_3 \sqcup \{c\} \circ x_5 & \sqcup\ 0 & = & \{a\} \circ \{d,e,g\} \sqcup \{c\} \circ \{g\} & = \{g\}
\end{array}$$

The solution shows that there is at least one path from each vertex of the graph to the vertex 6, since $x_i \neq 0$ for $i = 1,\ldots,6$. The paths from vertex 2 to vertex 6 have no edge in common, since $x_2 = 1 = \{\,\}$. The paths from vertex 1 to vertex 6 have the edge g in common.

8.5.5.3 Simple paths

Problem : Let the edges of a directed graph be uniquely labeled by the characters of an alphabet $\mathbb{A}$. A path is an edge sequence and is designated by a word, which is a sequence of the labels of its edges. A path in the graph is simple if it does not contain any edge more than once. A simple path is therefore labeled by a simple word which does not contain any character more than once. The words for all simple paths from vertex i to vertex k are to be determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of all simple words for the simple paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. If the path set a_{ik} is the zero set 0_W, then $z_{ik} = 0_Z = \{\,\}$. If the path set a_{ik} is the unit set 1_W, then $z_{ik} = 1_W = \{\lambda\}$ with the empty word λ. If the path set a_{ik} is neither the zero set 0_W nor the unit set 1_W, then z_{ik} is a set of simple words. Let the set of all simple words over the alphabet $\mathbb{A}$ including the empty word λ be $\mathbb{S}$. Then z_{ik} is a subset of $\mathbb{S}$, and hence an element of the power set $P(\mathbb{S})$. Thus the weight mapping is defined as follows :

$$
\begin{aligned}
f(0_W) &= 0_Z = \{\,\} \\
f(1_W) &= 1_Z = \{\lambda\} \\
f(a_{ik}) &= z_{ik} \in P(\mathbb{S}) \qquad \text{for} \qquad a_{ik} \notin \{0_W, 1_W\}
\end{aligned}
$$

Operations : The operations $\sqcup$ and $\circ$ are defined for the weight set $Z = P(\mathbb{S})$. Let the path sets a_{ik}, b_{ik} be weighted with sets x_{ik}, y_{ik} of simple words. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the union $x_{ik} \cup y_{ik}$ of the two sets of simple words. The weight $x_{ik} \circ y_{km}$ of the concatenation $a_{ik} \circ b_{km}$ is the set of all simple words formed by concatenating a simple word from x_{ik} with a simple word from y_{km}.

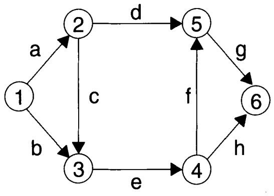
$$
\begin{array}{lll}
\text{union} & : & x_{ik} \sqcup y_{ik} := x_{ik} \cup y_{ik} \\
\text{concatenation} & : & x_{ik} \circ y_{km} := \{x \circ y \in \mathbb{S} \mid x \in x_{ik} \ \wedge \ y \in y_{km}\}
\end{array}
$$

The domain $(P(\mathbb{S})\,;\sqcup,\circ)$ is a literal path algebra with the zero element $0_Z = \{\,\}$ and the unit element $1_Z = \{\lambda\}$. The operations have the properties required in Section 8.5.2.

Weight matrices : Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix $\mathbf{Z}$ is a one-element set containing the simple word consisting for the character of the edge from i to k. Otherwise z_{ik} is the zero element. The matrix $\mathbf{Z}$ is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of words for the simple paths from i to k. Otherwise z^*_{ik} is the zero element. If the graph contains cycles through the vertex k, then the element z^*_{kk} of the closure $\mathbf{Z}^*$ contains the empty word λ as well as all words for the simple cycles through k. Otherwise z^*_{kk} is the unit element.

Example : Simple paths

Let the illustrated acyclic graph with literal edge labels be given. The words for the simple paths from each vertex of the graph to vertex 6 are to be determined.



Since the closure $\mathbf{Z}^*$ is not known, its sixth column is determined according to Section 8.5.2 as the solution of the following system of equations.

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_6$$

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline\end{array} = \begin{array}{|c|c|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 & 0 & 0 \\ \hline 0 & 0 & \{c\} & 0 & \{d\} & 0 \\ \hline 0 & 0 & 0 & \{e\} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \{f\} & \{h\} \\ \hline 0 & 0 & 0 & 0 & 0 & \{g\} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline\end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline\end{array} \sqcup \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline\end{array} \qquad \begin{array}{l} 0 := 0_Z = \{\ \} \\ 1 := 1_Z = \{\lambda\} \end{array}$$

The simple paths are determined by back substitution :

$$\begin{array}{lcllcl}
x_6 & = & 0 \circ x_6 & \sqcup\ 1 & = & 0 \sqcup 1 \quad = \quad 1 \\
x_5 & = & \{g\} \circ x_6 & \sqcup\ 0 & = & \{g\} \circ 1 \quad = \quad \{g\} \\
x_4 & = & \{f\} \circ x_5 \sqcup \{h\} \circ x_6 & \sqcup\ 0 & = & \{f\} \circ \{g\} \sqcup \{h\} \circ 1 \ = \ \{fg, h\} \\
x_3 & = & \{e\} \circ x_4 & \sqcup\ 0 & = & \{e\} \circ \{fg, h\} \quad = \ \{efg, eh\} \\
x_2 & = & \{c\} \circ x_3 \sqcup \{d\} \circ x_5 & \sqcup\ 0 & = & \{c\} \circ \{efg, eh\} \sqcup \{d\} \circ \{g\} \\
 & & & & = & \{cefg, ceh, dg\} \\
x_1 & = & \{a\} \circ x_2 \sqcup \{b\} \circ x_3 & \sqcup\ 0 & = & \{a\} \circ \{cefg, ceh, dg\} \sqcup \{b\} \circ \{efg, eh\} \\
 & & & & = & \{acefg, aceh, adg, befg, beh\}
\end{array}$$

The solution shows that there is at least one simple path from each vertex of the graph to vertex 6. Two simple paths, efg and eh, lead from vertex 3 to vertex 6.

8.5.5.4 Extreme simple paths

Problem : Let the edges of a directed graph be uniquely labeled by the characters of an alphabet $\mathbb{A}$. A simple path from vertex i to vertex k does not contain any edge more than once. It is called a shortest path from i to k if it does not contain more edges than any other path from i to k. It is called a longest path from i to k if it does not contain fewer edges than any other path from i to k. The words for all shortest or all longest paths from vertex i to vertex k are to determined.

Weights : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of all extreme simple words for the shortest or longest paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. The weight mapping has the same form as for simple paths :

$$f(0_W) = 0_Z = \{\,\}$$

$$f(1_W) = 1_Z = \{\lambda\}$$

$$f(a_{ik}) = z_{ik} \in P(\mathbb{S}) \qquad \text{for} \qquad a_{ik} \notin \{0_W, 1_W\}$$

Operations : The operations $\sqcup$ and $\circ$ are defined for the weight set $P(\mathbb{S})$. Let the path sets a_{ik}, b_{ik} be weighted by sets x_{ik}, y_{ik} of extreme simple words. The weight $x_{ik} \sqcup y_{ik}$ of the union $a_{ik} \sqcup b_{ik}$ is the reduction $\text{extr}(x_{ik} \cup y_{ik})$ of the union $x_{ik} \cup y_{ik}$ to the set of extreme simple words. The concatenation $\circ$ is defined as for simple paths :

$$\begin{array}{lll} \text{union} & : & x_{ik} \sqcup y_{ik} := \text{extr}\,(x_{ik} \cup y_{ik}) \\ \text{concatenation} & : & x_{ik} \circ y_{km} := \{x \circ y \in \mathbb{S} \mid x \in x_{ik} \ \wedge \ y \in y_{km}\} \end{array}$$

As in the case of simple paths, the domain $(P(\mathbb{S})\,;\sqcup, \circ)$ is a literal path algebra with the zero element $0_Z = \{\,\}$ and the unit element $1_Z = \{\lambda\}$.

Weight matrices : The elementary weight matrices **Z** of the literal path algebra for simple paths and for extreme simple paths coincide. The matrix **Z** is stable both for shortest and for longest simple paths. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of words for the extreme simple paths from i to k. Otherwise z^*_{ik} is the zero element. If the graph contains cycles through the vertex k, then the element z^*_{kk} of the closure $\mathbf{Z}^*$ contains all words for the extreme simple cycles through k. Otherwise z^*_{kk} is the unit element. Since the shortest cycle through a vertex k is always the empty path λ, z^*_{kk} is always the unit element in the case of shortest simple paths.

Example : Shortest paths

Let the acyclic graph with the literal edge labels from the example in the preceding section be given, for which the simple paths from each vertex of the graph to the vertex 6 were determined. In this example, the shortest paths from each vertex of the graph to the vertex 6 are to be determined. The system of equations for the shortest paths is formally identical with the system of equations for the simple paths. It is solved by back substitution using the operations defined for shortest paths :

$$
\begin{aligned}
x_6 &= 0 \quad \circ x_6 \sqcup 1 &&= 0 \sqcup 1 &&= 1 \\
x_5 &= \{g\} \circ x_6 \sqcup 0 &&= \{g\} \circ 1 &&= \{g\} \\
x_4 &= \{f\} \circ x_5 \sqcup \{h\} \circ x_6 \sqcup 0 &&= \text{extr}\,(\{f\} \circ \{g\} \cup \{h\} \circ 1) &&= \{h\} \\
x_3 &= \{e\} \circ x_4 \sqcup 0 &&= \{e\} \circ \{h\} &&= \{eh\} \\
x_2 &= \{c\} \circ x_3 \sqcup \{d\} \circ x_5 \sqcup 0 &&= \text{extr}\,(\{c\} \circ \{eh\} \cup \{d\} \circ \{g\}) &&= \{dg\} \\
x_1 &= \{a\} \circ x_2 \sqcup \{b\} \circ x_3 \sqcup 0 &&= \text{extr}\,(\{a\} \circ \{dg\} \cup \{b\} \circ \{eh\}) &&= \{adg, beh\}
\end{aligned}
$$

The solution shows that there is at least one simple path from each vertex of the graph to vertex 6. Two shortest paths, adg and beh, lead from vertex 1 to vertex 6.

8.5.5.5 Literal vertex labels

Introduction : Let every vertex of a directed graph be labeled by a character from an alphabet $\mathbb{A}$. Let any two vertices be labeled by different characters, so that the vertex labels are unique. Every edge of the directed graph is labeled by the characters of the start and the end vertex, so that the literal path algebras treated in the preceding sections may be used to determine path vertices, common path vertices, elementary paths and extreme elementary paths.

Path vertices : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of characters for the vertices contained in the paths of a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. The rules for the union and concatenation of weighted path sets and the definitions for the zero and unit element are identical with the rules and definitions in Section 8.5.5.1.

Let a directed graph be given. If the graph contains an edge from vertex i to vertex k, then the element z_{ik} of the elementary weight matrix **Z** is a set containing the character for the vertex i and the character for the vertex k. Otherwise z_{ik} is the zero element. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of characters for the vertices contained in at least one of the paths from i to k. Otherwise z^*_{ik} is the zero element.

Common path vertices : Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of characters for the vertices contained in every path of a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. The rules for the union and concatenation of weighted path sets and the definitions for the zero and unit element are identical with the rules and definitions in Section 8.5.5.2.

The elementary weight matrices **Z** of a directed graph for path vertices and common path vertices coincide. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex k, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of characters for the common vertices contained in every path from i to k. Otherwise z^*_{ik} is the zero element.

Elementary paths : Every path in the directed graph with literal vertex labels is associated with a word. In this word, the characters occur in the order in which the associated vertices occur in the path. An elementary path does not contain any vertex more than once and is therefore designated by a simple word. Let two paths in the directed graph be labeled by the words a and b. The word a can be concatenated with the word b to form a word $c = a \circ b$ only if the last character of a and the first character of b coincide. The concatenated word c is formed by appending the word b without its first character to the word a.

Let the path set a_{ik} containing paths from vertex i to vertex k be given. The set of simple words for the elementary paths contained in a_{ik} is chosen as the weight z_{ik} of the path set a_{ik}. The rules for the union and concatenation of weighted path sets and the definitions for the zero and unit element are identical with the rules and definitions in Section 8.5.5.3.

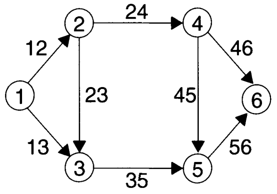
Let a directed graph be given. If the graph contains an edge from vertex i to vertex $k \neq i$, then the element z_{ik} of the elementary weight matrix **Z** is a one-element set containing the simple word with the characters of the vertices i and k. Otherwise z_{ik} is the zero element. The matrix **Z** is stable. If the graph contains paths from vertex i to vertex $k \neq i$, then the element z^*_{ik} of the closure $\mathbf{Z}^*$ is equal to the set of words for the elementary paths from i to k. Otherwise z^*_{ik} is the zero element. Every diagonal element z^*_{kk} of the closure $\mathbf{Z}^*$ is the unit element.

Elementary cycles through a vertex k cannot be determined using this path algebra, since the word for such a cycle contains the character for the vertex k at the beginning and at the end and is therefore not simple. If the set of simple words is extended to include words with identical first and last characters, elementary paths including elementary cycles may be determined using this extended set of words.

Extreme elementary paths : On the basis of the path algebra for elementary paths, the path algebra for the shortest or longest elementary paths is defined according to Section 8.5.5.4.

Example : Elementary paths

Let the illustrated acyclic graph with digits as literal vertex labels be given. Let every edge be weighted by the simple word consisting of the digits of its start and end vertex. The elementary paths from each vertex of the graph to vertex 6 are to be determined.



Since the closure $\mathbf{Z}^*$ is not known, its sixth column is determined according to Section 8.5.2 as the solution of the following system of equations :

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_6$$

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|c|}\hline 0 & \{12\} & \{13\} & 0 & 0 & 0 \\ \hline 0 & 0 & \{23\} & \{24\} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & \{35\} & 0 \\ \hline 0 & 0 & 0 & 0 & \{45\} & \{46\} \\ \hline 0 & 0 & 0 & 0 & 0 & \{56\} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 \\ \hline \end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} \sqcup \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline \end{array} \qquad \begin{array}{l} 0 := \{\ \} \\ 1 := \{\lambda\} \end{array}$$

The elementary paths are determined by back substitution :

$$\begin{array}{lclcl}
x_6 & = & 0 \circ x_6 \sqcup 1 & = & 0 \sqcup 1 \quad = \quad 1 \\
x_5 & = & \{56\} \circ x_6 \sqcup 0 & = & \{56\} \circ 1 \quad = \quad \{56\} \\
x_4 & = & \{46\} \circ x_6 \sqcup \{45\} \circ x_5 \sqcup 0 & = & \{46\} \circ 1 \sqcup \{45\} \circ \{56\} \quad = \quad \{46, 456\} \\
x_3 & = & \{35\} \circ x_5 \sqcup 0 & = & \{35\} \circ \{56\} \quad = \quad \{356\} \\
x_2 & = & \{23\} \circ x_3 \sqcup \{24\} \circ x_4 \sqcup 0 & = & \{23\} \circ \{356\} \sqcup \{24\} \circ \{46{,}456\} \\
 & & & = & \{2356, 246, 2456\} \\
x_1 & = & \{12\} \circ x_2 \sqcup \{13\} \circ x_3 \sqcup 0 & = & \{12\} \circ \{2356{,}246{,}2456\} \sqcup \{13\} \circ \{356\} \\
 & & & = & \{12356, 1246, 12456, 1356\}
\end{array}$$

The solution shows that there is at least one elementary path from each vertex of the graph to vertex 6. The two elementary paths 46 and 456 lead from vertex 4 to vertex 6. Due to the chosen weighting of the edges, the elementary paths are described as a vertex sequence with the start vertex, all intermediate vertices and the end vertex.

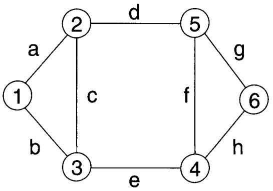
8.5.5.6 Literal edge labels for simple graphs

Introduction : A simple graph consists of vertices and undirected edges between pairs of different neighboring vertices. Let every undirected edge be labeled by a character from an alphabet $\mathbb{A}$. Let any two undirected edges of the simple graph be labeled by different characters, so that the labels of the undirected edges are unique. A simple graph is treated as a symmetric graph by replacing each undirected edge with its label by a pair of edges in opposite directions with the same label. Thus the literal path algebras for directed graphs treated in Sections 8.5.6.1 to 8.5.6.4 may also be used directly for simple graphs.

Symmetry : Since a simple graph is treated as a symmetric directed graph with symmetric labels, the elementary weight matrix **Z** is symmetric. The closure $\mathbf{Z}^*$ is also symmetric, since for every path from vertex i to vertex k via an edge sequence there is also a corresponding path from vertex k to vertex i via the reverse edge sequence. This symmetry property may be exploited in the algorithms for the path algebra.

Example : Simple paths in a simple graph

Let the illustrated simple graph with literal edge labels be given. The simple paths from each vertex of the graph to vertex 6 are to be determined.



Since the closure $\mathbf{Z}^*$ is not known, its sixth column is determined according to Section 8.5.2 as the solution of the following system of equations :

$$\mathbf{x} = \mathbf{Z} \circ \mathbf{x} \sqcup \mathbf{e}_6$$

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|c|}\hline 0 & \{a\} & \{b\} & 0 & 0 & 0 \\ \hline \{a\} & 0 & \{c\} & 0 & \{d\} & 0 \\ \hline \{b\} & \{c\} & 0 & \{e\} & 0 & 0 \\ \hline 0 & 0 & \{e\} & 0 & \{f\} & \{h\} \\ \hline 0 & \{d\} & 0 & \{f\} & 0 & \{g\} \\ \hline 0 & 0 & 0 & \{h\} & \{g\} & 0 \\ \hline \end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline x_6 \\ \hline \end{array} \sqcup \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 0 \\ \hline 1 \\ \hline \end{array} \qquad \begin{array}{l} 0 := \{\ \} \\ 1 := \{\lambda\} \end{array}$$

The matrix **Z** of this system of equations is symmetric. Its solutions may be determined using one of the procedures in Section 8.5.7.

8.5.5.7 Applications in structural analysis

Literal path algebras may be employed in the structural analysis of directed and simple graphs. Some typical examples are listed in the following.

Subgraphs and components : A subgraph for a vertex pair (i, k) consists of the vertices and edges contained in at least one of the paths in the graph from vertex i to vertex k. The edge set of the subgraph may be determined using the path algebra for path edges in Section 8.5.5.1. The vertex set of the subgraph consists of the vertex k and the start vertices of all edges of the edge set. The subgraph for the vertex pair (k, k) is a strongly connected component in the case of directed graphs and a simply connected component in the case of simple graphs.

Basic edges and bridges : A path edge which occurs in every path from vertex i to vertex k is a basic edge in the case of directed graphs and a bridge in the case of simple graphs. The path algebra for common path edges in Section 8.5.5.2 may therefore be used to determine basic edges or bridges. If for a connected vertex pair (i, k) there is no path edge which occurs in all paths from i to k, then there are at least two edge-disjoint paths from i to k.

Separating vertices and articulation vertices : A path vertex a which occurs in every path from vertex $i \neq a$ to vertex $k \neq a$ is a separating vertex in the case of directed graphs and an articulation vertex in the case of simple graphs. The path algebra for common path vertices in Section 8.5.5.5 may therefore be used to determine separating vertices or articulation vertices. If for a connected vertex pair (i, k) there is no vertex other than i and k which occurs in every path from i to k, then there are at least two vertex-disjoint paths from i to k.

Eulerian and Hamiltonian paths : If a longest simple path from vertex i to vertex k contains all edges of the graph, then this path is an Eulerian path for $i \neq k$ and an Eulerian cycle for $i = k$. The path algebra for the longest simple paths in Section 8.5.5.4 may therefore be used to determine Eulerian paths and cycles. If a longest elementary path from vertex i to vertex k contains all vertices of the graph, then this path is a Hamiltonian path for $i \neq k$ and a Hamiltonian cycle for $i = k$. The path algebra for the longest elementary paths in Section 8.5.5.5 may therefore be used to determine Hamiltonian paths and cycles.

8.5.6 PROPERTIES OF PATH ALGEBRAS

Introduction : The path algebras treated in the preceding sections have the same algebraic structure. They differ in the definition of the weight sets and in the rules for union and concatenation. This results in different properties, which lead to a further classification of path algebras and are of fundamental importance for the general solution of the systems of equations for path problems.

Compilation : The basic definitions of the boolean, real and literal path algebras are compiled in the following table.

Path algebra	Weight set Z	Union $x \sqcup y$	Concatenation $x \circ y$
path existence	$\{0,1\}$	$x \vee y$	$x \wedge y$
min. path length	$\mathbb{R}_0^+ \cup \{0_Z\}$	min (x,y)	$x + y$
max. path length	$\mathbb{R}_0^+ \cup \{0_Z\}$	max (x,y)	$x + y$
max. path reliability	$[0.0, 1.0]$	max (x,y)	$x * y$
max. path capacity	$\mathbb{R}_0^+ \cup \{1_Z\}$	max (x,y)	min (x,y)
path edges	$P(\mathbb{A}) \cup \{0_Z\}$	$x \cup y$	$x \cup y$
common path edges	$P(\mathbb{A}) \cup \{0_Z\}$	$x \cap y$	$x \cup y$
simple paths	$P(\mathbb{S})$	$x \cup y$	$x \circ y$
extreme simple paths	$P(\mathbb{S})$	$\text{extr}(x \cup y)$	$x \circ y$

Inclusion : As in set theory, the inclusion $x \sqsubseteq y$ of the elements $x, y \in Z$ of a weight set Z is defined in terms of union and equality. Its result is the logical constant true or false.

inclusion $\quad x \sqsubseteq y \quad :\Leftrightarrow \quad x \sqcup y \; = \; y$

The following rules of calculation hold for inclusion with respect to union and concatenation for $x, y, z \in Z$:

union

$$x \sqsubseteq y \quad \Rightarrow \quad x \sqcup z \;\sqsubseteq\; y \sqcup z$$
$$x \sqsubseteq y \quad \Rightarrow \quad z \sqcup x \;\sqsubseteq\; z \sqcup y$$

concatenation

$$x \sqsubseteq y \quad \Rightarrow \quad x \circ z \;\sqsubseteq\; y \circ z$$
$$x \sqsubseteq y \quad \Rightarrow \quad z \circ x \;\sqsubseteq\; z \circ y$$

Order structure : The inclusion $u \sqsubseteq v$ in a weight set Z is reflexive, antisymmetric and transitive. It is therefore a partial order relation. The weight set Z is ordered.

reflexive	$x \sqsubseteq x$
antisymmetric	$x \sqsubseteq y \ \wedge \ y \sqsubseteq x \quad \Rightarrow \quad x = y$
transitive	$x \sqsubseteq y \ \wedge \ y \sqsubseteq z \quad \Rightarrow \quad x \sqsubseteq z$

If an inclusion is also linear, then it is a total order relation. The weight set Z is totally ordered.

linear	$x \sqsubseteq y \ \vee \ y \sqsubseteq x$

Least and greatest element : The least element in a weight set Z is the zero element 0_Z. A greatest element need not exist. If a greatest element does exist, it may be the unit element 1_Z or another element of the weight set. Correspondingly, weight sets with the following properties are distinguished :

weight sets without greatest element	$0_Z \sqsubseteq x$
weight sets with the greatest element 1_Z	$0_Z \sqsubseteq x \sqsubseteq 1_Z$
weight sets with the greatest element g_Z	$0_Z \sqsubseteq x \sqsubseteq g_Z$

The ordinal property and the least and greatest element are compiled for the weight sets of the different path algebras :

Path algebra	Inclusion $\sqsubseteq$	Least element	Greatest element
path existence	total	0_Z	1_Z
min. path length	total	0_Z	1_Z
max. path length	total	0_Z	–
max. path reliability	total	0_Z	1_Z
max. path capacity	total	0_Z	1_Z
path edges	partial	0_Z	$\mathbb{A}$
common path edges	partial	0_Z	1_Z
simple paths	partial	0_Z	$\mathbb{S}$
extreme simple paths	partial	0_Z	$\mathbb{S}$

Sub- and superunitary elements : The elements $x \in Z$ of a weight set Z are considered with respect to the unit element 1_Z. An element $x \sqsubseteq 1_Z$ is said to be subunitary. An element $x \sqsupseteq 1_Z$ is said to be superunitary. If the unit element is the greatest element, then all elements are subunitary.

subunitary	$x \sqsubseteq 1_Z$
superunitary	$x \sqsupseteq 1_Z$

Powers of an element : The 0-th power x^0 of an element $x \in Z$ of the weight set Z is defined to be the unit element 1_Z. The m-th power x^m is defined as the concatenation of x^{m-1} and x.

$$\text{power} \qquad x^0 := 1_Z \qquad\qquad x^m := x^{m-1} \circ x$$

An element x is said to be idempotent if $x^2 = x$. Every power x^m of an idempotent element x is equal to x for $m \geq 1$. An element x is said to be nilpotent of degree q if $x^q = 0_Z$. Every power x^m of a nilpotent element x is equal to 0_Z for $m \geq q$.

$$\begin{array}{ll} \text{idempotent} & x^2 = x \\ \text{nilpotent} & x^q = 0_Z \end{array}$$

Closure of an element : The reflexive transitive closure $\hat{x}$ of an element $x \in Z$ of a weight set Z is calculated as the union of the powers of x. If the union does not change beyond a certain power x^p, then the element x is stable and the closure $\hat{x}$ exists. The positive integer p is called the stability index of the element.

$$\hat{x} = 1_Z \sqcup x \sqcup x^2 \sqcup \ldots = \bigsqcup_{m \geq 0} x^m = \bigsqcup_{m=0}^{p} x^m$$

If an element is nilpotent, idempotent or subunitary, then it is stable.

$$\begin{array}{ll} \text{nilpotent} & \hat{x} = 1_Z \sqcup x \sqcup x^2 \sqcup \ldots \sqcup x^{q-1} \\ \text{idempotent} & \hat{x} = 1_Z \sqcup x \\ \text{subunitary} & \hat{x} = 1_Z \end{array}$$

Stability : Path algebras are classified with respect to stability. A path algebra is said to be conditionally stable if at least one element of the weight set is not stable. It is said to be unconditionally stable if every element of the weight set is stable. It is said to be unitarily stable if the reflexive transitive closure of every element of the weight set is the unit element.

Path algebra	Stability	Closure $\hat{x}$
path existence	unitary	$\hat{x} = 1_Z$
min. path length	unitary	$\hat{x} = 1_Z$
max. path length	conditional	$\hat{x} = 1_Z$ for $x \in \{0_Z, 1_Z\}$
max. path reliability	unitary	$\hat{x} = 1_Z$
max. path capacity	unitary	$\hat{x} = 1_Z$
path edges	unconditional	$\hat{x} = 1_Z \sqcup x$
common path edges	unitary	$\hat{x} = 1_Z$
simple paths	unconditional	$\hat{x} = \bigsqcup_{m=0}^{p} x^m \qquad p \leq \lvert x \rvert$
extreme simple paths	unconditional	

Example 1 : Real path algebra for minimal path lengths

The weight set of the real path algebra for minimal path lengths is $Z = \mathbb{R}_0^+ \cup \{0_Z\}$ with the zero element $0_Z = \infty$ and the unit element $1_Z = 0$. The inclusion $x \sqsubseteq y$ of the elements $x, y \in Z$ is equivalent to $x \geq y$.

$$x \sqsubseteq y \quad \Leftrightarrow \quad x \sqcup y = y \quad \Leftrightarrow \quad \min(x, y) = y \quad \Leftrightarrow \quad x \geq y$$

The zero element $0_Z = \infty$ is the least element in Z. The unit element $1_Z = 0$ is the greatest element in Z.

$$0_Z \sqsubseteq x \sqsubseteq 1_Z \quad \Leftrightarrow \quad \infty \geq x \geq 0$$

Every element $x \in Z$ is subunitary.

$$x \in Z \quad \Leftrightarrow \quad x \geq 0 \quad \Leftrightarrow \quad x \sqsubseteq 1_Z$$

The closure $\hat{x}$ of an element $x \geq 0$ is the unit element 1_Z.

$$\hat{x} = 1_Z \sqcup x \sqcup x^2 \sqcup x^3 \sqcup \ldots = \min\{0, x, x+x, x+x+x, \ldots\} = 0 = 1_Z$$

Since the reflexive transitive closure $\hat{x}$ of every element $x \in Z$ is equal to the unit element 1_Z, the path algebra for minimal path lengths is unitarily stable.

Example 2 : Real path algebra for maximal path lengths

The weight set of the real path algebra for maximal path lengths is $Z = \mathbb{R}_0^+ \cup \{0_Z\}$ with the zero element $0_Z = -\infty$ and the unit element $1_Z = 0$. The inclusion $x \sqsubseteq y$ of the elements $x, y \in z$ is equivalent to $x \leq y$.

$$x \sqsubseteq y \quad \Leftrightarrow \quad x \sqcup y = y \quad \Leftrightarrow \quad \max\{x, y\} = y \quad \Leftrightarrow \quad x \leq y$$

The zero element $0_Z = -\infty$ is the least element in Z. There is no greatest element.

$$0_Z \sqsubseteq x \quad \Leftrightarrow \quad -\infty \leq x$$

The closure $\hat{x}$ cannot be formed in the general case, since the union of the powers of every $x \in \mathbb{R}^+$ tends to infinity.

$$\hat{x} = 1_Z \sqcup x \sqcup x^2 \sqcup x^3 \sqcup \ldots = \max\{0, x, x+x, x+x+x, \ldots\}$$

$$\hat{x} = 0 \qquad \text{for} \qquad x \in \{-\infty, 0\}$$

$$\hat{x} \to \infty \qquad \text{for} \qquad x \in \mathbb{R}^+$$

Example 3 : Literal path algebra for path edges

The weight set of the literal path algebra for path edges is $Z = P(\mathbb{A}) \cup \{0_Z\}$ with the alphabet $\mathbb{A}$, the zero element $0_Z = \{\infty\}$ and the unit element $1_Z = \{\ \}$. The inclusion $x \sqsubseteq y$ of the elements $x, y \in P(\mathbb{A})$ is equivalent to $x \subseteq y$.

$$x \sqsubseteq y \quad \Leftrightarrow \quad x \sqcup y = y \quad \Leftrightarrow \quad x \cup y = y \quad \Leftrightarrow \quad x \subseteq y$$

The zero element $0_Z = \{\infty\}$ is the least element in Z. The element $\mathbb{A}$ is the greatest element in Z.

$$0_Z \sqsubseteq x \sqsubseteq \mathbb{A}$$

Every element $x \in Z$ is idempotent and therefore stable.

$$x^2 = x \circ x = x \cup x = x$$

Due to idempotency, the closure $\hat{x}$ of an element $x \in Z$ is $1_Z \sqcup x$.

$$\begin{aligned} \hat{x} &= 1_Z \sqcup x \\ \hat{x} &= 1_Z \quad \text{for} \quad x = 0_Z \\ \hat{x} &= x \quad \text{for} \quad x \in P(\mathbb{A}) \end{aligned}$$

Example 4 : Literal path algebra for simple paths

The weight set of the literal path algebra for simple paths is $Z = P(\mathbb{S})$ with the set $\mathbb{S}$ of all simple words, the zero element $0_Z = \{\ \}$ and the unit element $1_Z = \{\lambda\}$. Let an element $x \in Z$ be a set of p simple words. Assume first that no character of a simple word occurs in another simple word.

$$x = \{x_1, x_2, \ldots, x_p\} \qquad\qquad p = |x|$$

In forming the power x^p, all simple words are concatenated to form simple words. Every element of x^p is one possible concatenation of the p simple words of x.

$$x^p = \{x_1 x_2 \cdots x_p, x_2 x_1 \cdots x_p, \ldots\}$$

The power x^{p+1} is the concatenation of x^p with x. If a simple word of x^p is concatenated with a simple word x_j of x, the result is not a simple word, since x_j occurs twice in the concatenation. This concatenation is therefore not an element of x^{p+1}. Hence the power x^{p+1} is the zero set $\emptyset$. The element x is nilpotent of degree $p + 1$. It is stable with the stability index p. The closure $\hat{x}$ is given by

$$\hat{x} = \bigsqcup_{m=0}^{p} x^m$$

If the above assumption does not hold, the stability index p is less than the number of simple words in x.

8.5.7 SYSTEMS OF EQUATIONS

8.5.7.1 Solutions of systems of equations

Introduction : In the preceding sections different problems of path determination in graphs are shown to be reducible to a common path algebra. This path algebra is very similar to linear algebra with real numbers. Methods of linear algebra may therefore be transferred to the path algebra for graphs.

The formulation of problems in path algebra leads to systems of equations. If the coefficient matrix is stable, a system of equations has at least one solution. The theoretical foundations for systems of equations in path algebra are treated in this section.

System of equations : Let a directed graph with n vertices be given. The edge weights of the graph are arranged in a quadratic matrix **A**. A path algebra for a path problem in this graph leads to a system of n equations with the solution vector **x** depending on the vector **b** on the right-hand side. The notation for path algebra is simplified by omitting the operator $\circ$ for concatenations. With this simplification, the system of n equations with n variables is formulated as follows in matrix and vector notation and in element notation :

$$\mathbf{x} \;=\; \mathbf{A\,x} \sqcup \mathbf{b}$$

$$x_i \;=\; a_{i1}\,x_1 \sqcup a_{i2}\,x_2 \sqcup \ldots \sqcup a_{in}\,x_n \sqcup b_i \qquad\qquad i = 1,\ldots,n$$

Solutions : Let the matrix **A** of a system of equations $\mathbf{x} = \mathbf{Ax} \sqcup \mathbf{b}$ be stable. Then the system of equations has a solution $\mathbf{x} = \mathbf{A}^*\,\mathbf{b}$ with the closure $\mathbf{A}^*$ of **A**. If the matrix **A** is nilpotent, then the solution **x** is unique. If the matrix **A** is not nilpotent, several solutions may exist. If several solutions exist, then $\mathbf{x} = \mathbf{A}^*\,\mathbf{b}$ is the least solution.

Proof : Properties of solutions

The closure $\mathbf{A}^*$ of a matrix **A** with stability index p is $\mathbf{A}^* = \mathbf{I} \sqcup \mathbf{A} \sqcup \ldots \sqcup \mathbf{A}^p$. The system of equations $\mathbf{x} = \mathbf{A\,x} \sqcup \mathbf{b}$ has a solution $\mathbf{x} = \mathbf{A}^*\mathbf{b}$. This is proved by substituting **x** into $\mathbf{A\,x} \sqcup \mathbf{b}$ and using $\mathbf{A}^* = \mathbf{I} \sqcup \mathbf{AA}^*$:

$$\mathbf{A\,x} \sqcup \mathbf{b} \;=\; \mathbf{A\,A^*b} \sqcup \mathbf{b} \;=\; (\mathbf{I} \sqcup \mathbf{A\,A^*})\mathbf{b} \;=\; \mathbf{A^*b} \;=\; \mathbf{x}$$

Let $\mathbf{x}_0$ be a general solution of the system of equations. Then the following equation is obtained by repeatedly substituting $\mathbf{A\,x}_0 \sqcup \mathbf{b}$ for $\mathbf{x}_0$ on the right-hand side of the system of equations $\mathbf{x} = \mathbf{A\,x} \sqcup \mathbf{b}$:

$$\begin{aligned}
\mathbf{x}_0 &= \mathbf{A}\,\mathbf{x}_0 \sqcup \mathbf{b} \\
\mathbf{x}_0 &= \mathbf{A}\ \ (\mathbf{A}\,\mathbf{x}_0 \sqcup \mathbf{b}) \sqcup \mathbf{b} &&= \mathbf{A}^2\,\mathbf{x}_0 \sqcup (\mathbf{I} \sqcup \mathbf{A})\,\mathbf{b} \\
\mathbf{x}_0 &= \mathbf{A}^2\,(\mathbf{A}\,\mathbf{x}_0 \sqcup \mathbf{b}) \sqcup (\mathbf{I} \sqcup \mathbf{A})\,\mathbf{b} &&= \mathbf{A}^3\,\mathbf{x}_0 \sqcup (\mathbf{I} \sqcup \mathbf{A} \sqcup \mathbf{A}^2)\,\mathbf{b} \\
&\vdots \\
\mathbf{x}_0 &= \mathbf{A}^k\,\mathbf{x}_0 \sqcup \mathbf{A}^*\,\mathbf{b} = \mathbf{A}^k\,\mathbf{x}_0 \sqcup \mathbf{x} && \qquad k > p
\end{aligned}$$

If $\mathbf{A}$ is nilpotent, then $\mathbf{A}^k = \mathbf{0}$ for some $k > p$. Hence the solution $\mathbf{x}_0 = \mathbf{A}^*\mathbf{b} = \mathbf{x}$ is unique. If $\mathbf{A}$ is not nilpotent, then a solution $\mathbf{x} \neq \mathbf{x}_0$ may exist. Then $\mathbf{x}_0 = \mathbf{A}^k\mathbf{x}_0 \sqcup \mathbf{x}$ implies $\mathbf{x}_0 \sqsupseteq \mathbf{x}$, so that the solution $\mathbf{x}$ is contained in the general solution $\mathbf{x}_0$. Hence $\mathbf{x}$ is the least solution among all possible solutions $\mathbf{x}_0$ of the system of equations.

Example 1 : Solution of an equation with one variable

Let an equation with the boolean values a, b and the boolean variable x as well as the boolean operators $\vee$ and $\wedge$ for union and concatenation be given.

$$x = ax \sqcup b \qquad \Leftrightarrow \qquad x = (a \wedge x) \vee b$$

The closure $\hat{a}$ of the element a is equal to the unit element 1. The least solution x of the equation is given by :

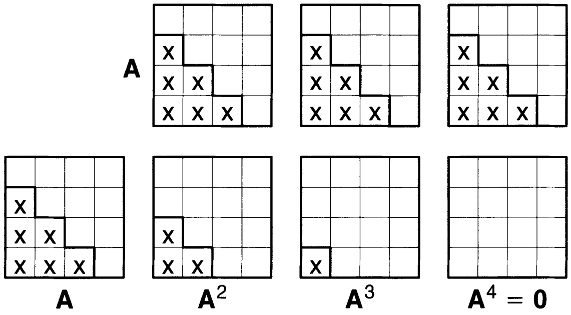
$$x = \hat{a}\,b \qquad \Leftrightarrow \qquad x = \hat{a} \wedge b = 1 \wedge b = b$$

If $b = 1$, then the least solution is $x = 1$. This solution is unique, since $x_0 = 0$ does not satisfy the equation. If $b = 0$, then the least solution is $x = 0$. This solution is not unique for $a = 1$, since $x_0 = 1$ also satisfies the equation.

Staggered system of equations : A system of equations $\mathbf{x} = \mathbf{A}\,\mathbf{x} \sqcup \mathbf{b}$ is said to be staggered if the matrix $\mathbf{A}$ is a lower triangular matrix with zero elements on and above the diagonal or an upper triangular matrix with zero elements on and below the diagonal. The solution $\mathbf{x}$ of a staggered system of equations is unique.

Proof : Unique solution of a staggered system of equations

Let the matrix $\mathbf{A}$ be a lower triangular matrix with n rows and n columns. It contains zero elements in all positions (i, j) with $j > i - 1$. The power $\mathbf{A}^s$ of $\mathbf{A}$ for $s \geq 1$ contains zero elements in all positions (i, j) with $j > i - s$. The power $\mathbf{A}^n$ of $\mathbf{A}$ is the zero matrix $\mathbf{0}$. Hence the matrix $\mathbf{A}$ is nilpotent. The properties of the powers of a lower triangular matrix $\mathbf{A}$ are illustrated for $n = 4$.



Since the lower triangular matrix **A** is nilpotent, the solution of the staggered system of equations $\mathbf{x} = \mathbf{A x} \sqcup \mathbf{b}$ is unique. The uniqueness of the solution for a staggered system of equations with an upper triangular matrix is proved analogously.

Example 2 : Staggered systems of equations

Staggered systems of equations for various path problems for acyclic graphs are treated in Sections 8.5.3 to 8.5.5. For a suitable numbering of the vertices of an acyclic graph, the system of equations is always staggered. A suitable numbering of the vertices may be determined by topological sorting using the method in Section 8.4.8. Thus the solutions of path problems for acyclic graphs are unique.

Equivalent systems of equations : Two systems of equations $\mathbf{x} = \mathbf{A x} \sqcup \mathbf{b}$ and $\mathbf{x} = \mathbf{C x} \sqcup \mathbf{d}$ with stable matrices **A** and **C** are said to be equivalent if their least solutions $\mathbf{A}^*\mathbf{b}$ and $\mathbf{C}^*\mathbf{d}$ are identical. The system $\mathbf{x} = \mathbf{A x} \sqcup \mathbf{b}$ may be transformed into an equivalent system $\mathbf{x} = \mathbf{C x} \sqcup \mathbf{d}$ by splitting the matrix $\mathbf{A} = \mathbf{Q} \sqcup \mathbf{S}$ into matrices **Q** and **S**.

$$\begin{aligned}
\mathbf{x} &= \mathbf{A x} \sqcup \mathbf{b} \quad \text{with} \quad \mathbf{A} = \mathbf{Q} \sqcup \mathbf{S} && \rightarrow \\
\mathbf{x} &= (\mathbf{Q} \sqcup \mathbf{S})\, \mathbf{x} \sqcup \mathbf{b} && \rightarrow \\
\mathbf{x} &= \mathbf{Q x} \sqcup (\mathbf{S x} \sqcup \mathbf{b}) && \rightarrow \\
\mathbf{x} &= \mathbf{Q}^* (\mathbf{S x} \sqcup \mathbf{b}) && \rightarrow \\
\mathbf{x} &= \mathbf{C x} \sqcup \mathbf{d} \quad \text{with} \quad \mathbf{C} = \mathbf{Q}^*\mathbf{S} \quad \text{and} \quad \mathbf{d} = \mathbf{Q}^*\mathbf{b}
\end{aligned}$$

Proof : Equivalence of systems of equations

The system $\mathbf{x} = \mathbf{A x} \sqcup \mathbf{b}$ has the least solution $\mathbf{A}^*\mathbf{b}$. The system $\mathbf{x} = \mathbf{C x} \sqcup \mathbf{d}$ has the least solution $\mathbf{C}^*\mathbf{d}$. With $\mathbf{A} = \mathbf{Q} \sqcup \mathbf{S}$, $\mathbf{C} = \mathbf{Q}^*\mathbf{S}$ and $\mathbf{d} = \mathbf{Q}^*\mathbf{b}$, the rules for closures in Section 8.2.6 yield the equivalence of the two systems :

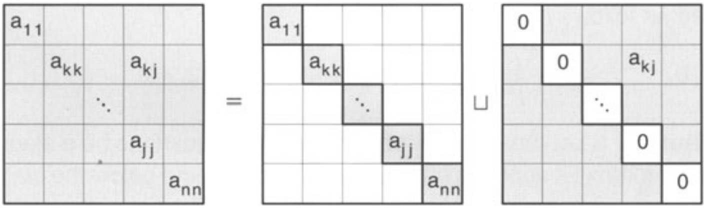
$$\mathbf{A}^*\mathbf{b} = (\mathbf{Q} \sqcup \mathbf{S})^*\mathbf{b} = (\mathbf{Q}^*\mathbf{S})^*\mathbf{Q}^*\mathbf{b} = \mathbf{C}^*\mathbf{d}$$

Example 3 : Construction of matrices with zero diagonal elements

Let a system of equations $\mathbf{x} = \mathbf{Ax} \sqcup \mathbf{b}$ be given. Let $\mathbf{A}$ be a stable matrix whose diagonal elements are not all zero. The matrix $\mathbf{A}$ is split into the diagonal matrix $\mathbf{D}$ and the matrix $\mathbf{S}$. The diagonal matrix $\mathbf{D}$ contains the diagonal elements a_{kk} of $\mathbf{A}$; all other elements of $\mathbf{A}$ are zero. The matrix $\mathbf{S}$ contains zero elements on the diagonal and the elements a_{kj} of $\mathbf{A}$ for $k \neq j$. The closure $\mathbf{D}^*$ of the diagonal matrix $\mathbf{D}$ is a diagonal matrix with the closures $\hat{a}_{kk}$ of the diagonal elements of $\mathbf{A}$. The decomposition $\mathbf{A} = \mathbf{D} \sqcup \mathbf{S}$ leads to an equivalent system of equations $\mathbf{x} = \mathbf{Cx} \sqcup \mathbf{d}$ with $\mathbf{C} = \mathbf{D}^*\mathbf{S}$ and $\mathbf{d} = \mathbf{D}^*\mathbf{b}$. The matrix $\mathbf{C}$ contains the elements $c_{kj} = \hat{a}_{kk}\, a_{kj}$ for $k \neq j$ and $c_{kk} = 0$. The vector $\mathbf{d}$ contains the elements $d_k = \hat{a}_{kk}\, b_k$.

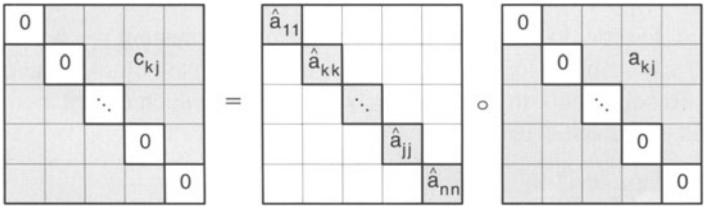
$$\mathbf{x} = \mathbf{Ax} \sqcup \mathbf{b}$$

$$\mathbf{A} = \mathbf{D} \sqcup \mathbf{S}$$



$$\mathbf{x} = \mathbf{Cx} \sqcup \mathbf{d}$$

$$\mathbf{C} = \mathbf{D}^*\mathbf{S}$$



$c_{kj} = \hat{a}_{kk}\, a_{kj}$ for $k \neq j$

The iterative methods of solution in Section 8.5.7.3 assume that all diagonal elements of the matrix $\mathbf{A}$ of a system of equations $\mathbf{x} = \mathbf{Ax} \sqcup \mathbf{b}$ are zero. If this condition is not satisfied, then this procedure is used to construct an equivalent system of equations $\mathbf{x} = \mathbf{Cx} \sqcup \mathbf{d}$ whose matrix $\mathbf{C}$ satisfies the condition.

8.5.7.2 Direct methods of solution

Introduction : The least solution of a system of equations may be determined directly if the system of equations is staggered. The solutions of the staggered system of equations are determined by forward or back substitution. If the system of equations is not staggered, it is transformed into an equivalent staggered system of equations by elimination. The best-known elimination method is the one due to Gauss. The reduction method due to Dijkstra is used for special path problems. The fundamentals of the direct methods of solution are treated in the following.

Forward substitution : Let the matrix **A** of the system of equations be a staggered lower triangular matrix : It contains only zero elements on and above the diagonal. The system of equations is solved by forward substitution. The variables are calculated as follows :

$$x_1 = b_1 \qquad x_k = \bigsqcup_{j=1}^{k-1} a_{kj}\, x_j \sqcup b_k \qquad k=2,\ldots,n$$

Back substitution : Let the matrix **A** of the system of equations be a staggered upper triangular matrix : It contains only zero elements on and below the diagonal. The system of equations is solved by back substitution. The variables are calculated as follows :

$$x_n = b_n \qquad x_k = \bigsqcup_{j=k+1}^{n} a_{kj}\, x_j \sqcup b_k \qquad k=n-1,\ldots,1$$

Elimination : In order to eliminate a variable x_k from the system $\mathbf{x} = \mathbf{Ax} \sqcup \mathbf{b}$, the k-th equation is first solved for x_k. Then x_k is eliminated in the other equations by substitution. To solve the k-th equation for x_k, the terms which do not involve x_k are combined into a value c_k.

$$\begin{aligned} x_k &= \bigsqcup_{j} a_{kj}\, x_j \sqcup b_k \qquad \Leftrightarrow \\ x_k &= a_{kk}\, x_k \sqcup c_k \qquad \text{with} \qquad c_k = \bigsqcup_{j \neq k} a_{kj}\, x_j \sqcup b_k \end{aligned}$$

The equation $x_k = a_{kk}\, x_k \sqcup c_k$ has a least solution if the element a_{kk} is stable, so that the closure $\hat{a}_{kk}$ exists. The least solution is :

$$\begin{aligned} x_k &= \hat{a}_{kk}\, c_k \qquad \Leftrightarrow \\ x_k &= \bigsqcup_{j \neq k} \hat{a}_{kk}\, a_{kj}\, x_j \sqcup \hat{a}_{kk}\, b_k \end{aligned} \tag{1}$$

To eliminate the variable x_k in the i-th equation, the terms which do not involve x_k are combined into a value c_i :

$$\begin{aligned} x_i &= \bigsqcup_{j} a_{ij}\, x_j \sqcup b_i \qquad \Leftrightarrow \\ x_i &= a_{ik}\, x_k \sqcup c_i \qquad \text{with} \qquad c_i = \bigsqcup_{j \neq k} a_{ij}\, x_j \sqcup b_i \end{aligned}$$

The solution for x_k is substituted into the i-th equation $x_i = a_{ik} x_k \sqcup c_i$. This substitution eliminates x_k in the i-th equation :

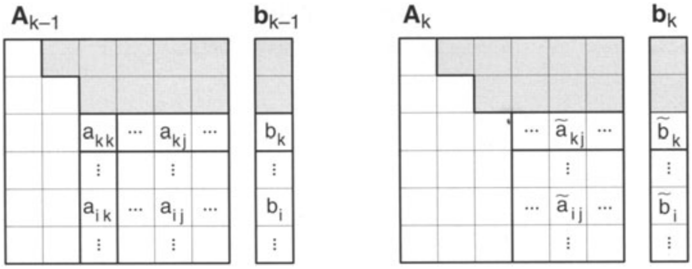
$$\begin{aligned} x_i &= a_{ik}\ \hat{a}_{kk}\ c_k \sqcup c_i \qquad \Leftrightarrow \\ x_i &= \bigsqcup_{j \neq k} (a_{ij} \sqcup a_{ik}\ \hat{a}_{kk}\ a_{kj})\ x_j \sqcup (b_i \sqcup a_{ik}\ \hat{a}_{kk}\ b_k) \end{aligned} \tag{2}$$

In performing the elimination, it is assumed that the element a_{kk} is stable, so that the closure $\hat{a}_{kk}$ exists. If this is not the case, the elimination cannot be performed. The closure $\hat{a}_{kk}$ of the element a_{kk} is calculated as a union of powers of a_{kk} according to the definition in Section 8.5.6. For various path algebras, the closure $\hat{a}_{kk}$ is known a priori and need not be calculated explicitly.

Gaussian elimination method : Let a system $\mathbf{x} = \mathbf{A}_0\,\mathbf{x} \sqcup \mathbf{b}_0$ with n variables be given. It is transformed into a staggered system of equations with an upper triangular matrix in n consecutive steps.

$$\mathbf{x} \ = \ \mathbf{A}_k\,\mathbf{x} \sqcup \mathbf{b}_k \qquad\qquad k = 1,...,n$$

In every step $k = 1,...,n$, the variable x_k is eliminated in the equations $i = k,...,n$ of the system $\mathbf{x} = \mathbf{A}_{k-1}\,\mathbf{x} \sqcup \mathbf{b}_{k-1}$. The formulas for the elements of the matrix $\mathbf{A}_k$ and the vector $\mathbf{b}_k$ are compiled below.



$$\begin{aligned} \tilde{a}_{kj} &= \hat{a}_{kk}\ a_{kj} && & j &= k+1,...,n \\ \tilde{a}_{ij} &= a_{ij} \sqcup a_{ik}\ \hat{a}_{kk}\ a_{kj} &&= a_{ij} \sqcup a_{ik}\ \tilde{a}_{kj} & i, j &= k+1,...,n \\ \tilde{b}_k &= \hat{a}_{kk}\ b_k && & & \\ \tilde{b}_i &= b_i \sqcup a_{ik}\ \hat{a}_{kk}\ b_k &&= b_i \sqcup a_{ik}\ \tilde{b}_k & i &= k+1,...,n \end{aligned}$$

The matrices $\mathbf{A}_k$ and the vectors $\mathbf{b}_k$ in the steps $k = 1,...,n$ are not explicitly constructed in the algorithms. Instead, the matrix and the vector of the original system of equations are repeatedly overwritten. In the k-th step, the elements are overwritten as follows :

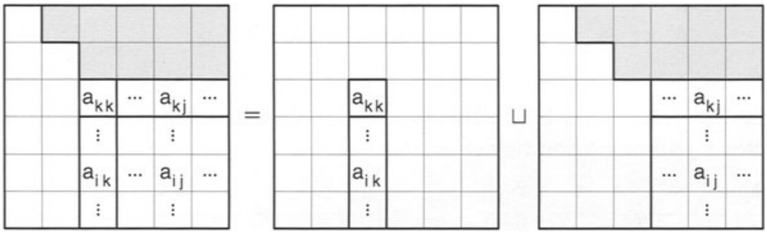
$$\begin{aligned} a_{kj} &\leftarrow \hat{a}_{kk}\ a_{kj} & j &= k+1,...,n \\ a_{ij} &\leftarrow a_{ij} \sqcup a_{ik}\ a_{kj} & i, j &= k+1,...,n \\ b_k &\leftarrow \hat{a}_{kk}\ b_k & & \\ b_i &\leftarrow b_i \sqcup a_{ik}\ b_k & i &= k+1,...,n \end{aligned}$$

The Gaussian elimination method assumes that in each step the diagonal element a_{kk} is stable, so that the closure $\hat{a}_{kk}$ exists. If this is not the case, the elimination process fails. Upon successful completion of the elimination process, the system of equations is staggered, and the variables may be determined by back substitution. The solution reached by Gaussian elimination is always the least solution.

Proof : Gaussian elimination leads to the least solution.

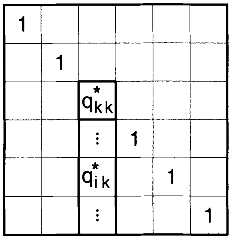
According to Section 8.5.7.1, a system of equations $\mathbf{x} = \mathbf{A}_{k-1}\,\mathbf{x} \sqcup \mathbf{b}_{k-1}$ is transformed into an equivalent system of equations $\mathbf{x} = \mathbf{A}_k \mathbf{x} \sqcup \mathbf{b}_k$ by splitting the matrix $\mathbf{A}_{k-1}$ into $\mathbf{Q}_k$ and $\mathbf{S}_k$. The matrix $\mathbf{A}_k = \mathbf{Q}_k^* \, \mathbf{S}_k$ and the vector $\mathbf{b}_k = \mathbf{Q}_k^* \, \mathbf{b}_{k-1}$ are calculated. For the Gaussian elimination method, the matrix $\mathbf{A}_{k-1}$ is split as follows in step k :

$$\mathbf{A}_{k-1} = \mathbf{Q}_k \sqcup \mathbf{S}_k$$



The matrix $\mathbf{Q}_k$ contains the elements of $\mathbf{A}_{k-1}$ on and below the diagonal in column k and zero elements everywhere else. A power $\mathbf{Q}_k^s$ for $s \geq 1$ has the same structure as the matrix $\mathbf{Q}_k$. The element in the diagonal position (k,k) is a_{kk}^s. The element in position (i, k) with $i > k$ is $a_{ik}\,a_{kk}^{s-1}$. All remaining elements are zero elements. The closure $\mathbf{Q}_k^*$ exists if the diagonal element a_{kk} is stable, so that its closure $\hat{a}_{kk}$ exists.

$$\mathbf{Q}_k^* = \mathbf{I} \sqcup \mathbf{Q}_k \sqcup \mathbf{Q}_k^2 \sqcup \ldots$$



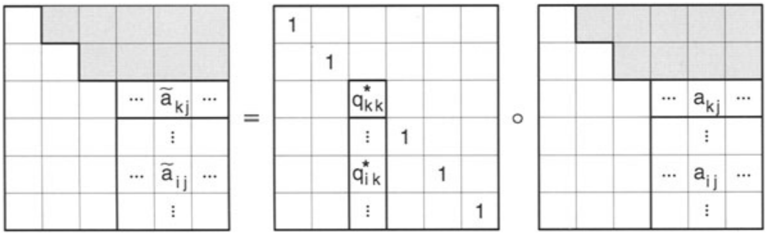
1 := unit element

$$q_{kk}^* = \hat{a}_{kk}$$

$$q_{ik}^* = a_{ik}\,\hat{a}_{kk}$$

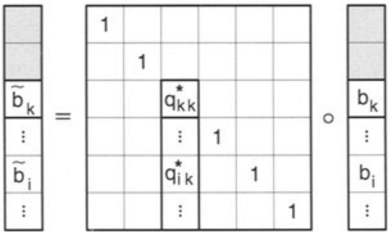
The matrix $\mathbf{A}_k$ is the product $\mathbf{Q}^*_k\ \mathbf{S}_k$. The vector $\mathbf{b}_k$ is the product $\mathbf{Q}^*_k\ \mathbf{b}_{k-1}$. The elements of the matrix $\mathbf{A}_k$ in column k on and below the diagonal are zero elements. This corresponds to the elimination of the variable x_k in equations k to n of the system of equations $\mathbf{x} = \mathbf{A}_{k-1}\,\mathbf{x} \sqcup \mathbf{b}_{k-1}$.

$$\mathbf{A}_k = \mathbf{Q}^*_k\ \mathbf{S}_k$$



$$\begin{array}{llllll} \tilde{a}_{kj} & = q^*_{kk}\ a_{kj} & = \hat{a}_{kk}\ a_{kj} & j & = & k+1,\ldots,n \\ \tilde{a}_{ij} & = a_{ij} \sqcup q^*_{ik}\ a_{kj} & = a_{ij} \sqcup a_{ik}\,\hat{a}_{kk} a_{kj} & i,j & = & k+1,\ldots,n \end{array}$$

$$\mathbf{b}_k = \mathbf{Q}^*_k\ \mathbf{b}_{k-1}$$



$$\begin{array}{lllll} \tilde{b}_k & = q^*_{kk}\ b_{kj} & = \hat{a}_{kk}\ b_{kj} & & \\ \tilde{b}_i & = b_i \sqcup q^*_{ik}\ b_k & = b_i \sqcup a_{ik}\,\hat{a}_{kk} b_{kj} & i = & k+1,\ldots,n \end{array}$$

The systems of equations $\mathbf{x} = \mathbf{A}_k \mathbf{x} \sqcup \mathbf{b}_k$ for $k = 0,1,\ldots,n$ in the Gaussian elimination method are equivalent and therefore have the same least solution. Upon completion of the elimination process, the system of equations is staggered. According to Section 8.5.7.1, the solution of a staggered system of equations is unique. Thus Gaussian elimination yields the least solution of a given system of equations.

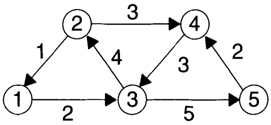
Simplified Gaussian elimination method : For a unitarily stable path algebra, every closure of an element is equal to the unit element. The Gaussian elimination method may therefore be simplified. It is carried out in the steps $k = 1,\ldots,n-1$. In the k-th step, the elements of the matrix and the vector are overwritten as follows :

$$\begin{array}{llll} a_{ij} \leftarrow a_{ij} \sqcup a_{ik}\ a_{kj} & & i,j & = k+1,\ldots,n \\ b_i \leftarrow b_i \sqcup a_{ik}\ b_k & & i & = k+1,\ldots,n \end{array}$$

Example 1 : Minimal path lengths

Let the illustrated cyclic graph with positive integer edge lengths be given. The minimal path lengths from each vertex of the graph to vertex 3 are to be determined. The system of equations for this path problem is specified.

cyclic graph



system of equations $\mathbf{x} = \mathbf{A}\mathbf{x} \sqcup \mathbf{b}$ with $\mathbf{b} = \mathbf{e}_3$

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|}\hline n & n & 2 & n & n \\ \hline 1 & n & n & 3 & n \\ \hline n & 4 & n & n & 5 \\ \hline n & n & 3 & n & n \\ \hline n & n & n & 2 & n \\ \hline \end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline \end{array} \sqcup \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline n \\ \hline n \\ \hline \end{array} \qquad \begin{array}{l} n := \infty \\ e := 0 \end{array}$$

The path algebra for minimal path lengths is unitarily stable, so that the simplified elimination method may be applied. The individual elimination steps are illustrated. The binary operations min and + from Section 8.5.4.1 are used for union and concatenation. For selected elements, the necessary operations are shown. The zero elements generated in the course of the elimination are not explicitly shown. Instead, the old elements are shaded.

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 2 & n & n \\ \hline 1 & n & \textcircled{3} & 3 & n \\ \hline n & 4 & n & n & 5 \\ \hline n & n & 3 & n & n \\ \hline n & n & n & 2 & n \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline n \\ \hline n \\ \hline \end{array}$$

elimination k = 1

$$a_{23} \leftarrow a_{23} \sqcup a_{21}\ a_{13}$$

$$a_{23} \leftarrow \min\{n, 1 + 2\} = 3$$

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 2 & n & n \\ \hline 1 & n & 3 & 3 & n \\ \hline n & 4 & 7 & \textcircled{7} & 5 \\ \hline n & n & 3 & n & n \\ \hline n & n & n & 2 & n \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline n \\ \hline n \\ \hline \end{array}$$

elimination k = 2

$$a_{34} \leftarrow a_{34} \sqcup a_{32}\ a_{24}$$

$$a_{34} \leftarrow \min\{n, 4 + 3\} = 7$$

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 2 & n & n \\ \hline 1 & n & 3 & 3 & n \\ \hline n & 4 & 7 & 7 & 5 \\ \hline n & n & 3 & 10 & 8 \\ \hline n & n & n & 2 & n \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline 3 \\ \hline n \\ \hline \end{array}$$

elimination k = 3

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 2 & n & n \\ \hline 1 & n & 3 & 3 & n \\ \hline n & 4 & 7 & 7 & 5 \\ \hline n & n & 3 & 10 & 8 \\ \hline n & n & n & 2 & 10 \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline 3 \\ \hline 5 \\ \hline \end{array}$$

elimination k = 4

back substitution of the staggered equations

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|}\hline & n & 2 & n & n \\ \hline & & 3 & 3 & n \\ \hline & & & 7 & 5 \\ \hline & & & & 8 \\ \hline & & & & \\ \hline \end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline x_5 \\ \hline \end{array} \sqcup \begin{array}{|c|}\hline n \\ \hline n \\ \hline e \\ \hline 3 \\ \hline 5 \\ \hline \end{array}$$

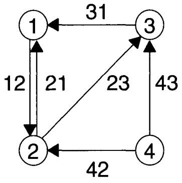
solution

$$\mathbf{x} = \begin{array}{|c|}\hline 2 \\ \hline 3 \\ \hline e \\ \hline 3 \\ \hline 5 \\ \hline \end{array}$$

The calculation shows that the vertex 3 is reachable from each vertex of the graph.

Example 2 : Elementary paths and cycles

Let the illustrated graph with vertices labeled by digits be given, and let the edges be labeled accordingly, as in Section 8.5.5.5. The elementary paths and cycles from each vertex of the graph to vertex 2 are to be determined. The system of equations for this path problem is shown.



system of equations $\mathbf{x} = \mathbf{A}\,\mathbf{x} \sqcup \mathbf{b}$ with $\mathbf{b} = \mathbf{e}_2$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 & \{12\} & 0 & 0 \\ \{21\} & 0 & \{23\} & 0 \\ \{31\} & 0 & 0 & 0 \\ 0 & \{42\} & \{43\} & 0 \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} \sqcup \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \begin{aligned} 0 &:= \emptyset \\ 1 &:= \{\lambda\} \end{aligned}$$

The path problem for elementary paths and cycles is solved using the literal algebra for simple words from Section 8.5.5.5. In order to determine not only elementary paths but also elementary cycles, the set of simple words is extended to include words whose first and last characters coincide. If an element a contains words with the same character z as the first and last character, the concatenation $a^2 = a \circ a$ would lead to a word with three characters z which does not belong to the extended set of simple words. The element a is therefore nilpotent of degree 2. According to Section 8.5.6 its closure is $\hat{a} = \{\lambda\} \sqcup a$. The path algebra for elementary paths and cycles is not unitarily stable, so that the general elimination method must be applied. The individual elimination steps are illustrated. For selected elements, the necessary operations are specified. The zero elements generated in the course of the elimination are not explicitly shown. Instead, the old elements are shaded.

elimination $k = 1 :\ a_{11} = 0 \quad \hat{a}_{11} = \{\lambda\} = 1$

$$\mathbf{A} = \begin{bmatrix} 0 & \{12\} & 0 & 0 \\ \{21\} & \{212\} & \{23\} & 0 \\ \{31\} & \{312\} & 0 & 0 \\ 0 & \{42\} & \{43\} & 0 \end{bmatrix} \qquad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$$

$$a_{22} \leftarrow a_{22} \sqcup a_{21}\ a_{12} = 0 \cup \{21\} \circ \{12\} = \{212\}$$

elimination $k = 2 : \; a_{22} = \{212\} \quad \hat{a}_{22} = \{\lambda, 212\}$

$$\begin{array}{lcl} a_{23} \leftarrow \hat{a}_{22}\, a_{23} & = \{\lambda, 212\} & \circ \{23\} = \{23\} \\ a_{33} \leftarrow a_{33} \sqcup a_{32}\, a_{23} & = \emptyset \cup \{312\} & \circ \{23\} = \{3123\} \\ a_{43} \leftarrow a_{43} \sqcup a_{42}\, a_{23} & = \{43\} \cup \{42\} & \circ \{23\} = \{43,423\} \end{array}$$

$$\mathbf{A} = \begin{array}{|c|c|c|c|}\hline 0 & \{12\} & 0 & 0 \\ \hline \{21\} & \{212\} & \{23\} & 0 \\ \hline \{31\} & \{312\} & \{3123\} & 0 \\ \hline 0 & \{42\} & \{43,423\} & 0 \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline 0 \\ \hline \{\lambda,212\} \\ \hline \{312\} \\ \hline \{42\} \\ \hline \end{array}$$

elimination $k = 3 : \; a_{33} = \{3123\} \quad \hat{a}_{33} = \{\lambda, 3123\}$

$$b_3 \leftarrow \hat{a}_{33}\, b_3 = \{\lambda,3123\} \circ \{312\} = \{312\}$$

$$\mathbf{A} = \begin{array}{|c|c|c|c|}\hline 0 & \{12\} & 0 & 0 \\ \hline \{21\} & \{212\} & \{23\} & 0 \\ \hline \{31\} & \{312\} & \{3123\} & 0 \\ \hline 0 & \{42\} & \{43,423\} & 0 \\ \hline \end{array} \qquad \mathbf{b} = \begin{array}{|c|}\hline 0 \\ \hline \{\lambda,212\} \\ \hline \{312\} \\ \hline \{42,4312\} \\ \hline \end{array}$$

elimination $k = 4 : \; a_{44} = 0 \qquad \hat{a}_{44} = \{\lambda\} = 1$ no change

back substitution of the staggered equations

$$\begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline & \{12\} & 0 & 0 \\ \hline & & \{23\} & 0 \\ \hline & & & 0 \\ \hline & & & \\ \hline \end{array} \circ \begin{array}{|c|}\hline x_1 \\ \hline x_2 \\ \hline x_3 \\ \hline x_4 \\ \hline \end{array} \sqcup \begin{array}{|c|}\hline 0 \\ \hline \{\lambda,212\} \\ \hline \{312\} \\ \hline \{42,4312\} \\ \hline \end{array}$$

$$\mathbf{x} = \begin{array}{|c|}\hline \{12\} \\ \hline \{\lambda,212,2312\} \\ \hline \{312\} \\ \hline \{42,4312\} \\ \hline \end{array} \qquad \begin{array}{l} x_2 = b_2 \sqcup a_{23}\, x_3 \\ x_2 = \{\lambda, 212\} \cup \{23\} \circ \{312\} \\ x_2 = \{\lambda, 212, 2312\} \end{array}$$

The calculation shows that the vertex 2 is reachable from each vertex of the graph. The vertex 2 itself lies on two elementary cycles with the vertex sequences 2,1,2 and 2,3,1,2.

Dijkstra's reduction method : Let a system of equations $\mathbf{x} = \mathbf{A}\,\mathbf{x} \sqcup \mathbf{b}$ with n variables for a unitarily stable path algebra with a totally ordered weight set be given. By virtue of the total ordering, the vector **b** of the system of equations contains a greatest element b_k. The variable x_k of the solution vector **x** is identical with the greatest element b_k (see the following proof). Substituting $x_k = b_k$ into the system of equations and deleting the k-th equation leads to a reduced system of equations. By repeated reduction, all variables of the system of equations are determined. This method, which is due to Dijkstra, is executed in the following steps :

1. Set an index set for the n variables to $M = \{1,2,...,n\}$.

2. Find a greatest element b_k in the vector **b**.

   $b_i \sqsubseteq b_k$ for all $i \in M$ and $k \in M$

3. Remove k from the index set M and calculate the vector **b** of the reduced system of equations.

   $b_i \leftarrow b_i \sqcup a_{ik}\, b_k$ for all $i \in M$

4. Repeat steps 2 and 3 if the index set M is not empty. If it is empty, the vector **b** is equal to the solution vector **x**.

Dijkstra's method yields the least solution of the system of equations.

Proof : Dijkstra's reduction method yields the least solution.

The least solution of the system of equations $\mathbf{x} = \mathbf{A}\,\mathbf{x} \sqcup \mathbf{b}$ is $\mathbf{x} = \mathbf{A}^*\mathbf{b}$. The variable x_k is calculated with the elements a^*_{ki} of the closure $\mathbf{A}^*$ as follows :

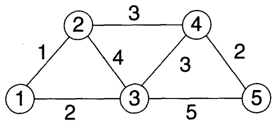
$$x_k = \bigsqcup_{i=1}^{n} a^*_{ki}\, b_i = a^*_{kk}\, b_k \sqcup \bigsqcup_{i \neq k} a^*_{ki}\, b_i$$

In a unitarily stable path algebra, the unit element 1_Z is the greatest element in a totally ordered weight set Z. Hence every diagonal element of the closure $\mathbf{A}^*$ is $a^*_{kk} = 1_Z$, and for every non-diagonal element $a^*_{ki} \sqsubseteq 1_Z$. This implies $a^*_{kk}\, b_k = b_k$ and $a^*_{ki}\, b_i \sqsubseteq b_i$. If b_k is the greatest element of **b**, then $b_i \sqsubseteq b_k$, and therefore $a^*_{ki}\, b_i \sqsubseteq a^*_{kk}\, b_k$. This implies $x_k = a^*_{kk}\, b_k = b_k$. Hence the greatest element b_k of **b** is identical with the variable x_k of the least solution **x**.

Example 3 : Minimal path lengths in a simple graph

A simple graph with positive integer edge lengths is illustrated below. The minimal path lengths between each vertex of the graph and the vertex 4 are to be determined. The system of equations for this path problem is shown. Since the graph is simple, its matrix is symmetric.

simple graph



system of equations $\mathbf{x} = \mathbf{A}\mathbf{x} \sqcup \mathbf{b}$ with $\mathbf{b} = \mathbf{e}_4$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} n & 1 & 2 & n & n \\ 1 & n & 4 & 3 & n \\ 2 & 4 & n & 3 & 5 \\ n & 3 & 3 & n & 2 \\ n & n & 5 & 2 & \infty \end{bmatrix} \circ \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} \sqcup \begin{bmatrix} n \\ n \\ n \\ e \\ n \end{bmatrix} \qquad \begin{aligned} n &:= \infty \\ e &:= 0 \end{aligned}$$

The path algebra for minimal path lengths is unitarily stable, and its weight set is totally ordered. Thus Dijkstra's method may be applied. The individual steps of the method are shown. The binary operations min and + from Section 8.5.4.1 are used for union and concatenation. In Example 1 of Section 8.5.6, the inclusion $\sqsubseteq$ in the weight set for minimal path lengths is shown to correspond to the relation $\geq$. The greatest element b_k in a vector **b** is therefore a minimal value. Thus a step k consists of the following operations :

greatest element	:	$b_k = \min\{\ldots, b_i, \ldots\}$
reduction	:	$b_i \leftarrow \min\{b_i, a_{ik} + b_k\}$

The iterative calculation of the vector $\mathbf{b}^T$ is shown. For each step k, the greatest elements determined in this and the preceding steps are shaded. The shaded elements of the vector are no longer considered in subsequent steps.

∞	∞	∞	0	∞	$k = 4$	$b_4 = 0$	$b_i \leftarrow \min\{b_i, a_{i4} + 0\}$
∞	3	3	0	2	$k = 5$	$b_5 = 2$	$b_i \leftarrow \min\{b_i, a_{i5} + 2\}$
∞	3	3	0	2	$k = 3$	$b_3 = 3$	$b_i \leftarrow \min\{b_i, a_{i3} + 3\}$
5	3	3	0	2	$k = 2$	$b_2 = 3$	$b_i \leftarrow \min\{b_i, a_{i2} + 3\}$
4	3	3	0	2	$k = 1$	$b_1 = 4$	solution vector $\mathbf{x}^T$

The calculation shows that the simple graph contains paths between each of its vertices and the vertex 4. Between vertices 1 and 4, there is a path of minimal length 4.

8.5.7.3 Iterative methods of solution

Introduction : Various iterative methods have been developed for solving systems of equations. Such methods form the basis for powerful algorithms in graph theory. In formulating these methods, it is assumed that the matrix of the system of equations contains zero elements on the diagonal. The simplest iterative methods are the Jacobi method, the Gauss-Seidel method and the forward and back substitution method. They form a class of methods and are treated in the following in generalized form.

General iteration : The general iteration for solving a system of equations $\mathbf{x} = \mathbf{A}\mathbf{x} \sqcup \mathbf{b}$ consists of the following steps :

$$\begin{array}{lllll} \text{initial values} & \mathbf{x}_0 & = & \mathbf{b} & \\ \text{iteration} & \mathbf{x}_{k+1} & = & \mathbf{M}\mathbf{x}_k \sqcup \mathbf{N}\mathbf{b} & k = 0,1,... \\ \text{termination} & \mathbf{x}_{k+1} & = & \mathbf{x}_k & \end{array}$$

The vector $\mathbf{b}$ is conveniently chosen as the initial vector $\mathbf{x}_0$ for the iteration, since every solution $\mathbf{x}$ of the system of equations $\mathbf{x} = \mathbf{A}\mathbf{x} \sqcup \mathbf{b}$ contains the vector $\mathbf{b}$. In each iteration $k = 0,1,...$ an iterated vector $\mathbf{x}_{k+1}$ is calculated from the vector $\mathbf{x}_k$ and the vector $\mathbf{b}$ using the matrices $\mathbf{M}$ and $\mathbf{N}$. The iteration is terminated if two consecutive iterated vectors $\mathbf{x}_{k+1}$ and $\mathbf{x}_k$ coincide. The matrices $\mathbf{M}$ and $\mathbf{N}$ of the iteration procedure must be chosen such that the iteration yields the least solution $\mathbf{x} = \mathbf{A}^* \mathbf{b}$ of the system of equations. The relevant conditions are derived in the following.

Conditions : The iteration with the general rule defined above yields iterated vectors of the following form :

$$\begin{array}{lllll} \mathbf{x}_0 & = & \mathbf{b} & & \\ \mathbf{x}_1 & = & \mathbf{M}\mathbf{x}_0 \sqcup \mathbf{N}\mathbf{b} & = & \mathbf{M}\mathbf{b} \sqcup \mathbf{N}\mathbf{b} \\ \mathbf{x}_2 & = & \mathbf{M}\mathbf{x}_1 \sqcup \mathbf{N}\mathbf{b} & = & \mathbf{M}^2\mathbf{b} \sqcup (\mathbf{I} \sqcup \mathbf{M})\mathbf{N}\mathbf{b} \\ \mathbf{x}_{k+1} & = & \mathbf{M}\mathbf{x}_k \sqcup \mathbf{N}\mathbf{b} & = & \mathbf{M}^{k+1}\mathbf{b} \sqcup (\mathbf{I} \sqcup \mathbf{M} \sqcup \mathbf{M}^2 \sqcup ... \sqcup \mathbf{M}^k)\mathbf{N}\mathbf{b} \end{array}$$

The iteration can only yield a solution if the matrix $\mathbf{M}$ is stable. If the stability index of the matrix $\mathbf{M}$ is p, the vector $\mathbf{x}_{p+1}$ is obtained as :

$$\mathbf{x}_{p+1} = \mathbf{M}^{p+1}\mathbf{b} \sqcup \mathbf{M}^*\mathbf{N}\mathbf{b}$$

The vector $\mathbf{x}_{p+1}$ contains the least solution $\mathbf{x} = \mathbf{A}^* \mathbf{b}$ of the system of equations if the product $\mathbf{M}^*\mathbf{N}$ is equal to the closure $\mathbf{A}^*$.

$$\mathbf{x}_{p+1} = \mathbf{M}^{p+1}\mathbf{b} \sqcup \mathbf{A}^*\mathbf{b} \quad \text{with} \quad \mathbf{M}^*\mathbf{N} = \mathbf{A}^*$$

The vector $\mathbf{x}_{p+1}$ is the least solution $\mathbf{x} = \mathbf{A}^* \mathbf{b}$ of the system of equations only if $\mathbf{M}^{p+1}\,\mathbf{b} \sqsubseteq \mathbf{A}^*\,\mathbf{b}$. Since $\mathbf{b} \sqsubseteq \mathbf{A}^*\,\mathbf{b}$, the rules for the inclusion and the closure lead to :

$$\mathbf{M}^{p+1}\,\mathbf{b} \sqsubseteq \mathbf{M}^{p+1}\mathbf{A}^*\,\mathbf{b} = \mathbf{M}^{p+1}\,\mathbf{M}^*\mathbf{N}\mathbf{b} \sqsubseteq \mathbf{M}^*\mathbf{N}\mathbf{b} = \mathbf{A}^*\,\mathbf{b}$$

Hence the general iteration procedure yields the least solution $\mathbf{x} = \mathbf{A}^*\,\mathbf{b}$ of the system of equations if the matrix $\mathbf{M}$ is stable and the product $\mathbf{M}^*\mathbf{N}$ is identical with the closure $\mathbf{A}^*$. If the stability index of the matrix $\mathbf{M}$ is p, then p + 1 iterations are required to determine the least solution.

Jacobi method : The Jacobi method is the simplest method for solving a system of equations. The iteration is carried out according to the following rule :

iteration $\qquad \mathbf{x}_{k+1} = \mathbf{A}\,\mathbf{x}_k \sqcup \mathbf{b}$

The iteration procedure is a special case of the general iteration procedure and satisfies the conditions for the least solution of the system of equations :

matrices $\qquad \mathbf{M} = \mathbf{A} \qquad \mathbf{N} = \mathbf{I}$

condition $\qquad \mathbf{M}^*\mathbf{N} = \mathbf{A}^*\mathbf{I} = \mathbf{A}^*$

Gauss-Seidel method : In the Gauss-Seidel method, the matrix **A** of the system of equations is represented as the union of a lower triangular matrix **L** and an upper triangular matrix **R**. The lower triangular matrix **L** contains zero elements on and above the diagonal. The upper triangular matrix **R** contains zero elements on and below the diagonals. The system of equations to be solved may thus be formulated as follows :

$$\mathbf{x} = \mathbf{A}\mathbf{x} \sqcup \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x} = (\mathbf{L} \sqcup \mathbf{R})\mathbf{x} \sqcup \mathbf{b} \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{L}\mathbf{x} \sqcup \mathbf{R}\mathbf{x} \sqcup \mathbf{b}$$

The Gauss-Seidel iteration is carried out according to the following rule :

iteration $\qquad \mathbf{x}_{k+1} = \mathbf{L}\mathbf{x}_{k+1} \sqcup \mathbf{R}\mathbf{x}_k \sqcup \mathbf{b}$

This iteration procedure corresponds to a staggered system of equations with the matrix **L** and the solution vector $\mathbf{x}_{k+1}$. To reduce it to the general iteration procedure, the solution vector $\mathbf{x}_{k+1}$ is written as a function of $\mathbf{x}_k$ and $\mathbf{b}$ using the closure $\mathbf{L}^*$. With the rules for closures in Section 8.2.5, the iteration procedure is shown to satisfy the conditions for the least solution of the system of equations :

iteration $\qquad \mathbf{x}_{k+1} = \mathbf{L}^*\,(\mathbf{R}\mathbf{x}_k \sqcup \mathbf{b}) = \mathbf{L}^*\,\mathbf{R}\mathbf{x}_k \sqcup \mathbf{L}^*\mathbf{b}$

matrices $\qquad \mathbf{M} = \mathbf{L}^*\mathbf{R} \qquad \mathbf{N} = \mathbf{L}^*$

condition $\qquad \mathbf{M}^*\mathbf{N} = (\mathbf{L}^*\mathbf{R})^*\,\mathbf{L}^* = (\mathbf{L} \sqcup \mathbf{R})^* = \mathbf{A}^*$

Forward and back substitution method : Like the Gauss-Seidel method, this method uses a decomposition of the matrix **A** of the system of equations into a union of a lower triangular matrix **L** and an upper triangular matrix **R**. The iteration is carried out according to the following rules :

$$\begin{array}{lll} \text{iteration} & \mathbf{y}_{k+1} & = \mathbf{R}\,\mathbf{y}_{k+1} \sqcup \mathbf{x}_k \sqcup \mathbf{b} \\ & \mathbf{x}_{k+1} & = \mathbf{L}\,\mathbf{x}_{k+1} \sqcup \mathbf{y}_{k+1} \end{array}$$

The first equation corresponds to a system of equations with the matrix **R** and the solution vector $\mathbf{y}_{k+1}$, which is solved by back substitution. The second equation corresponds to a system of equations with the matrix **L** and the solution vector $\mathbf{x}_{k+1}$, which is solved by forward substitution. In order to reduce the iteration procedure to the general iteration procedure, the solution vectors $\mathbf{y}_{k+1}$ and $\mathbf{x}_{k+1}$ are specified using the closures $\mathbf{R}^*$ and $\mathbf{L}^*$, and the first equation is substituted into the second equation. By the rules for closures in Section 8.2.5, the iteration procedure satisfies the required condition.

$$\begin{array}{lll} \text{iteration} & \mathbf{y}_{k+1} & = \mathbf{R}^*(\mathbf{x}_k \sqcup \mathbf{b}) \\ & \mathbf{x}_{k+1} & = \mathbf{L}^*\,\mathbf{y}_{k+1} \\ & \mathbf{x}_{k+1} & = \mathbf{L}^*\,\mathbf{R}^*(\mathbf{x}_k \sqcup \mathbf{b}) \;=\; \mathbf{L}^*\,\mathbf{R}^*\,\mathbf{x}_k \sqcup \mathbf{L}^*\,\mathbf{R}^*\,\mathbf{b} \\ \text{matrices} & \mathbf{M} & = \mathbf{N} = \mathbf{L}^*\,\mathbf{R}^* \\ \text{condition} & \mathbf{M}^*\mathbf{N} & = (\mathbf{L}^*\,\mathbf{R}^*)^*(\mathbf{L}^*\,\mathbf{R}^*) = \;(\mathbf{L}^*\,\mathbf{R}^*)^* \;=(\mathbf{L} \sqcup \mathbf{R})^* \;= \mathbf{A}^* \end{array}$$

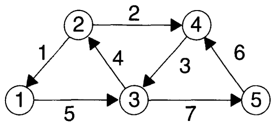
Number of iterations : Every iterative method yields the least solution $\mathbf{x} = \mathbf{A}^*\mathbf{b}$ of the system of equations after at most $p+1$ iterations, where p is the stability index of the matrix **M**. An upper bound for the stability index p of **M** is given by the stability index q of the matrix **A**. The quadratic matrix **A** with n rows and columns has a stability index $q < n$ if the path algebra is stable. In this case, the iterative methods require at most n iterations.

A stronger upper bound may be derived for the matrix **M** of the forward and back substitution method assuming a unitarily stable path algebra. The derivation leads to a stability index $p \le q/2+1$. This method thus requires roughly half as many iterations as the Jacobi method and the Gauss-Seidel method do in the worst case. Since the calculational cost per iteration is the same for all iterative methods, the calculational cost of this method is roughly half that of the Jacobi and Gauss-Seidel methods in the worst case.

Knowledge of an upper bound on the number of iterations in the case of stable matrices is of fundamental importance for algorithms. If the upper bound is exceeded in the course of the iteration process, then the matrix **A** of the system of linear equations is not stable, and the iteration is aborted without a result.

Example : Iterative methods

Let the illustrated cyclic graph with positive integer edge capacities be given. The maximal capacities of the paths from each vertex of the graph to vertex 5 are to be determined. The matrix **A** for the edge capacities is stable. The binary operations max and min from Section 8.5.4.4 are used for union and concatenation in the path algebra for path capacities. The desired path capacities are determined iteratively according to the methods described.



system of equations $\mathbf{x} = \mathbf{A}\,\mathbf{x} \sqcup \mathbf{b}$ with $\mathbf{b} = \mathbf{e}_5$

$$\begin{array}{|c|}\hline x_1\\ \hline x_2\\ \hline x_3\\ \hline x_4\\ \hline x_5\\ \hline\end{array} = \begin{array}{|c|c|c|c|c|}\hline n & n & 5 & n & n\\ \hline 1 & n & n & 2 & n\\ \hline n & 4 & n & n & 7\\ \hline n & n & 3 & n & n\\ \hline n & n & n & 6 & n\\ \hline\end{array} \circ \begin{array}{|c|}\hline x_1\\ \hline x_2\\ \hline x_3\\ \hline x_4\\ \hline x_5\\ \hline\end{array} \sqcup \begin{array}{|c|}\hline n\\ \hline n\\ \hline n\\ \hline n\\ \hline e\\ \hline\end{array} \qquad \begin{array}{l} n := 0\\ e := \infty\end{array}$$

Jacobi method : $\mathbf{y} = \mathbf{A}\,\mathbf{x}_k \sqcup \mathbf{b}$ and $\mathbf{x}_{k+1} = \mathbf{y}$

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 5 & n & n\\ \hline 1 & n & n & 2 & n\\ \hline n & 4 & n & n & 7\\ \hline n & n & 3 & n & n\\ \hline n & n & n & 6 & n\\ \hline\end{array} \qquad \begin{array}{ccccc} \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 & \mathbf{x}_4\\ n & n & 5 & 5 & 5\\ n & n & n & 2 & 2\\ n & 7 & 7 & 7 & 7\\ n & n & 3 & 3 & 3\\ e & e & e & e & e\end{array}$$

$$y_i = \bigsqcup_{j=1}^{n} a_{ij}\, x_j \sqcup b_i$$

$$y_i = \max\{\min\{a_{i1}, x_1\}, \ldots, \min\{a_{in}, x_n\}, b_i\}$$

In the Gauss-Seidel algorithm it is efficient to overwrite the vector **x** elementwise in each iteration. This is indicated by the symbol ← in the formulation of the iteration procedures.

Gauss-Seidel method : $\mathbf{x}_{k+1} = \mathbf{L}\,\mathbf{x}_{k+1} \sqcup \mathbf{R}\,\mathbf{x}_k \sqcup \mathbf{b}$

$$\mathbf{A} = \begin{array}{|c|c|c|c|c|}\hline n & n & 5 & n & n \\ \hline 1 & n & n & 2 & n \\ \hline n & 4 & n & n & 7 \\ \hline n & n & 3 & n & n \\ \hline n & n & n & 6 & n \\ \hline \end{array} \qquad \begin{array}{cccc} \mathbf{x}_0 & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \\ n & n & 5 & 5 \\ n & n & 2 & 2 \\ n & 7 & 7 & 7 \\ n & 3 & 3 & 3 \\ e & e & e & e \end{array}$$

$$x_i \leftarrow \bigsqcup_{j=1}^{n} a_{ij}\, x_j \sqcup b_i \qquad\qquad i = 1,\ldots,n$$

$$x_i \leftarrow \max\{\min\{a_{i1}, x_1\},\ldots, \min\{a_{in}, x_n\}, b_i\}$$

Forward and back substitution method :

$$\mathbf{y}_{k+1} = \mathbf{R}\,\mathbf{y}_{k+1} \sqcup \mathbf{x}_k \sqcup \mathbf{b} \quad \text{and} \quad \mathbf{x}_{k+1} = \mathbf{L}\mathbf{x}_{k+1} \sqcup \mathbf{y}_{k+1}$$

$$\mathbf{A} = \begin{array}{c|c|c|c|c|c|c} \multicolumn{6}{c}{} & \mathbf{R} \\ \cline{2-6} & n & n & 5 & n & n & \\ \cline{2-6} & 1 & n & n & 2 & n & \\ \cline{2-6} & n & 4 & n & n & 7 & \\ \cline{2-6} & n & n & 3 & n & n & \\ \cline{2-6} & n & n & n & 6 & n & \\ \cline{2-6} \multicolumn{2}{c}{\mathbf{L}} \end{array} \qquad \begin{array}{ccccc} \mathbf{x}_0 & \mathbf{y}_1 & \mathbf{x}_1 & \mathbf{y}_2 & \mathbf{x}_2 \\ n & 5 & 5 & 5 & 5 \\ n & n & 1 & 2 & 2 \\ n & 7 & 7 & 7 & 7 \\ n & n & 3 & 3 & 3 \\ e & e & e & e & e \end{array}$$

$$y_n = x_n \sqcup b_n = \max\{x_n, b_n\}$$

$$y_i = \bigsqcup_{j=i+1}^{n} a_{ij}\, y_j \sqcup x_i \sqcup b_i \qquad\qquad i = n-1,\ldots,1$$

$$y_i = \max\{\min\{a_{i\,i+1}, y_{i+1}\},\ldots, \min\{a_{in}, y_n\}, x_i, b_i\}$$

$$x_1 = y_1$$

$$x_i = \bigsqcup_{j=1}^{i-1} a_{ij}\, x_j \sqcup y_i \qquad\qquad i = 2,\ldots,n$$

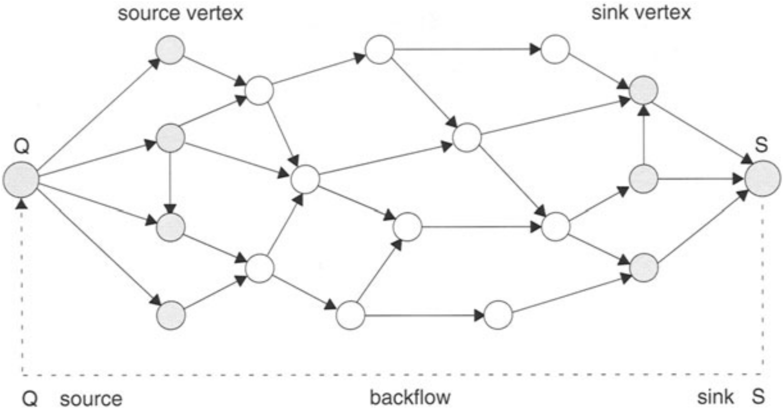
$$x_i = \max\{\min\{a_{i1}, x_1\},\ldots, \min\{a_{i\,i-1}, x_{i-1}\}, y_i\}$$

8.6 NETWORK FLOWS

8.6.1 INTRODUCTION

The determination of flows in networks is a problem of graph theory and optimization theory. Flow models are formulated on the basis of the principles of fluid mechanics. However, the application of flow models is not restricted to flow problems in the physical sense, but rather comprises a wide range of problems, particularly in the field of logistics.

Flow model : A simple flow model consists of a network with a source and a sink. Mass flows originate at the source, flow through the network and disappear in the sink. The network vertices with a direct mass input from the source are called source vertices. The network vertices with a direct mass output to the sink are called sink vertices. The mass flows in the network are assumed to be stationary (time-independent). They are bounded by capacities, and they may incur a cost. A flow model is described by a weighted graph. Each edge is associated with nonnegative numbers for the flow, the capacity restriction and the cost per unit flow.



Flow conservation : The mass flows in the network are stationary. By the law of conservation of mass, at each vertex of the network the combined mass input must be equal to the combined mass output. The combined mass output from the source must be equal to the combined input to the sink. This condition is taken into account in the weighted graph of the flow model by introducing a return edge for a backflow from the sink to the source. This turns the network into a closed system.

Elementary flow : A constant unit flow in an elementary cycle of the graph is called an elementary flow. It satisfies the condition of conservation of mass at every vertex of the cycle, since the mass input to the vertex is equal to the mass output from the vertex. Every mass-conserving flow in the network is a linear combination of elementary flows. Thus the determination of flows in networks is reduced to the determination of elementary cycles in graphs.

Optimal flow : An admissible flow in a network is a mass-conserving flow which satisfies the given capacity restrictions. A maximal flow in a network is an admissible flow for which the total mass flow from the source through the network to the sink is maximal. The total mass flow through the network is a measure for the capacity of the network. Often there are different maximal flows in the network with the same mass flow through the network. In these cases, a maximal flow with minimal cost may be determined.

Methods : The theoretical foundations for the determination of admissible flows in networks are treated in Sections 8.6.2 to 8.6.4; the basic methods for the determination of optimal flows in networks are treated in Sections 8.6.5 to 8.6.7. Matrix and vector algebra with real numbers are used in the formulations. There are close relationships between graph-theoretical methods, linear optimization methods and dynamic programming methods.

8.6.2 NETWORKS AND FLOWS

Introduction : The basic definitions for flows in networks and the law of conservation of mass are treated in the following.

Network : A network is a multigraph $G = (V, K ; A, E)$ with a vertex set V, an edge set K and the initial and terminal incidences A and E. The multigraph G has the following properties :

- There are no loops on the vertices.
- The graph does not contain partial edges.
- The graph is weakly connected.

Matrix representation : A network with n vertices and m edges is conveniently described by an incidence matrix **S** with n rows and m columns. The elements of the matrix take the values –1, 0, 1.

$$\text{incidence} \qquad \mathbf{S} = [s_{ij}] \qquad s_{ij} \in \{-1, 0, +1\}$$

$$s_{ij} = \begin{cases} -1 & \text{vertex } i \text{ is the end vertex of edge } j \\ 0 & \text{vertex } i \text{ is neither start nor end vertex of edge } j \\ +1 & \text{vertex } i \text{ is the start vertex of edge } j \end{cases}$$

Flow : A flow in a network is a function f which assigns every edge a non-negative real number. A flow along a directed edge may be regarded as mass transport per unit of time from the start vertex to the end vertex. The flow along edge j is designated by f_j. The flows for all edges are conveniently combined into a vector **f**.

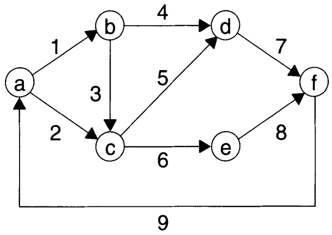
$$\text{flow} \qquad \mathbf{f} = [f_j] \qquad \mathbf{f} \geq \mathbf{0}$$

Conservation law : By the physical law of conservation of mass, the input is equal to the output at each vertex of the network. This condition is formulated as follows in terms of the incidence **S** and the flow **f** :

$$\text{conservation law} \qquad \mathbf{S}\,\mathbf{f} = \mathbf{0}$$

Example : Description of a network

The incidence matrix **S** for the illustrated network with letters as vertex labels and numbers as edge labels is shown.



$$\mathbf{S} = \begin{array}{ccccccccc|c} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & \\ \hline +1 & +1 & 0 & 0 & 0 & 0 & 0 & 0 & -1 & a \\ -1 & 0 & +1 & +1 & 0 & 0 & 0 & 0 & 0 & b \\ 0 & -1 & -1 & 0 & +1 & +1 & 0 & 0 & 0 & c \\ 0 & 0 & 0 & -1 & -1 & 0 & +1 & 0 & 0 & d \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & +1 & 0 & e \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & -1 & +1 & f \end{array}$$

At the vertex c, the flows f_2 and f_3 form the input to the vertex and the flows f_5 and f_6 form the output from the vertex. By the conservation law, the input must be equal to the output :

$$-f_2 - f_3 + f_5 + f_6 = 0$$

The product of the row for vertex c in the matrix **S** with the vector **f** leads to this equation.

8.6.3 UNRESTRICTED FLOW

Introduction : Let arbitrary flows be allowed along the edges of a network. Such flows are called unrestricted flows. Admissible non-zero flows can only exist if the network contains cycles. The flows in elementary cycles of networks are of fundamental importance for the solution of flow problems in networks. The theoretical foundations for unrestricted flows in networks are treated in the following.

Unrestricted flow : An unrestricted flow **f** in a network is said to be admissible if for each edge j the flow f_j is non-negative and the conservation law is satisfied at the vertices. Thus the following conditions hold :

non-negativity $\quad \mathbf{f} \geq \mathbf{0}$

conservation law $\quad \mathbf{Sf} = \mathbf{0}$

Zero flow : The zero flow $\mathbf{f} = \mathbf{0}$ satisfies the conditions for an admissible unrestricted flow. If the graph of the network is acyclic, there are no admissible flows other than the zero flow.

zero flow $\quad \mathbf{f} = \mathbf{0}$

The following proof shows that flows other than $\mathbf{f} = \mathbf{0}$ cannot exist in an acyclic graph. Consider every path through an acyclic graph which starts at a vertex a without predecessors and ends at a vertex b without successors. If there were a flow $f > 0$ along this path, the conservation law for the vertex a and the vertex b would be violated. The flow from a to b must therefore be zero. Thus the zero flow is the only admissible flow in the acyclic graph.

Elementary flow : There are admissible flows $\mathbf{f} \neq \mathbf{0}$ in a network only if the graph contains cycles. A unit flow in an elementary cycle is called an elementary flow and is designated by a vector **v**. This vector contains the value 1 for each edge which belongs to the elementary cycle and the value 0 for each edge which does not belong to the elementary cycle. Every elementary flow **v** satisfies the conditions for an admissible flow.

elementary flow $\quad \mathbf{v} = [v_j] \qquad v_j \in \{0, 1\}$

$$v_j = \begin{cases} 1 & \text{edge j belongs to the elementary cycle} \\ 0 & \text{edge j does not belong to the elementary cycle} \end{cases}$$

Admissible flow : Every linear combination of the elementary flows $\mathbf{v}_k$ with non-negative coefficients λ_k is an admissible flow $\mathbf{f}$ in a network.

admissible flow $$\mathbf{f} = \sum_{k=1}^{n} \lambda_k \mathbf{v}_k \qquad \lambda_k \geq 0$$

If an admissible flow $\mathbf{f}$ is given, it may be decomposed into elementary flows. The decomposition is performed by iteratively reducing $\mathbf{f}$ to the zero flow $\mathbf{0}$. In step k, $\mathbf{f}$ is reduced by the greatest possible contribution $\lambda_k \mathbf{v}_k$ of an elementary flow $\mathbf{v}_k$. The coefficient λ_k is the minimum of all edge flows f_j for the edges contained in the elementary cycle $\mathbf{v}_k$. The reduced flow $\mathbf{f} - \lambda_k \mathbf{v}_k$ is also an admissible flow.

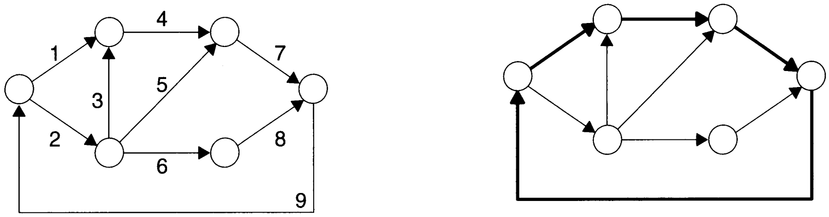
decomposition loop $k = 1,...,n$

coefficient $\lambda_k = \min_j \{f_j \mid v_{kj} = 1\}$

reduction $\mathbf{f} \leftarrow \mathbf{f} - \lambda_k \mathbf{v}_k$

Example 1 : Admissible flows

Let the illustrated network with numbers as edge labels be given. It contains four elementary cycles. One of the elementary cycles is highlighted by thick lines.



Let an admissible flow $\mathbf{f}$ in the network be given. It is to be decomposed into the elementary flows $\mathbf{v}_i$. The iterative decomposition is shown.

1	2	3	4	5	6	7	8	9	edges
4	6	1	5	2	3	7	3	10	flow $\mathbf{f}^T$
1	0	0	1	0	0	1	0	1	flow $\mathbf{v}_1^T$: $\lambda_1 = \min\{4,5,7,10\} = 4$
0	6	1	1	2	3	3	3	6	reduced flow $\mathbf{f}^T \leftarrow \mathbf{f}^T - \lambda_1 \mathbf{v}_1^T$
0	1	0	0	0	1	0	1	1	flow $\mathbf{v}_2^T$: $\lambda_2 = \min\{6,3,3,6\} = 3$
0	3	1	1	2	0	3	0	3	reduced flow $\mathbf{f}^T \leftarrow \mathbf{f}^T - \lambda_2 \mathbf{v}_2^T$
0	1	1	1	0	0	1	0	1	flow $\mathbf{v}_3^T$: $\lambda_3 = \min\{3,1,1,3,3\} = 1$
0	2	0	0	2	0	2	0	2	reduced flow $\mathbf{f}^T \leftarrow \mathbf{f}^T - \lambda_3 \mathbf{v}_3^T$
0	1	0	0	1	0	1	0	1	flow $\mathbf{v}_4^T$: $\lambda_4 = \min\{2,2,2,2\} = 2$
0	0	0	0	0	0	0	0	0	reduced flow $\mathbf{f}^T \leftarrow \mathbf{f}^T - \lambda_4 \mathbf{v}_4^T$ zero flow

Cut : The vertex set V of a network is partitioned into two disjoint subsets X and Y. The set of all edges with a start vertex $x \in X$ and an end vertex $y \in Y$ is called the corresponding cut set. It is conveniently described by a cut vector $\mathbf{s}(X,Y)$, which is formed according to the following rules for all edges j :

$$\text{cut vector} \qquad \mathbf{s}(X,Y) = [s_j] \qquad s_j \in \{0, 1\}$$

$$s_j = \begin{cases} 1 & \text{edge } j \text{ with start vertex } x \in X \text{ and end vertex } y \in Y \\ 0 & \text{edge } j \text{ with start vertex } a \notin X \text{ or end vertex } b \notin Y \end{cases}$$

Cut flow : Let a cut vector $\mathbf{s}(X,Y)$ for the edges separating the vertex set X from the vertex set Y be given. The resulting flow in the cut is the sum of the flows along all cut edges. It is called the cut flow and is designated by f(X,Y). The cut flow is calculated as follows as a function of the flow **f** in the network :

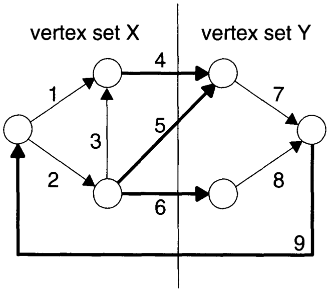
$$\text{cut flow} \qquad f(X,Y) = \mathbf{f}^T \mathbf{s}(X,Y)$$

If the vertex set X contains only one vertex x and the vertex set Y contains all remaining vertices, then f(X,Y) is the combined output from the vertex x and f(Y,X) is the combined input to the vertex x. By the conservation law, the combined flow from the vertices $x \in X$ to the vertices $y \in Y$ is equal to the combined flow from the vertices $y \in Y$ to the vertices $x \in X$:

$$\text{conservation law} \qquad f(X,Y) = f(Y,X)$$

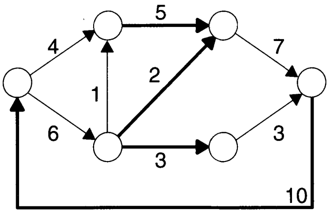
Example 2 : Cut flows

Let the illustrated network with numbers as edge labels be given. A cut is performed in the network by partitioning the vertex set into two disjoint subsets X and Y. The cut is illustrated by a line separating the two sets. The corresponding cut vectors s(X,Y) and s(Y,X) are specified.



$$\begin{array}{r|c|c|c|c|c|c|c|c|c|} & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ \mathbf{s}^T(X,Y) = & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ \mathbf{s}^T(Y,X) = & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}$$

Let an admissible flow **f** in the network be given. The cut flows f(X,Y) and f(Y,X) are calculated. They are equal.



$$\mathbf{f}^T = \begin{array}{|c|c|c|c|c|c|c|c|c|}\hline 4 & 6 & 1 & 5 & 2 & 3 & 7 & 3 & 10 \\ \hline \end{array}$$

$$f(X,Y) = \mathbf{f}^T\mathbf{s}(X,Y) = 5 + 2 + 3 = 10$$

$$f(Y,X) = \mathbf{f}^T\mathbf{s}(Y,X) = 10$$

8.6.4 RESTRICTED FLOW

Introduction : Let the flows along the edges of a network be bounded by a given capacity. Such flows are called restricted flows. The restriction leads to the concept of the residual graph, in which the extent to which a flow exhausts the capacity of the network is represented. The theoretical foundations for restricted flows in networks are based on the foundations for unrestricted flows and are treated in the following.

Capacity : Let the flow along each edge in the network be bounded from above. Every edge is thus assigned a non-negative real number which represents the capacity, that is the maximal admissible flow. The capacities for all edges j of the network are conveniently combined into a vector $\mathbf{c}$.

capacity $\qquad \mathbf{c} = [c_j] \qquad \mathbf{c} \geq \mathbf{0}$

Restricted flow : A restricted flow $\mathbf{f}$ in a network is said to be admissible if for each edge j the flow f_j is non-negative and does not exceed the given capacity c_j. The conservation law must be satisfied at the vertices. Thus the following conditions apply :

non-negativity $\qquad \mathbf{f} \geq \mathbf{0}$
capacity restriction $\qquad \mathbf{f} \leq \mathbf{c}$
conservation law $\qquad \mathbf{S}\mathbf{f} = \mathbf{0}$

Residual graph : Let a network with the capacity $\mathbf{c}$ and an admissible flow $\mathbf{f}$ be given. To assess the extent to which the flow exhausts the capacity of the network, a residual graph is constructed according to the following rules :

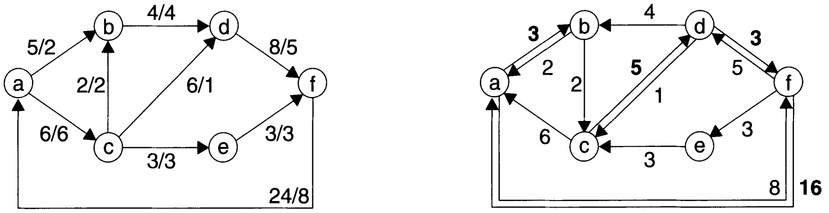
- The residual graph has the same vertices as the network.
- For each edge j of the network, a forward edge j^+ and a backward edge j^- are introduced. The forward edge has the same direction as the edge j, the backward edge has the opposite direction.
- The forward edge j^+ is assigned the unused capacity $c_j^+ = c_j - f_j$. The residual graph contains only the forward edges with $c_j^+ > 0$.
- The backward edge j^- is assigned the used capacity $c_j^- = f_j$. The residual graph contains only the backward edges with $c_j^- > 0$.
- An edge pair (j^+, j^-) of the residual graph is not regarded as a cycle with respect to the network flow.

The capacity components c_j^+ and c_j^- for all edges j are conveniently combined into vectors $\mathbf{c}^+$ and $\mathbf{c}^-$, whose sum equals the capacity $\mathbf{c}$.

capacity components $\qquad \mathbf{c}^+ = \mathbf{c} - \mathbf{f} \geq \mathbf{0}$
$\qquad \mathbf{c}^- = \mathbf{f} \geq \mathbf{0}$
sum $\qquad \mathbf{c}^+ + \mathbf{c}^- = \mathbf{c} \geq \mathbf{0}$

Example 1 : Residual graph

Let the illustrated network with admissible values c_j / f_j of the capacity and the flow for every edge j be given. The corresponding residual graph is shown. The forward edges are labeled with the boldface values $c_j^+ = c_j - f_j$, the backward edges are labeled with the values $c_j^- = f_j$ in plain face.



Zero flow : The zero flow $\mathbf{f} = \mathbf{0}$ satisfies the conditions for an admissible unrestricted flow. The residual graph for $\mathbf{f} = \mathbf{0}$ contains a forward edge and no backward edge for every edge j with capacity $c_j > 0$. It coincides with the graph of the network if no edge j has the capacity $c_j = 0$. If the residual graph is acyclic, then there is no admissible flow other than the zero flow $\mathbf{f} = \mathbf{0}$.

Elementary flow : There are admissible flows $\mathbf{f} \neq \mathbf{0}$ in a network only if the residual graph for $\mathbf{f} = \mathbf{0}$ contains cycles. The elementary flows are determined as in the case of unrestricted flow problems. An admissible restricted flow **f** is decomposed into the component elementary flows in the same way as an admissible unrestricted flow.

Flow increment : Let a network with an admissible flow **f** be given. If the residual graph for **f** is not acyclic, the flow **f** may be increased by a flow increment $\Delta\mathbf{f}$. A flow increment $\Delta\mathbf{f}$ is admissible if $\mathbf{f} + \Delta\mathbf{f}$ is an admissible flow. This leads to the following conditions for an admissible flow increment :

non-negativity	$\mathbf{f} + \Delta\mathbf{f} \ \geq \ \mathbf{0}$	$\Leftrightarrow$	$\Delta\mathbf{f} \ \geq -\mathbf{c}^-$
capacity restriction	$\mathbf{f} + \Delta\mathbf{f} \ \leq \ \mathbf{c}$	$\Leftrightarrow$	$\Delta\mathbf{f} \ \leq \ \mathbf{c}^+$
conservation law	$\mathbf{S}(\mathbf{f} + \Delta\mathbf{f}) = \mathbf{0}$	$\Leftrightarrow$	$\mathbf{S}\Delta\mathbf{f} = \ \mathbf{c}$

The flow increment $\Delta\mathbf{f}$ must satisfy the conservation law. The capacity components $-\mathbf{c}^-$ and $\mathbf{c}^+$ determined for the residual graph are a lower and upper bound for the flow increment, respectively. If the flow **f** is increased by $\Delta\mathbf{f}$, the capacity components $\mathbf{c}^+$ and $\mathbf{c}^-$ change as follows :

capacity components	$\mathbf{c}^- \leftarrow \mathbf{c}^- + \Delta\mathbf{f}$	$\mathbf{c}^+ \leftarrow \mathbf{c}^+ - \Delta\mathbf{f}$

Elementary flow increment : Let an admissible flow **f** in a network and the corresponding residual graph be given. If there is an elementary cycle in the residual graph, then there is also an elementary flow increment $\Delta\mathbf{v}$. An elementary cycle in the residual graph may consist of both forward and backward edges. If both the forward and the backward edge for a vertex pair exist, then only one of these two edges can belong to the elementary cycle. Accordingly, the vector $\Delta\mathbf{v}$ for an elementary flow increment contains the values –1, 0, +1. It is formed according to the following rule :

elementary flow increment $\qquad \Delta\mathbf{v} = [\Delta v_j] \qquad \Delta v_j \in \{-1, 0, +1\}$

$$\Delta v_j = \begin{cases} -1 & \text{backward edge } j^- \text{ belongs to the elementary cycle} \\ 0 & \text{edges } j^+ \text{ and } j^- \text{ do not belong to the elementary cycle} \\ +1 & \text{forward edge } j^+ \text{ belongs to the elementary cycle} \end{cases}$$

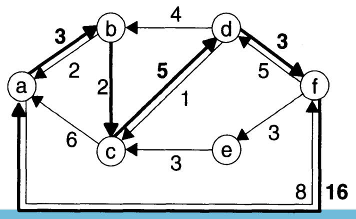
The elementary flow increment $\Delta\mathbf{v}$ satisfies the conservation of mass. However, it may violate the bound from above or below. An admissible flow increment $\Delta\mathbf{f}$ is formed by multiplying the elementary flow increment $\Delta\mathbf{v}$ by a non-negative coefficient λ. The coefficient is determined such that the bounds from above and below are not violated. It is less than or equal to the minimum of all capacity components of the forward and backward edges contained in the elementary cycle.

admissible flow increment $\qquad \Delta\mathbf{f} = \lambda\, \Delta\mathbf{v} \qquad \lambda \geq 0$

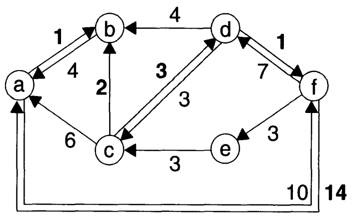
coefficient $\qquad \lambda \leq \min\limits_j \{c_j^+ \mid \Delta v_j = 1, \;\; c_j^- \mid \Delta v_j = -1\}$

Example 2 : Elementary flow increment

Let the illustrated residual graph for a network with an admissible flow **f** be given. The forward edges are labeled with the boldface values $c_j^+ = c_j - f_j$, the backward edges are labeled with the values $c_j^- = f_j$ in plain face. The residual graph contains an elementary cycle with the vertex sequence < a, b, c, d, f, a >, which consists of four forward edges and one backward edge. The greatest possible incremental flow in the elementary cycle is $\lambda = 2$. The flow **f** is increased by the corresponding flow increment $\Delta\mathbf{f}$. The corresponding residual graph with the modified capacity components is shown.



given residual graph



incremented residual graph

Cut : As in the case of unrestricted flow, a cut may be performed in a network with restricted flow by partitioning the vertex set of the network into two disjoint subsets X and Y. In the case of restricted flow, the cut capacities are considered in addition to the cut flows.

Cut flow and cut capacity : Let the cut vectors **s**(X,Y) and **s**(Y,X) for the cut sets be given. The corresponding cut flows are determined in terms of the flow **f**, the corresponding cut capacities are determined in terms of the capacity **c** :

cut flow	$f(X,Y) = \mathbf{f}^T\mathbf{s}(X,Y)$	$f(Y,X) = \mathbf{f}^T\mathbf{s}(Y,X)$
cut capacity	$c(X,Y) = \mathbf{c}^T\mathbf{s}(X,Y)$	$c(Y,X) = \mathbf{c}^T\mathbf{s}(Y,X)$

Due to the capacity restrictions $f_j \le c_j$ for all edges j, the following restrictions hold :

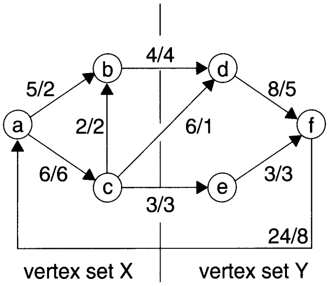
restrictions	$f(X,Y) \le c(X,Y)$	$f(Y,X) \le c(Y,X)$

Since the cut flows f(X,Y) and f(Y,X) are equal by the conservation law, the following stronger restriction holds :

restriction $\quad f(X,Y) = f(Y,X) \le \min\{c(X,Y), c(Y,X)\}$

Example 3 : Cut flows and cut capacities

Let the illustrated network with admissible values c_j / f_j of the capacity and the flow for every edge j be given. The cut capacities and cut flows for the illustrated cut are determined.



$$
\begin{aligned}
c(X,Y) &= 4+6+3 = 13 \\
c(Y,X) &= 24 \\
f(X,Y) &= 4+1+3 = 8 \quad \le 13 \\
f(Y,X) &= 8 \quad \le 24 \\
f(X,Y) &= f(Y,X) \quad \le 13
\end{aligned}
$$

8.6.5 MAXIMAL FLOW

Introduction : In determining the maximal flow in a network, it is assumed that the network has exactly one source and one sink. The flows originating at the source flow through the network to the sink. The flows along the directed edges are bounded by given capacities. In order to satisfy the conservation law for the network flows, a backflow from the sink to the source is introduced, which is equal to the combined output from the source and, equivalently, to the combined input to the sink. The backflow may be bounded or unbounded. The maximal flow from the source through the network to the sink is a measure of the total capacity of the network. It is determined by solving a linear optimization problem. The theoretical foundations for maximal flows in networks are based on the foundations for restricted flows and are treated in the following.

Source and sink : A restricted flow **f** in the network is said to be maximal if the flow **f** is admissible and the resulting flow f_R from the source q to the sink s is maximal. This definition leads to the following linear optimization problem :

non-negativity	$\mathbf{f} \geq \mathbf{0}$
capacity restriction	$\mathbf{f} \leq \mathbf{c}$
conservation law	$\mathbf{S}\,\mathbf{f} = \mathbf{0}$
resulting flow	$f_R \rightarrow \max$

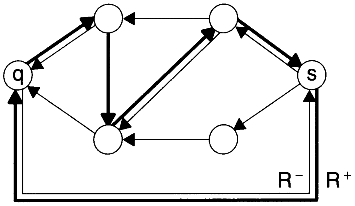
Elementary cycles in the residual graph : Consider the residual graph for a network with an admissible flow **f**. The residual graph may contain elementary cycles, which are classified as follows with respect to the resulting flow f_R :

- An elementary cycle is called an augmenting cycle if it contributes to an increase of the resulting flow f_R. It consists of an elementary path from the source to the sink and the edge R^+.

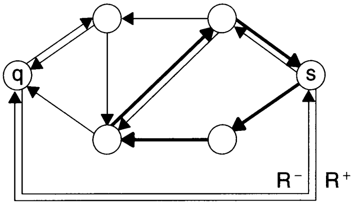
- An elementary cycle is called a diminishing cycle if it contributes to a decrease of the resulting flow f_R. It consists of an elementary path from the sink to the source and the edge R^-.

- An elementary cycle is called a preserving cycle if it does not contribute to a change in the resulting flow f_R. It contains neither the edge R^+ nor the edge R^-.

The augmenting and preserving cycles of the residual graph are of fundamental importance for determining the flow **f** with the maximal resulting flow max f_R.



augmenting cycle



preserving cycle

Optimality condition : An admissible flow $\mathbf{f}$ is maximal if and only if the residual graph for $\mathbf{f}$ does not contain an augmenting cycle. A maximal flow $\mathbf{f}$ is unique if and only if the residual graph for $\mathbf{f}$ does not contain a preserving cycle.

Proof : Optimality condition

The condition for a maximal flow is proved by showing that the flow cannot be maximal if there is an augmenting cycle and that in case of a non-maximal flow there necessarily exists an augmenting cycle. If the residual graph of $\mathbf{f}$ contains an augmenting cycle, then there is an admissible flow increment $\Delta\mathbf{f} \neq \mathbf{0}$ with $\Delta f_R > 0$. Then $\mathbf{g} = \mathbf{f} + \Delta\mathbf{f}$ is an admissible flow with $g_R > f_R$. In this case the flow $\mathbf{f}$ is not maximal. Conversely, if the flow $\mathbf{f}$ is not maximal, then there must exist an admissible flow $\mathbf{g} \neq \mathbf{f}$ with $g_R > f_R$. Then $\Delta\mathbf{f} = \mathbf{g} - \mathbf{f} \neq \mathbf{0}$ is an admissible flow increment with $\Delta f_R > 0$. In this case there is at least one augmenting cycle in the residual graph for $\mathbf{f}$.

The condition for the uniqueness of a maximal flow $\mathbf{f}$ is proved in the same manner. There is no augmenting cycle in the residual graph for $\mathbf{f}$. If there is a preserving cycle in the residual graph for $\mathbf{f}$, then there is an admissible flow increment $\Delta\mathbf{f} \neq \mathbf{0}$ with $\Delta f_R = 0$. Then $\mathbf{g} = \mathbf{f} + \Delta\mathbf{f}$ is an admissible flow with $g_R = f_R$. In this case the maximal flow $\mathbf{f}$ is not unique. Conversely, if the maximal flow is not unique, there must be an admissible flow $\mathbf{g} \neq \mathbf{f}$ with $g_R = f_R$. Then $\Delta\mathbf{f} = \mathbf{g} - \mathbf{f} \neq \mathbf{0}$ is an admissible flow increment with $\Delta f_R = 0$. In this case there is at least one preserving cycle in the residual graph for $\mathbf{f}$.

Optimization procedure : A maximal flow $\mathbf{f}$ in a network is determined in the following steps :

1. Choose an admissible flow $\mathbf{f}$ as an initial flow. The zero flow $\mathbf{f} = \mathbf{0}$ is a possible choice.
2. Determine the residual graph for the flow $\mathbf{f}$ and find an augmenting cycle in the residual graph.
3. If there is an augmenting cycle, determine the greatest possible flow increment $\Delta\mathbf{f}$, increase the flow $\mathbf{f}$ by $\Delta\mathbf{f}$ and repeat step 2.
4. If there is no augmenting cycle, the flow $\mathbf{f}$ is maximal.

If the capacities are integers, then the edge flows of a maximal flow are also integers. The procedure requires a finite number of steps to calculate the maximal flow.

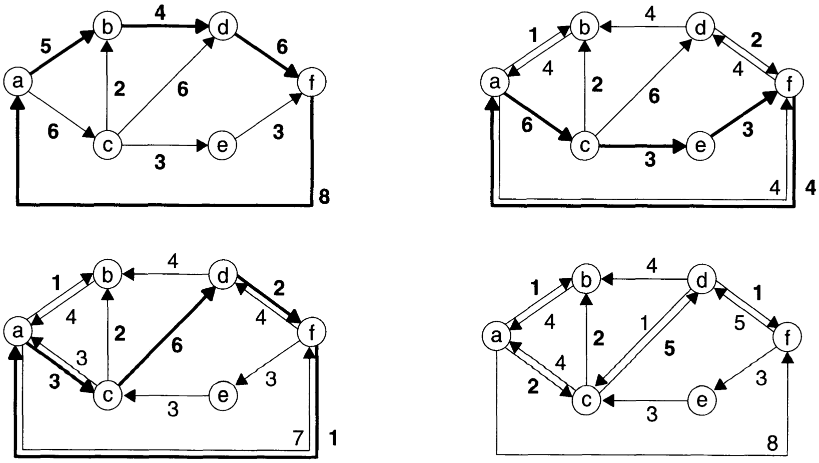
Proof : Properties of the optimization procedure

To prove the integer property, the zero flow $\mathbf{f} = \mathbf{0}$ is taken as the initial flow. The residual graph for **f** possesses integral edge capacities. The greatest possible flow increment $\Delta\mathbf{f}$ for an augmenting cycle is therefore also integral. Hence the flow **f** remains integral when increased by $\Delta\mathbf{f}$. The same is true for all subsequent augmenting cycles. Hence the procedure yields an integral maximal flow **f** with the integral resulting flow max f_R. Since for each augmenting cycle the resulting flow f_R is increased by an integral component $\Delta\mathbf{f}$ with $\Delta f_R \geq 1$, at most max f_R augmenting cycles need to be considered in the course of the procedure. The number of steps required for calculating the maximal flow is therefore finite.

If the given edge capacities are rational numbers, the optimization problem may be transformed into an integer optimization problem by multiplying all edge capacities by their common denominator. It follows that the procedure yields a maximal flow with rational edge flows after a finite number of steps.

If some edge capacities are irrational numbers, the procedure may require an infinite number of steps and may even converge to incorrect edge flows.

Example 1 : Let the illustrated network with the source a, the sink f and the indicated edge capacities be given. Let the backflow from the sink to the source be bounded. The graph corresponds to the residual graph for the zero flow $\mathbf{f} = \mathbf{0}$. A maximal flow **f** in the network is determined iteratively. In each step an augmenting cycle, represented by thick edges, is determined. In the residual graph, the forward edges j^+ are labeled with the boldface values $c_i^+ = c_j - f_j$, and the backward edges j^- are labeled with the values $c_j^- = f_j$ in plain face. The residual graphs for the individual steps are shown.



The illustrated flow in the residual graph with the resulting flow $f_R = 8$ is maximal, since there is no augmenting cycle in the residual graph. However, it is not unique, since the residual graph contains several preserving cycles.

Minimal cut : Cuts in networks with restricted flows are treated in Section 8.6.4. They are applied to maximal flows as follows. The vertex set of the network is partitioned into two disjoint subsets Q and S. The vertex set Q contains at least the source q, and the vertex set S contains at least the sink s. The partition consisting of Q and S is called a minimal cut if the cut capacity is minimal among all such partitions.

$$\text{minimal cut} \qquad \min c \;=\; \min_{Q,S} \{c(Q,S)\}$$

Maximal flow and minimal cut : The maximal flow max f_R which flows from the source q through the network to the sink s is equal to the minimal capacity min c of a cut which separates the source q from the sink s. The backflow from the sink s to the source q is assumed to be unbounded.

$$\max f_R \;=\; \min c$$

The proof of this theorem is contained in the following procedure for constructing a minimal cut with maximal flow.

Construction of a minimal cut with maximal flow : Let an arbitrary cut through the network with the vertex sets Q and S be given. For a maximal flow **f** the resulting flow max f_R along the return edge R contributes to the cut flow f(S, Q). By the conservation law, the cut flows f(S, Q) and f(Q, S) are equal. Due to the capacity restriction, the cut flow f(Q, S) is less than or equal to the cut capacity c(Q, S) :

$$\max f_R \;\le\; f(Q,S) \;=\; f(Q,S) \;\le\; c(Q,S)$$

If there is a cut with c(Q, S) = max f_R, then this cut is minimal. A minimal cut may be constructed from the residual graph for a maximal flow as follows :

1. The residual graph is reduced by removing the backward edge R^- for the backflow.

2. In the reduced residual graph, all vertices reachable from the source q are determined and combined into the vertex set Q. The source q belongs to Q, since q is reachable from itself. The sink s does not belong to Q, since s is not reachable from q. If s were reachable from q, there would be an augmenting cycle and the flow **f** would not be maximal.

3. All vertices which do not belong to the vertex set Q are collected in the vertex set S. The sink s belongs to S.

4. There are no edges from $x \in Q$ to $y \in S$ in the reduced residual graph. If an edge from x to y existed, y would be reachable from q via x, and hence by point 2 the vertex y would have to belong to the vertex set Q.

5. There are only edges from $y \in S$ to $x \in Q$ in the reduced residual graph. Each of these edges is either a forward edge j^+ or a backward edge j^-:

   Every forward edge $j^+ \neq R^+$ from $y \in S$ to $x \in Q$ in the reduced residual graph corresponds to an edge j from y to x with the flow $f_j = 0$ in the graph. The forward edge $j^+ = R^+$ from the sink $s \in S$ to the source $q \in Q$ in the reduced residual graph corresponds to the return edge R from s to q with the resulting maximal flow max f_R in the graph. The cut flow f(S, Q) in the graph is therefore equal to max f_R.

   Every backward edge j^- from $y \in S$ to $x \in Q$ in the reduced residual graph corresponds to an edge j from x to y with the flow $f_j = c_j$ in the graph. The cut flow f(Q, S) in the graph is therefore equal to the cut capacity c(Q, S).

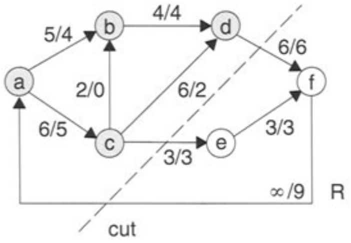
6. Together, max f_R = f(Q, S) and f(S, Q) = f(Q, S) and f(Q, S) = c(Q, S) imply max f_R = c(Q, S). Hence the cut with the vertex sets Q and S is minimal.

Example 2 : Maximal flow and minimal cut

Let the illustrated graph, a maximal flow **f** and the corresponding residual graph be given. The edges of the graph are labeled with the value pairs capacity / flow. The forward edges of the residual graph are labeled with boldface capacities, the backward edges are labeled with capacities in plain face. The reduced residual graph is the residual graph without the edge R^-.

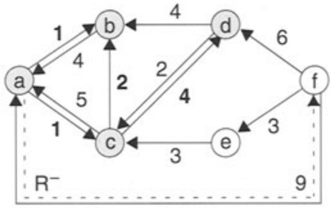
In the residual graph without the edge R^-, the vertices a,b,c,d are reachable from the source a. They form the vertex set Q. The remaining vertices e,f form the vertex set S. The partition of the vertex set into Q and S is a minimal cut. The maximal resulting flow max f_R is equal to the minimal cut capacity.

graph



min c = max f_R = 6 + 3 = 9

residual graph



Q = {a, b, c, d}

S = {e, f}

8.6.6 MAXIMAL FLOW AND MINIMAL COST

Introduction : Let the maximal resulting flow though a network from the source to the sink be calculated. There may be several different flow states which lead to the same maximal resulting flow through the network. Among these flow states, a flow with minimal cost is to be determined. The cost is assumed to be proportional to the flow. Cost values per unit flow are therefore introduced for the edges. Starting from a calculated maximal flow in the network, a minimum cost maximal flow is determined by solving a linear optimization problem. The theoretical foundations for minimum cost maximal flows in networks are treated in the following.

Cost : Let the flow along every edge in the network incur a cost proportional to the flow. Every edge is thus assigned a non-negative real number which represents the cost per unit flow. The costs for all edges j of the network are conveniently combined into a vector **k**.

cost	$\mathbf{k} = [\,k_j\,]$	$\mathbf{k} \geq \mathbf{0}$

Maximal flow with minimal cost : There may be several different maximal flows **f** in a network with the same resulting flow f_R from the source through the network to the sink. A maximal flow **f** with minimal cost is to be determined. This is a linear optimization problem :

non-negativity	$\mathbf{f}$	$\geq$	$\mathbf{0}$
capacity restriction	$\mathbf{f}$	$\leq$	$\mathbf{c}$
conservation law	$\mathbf{S}\mathbf{f}$	$=$	$\mathbf{0}$
maximal flow	f_R	$\rightarrow$	max
minimal cost	$\mathbf{k}^T\mathbf{f}$	$\rightarrow$	min

Preserving cycles in the residual graph : Consider the residual graph for a maximal flow **f**. The residual graph may contain preserving cycles with the admissible flow increments $\Delta\mathbf{f}$ and $\Delta f_R = 0$. They are classified as follows with respect to the cost :

cost-raising cycles	$\mathbf{k}^T\Delta\mathbf{f}$	$>$	0
cost-preserving cycles	$\mathbf{k}^T\Delta\mathbf{f}$	$=$	0
cost-reducing cycles	$\mathbf{k}^T\Delta\mathbf{f}$	$<$	0

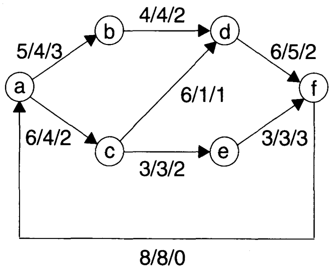
Optimality conditions : A maximal flow **f** incurs minimal cost if and only if the residual graph for **f** contains no cost-reducing cycle. A maximal flow **f** with minimal cost is unique if and only if the residual graph for **f** contains no cost-preserving cycle. The proofs are analogous to the ones in Section 8.6.5.

Optimization procedure : A maximal flow **f** with minimal cost K is determined in the following steps :

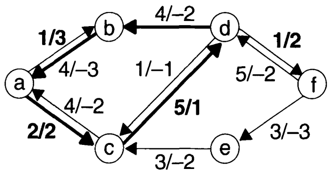
1. Determine a maximal flow **f** with cost $K = \mathbf{k}^T\mathbf{f}$ as an initial flow.
2. Determine the residual graph for the flow **f** and find a cost-reducing cycle in the residual graph.
3. If there is a cost-reducing cycle, determine the greatest possible flow increment $\Delta\mathbf{f}$, increase the flow **f** by $\Delta\mathbf{f}$, reduce the cost K by $\mathbf{k}^T\Delta\mathbf{f}$ and repeat step 2.
4. If there is no cost-reducing cycle, the maximal flow **f** incurs minimal cost.

Example 1 : Let the illustrated network with the source a, the sink f, a maximal flow **f**, the edge capacities **c** and edge costs **k** per unit flow be given. Each edge j is labeled by a triple $c_j / f_j / k_j$. A maximal flow **f** with minimal cost is determined iteratively using the residual graphs. In each step a cost-reducing cycle, represented by thick edges, is determined. In the residual graph, forward edges are labeled with the boldface values $c_j^+ = c_j - f_j$ and $k_j^+ = k_j$, and backward edges are labeled with the values $c_j^- = f_j$ and $k_j^- = -k_j$ in plain face. Labeling the edges with k_j^+ and k_j^- makes it easier to find cost-reducing cycles in the graphical representation. The residual graphs for the individual steps are shown.

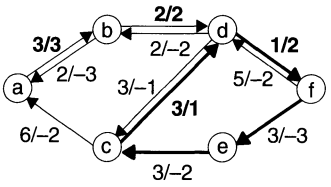
network with capacity / flow / cost



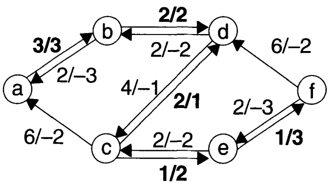
residual graph



$$\text{cost} = -\sum_j c_j^- k_j^- = 54$$



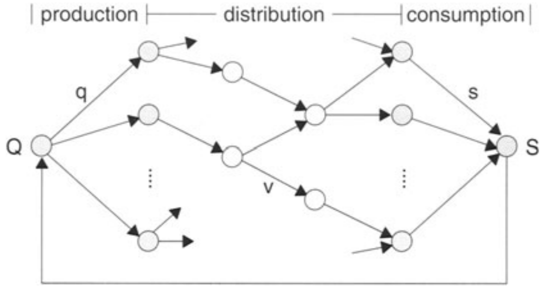
cost $= 54 + 2(-2) = 50$



cost $= 50 + 1(-2) = 48$

Example 2 : Distribution and assignment problems

Distribution and assignment problems occur in various forms in the field of logistics. They are exemplified by the production and consumption of goods. Goods are distributed from production sites to consumption sites in a network. The production and the consumption are bounded by given capacities. The goods are distributed by transporting them through the network; this transport is bounded by given transport capacities and incurs a cost per unit of goods. The most cost-effective distribution of goods at maximal consumption is to be determined. This distribution problem is a problem of maximal flow with minimal cost. The weighted graph for the distribution problem is shown schematically.

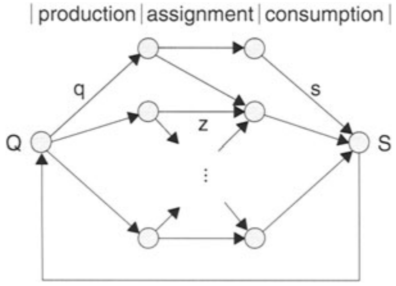


q production edge with capacity c > 0 and cost k = 0

s consumption edge with capacity c > 0 and cost k = 0

v distribution edge with capacity c > 0 and cost k

The problem of the distribution of goods may be reduced to an assignment problem if the transport path from each production site to each consumption site is predetermined and the transport capacity in the network is sufficient. In this case, the production sites are directly assigned to consumption sites. This assignment problem is a special problem of maximal flow with minimal cost. The weighted graph for the assignment problem is shown schematically.



q production edge with capacity c > 0 and cost k = 0

s consumption edge with capacity c > 0 and cost k = 0

z assignment edge with capacity c = ∞ and cost k

8.6.7 CIRCULATION

Introduction : Circulations are composed of cyclic flows in a network without a source or a sink. The flow along each edge of the network may be bounded from above and below. Depending on the structure of the network and the restrictions on the flows, a circulation may or may not exist. The determination of a circulation in the network is reduced to the determination of a maximal flow in a substitute network with a source and a sink. The procedures described in the preceding sections may then be applied. The theoretical foundations for circulations in networks are treated in the following.

Circulation : A flow **f** in a network is called a circulation if it is bounded from above and below and the conservation law is satisfied. Thus the following conditions hold for a circulation **f**.

restriction $\mathbf{b} \le \mathbf{f} \le \mathbf{c}$ $\qquad \mathbf{0} \le \mathbf{b} \le \mathbf{c}$

conservation law $\mathbf{S}\mathbf{f} = \mathbf{0}$

For a given network with the restrictions **b** and **c**, a circulation may or may not exist. The existence depends on the values of the restrictions.

Admissible restricted flows : The determination of a circulation **f** is reduced to the determination of an admissible restricted network flow $\mathbf{f}_N$. The following ansatz is used for this purpose :

circulation $\mathbf{f} = \mathbf{f}_N + \mathbf{b}$

network flow $\mathbf{f}_N = \mathbf{f} - \mathbf{b}$

Substituting the ansatz into the restriction for the circulation **f** yields the following restriction for the network flow $\mathbf{f}_N$:

restriction $\mathbf{b} \le \mathbf{f} \le \mathbf{c} \quad \Rightarrow$

$$\mathbf{0} \le \mathbf{f}_N \le \mathbf{c} - \mathbf{b}$$

Substituting the ansatz into the conservation law for the circulation **f** yields the following restriction for the network flow $\mathbf{f}_N$:

conservation law $\mathbf{S}\mathbf{f} = \mathbf{0} \quad \Rightarrow$

$$\mathbf{S}\mathbf{f}_N + \mathbf{S}\mathbf{b} = \mathbf{0}$$

The expression $\mathbf{Sb}$ in the conservation law is decomposed into the components $(\mathbf{Sb})^+ \geq \mathbf{0}$ and $(\mathbf{Sb})^- \geq \mathbf{0}$ with $\mathbf{Sb} = (\mathbf{Sb})^+ - (\mathbf{Sb})^-$. The component $\mathbf{f}_S = (\mathbf{Sb})^+$ corresponds to outputs from the vertices, the component $\mathbf{f}_Q = (\mathbf{Sb})^-$ corresponds to inputs to the vertices. This decomposition leads to the following conservation law :

$$\begin{array}{lll} \text{conservation law} & \mathbf{S}\mathbf{f}_N + \mathbf{f}_S - \mathbf{f}_Q & = \mathbf{0} \\ \text{input} & \mathbf{f}_Q = (\mathbf{Sb})^- & \geq \mathbf{0} \\ \text{output} & \mathbf{f}_S = (\mathbf{Sb})^+ & \geq \mathbf{0} \end{array}$$

Multiplying the conservation law $\mathbf{S}\mathbf{f}_N + \mathbf{f}_S - \mathbf{f}_Q = \mathbf{0}$ by the transpose $\mathbf{e}^T = [1,...,1]$ of the one vector from the left and using the special property $\mathbf{e}^T\mathbf{S} = \mathbf{0}^T$ of the incidence matrix, one obtains the result that the sum $\mathbf{e}^T\mathbf{f}_Q$ of the vertex inputs is equal to the sum $\mathbf{e}^T\mathbf{f}_S$ of the vertex outputs.

$$\begin{array}{lll} \text{conservation law} & \mathbf{e}^T\mathbf{f}_Q = \mathbf{e}^T\mathbf{f}_S & \mathbf{e}^T = [1,...,1] \end{array}$$

Thus the conditions for a circulation $\mathbf{f}$ are reduced to the conditions for the restricted admissible flows $\mathbf{f}_N$, $\mathbf{f}_Q$, $\mathbf{f}_S$.

Substitute network : In order to determine the restricted admissible flows for a circulation, a substitute network is constructed. The vertex set of the substitute network contains all vertices of the network as well as a source q and a sink s. The edge set of the substitute network contains the set N of all edges of the network, the set Q of input edges from the source q to each vertex of the network, the set S of output edges from each vertex of the network to the sink s, and the return edge R from the sink s to the source q. Non-negative flows and capacities are introduced for all edges of the substitute network. The following conditions hold for the admissible restricted flows in the substitute network :

$$\begin{array}{llll} \text{network flow} & \mathbf{f}_N \geq \mathbf{0} & \mathbf{f}_N \leq \mathbf{c}_N & = \mathbf{c} - \mathbf{b} \\ \text{input} & \mathbf{f}_Q \geq \mathbf{0} & \mathbf{f}_Q = \mathbf{c}_Q & = (\mathbf{Sb})^- \\ \text{output} & \mathbf{f}_S \geq \mathbf{0} & \mathbf{f}_S = \mathbf{c}_S & = (\mathbf{Sb})^+ \\ \text{backflow} & f_R \geq 0 & f_R \leq c_R & = \infty \\ & & & \\ \text{conservation law} & \multicolumn{3}{l}{\mathbf{S}\mathbf{f}_N - \mathbf{f}_Q + \mathbf{f}_S = \mathbf{0}} \\ & \multicolumn{3}{l}{f_R = \mathbf{e}^T\mathbf{f}_Q = \mathbf{e}^T\mathbf{f}_S} \end{array}$$

Maximal flow in the substitute network : The existence of a circulation is determined by calculating a maximal flow max f_R in the substitute network. The equations $\mathbf{f}_Q = \mathbf{c}_Q$ and $\mathbf{f}_S = \mathbf{c}_S$ are replaced by the inequalities $\mathbf{f}_Q \le \mathbf{c}_Q$ and $\mathbf{f}_S \le \mathbf{c}_S$, so that the following conditions are imposed :

network flow	$\mathbf{f}_N \ge \mathbf{0}$	$\mathbf{f}_N \le \mathbf{c}_N = \mathbf{c} - \mathbf{b}$
input	$\mathbf{f}_Q \ge \mathbf{0}$	$\mathbf{f}_Q \le \mathbf{c}_Q = (\mathbf{Sb})^-$
output	$\mathbf{f}_S \ge \mathbf{0}$	$\mathbf{f}_S \le \mathbf{c}_S = (\mathbf{Sb})^+$
backflow	$f_R \ge 0$	$f_R \le c_R = \infty$
conservation law	$\mathbf{Sf}_N - \mathbf{f}_Q + \mathbf{f}_S = \mathbf{0}$	$f_R = \mathbf{e}^T\mathbf{f}_Q = \mathbf{e}^T\mathbf{f}_S$
resulting flow	$f_R \rightarrow \max$	

There exists a circulation in the network if and only if the following conditions are satisfied for the maximal flow in the substitute network :

conditions	$\mathbf{f}_Q = \mathbf{c}_Q$	$\mathbf{f}_S = \mathbf{c}_S$

A circulation in the network is calculated from the maximal flow in the substitute network by increasing the network flow $\mathbf{f}_N$ by the lower bound $\mathbf{b}$ of the circulation.

circulation	$\mathbf{f} = \mathbf{f}_N + \mathbf{b}$

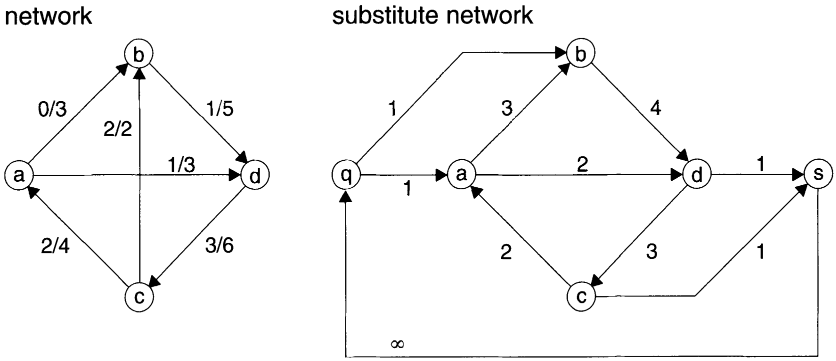
The substitute network for the maximal flow may be simplified as follows. The edges with the capacity restriction $c = 0$ are not included in the substitute network, since there can be no flow $f > 0$ along these edges. Then each vertex of the network has either an input edge from the source or an output edge to the sink or neither of these edges. The maximal flow in the substitute network is calculated according to the procedure in Section 8.6.5.

Minimal cost of a circulation : If there is a circulation in a network and the costs per unit flow for the network edges are specified, a circulation with minimal cost may be calculated. This calculation may be carried out for the substitute network using the procedure in Section 8.6.6 for the maximal flow with minimal cost.

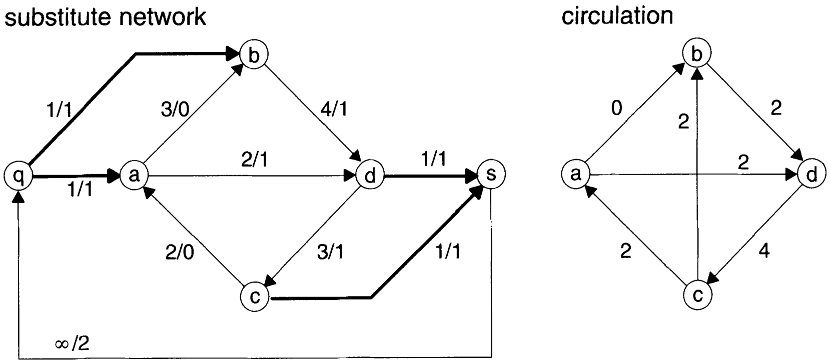
Example 1 : Circulation

Let the network illustrated below with values b_j / c_j for the lower / upper bound of the flow along each edge j be given. The corresponding substitute network contains the network vertices as well as a source q and a sink s. The edges and their capacities in the substitute network are determined as follows :

- Each network edge j is assigned a capacity $c_{Nj} = c_j - b_j$ in the substitute network. If $c_{Nj} = 0$, then the edge j is not included in the substitute network.
- For each vertex x of the network, the combined output a due to the minimal flows $f_j = b_j$ of all network edges j incident at the vertex x is calculated. If $a < 0$, an input edge from q to x with the capacity $-a$ is introduced. If $a > 0$, an output edge from x to s with the capacity a is introduced.
- A return edge R with the capacity $c_R = \infty$ is introduced from the sink s to the source q.



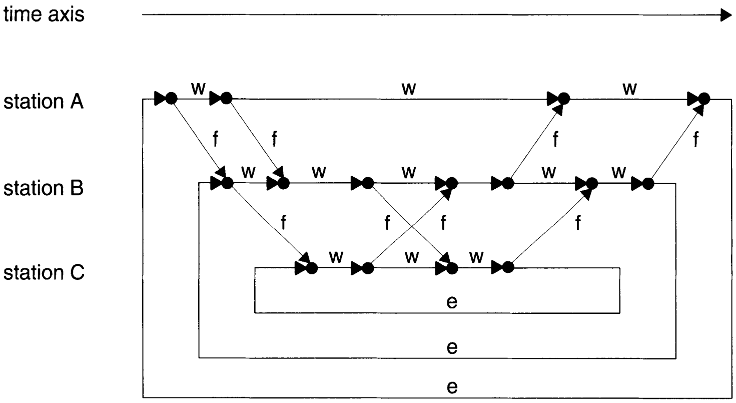
The maximal flow in the substitute network is calculated according to the procedure in Section 8.6.5. The substitute network with the values c_j / f_j for the capacity and the flow along each edge j is shown. For the input and output edges introduced, the flow is equal to the capacity. Hence the conditions for the existence of a circulation in the network are satisfied. A circulation in the network is calculated by increasing the calculated flow f_j in the substitute network by the lower bound b_j for each network edge j.



Example 2 : Disposition problem

In various application areas the planning of the deployment of resources leads to disposition problems. The optimal deployment of trains according to a specified schedule is a typical example. For each train connection, the schedule specifies the initial station with the time of departure and the destination with the time of arrival. The schedule is constructed for a given period of time and then repeated periodically. The minimal number of trains required at each station to implement the schedule is to be determined. It is assumed that a train can wait at a station for an arbitrary amount of time. This disposition problem corresponds to a circulation in a network with minimal cost.

The weighted graph for this disposition problem is constructed as follows. The vertices of the graph are the times of departure and arrival of trains at each station, which are represented as an ordered sequence along a time axis. The edges of the graph are waiting edges at every station, journey edges for train journeys between two stations and deployment edges for each train. Since the schedule is periodic, every deployment of a train must end at the station at which it began. Deploying a train incurs one unit of cost, so that the minimal cost is given by the minimal number of deployed trains. The graph with the edge weighting for the lower and upper capacity restrictions and for the costs is shown below for a railway system with three stations.



		lower bound	upper bound	cost
waiting edge	w :	$b = 0$	$c = \infty$	$k = 0$
journey edge	f :	$b = 1$	$c = \infty$	$k = 0$
deployment edge	e :	$b = 0$	$c = \infty$	$k = 1$

The solution of this simple disposition problem is the deployment of at least two trains at station A. Each train travels from A via B to C and returns from C via B to A.

9 TENSORS

9.1 INTRODUCTION

Physical quantities are independent of the coordinate systems which are used to describe the physical problem. Variables which represent physical quantities are called tensors. The mathematical properties of tensors are of fundamental importance for the mathematical formulation of physical phenomena.

Tensors may be defined by their mathematical properties without reference to their physical significance. The special property of every tensor is that the tensor is a linear scalar mapping of a vector m-tuple (tensor of rank m). Tensors are therefore represented using the vector algebra introduced in Chapter 3. Real vector spaces with the euclidean metric are particularly important for physical problems. Tensor algebra is based on the rules of transformation for basis vectors and for vector coordinates.

A tensor is completely described by its coordinates in an arbitrary basis. Each of these coordinates is the image of one of the possible m-tuples of basis vectors. Since the mapping which defines the tensor is linear, the image of an arbitrary m-tuple of vectors may be represented as a linear combination of the tensor coordinates. Tensor algebra deals with the transformation rules for tensor coordinates under transformations of the basis of the vector space. These rules may be used to establish the tensor character of a variable. There are tensors with special properties, such as symmetric, antisymmetric and isotropic tensors. Tensors of rank 1 (vectors) and tensors of rank 2 (dyads) are often used to solve physical problems.

A tensor with space-dependent coordinates is called a tensor field. The properties of tensor fields are treated in tensor analysis. Tensor analysis deals with a point space. A vector space is associated with this point space by regarding the difference of two points of the point space as a vector of the vector space. Depending on the type of problem considered, either the same (global) basis for the associated vector space is chosen at all points of the space, or a different (local) basis is chosen at every point. Christoffel symbols are defined to describe the space dependence of the basis vectors. The coordinates of a tensor in a global basis are called rectilinear coordinates : Their space dependence is described by partial derivatives. The coordinates of a tensor in local bases are called curvilinear coordinates : Their space dependence is described by covariant derivatives. Tensor integrals and field operations are defined for tensor fields.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

9.2 VECTOR ALGEBRA

9.2.1 VECTOR SPACES

Introduction : General vector spaces are treated in Section 3.5. The concepts of vector, linear combination, independence, basis and dimension are defined there. The finite-dimensional real vector space $(\mathbb{R}, \mathbb{R}^n \,;\, +\,,\circ)$ is particularly important for physical problems. The study of the algebraic structure of vector spaces with linear mappings is treated in Section 3.6. The definition of the topological structure of real vector spaces with a metric is treated in Section 5.4.

The relationship between real metric vector spaces, vector spaces with a scalar product and euclidean vector spaces is treated in the following. The concept of dual bases is introduced to allow a convenient representation of the scalar product in non-orthonormal bases. These bases are distinguished by the terms covariant and contravariant. Accordingly, vectors have covariant and contravariant coordinates.

Real metric vector space : A vector $\mathbf{u}$ of the real vector space $\mathbb{R}^n$ is an n-tuple $(u_1,\ldots,u_n)$ of real numbers with the coordinates $u_i \in \mathbb{R}$. The vectors $\mathbf{u},\mathbf{v},\mathbf{w} \in \mathbb{R}^n$ have the following general properties :

associative	:	$\mathbf{u} + (\mathbf{v} + \mathbf{w})$	$=$	$(\mathbf{u} + \mathbf{v}) + \mathbf{w}$
commutative	:	$\mathbf{u} + \mathbf{w}$	$=$	$\mathbf{w} + \mathbf{u}$
identity element	:	$\mathbf{u} + \mathbf{0}$	$=$	$\mathbf{u}$
inverse element	:	$\mathbf{u} + (-\mathbf{u})$	$=$	$\mathbf{0}$

The product of vectors $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$ with scalars $a, b \in \mathbb{R}$ has the following properties :

associative	:	$a(b\,\mathbf{u})$	$=$	$(ab)\mathbf{u}$
distributive	:	$(a + b)\mathbf{u}$	$=$	$a\mathbf{u} + b\mathbf{u}$
		$a(\mathbf{u} + \mathbf{w})$	$=$	$a\mathbf{u} + a\mathbf{w}$
identity element	:	$1\,\mathbf{u}$	$=$	$\mathbf{u}$

A real space $\mathbb{R}^n$ is called a real metric space if a metric $d(\mathbf{u},\mathbf{w})$ with the following properties is defined for any two vectors $\mathbf{u},\mathbf{w} \in \mathbb{R}^n$:

(M1)	$d(\mathbf{u},\mathbf{u}) = 0$	
(M2)	$d(\mathbf{u},\mathbf{w}) > 0$	for $\mathbf{u} \neq \mathbf{w}$
(M3)	$d(\mathbf{u},\mathbf{w}) = d(\mathbf{w},\mathbf{u})$	
(M4)	$d(\mathbf{u},\mathbf{w}) \leq d(\mathbf{u},\mathbf{v}) + d(\mathbf{v},\mathbf{w})$	for every $\mathbf{v} \in \mathbb{R}^n$

Scalar product : A mapping $s : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ with $s(\mathbf{u}, \mathbf{w}) = \mathbf{u} \cdot \mathbf{w}$ is called a scalar product of the vectors $\mathbf{u}$ and $\mathbf{w}$ in the real space $\mathbb{R}^n$ if the following conditions are satisfied for vectors $\mathbf{u}, \mathbf{v}, \mathbf{w} \in \mathbb{R}^n$ and every scalar $c \in \mathbb{R}$:

$$\begin{array}{llcl}
\text{(S1)} & \mathbf{u} \cdot \mathbf{u} > 0 \quad \text{for} \quad \mathbf{u} \neq \mathbf{0} & \text{and} & \mathbf{u} \cdot \mathbf{u} = 0 \quad \text{for} \quad \mathbf{u} = \mathbf{0} \\
\text{(S2)} & \mathbf{u} \cdot \mathbf{w} = \mathbf{w} \cdot \mathbf{u} & & \\
\text{(S3)} & (c\mathbf{u}) \cdot \mathbf{w} = \mathbf{u} \cdot (c\mathbf{w}) = c(\mathbf{u} \cdot \mathbf{w}) & & \\
\text{(S4)} & \mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w} & &
\end{array}$$

Magnitude of a vector : The positive square root of the scalar product $\mathbf{u} \cdot \mathbf{u}$ of a vector $\mathbf{u} \in \mathbb{R}^n$ with itself is called the magnitude of the vector $\mathbf{u}$ and is designated by $|\mathbf{u}|$. The absolute value of the scalar product $\mathbf{u} \cdot \mathbf{w}$ is less than or equal to the product of the magnitudes of the vectors $\mathbf{u}$ and $\mathbf{w}$ (Schwarz inequality).

$$|\mathbf{u}| = \sqrt{\mathbf{u} \cdot \mathbf{u}}$$

$$|\mathbf{u} \cdot \mathbf{w}| \leq |\mathbf{u}||\mathbf{w}|$$

Proof : Schwarz inequality

A linear combination $\lambda\mathbf{u} + \mathbf{w}$ is formed using the vectors $\mathbf{u}$ and $\mathbf{w}$ and a scalar $\lambda \in \mathbb{R}$. Properties (S1) and (S2) of the scalar product lead to the following inequality :

$$(\lambda\mathbf{u} + \mathbf{w}) \cdot (\lambda\mathbf{u} + \mathbf{w}) = \lambda^2 \, \mathbf{u} \cdot \mathbf{u} + 2\lambda \, \mathbf{u} \cdot \mathbf{w} + \mathbf{w} \cdot \mathbf{w} \geq 0$$

This condition is satisfied if the discriminant of the quadratic equation in λ is not positive. Since both sides of the resulting inequality are non-negative by property (S1), the square root may be taken on both sides :

$$4(\mathbf{u} \cdot \mathbf{w})^2 - 4(\mathbf{u} \cdot \mathbf{u})(\mathbf{w} \cdot \mathbf{w}) \leq 0$$

$$(\mathbf{u} \cdot \mathbf{w})^2 \leq (\mathbf{u} \cdot \mathbf{u})(\mathbf{w} \cdot \mathbf{w})$$

$$|\mathbf{u} \cdot \mathbf{w}| \leq |\mathbf{u}||\mathbf{w}|$$

Angle between vectors : By the Schwarz inequality, an angle α may be defined for arbitrary vectors $\mathbf{u}, \mathbf{w} \neq \mathbf{0}$ with the scalar product $\mathbf{u} \cdot \mathbf{w}$. The vectors $\mathbf{u}$ and $\mathbf{w}$ are said to be parallel if $\cos\alpha = 1$. The vectors are said to be orthogonal if $\cos\alpha = 0$. Orthogonal vectors are said to be orthonormal if their absolute value is 1.

$$\cos\alpha = \frac{\mathbf{u} \cdot \mathbf{w}}{|\mathbf{u}||\mathbf{w}|}$$

$$\cos\alpha = 1 \ : \ \mathbf{u} \text{ and } \mathbf{w} \text{ are parallel}$$

$$\cos\alpha = 0 \ : \ \mathbf{u} \text{ and } \mathbf{w} \text{ are orthogonal}$$

$$\cos\alpha = 0 \wedge |\mathrm{u}| = |\mathrm{w}| = 1 \ : \ \mathbf{u} \text{ and } \mathbf{w} \text{ are orthonormal}$$

Euclidean vector space : A real vector space $\mathbb{R}^n$ is said to be euclidean if a scalar product $\mathbf{u} \cdot \mathbf{w}$ is defined for any two vectors $\mathbf{u}, \mathbf{w}$ of the space. The metric $\mathrm{d}(\mathbf{u},\mathbf{w})$ of the vectors $\mathbf{u}$ and $\mathbf{w}$ in the euclidean space is the magnitude of their difference $\mathbf{u} - \mathbf{w}$.

$$\mathrm{d}(\mathbf{u},\mathbf{w}) \quad = \quad \sqrt{(\mathbf{u}-\mathbf{w}) \cdot (\mathbf{u}-\mathbf{w})}$$

Proof : Properties of the metric in a euclidean space

Properties (M1) and (M2) of the euclidean metric follow from its definition by property (S1) of the scalar product. Property (M3) follows from (S2). To prove property (M4), the distance $\mathrm{d}(\mathbf{u},\mathbf{w})$ of the points $\mathbf{u}$, $\mathbf{w} \in \mathbb{R}^n$ is expressed as a function of a third point $\mathbf{v} \in \mathbb{R}^n$:

$$\begin{aligned} \mathrm{d}^2(\mathbf{u},\mathbf{w}) &= (\mathbf{u}-\mathbf{w}) \cdot (\mathbf{u}-\mathbf{w}) \\ &= (\mathbf{u}-\mathbf{v}+\mathbf{v}-\mathbf{w}) \cdot (\mathbf{u}-\mathbf{v}+\mathbf{v}-\mathbf{w}) \\ &= (\mathbf{u}-\mathbf{v}) \cdot (\mathbf{u}-\mathbf{v}) + 2(\mathbf{u}-\mathbf{v}) \cdot (\mathbf{v}-\mathbf{w}) + (\mathbf{v}-\mathbf{w}) \cdot (\mathbf{v}-\mathbf{w}) \end{aligned}$$

By definition $(\mathbf{u}-\mathbf{v}) \cdot (\mathbf{u}-\mathbf{v}) = \mathrm{d}^2(\mathbf{u},\mathbf{v})$ and $(\mathbf{v}-\mathbf{w}) \cdot (\mathbf{v}-\mathbf{w}) = \mathrm{d}^2(\mathbf{v},\mathbf{w})$. Hence the Schwarz inequality yields $(\mathbf{u}-\mathbf{v}) \cdot (\mathbf{v}-\mathbf{w}) \le |(\mathbf{u}-\mathbf{v}) \cdot (\mathbf{v}-\mathbf{w})| \le |\mathbf{u}-\mathbf{v}|\,|\mathbf{u}-\mathbf{w}| = \mathrm{d}(\mathbf{u},\mathbf{v})\ \mathrm{d}(\mathbf{v},\mathbf{w})$. Property (M4) now follows by substitution:

$$\begin{aligned} \mathrm{d}^2(\mathbf{u},\mathbf{w}) &\le \mathrm{d}^2(\mathbf{u},\mathbf{v}) + 2\,\mathrm{d}(\mathbf{u},\mathbf{v})\ \mathrm{d}(\mathbf{v},\mathbf{w}) + \mathrm{d}^2(\mathbf{v},\mathbf{w}) \\ \mathrm{d}\ (\mathbf{u},\mathbf{w}) &\le \mathrm{d}\ (\mathbf{u},\mathbf{v}) + \mathrm{d}(\mathbf{v},\mathbf{w}) \end{aligned}$$

Components of a vector : Let a vector $\mathbf{u}$ of the euclidean space $\mathbb{R}^n$ be the sum of the vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$. Then the vectors $\mathbf{a}$ and $\mathbf{b}$ are called components of the vector $\mathbf{u}$. The vector $\mathbf{u}$ may be decomposed into components in different ways.

$$\mathbf{u} = \mathbf{a}+\mathbf{b} \qquad \wedge \qquad \mathbf{u}, \mathbf{a}, \mathbf{b} \in \mathbb{R}^n$$

9.2.2 BASES

Basis of a real vector space : Every set of n linearly independent vectors is a basis of the real vector space $\mathbb{R}^n$ (see Section 3.5). The basis vectors are designated by $\mathbf{b}_1,...,\mathbf{b}_n$. A basis is said to be orthogonal if the basis vectors are pairwise orthogonal. A basis is said to be orthonormal if the basis vectors have magnitude 1 and are pairwise orthogonal.

orthogonal basis : $i \neq m \quad \Rightarrow \quad \mathbf{b}_i \cdot \mathbf{b}_m = 0$

orthonormal basis : $i = m \quad \Rightarrow \quad \mathbf{b}_i \cdot \mathbf{b}_m = 1$

$i \neq m \quad \Rightarrow \quad \mathbf{b}_i \cdot \mathbf{b}_m = 0$

Canonical basis : A vector of the real vector space $\mathbb{R}^n$ is called a canonical unit vector if one coordinate has the value 1 and all other coordinates have the value 0. The canonical unit vector whose i-th coordinate is 1 is designated by $\mathbf{e}_i$. The basis vectors of an orthonormal basis are generally not canonical unit vectors, but their magnitude is 1. The special orthonormal basis $\mathbf{e}_1,...,\mathbf{e}_n$ whose basis vectors are the n canonical unit vectors of the space $\mathbb{R}^n$ is called the canonical basis of the vector space.

canonical basis of $\mathbb{R}^n$: $\mathbf{e}_i = (0,...,0,1,0,...,0)$

$i = m \quad \Rightarrow \quad \mathbf{e}_i \cdot \mathbf{e}_m = 1$

$i \neq m \quad \Rightarrow \quad \mathbf{e}_i \cdot \mathbf{e}_m = 0$

Dual bases : Two bases of a euclidean space $\mathbb{R}^n$ are said to be dual (reciprocal, contragredient) if their basis vectors with different indices are pairwise orthogonal and the scalar product of basis vectors with the same index is 1. One of these two bases is chosen arbitrarily. It is called the covariant basis and is associated with subscripts. Its basis vectors are designated by $\mathbf{b}_1,...,\mathbf{b}_n$. The other basis is called the contravariant basis and is associated with superscripts. Its basis vectors are designated by $\mathbf{b}^1,...,\mathbf{b}^n$.

covariant basis : $\mathbf{b}_1,...,\mathbf{b}_n$

contravariant basis : $\mathbf{b}^1,...,\mathbf{b}^n$

duality : $i = m \quad \Rightarrow \quad \mathbf{b}_i \cdot \mathbf{b}^m = \mathbf{b}^i \cdot \mathbf{b}_m = 1$

$i \neq m \quad \Rightarrow \quad \mathbf{b}_i \cdot \mathbf{b}^m = \mathbf{b}^i \cdot \mathbf{b}_m = 0$

Every orthonormal basis is self-dual; in particular, the canonical basis $\mathbf{e}_1,...,\mathbf{e}_n$ is self-dual. Its basis vectors may thus alternatively be designated by $\mathbf{e}^1,...,\mathbf{e}^n$ with $\mathbf{e}_m = \mathbf{e}^m$.

Coordinates of a basis : Let the vectors $\mathbf{b}_1,...,\mathbf{b}_n$ form a covariant basis of the euclidean space $\mathbb{R}^n$. By the definition of a basis, an arbitrary basis vector $\mathbf{b}_m$ and the vectors $\mathbf{e}^1,...,\mathbf{e}^n$ of the canonical basis of $\mathbb{R}^n$ are linearly dependent.

$$c_0 \mathbf{b}_m + c_1 \mathbf{e}^1 + ... + c_n \mathbf{e}^n = \mathbf{0} \qquad \text{with} \qquad c_0 \neq 0$$

This vector equation is solved for $\mathbf{b}_m$. The coefficients b_{im} of the vector equation are determined by forming the scalar products $\mathbf{e}_i \cdot \mathbf{b}_m = \mathbf{b}_m \cdot \mathbf{e}_i$ and exploiting the orthonormality of the canonical basis.

$$\mathbf{b}_m = b_{1m} \mathbf{e}^1 + ... + b_{nm} \mathbf{e}^n \qquad \text{with} \qquad b_{im} = -c_i / c_0$$

$$b_{im} = \mathbf{e}_i \cdot \mathbf{b}_m \qquad \text{coordinates of the basis vector } \mathbf{b}_m$$

Coordinates of dual bases : The coordinates of the vectors $\mathbf{b}_m$ of a covariant basis and of the vectors $\mathbf{b}^m$ of a contravariant basis are generally not the same. The indices of the coordinates of a covariant basis are written as subscripts, the indices of the coordinates of a contravariant basis are written as superscripts.

$$\mathbf{b}_m = b_{1m} \mathbf{e}^1 + ... + b_{nm} \mathbf{e}^n$$

$$\mathbf{b}^m = b^{1m} \mathbf{e}_1 + ... + b^{nm} \mathbf{e}_n$$

b_{im} coordinates of the covariant basis vector $\mathbf{b}_m$

b^{im} coordinates of the contravariant basis vector $\mathbf{b}^m$

Basis matrices : The coordinates of the basis vectors are arranged in a column vector according to Section 3.5. The basis vectors are arranged columnwise in a basis matrix. The canonical basis is designated by $\mathbf{E}$, the covariant basis by $\mathbf{B}_*$ and the contravariant basis by $\mathbf{B}^*$. The usual rule for the indices of the coefficients of the basis matrices applies : row index before column index.

$$\mathbf{E} = \begin{array}{|c|c|c|c|c|} \hline \\ \\ \mathbf{e}_1 & \cdots & \mathbf{e}_m & \cdots & \mathbf{e}_n \\ \\ \\ \hline \end{array} = \begin{array}{|c|c|c|c|c|} \hline 1 & & 0 & & 0 \\ \hline \vdots & & \vdots & & \vdots \\ \hline 0 & \cdots & 1 & \cdots & 0 \\ \hline \vdots & & \vdots & & \vdots \\ \hline 0 & & 0 & & 1 \\ \hline \end{array}$$

$$\mathbf{B}_* = \left[\begin{array}{|c|c|c|c|c|} \mathbf{b}_1 & \cdots & \mathbf{b}_m & \cdots & \mathbf{b}_n \end{array}\right] = \left[\begin{array}{|c|c|c|c|c|} \hline b_{11} & & b_{1m} & & b_{1n} \\ \hline \vdots & & \vdots & & \vdots \\ \hline b_{i1} & \cdots & b_{im} & \cdots & b_{in} \\ \hline \vdots & & \vdots & & \vdots \\ \hline b_{n1} & & b_{nm} & & b_{nn} \\ \hline \end{array}\right]$$

$$\mathbf{B}^* = \left[\begin{array}{|c|c|c|c|c|} \mathbf{b}^1 & \cdots & \mathbf{b}^m & \cdots & \mathbf{b}^n \end{array}\right] = \left[\begin{array}{|c|c|c|c|c|} \hline b^{11} & & b^{1m} & & b^{1n} \\ \hline \vdots & & \vdots & & \vdots \\ \hline b^{i1} & \cdots & b^{im} & \cdots & b^{in} \\ \hline \vdots & & \vdots & & \vdots \\ \hline b^{n1} & & b^{nm} & & b^{nn} \\ \hline \end{array}\right]$$

Kronecker symbols : The orthonormality of a basis $\mathbf{B}_*$ or $\mathbf{B}^*$ and the duality of the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ are conveniently expressed using the Kronecker symbols δ_{im}, δ^{im} and δ^i_m. Each of these symbols has the value 1 for i = m and the value 0 otherwise. The identity matrix is designated by $\mathbf{I}$.

orthonormal :	$\mathbf{b}^i \cdot \mathbf{b}^m = \delta^{im}$	$\Rightarrow \quad (\mathbf{B}^*)^T \mathbf{B}^* = \mathbf{I} = \mathbf{B}^*(\mathbf{B}^*)^T$
orthonormal :	$\mathbf{b}_i \cdot \mathbf{b}_m = \delta_{im}$	$\Rightarrow \quad (\mathbf{B}_*)^T \mathbf{B}_* = \mathbf{I} = \mathbf{B}_*(\mathbf{B}_*)^T$
dual :	$\mathbf{b}_i \cdot \mathbf{b}^m = \delta^m_i$	$\Rightarrow \quad (\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I} = \mathbf{B}^*(\mathbf{B}_*)^T$
dual :	$\mathbf{b}^i \cdot \mathbf{b}_m = \delta^i_m$	$\Rightarrow \quad (\mathbf{B}^*)^T \mathbf{B}_* = \mathbf{I} = \mathbf{B}_*(\mathbf{B}^*)^T$

$$i = m \quad \Rightarrow \quad \delta_{im} = \delta^{im} = \delta^i_m = 1$$
$$i \neq m \quad \Rightarrow \quad \delta_{im} = \delta^{im} = \delta^i_m = 0$$

Note : The symbolic conditions for orthonormal and dual bases using Kronecker symbols are similar. However, the properties of these bases are quite different. The vectors of an orthonormal basis are pairwise orthogonal, and their magnitude is 1. The vectors of one of two dual bases are generally neither pairwise orthogonal, nor is their magnitude 1.

Example 1 : Orthonormal basis

The basis **B** is orthonormal. Its vectors have magnitude 1 and are pairwise orthogonal : $\mathbf{B}^T\mathbf{B} = \mathbf{I}$.

				0.5185	0.6470	0.5591	**B**
				–0.2074	–0.5392	0.8163	
				0.8296	–0.5392	–0.1454	
	0.5185	–0.2074	0.8296	1.0000	0.0000	0.0000	
	0.6470	–0.5392	–0.5392	0.0000	1.0000	0.0000	
$\mathbf{B}^T$	0.5591	0.8163	–0.1454	0.0000	0.0000	1.0000	$\mathbf{I}$

Example 2 : Dual bases

The bases $\mathbf{B}_*$ and $\mathbf{B}^*$ are dual. Basis vectors with different indices are orthogonal, for example $\mathbf{b}_1$ and $\mathbf{b}^2$. The scalar product of vectors with the same index is 1, for example $\mathbf{b}_1 \cdot \mathbf{b}^1 = 1$. Hence $(\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I}$.

				1.2500	0.5000	0.0000	$\mathbf{B}^*$
				0.0000	1.0000	1.0000	
				–0.2500	0.5000	1.0000	
	1.0000	–1.0000	1.0000	1.0000	0.0000	0.0000	
	–0.5000	2.5000	–2.5000	0.0000	1.0000	0.0000	
$\mathbf{B}_*^T$	0.5000	–1.5000	2.5000	0.0000	0.0000	1.0000	$\mathbf{I}$

9.2.3 COORDINATES

Coordinates of a vector : By the definition of a basis, an arbitrary vector $\mathbf{u} \neq \mathbf{0}$ of the real vector space $\mathbb{R}^n$ and an arbitrary basis $\mathbf{b}^1,...,\mathbf{b}^n$ of this space are linearly dependent. The vector equation is solved for $\mathbf{u}$. The resulting linear combination of the basis vectors is called the representation of the vector $\mathbf{u}$ in the basis $\mathbf{B}^*$. The coefficients u_i of the linear combination are called the coordinates of the vector $\mathbf{u}$ in the basis $\mathbf{B}^*$.

$$a_0 \mathbf{u} + a_1 \mathbf{b}^1 + ... + a_n \mathbf{b}^n = \mathbf{0} \qquad \text{with} \qquad a_0 \neq 0$$

$$\mathbf{u} = u_1 \mathbf{b}^1 + ... + u_n \mathbf{b}^n \qquad \text{with} \qquad u_i = a_i / a_0$$

Covariant and contravariant coordinates : A vector $\mathbf{u} \in \mathbb{R}^n$ may alternatively be represented in the contravariant basis $\mathbf{B}^*$ or in the covariant basis $\mathbf{B}_*$. The coordinates of the vector $\mathbf{u}$ in the contravariant basis $\mathbf{B}^*$ are called the covariant coordinates of $\mathbf{u}$ and are written with a subscript. The coordinates of the vector $\mathbf{u}$ in the covariant basis $\mathbf{B}_*$ are called the contravariant coordinates of $\mathbf{u}$ and are written with a superscript.

$$\mathbf{u} = u_1 \mathbf{b}^1 + ... + u_n \mathbf{b}^n$$

$$\mathbf{u} = u^1 \mathbf{b}_1 + ... + u^n \mathbf{b}_n$$

u_i covariant coordinates of the vector $\mathbf{u}$

u^i contravariant coordinates of the vector $\mathbf{u}$

Representations of a vector : Vectors are designated by lowercase boldface letters. This designation does not contain a reference to the basis in which the coordinates of the vector are specified. However, the coordinates of a vector in different bases are different. The different representations of a vector are therefore distinguished as follows by diacritic marks on the vector symbol :

$\mathbf{u}$ coordinates of the vector in the canonical basis $\mathbf{E}$

$\mathbf{u}^*$ contravariant coordinates of the vector in the basis $\mathbf{B}_*$

$\mathbf{u}_*$ covariant coordinates of the vector in the basis $\mathbf{B}^*$

Matrix form of the coordinate relationships : The representation of a vector $\mathbf{u}$ in a basis $\mathbf{B}$ may be written as a matrix product. The coordinates of the vector in this basis are regarded as a vector. The relationships between the coordinates of the vector in the different bases are derived from the properties of dual bases.

$$\mathbf{u} = \begin{bmatrix} \mathbf{b}_1 & \cdots & \mathbf{b}_i & \cdots & \mathbf{b}_n \end{bmatrix} * \begin{bmatrix} u^1 \\ \vdots \\ u^i \\ \vdots \\ u^n \end{bmatrix} = \begin{bmatrix} \mathbf{b}^1 & \cdots & \mathbf{b}^i & \cdots & \mathbf{b}^n \end{bmatrix} * \begin{bmatrix} u_1 \\ \vdots \\ u_i \\ \vdots \\ u_n \end{bmatrix}$$

$$\mathbf{u} = \mathbf{B}_* \mathbf{u}^* \quad \wedge \quad (\mathbf{B}^*)^T \mathbf{B}_* = \mathbf{I} \quad \Rightarrow \quad \mathbf{u}^* = (\mathbf{B}^*)^T \mathbf{u}$$

$$\mathbf{u} = \mathbf{B}^* \mathbf{u}_* \quad \wedge \quad (\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I} \quad \Rightarrow \quad \mathbf{u}_* = (\mathbf{B}_*)^T \mathbf{u}$$

Summation convention : Linear combinations of vectors $\mathbf{b}^1,...,\mathbf{b}^n$ with coefficients $u_1,...,u_n$ often occur, as do linear combinations of vectors $\mathbf{b}_1,...,\mathbf{b}_n$ with coefficients $u^1,...,u^n$. To simplify the notation for such linear combinations, a summation convention is introduced : If an expression contains the same index once as a subscript and once as a superscript, a summation over this index is implied.

$$\mathbf{u} = u_i \mathbf{b}^i \quad := \quad u_1 \mathbf{b}^1 + ... + u_n \mathbf{b}^n$$

$$\mathbf{u} = u^i \mathbf{b}_i \quad := \quad u^1 \mathbf{b}_1 + ... + u^n \mathbf{b}_n$$

The range $i = 1,...,n$ of the index i must be implicitly known. The convention is suspended by enclosing the indices in parentheses. Thus the right-hand side of the equation $\mathbf{u} = u^{(i)} \mathbf{b}_{(i)}$ contains only one term, namely the product of the vector $\mathbf{b}_i$ with the scalar u^i.

Scalar form of the coordinate relationships : In the canonical basis **E** the covariant and the contravariant coordinates of a vector **u** are equal. These coordinates are marked by the symbol $*$ (asterisk) and are alternatively designated by $\overset{*}{u}_m$ or by $\overset{*}{u}^m$. The relationships between the coordinates of the vector **u** in the canonical basis $\mathbf{e}_1,...,\mathbf{e}_n$, in a covariant basis $\mathbf{b}_1,...,\mathbf{b}_n$ and in a contravariant basis $\mathbf{b}^1,...,\mathbf{b}^n$ are conveniently written using the summation convention :

$$\overset{*}{u}^m = u_i b^{mi} \quad := \quad u_1 b^{m1} + ... + u_n b^{mn}$$

$$\overset{*}{u}_m = u^i b_{mi} \quad := \quad u^1 b_{m1} + ... + u^n b_{mn}$$

Example 1 : Coordinates of a vector

Let the coordinates of the vectors **u** and **w** in the canonical basis **E** of the three-dimensional euclidean space $\mathbb{R}^3$ be given. Their coordinates in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of Example 2 in Section 9.2.2 are calculated.

$$(\mathbf{B}^*)^{\mathrm{T}}\mathbf{u} = \mathbf{u}^* \quad : \quad \begin{array}{|r|r|r|} \hline 1.2500 & 0.0000 & -0.2500 \\ \hline 0.5000 & 1.0000 & 0.5000 \\ \hline 0.0000 & 1.0000 & 1.0000 \\ \hline \end{array} \; * \; \begin{array}{|r|} \hline 3.0000 \\ \hline -6.0000 \\ \hline 7.0000 \\ \hline \end{array} \; = \; \begin{array}{|r|} \hline 2.0000 \\ \hline -1.0000 \\ \hline 1.0000 \\ \hline \end{array}$$

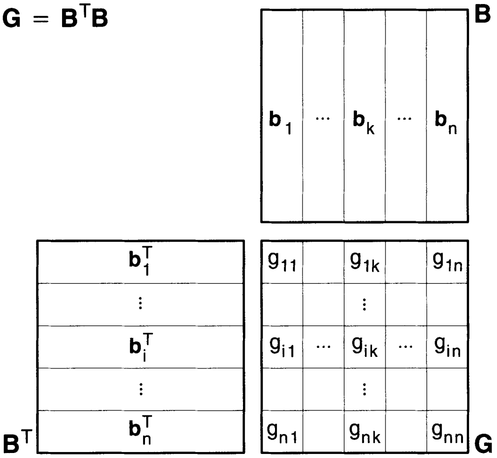
$$(\mathbf{B}_*)^{\mathrm{T}}\mathbf{u} = \mathbf{u}_* \quad : \quad \begin{array}{|r|r|r|} \hline 1.0000 & -1.0000 & 1.0000 \\ \hline -0.5000 & 2.5000 & -2.5000 \\ \hline 0.5000 & -1.5000 & 2.5000 \\ \hline \end{array} \; * \; \begin{array}{|r|} \hline 3.0000 \\ \hline -6.0000 \\ \hline 7.0000 \\ \hline \end{array} \; = \; \begin{array}{|r|} \hline 16.0000 \\ \hline -34.0000 \\ \hline 28.0000 \\ \hline \end{array}$$

$$(\mathbf{B}^*)^{\mathrm{T}}\mathbf{w} = \mathbf{w}^* \quad : \quad \begin{array}{|r|r|r|} \hline 1.2500 & 0.0000 & -0.2500 \\ \hline 0.5000 & 1.0000 & 0.5000 \\ \hline 0.0000 & 1.0000 & 1.0000 \\ \hline \end{array} \; * \; \begin{array}{|r|} \hline 0.5000 \\ \hline 0.5000 \\ \hline -1.5000 \\ \hline \end{array} \; = \; \begin{array}{|r|} \hline 1.0000 \\ \hline 0.0000 \\ \hline -1.0000 \\ \hline \end{array}$$

$$(\mathbf{B}_*)^{\mathrm{T}}\mathbf{w} = \mathbf{w}_* \quad : \quad \begin{array}{|r|r|r|} \hline 1.0000 & -1.0000 & 1.0000 \\ \hline -0.5000 & 2.5000 & -2.5000 \\ \hline 0.5000 & -1.5000 & 2.5000 \\ \hline \end{array} \; * \; \begin{array}{|r|} \hline 0.5000 \\ \hline 0.5000 \\ \hline -1.5000 \\ \hline \end{array} \; = \; \begin{array}{|r|} \hline -1.5000 \\ \hline 4.7500 \\ \hline -4.2500 \\ \hline \end{array}$$

9.2.4 METRICS

Metric of a basis : Let $\mathbf{b}_1,...,\mathbf{b}_n$ be an arbitrary basis of the euclidean vector space $\mathbb{R}^n$. For arbitrary vectors $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$ there is a scalar product $\mathbf{u} \cdot \mathbf{w}$. Thus there is also a scalar product $\mathbf{b}_i \cdot \mathbf{b}_k$ for any two basis vectors $\mathbf{b}_i$, $\mathbf{b}_k$. This scalar product is called a metric coefficient of the basis and is designated by g_{ik}. The metric coefficients are arranged in a quadratic matrix. This matrix is called the metric of the basis and is designated by $\mathbf{G}$. The symmetry $\mathbf{b}_i \cdot \mathbf{b}_k = \mathbf{b}_k \cdot \mathbf{b}_i$ of the scalar product implies $g_{ik} = g_{ki}$, so that the metric $\mathbf{G}$ is symmetric.



Dual metrics : Let the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of the euclidean vector space $\mathbb{R}^n$ be dual. The metric of the covariant basis $\mathbf{B}_*$ is called the covariant metric and is designated by $\mathbf{G}_*$. It contains the covariant metric coefficients g_{ik}. The metric of the contravariant basis $\mathbf{B}^*$ is called the contravariant metric and is designated by $\mathbf{G}^*$. It contains the contravariant metric coefficients g^{ik}.

covariant metric : $g_{ik} = \mathbf{b}_i \cdot \mathbf{b}_k$ $\qquad \mathbf{G}_* = (\mathbf{B}_*)^T \mathbf{B}_*$

contravariant metric : $g^{ik} = \mathbf{b}^i \cdot \mathbf{b}^k$ $\qquad \mathbf{G}^* = (\mathbf{B}^*)^T \mathbf{B}^*$

The metric and the transpose of the dual metric are inverse matrices :

$$(\mathbf{G}_*)^T \mathbf{G}^* = (\mathbf{B}_*)^T \mathbf{B}_* (\mathbf{B}^*)^T \mathbf{B}^* = (\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I}$$

Mixed metric : Let the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of the euclidean vector space $\mathbb{R}^n$ be dual. The scalar product of a basis vector $\mathbf{b}_i \in \mathbf{B}_*$ with a basis vector $\mathbf{b}^k \in \mathbf{B}^*$ is called a mixed metric coefficient and is designated by g_i^k. According to the definition of the dual basis, the coefficient g_i^k coincides with the Kronecker symbol δ_i^k.

mixed metric : $g_i^k = \mathbf{b}_i \cdot \mathbf{b}^k = \mathbf{b}^k \cdot \mathbf{b}_i = \delta_i^k$

$$\mathbf{I} = (\mathbf{B}_*)^T \mathbf{B}^* = (\mathbf{B}^*)^T \mathbf{B}_*$$

Dual basis vectors : The metric may be used to express the relationships between dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$. The relationship $(\mathbf{B}_*)^T\mathbf{B}^* = \mathbf{I}$ is substituted into the definition of the metric $\mathbf{G}_*$:

$$\begin{aligned} \mathbf{G}_* &= (\mathbf{B}_*)^T\mathbf{B}_* = (\mathbf{B}^*)^{-1}\,\mathbf{B}_* \quad \Rightarrow \quad \mathbf{B}_* = \mathbf{B}^*\mathbf{G}_* \\ \mathbf{b}_i &= g_{ik}\,\mathbf{b}^k \end{aligned}$$

The analogous relationship between the bases $\mathbf{B}^*$ and $\mathbf{B}_*$ is obtained by substituting the relationship $(\mathbf{B}_*)^T\mathbf{B}^* = \mathbf{I}$ into the definition of the metric $\mathbf{G}^*$:

$$\begin{aligned} \mathbf{G}^* &= (\mathbf{B}^*)^T\mathbf{B}^* = (\mathbf{B}_*)^{-1}\,\mathbf{B}^* \quad \Rightarrow \quad \mathbf{B}^* = \mathbf{B}_*\mathbf{G}^* \\ \mathbf{b}^i &= g^{ik}\,\mathbf{b}_k \end{aligned}$$

Dual coordinates of a vector : The metric may be used to express the relationships between the covariant form $\mathbf{u}_*$ and the contravariant form $\mathbf{u}^*$ of a vector $\mathbf{u} \in \mathbb{R}^n$. The matrix form of the relationships follows from the coordinate relationships in Section 9.2.3 :

$$\begin{aligned} \mathbf{u} &= \mathbf{B}_*\mathbf{u}^* \quad \wedge \quad \mathbf{u}_* = (\mathbf{B}_*)^T\mathbf{u} \quad \Rightarrow \quad \mathbf{u}_* = \mathbf{G}_*\mathbf{u}^* \\ \mathbf{u} &= \mathbf{B}^*\mathbf{u}_* \quad \wedge \quad \mathbf{u}^* = (\mathbf{B}^*)^T\mathbf{u} \quad \Rightarrow \quad \mathbf{u}^* = \mathbf{G}^*\mathbf{u}_* \end{aligned}$$

The relationships between the covariant coordinates u_i and the contravariant coordinates u^i of the vector $\mathbf{u}$ are expressed in scalar form using the summation convention :

$$\begin{aligned} u_i &= g_{ik}\,u^k \\ u^i &= g^{ik}u_k \end{aligned}$$

Coordinate form of the scalar product : In a euclidean vector space, let the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$, the associated metrics $\mathbf{G}_*$ and $\mathbf{G}^*$ as well as the covariant coordinates u_i, w_k and the contravariant coordinates u^i, w^k of the vectors $\mathbf{u}$ and $\mathbf{w}$ be given. Then the scalar product $\mathbf{u}\cdot\mathbf{w}$ may be expressed in various ways in terms of the covariant and contravariant coordinates of the vectors and the covariant and contravariant metric coefficients :

$$\begin{aligned} &\text{covariant} &&: \quad \mathbf{u}\cdot\mathbf{w} = (u_i\,\mathbf{b}^i)\cdot(w_k\,\mathbf{b}^k) = u_i\,w_k\,g^{ik} \\ &\text{contravariant} &&: \quad \mathbf{u}\cdot\mathbf{w} = (u^i\,\mathbf{b}_i)\cdot(w^k\,\mathbf{b}_k) = u^i\,w^k\,g_{ik} \\ &\text{mixed} &&: \quad \mathbf{u}\cdot\mathbf{w} = (u_i\,\mathbf{b}^i)\cdot(w^k\,\mathbf{b}_k) = u_i\,w^i \\ &\text{mixed} &&: \quad \mathbf{u}\cdot\mathbf{w} = (u^i\,\mathbf{b}_i)\cdot(w_k\,\mathbf{b}^k) = u^i\,w_i \end{aligned}$$

Example 1 : Dual coordinates

The metrics $\mathbf{G}_*$ and $\mathbf{G}^*$ are calculated for the dual bases in Example 2 of Section 9.2.2. The transformation between the dual coordinates $\mathbf{u}_*$ and $\mathbf{u}^*$ of the vector $\mathbf{u}$ in Example 1 of Section 9.2.3 is shown.

$$(\mathbf{B}_*)^T\mathbf{B}_* = \mathbf{G}_*$$

				1.0000	–0.5000	0.5000	$\mathbf{B}_*$
				–1.0000	2.5000	–1.5000	
				1.0000	–2.5000	2.5000	
	1.0000	–1.0000	1.0000	3.0000	–5.5000	4.5000	
	–0.5000	2.5000	–2.5000	–5.5000	12.7500	–10.2500	
$\mathbf{B}_*^T$	0.5000	–1.5000	2.5000	4.5000	–10.2500	8.7500	$\mathbf{G}_*$

$$(\mathbf{B}^*)^T\mathbf{B}^* = \mathbf{G}^*$$

				1.2500	0.5000	0.0000	$\mathbf{B}^*$
				0.0000	1.0000	1.0000	
				–0.2500	0.5000	1.0000	
	1.2500	0.0000	–0.2500	1.6250	0.5000	–0.2500	
	0.5000	1.0000	0.5000	0.5000	1.5000	1.5000	
$\mathbf{B}^{*T}$	0.0000	1.0000	1.0000	–0.2500	1.5000	2.0000	$\mathbf{G}^*$

$$\mathbf{G}^*\mathbf{u}_* = \mathbf{u}^*$$

				16.0000	$\mathbf{u}_*$
				–34.0000	
				28.0000	
	1.6250	0.5000	–0.2500	2.0000	
	0.5000	1.5000	1.5000	–1.0000	
$\mathbf{G}^*$	–0.2500	1.5000	2.0000	1.0000	$\mathbf{u}^*$

Example 2 : Calculation of the scalar product

The scalar product $\mathbf{u}\cdot\mathbf{w}$ of the vectors $\mathbf{u}$ and $\mathbf{w}$ in Example 1 of Section 9.2.3 is calculated using the coordinates in the canonical basis, in the covariant basis $\mathbf{B}_*$ and in the contravariant basis $\mathbf{B}^*$. All of these calculations lead to the same result.

canonical basis : $$\mathbf{u}\cdot\mathbf{w} = 3.0 * 0.5 - 6.0 * 0.5 - 7.0 * 1.5 = -12.00$$

contravariant : $$\begin{aligned} u^i\, w^k\, g_{ik} = {} & 2.00 * (\ 1.00 * 3.00 - 0.00 * 5.50 - 1.00 * 4.50) \\ & - 1.00 * (-1.00 * 5.50 + 0.00 * 12.75 + 1.00 * 10.25) \\ & + 1.00 * (\ 1.00 * 4.50 - 0.00 * 10.25 - 1.00 * 8.75) \\ = {} & -12.00 \end{aligned}$$

covariant : $$\begin{aligned} u_i\, w_k\, g^{ik} = {} & 16.00 * (-1.50 * 1.625 + 4.75 * 0.50 + 4.25 * 0.25) \\ & -34.00 * (-1.50 * 0.50 + 4.75 * 1.50 - 4.25 * 1.50) \\ & +28.00 * (\ 1.50 * 0.25 + 4.75 * 1.50 - 4.25 * 2.00) \\ = {} & -12.00 \end{aligned}$$

mixed : $$\begin{aligned} u_i\, w^i &= 16.00 * 1.00 - 34.00 * 0.00 - 28.00 * 1.00 = -12.00 \\ u^i\, w_i &= -\ 2.00 * 1.50 - 1.00 * 4.75 - 1.00 * 4.25 = -12.00 \end{aligned}$$

9.2.5 CONSTRUCTION OF BASES

Introduction : In a euclidean space $\mathbb{R}^n$ with a given basis **B**, one often seeks an orthonormal basis or the basis dual to **B**. The construction of such bases is treated in the following. Also, an orthonormal basis for the subspace of $\mathbb{R}^n$ spanned by given vectors $\mathbf{u}_1,...,\mathbf{u}_m$ is constructed and extended to an orthonormal basis for the entire space.

Construction of an orthonormal basis : Let a basis **B** in a euclidean space $\mathbb{R}^n$ be given. Then each basis vector of an arbitrary orthonormal basis of $\mathbb{R}^n$ may be represented as a linear combination of the vectors $\mathbf{b}_1,...,\mathbf{b}_n$ of the basis **B**. There is a special orthonormal basis **X** with basis vectors $\mathbf{x}_1,...,\mathbf{x}_n$ which may be constructed from the basis **B** with little effort.

The orthonormal basis **X** is determined in n steps using an orthogonalization procedure. In the m-th step, an auxiliary vector $\mathbf{w}_m$ and the basis vector $\mathbf{x}_m$ are constructed from the basis vector $\mathbf{b}_m$ using the basis vectors $\mathbf{x}_1,...,\mathbf{x}_{m-1}$ determined previously, the scalar products $\mathbf{x}_i \cdot \mathbf{b}_m$ and the magnitude $|\mathbf{w}_m|$. In the first step $\mathbf{w}_1 = \mathbf{b}_1$.

$$\mathbf{w}_m = \mathbf{b}_m - (\mathbf{x}_1 \cdot \mathbf{b}_m)\mathbf{x}_1 - ... - (\mathbf{x}_{m-1} \cdot \mathbf{b}_m)\mathbf{x}_{m-1}$$

$$\mathbf{x}_m = \mathbf{w}_m / |\mathbf{w}_m| \qquad m = 1,...,n$$

Proof : Construction of an orthonormal basis

The basis **X** is constructed by induction. The basis vector $\mathbf{x}_1$ can be determined according to $\mathbf{x}_1 = \mathbf{b}_1 / |\mathbf{b}_1|$, since $\mathbf{b}_1 \neq \mathbf{0}$. For step m the orthonormal basis vectors $\mathbf{x}_1,...,\mathbf{x}_{m-1}$ are assumed to be known.

$$\mathbf{x}_i \cdot \mathbf{x}_k = \delta_{ik} \quad \text{for} \quad i,k = 1,...,m-1$$

In step m the auxiliary vector $\mathbf{w}_m$ is constructed as a linear combination of the vectors $\mathbf{x}_1,...,\mathbf{x}_{m-1}$ and $\mathbf{b}_m$. The coefficients c_m^i are determined such that the auxiliary vector $\mathbf{w}_m$ is orthogonal to each of the vectors $\mathbf{x}_1,...,\mathbf{x}_{m-1}$.

$$\mathbf{w}_m = \mathbf{b}_m + c_m^1\, \mathbf{x}_1 + ... + c_m^{m-1}\, \mathbf{x}_{m-1}$$

$$\mathbf{x}_i \cdot \mathbf{w}_m = 0 \quad \Rightarrow \quad c_m^i = -\mathbf{x}_i \cdot \mathbf{b}_m$$

The basis vector $\mathbf{x}_i$ is a linear combination of the vectors $\mathbf{b}_1,...,\mathbf{b}_i$. Hence $\mathbf{w}_m$ is a linear combination of the vectors $\mathbf{b}_1,...,\mathbf{b}_m$. By the definition of the basis **B**, the auxiliary vector $\mathbf{w}_m$ cannot be the zero vector **0**, since $\mathbf{b}_m$ has the coefficient 1. Hence the vector $\mathbf{w}_m$ may be divided by its magnitude $|\mathbf{w}_m| \neq 0$. This yields the basis vector $\mathbf{x}_m$, which is orthogonal to $\mathbf{x}_1,...,\mathbf{x}_{m-1}$. By virtue of the symmetry $\mathbf{x}_i \cdot \mathbf{x}_m = \mathbf{x}_m \cdot \mathbf{x}_i$ of the scalar product, each of the vectors $\mathbf{x}_1,...,\mathbf{x}_{m-1}$ is also orthogonal to $\mathbf{x}_m$. The orthonormal basis **X** is determined in n steps.

$$\mathbf{x}_m = \mathbf{w}_m / |\mathbf{w}_m|$$

$$\mathbf{x}_i \cdot \mathbf{x}_k = \delta_{ik} \quad \text{for} \quad i,k = 1,...,m$$

Construction of a dual basis : Let a covariant basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ with the vectors $\mathbf{b}_1,...,\mathbf{b}_n$ be given. The basis $\mathbf{B}^*$ dual to $\mathbf{B}_*$ must satisfy the condition $\mathbf{b}_i \cdot \mathbf{b}^k = \delta_i^k$ for $i,k = 1,...,n$. For $i = 1,...,n$, this condition leads to a system of n linear equations. The variables of this system of equations are the coordinates of the covariant basis vector $\mathbf{b}^k$. In the following proof the system of equations is shown to have a unique solution. The n vectors of the basis $\mathbf{B}^*$ are obtained by solving the n systems of equations for $k = 1,...,n$.

$$(\mathbf{B}_*)^T \, \mathbf{b}^k = \mathbf{e}^k \qquad\qquad k = 1,...,n$$

$$\begin{array}{|c|}\hline \mathbf{b}_1^T \\ \hline \vdots \\ \hline \mathbf{b}_k^T \\ \hline \vdots \\ \hline \mathbf{b}_n^T \\ \hline \end{array} \; * \; \begin{array}{|c|}\hline \\ \\ \mathbf{b}^k \\ \\ \\ \hline \end{array} \; = \; \begin{array}{|c|l}\hline 0 & \leftarrow 1 \\ \hline \vdots & \\ \hline 1 & \leftarrow k \\ \hline \vdots & \\ \hline 0 & \leftarrow n \\ \hline \end{array}$$

Proof : Construction of a dual basis

An orthonormal basis $\mathbf{X}$ is constructed for the basis $\mathbf{B}_*$. According to the preceding proof, the basis vector $\mathbf{b}_n$ may be expressed as a linear combination of the basis vectors $\mathbf{x}_1,...,\mathbf{x}_{n-1}$ and the vector $\mathbf{w}_n$:

$$\mathbf{b}_n = \mathbf{w}_n - c_n^1 \, \mathbf{x}_1 - ... - c_n^{n-1} \, \mathbf{x}_{n-1}$$

$$\mathbf{x}_i \cdot \mathbf{w}_n = 0 \qquad \text{for} \qquad i = 1,...,n-1$$

Each of the basis vectors $\mathbf{b}_1,...,\mathbf{b}_{n-1}$ may be represented as a linear combination of the orthogonal basis vectors $\mathbf{x}_1,...,\mathbf{x}_{n-1}$. Hence $\mathbf{w}_n$ is orthogonal to each of the basis vectors $\mathbf{b}_1,...,\mathbf{b}_{n-1}$:

$$\mathbf{b}_i = a_1 \, \mathbf{x}_1 + ... + a_i \, \mathbf{x}_i$$

$$\mathbf{w}_n \cdot \mathbf{b}_i = a_1 \, (\mathbf{w}_n \cdot \mathbf{x}_1) + ... + a_i \, (\mathbf{w}_n \cdot \mathbf{x}_i) = 0$$

With the choice $\mathbf{b}^n = z \, \mathbf{w}_n$, the vector $\mathbf{b}^n$ is orthogonal to $\mathbf{b}_1,...,\mathbf{b}_{n-1}$. Substituting this choice into the condition $\mathbf{b}^n \cdot \mathbf{b}_n = 1$ yields :

$$z(\mathbf{w}_n \cdot \mathbf{w}_n - c_n^1 \, \mathbf{w}_n \cdot \mathbf{x}_1 - ... - c_n^{n-1} \, \mathbf{w}_n \cdot \mathbf{x}_{n-1}) = 1$$

$$\mathbf{b}^n = \frac{\mathbf{w}_n}{\mathbf{w}_n \cdot \mathbf{w}_n}$$

The vectors $\mathbf{b}^{n-1},...,\mathbf{b}^1$ may be constructed analogously by cyclically permuting the basis vectors $\mathbf{b}_1,...,\mathbf{b}_n$.

The vectors $\mathbf{b}^1,...,\mathbf{b}^n$ are linearly independent, since for any linear combination $c_i\,\mathbf{b}^i = \mathbf{0}$ scalar multiplication by $\mathbf{b}_1,...,\mathbf{b}_n$ leads to $c_1 = ... = c_n = 0$. The vectors $\mathbf{b}^1,...,\mathbf{b}^n$ are also unique, for if $\mathbf{y}^1,...,\mathbf{y}^n$ is another dual basis, then $\mathbf{y}^i$ is a linear combination of the basis vectors $\mathbf{b}^1,...,\mathbf{b}^n$. This implies $\mathbf{y}^i = \mathbf{b}^i$ for $i = 1,...,n$:

$$\begin{array}{llllllll} \mathbf{y}^i = c^i_m\,\mathbf{b}^m & \wedge & \mathbf{y}^i \cdot \mathbf{b}_k & = & \delta^i_k & \Rightarrow & c^i_m\,\mathbf{b}^m \cdot \mathbf{b}_k & = \delta^i_k \\ & \Rightarrow & c^i_k & = & \delta^i_k & \Rightarrow & \mathbf{y}^i & = \mathbf{b}^i \end{array}$$

Subspace : Let the vectors $\mathbf{u}_1,...,\mathbf{u}_m$ of the euclidean space $\mathbb{R}^n$ be given. These vectors may be linearly dependent. The set of vectors $\mathbf{s}_k$ which are linear combinations of the vectors $\mathbf{u}_1,...,\mathbf{u}_m$ is called a subspace of $\mathbb{R}^n$ and is designated by S^n. For every pair of vectors in S^n there is a scalar product, since these vectors also lie in the euclidean space $\mathbb{R}^n$. The space S^n is therefore euclidean. S^n is called the space spanned by the vectors $\mathbf{u}_1,...,\mathbf{u}_m$; it is designated by span($\mathbf{u}_1,...,\mathbf{u}_m$).

$$S^n := \text{span}(\mathbf{u}_1,...,\mathbf{u}_m) := \{\mathbf{s}_k \mid \mathbf{s}_k = c^i_k\,\mathbf{u}_i \;\wedge\; c^i_k \in \mathbb{R}\}$$

Basis of a subspace : Let S^n be the subspace spanned by the vectors $\mathbf{u}_1,...,\mathbf{u}_m$ in the euclidean space $\mathbb{R}^n$. Then there is a special orthonormal basis $\mathbf{Y}$ of S^n with the basis vectors $\mathbf{y}_1,...,\mathbf{y}_r$ which may be constructed from the vectors $\mathbf{u}_1,...,\mathbf{u}_m$ with little effort.

The orthonormal basis $\mathbf{Y}$ is determined in m steps using the orthogonalization procedure. In step k, an auxiliary vector $\mathbf{w}_k$ is constructed from the vector $\mathbf{u}_k$, using the basis vectors $\mathbf{y}_1,...,\mathbf{y}_p$ already determined and the scalar products $\mathbf{y}_i \cdot \mathbf{u}_k$. If $\mathbf{w}_k = \mathbf{0}$, the basis is not changed. Otherwise, the basis is augmented by the vector $\mathbf{y}_{p+1} = \mathbf{w}_k / |\mathbf{w}_k|$. In the first step $\mathbf{w}_1 = \mathbf{u}_1$.

$$\begin{array}{lll} \mathbf{w}_k = \mathbf{u}_k - (\mathbf{y}_1 \cdot \mathbf{u}_k)\mathbf{y}_1 - ... - (\mathbf{y}_p \cdot \mathbf{u}_k)\,\mathbf{y}_p & & \\ \mathbf{w}_k \neq \mathbf{0} \quad \Rightarrow \quad \mathbf{y}_{p+1} = \mathbf{w}_k / |\mathbf{w}_k| & & k = 1,...,m \end{array}$$

Proof : Construction of an orthonormal basis of a subspace
Before step k, the basis vectors $\mathbf{y}_1,...,\mathbf{y}_p$ with $p < k$ are determined. In step k, the auxiliary vector $\mathbf{w}_k$ is constructed as a linear combination of the vectors $\mathbf{y}_1,...,\mathbf{y}_p$, $\mathbf{u}_k$. The coefficients c^i_k are determined such that the auxiliary vector $\mathbf{w}_k$ is orthogonal to each of the vectors $\mathbf{y}_1,...,\mathbf{y}_p$.

$$\begin{array}{l} \mathbf{w}_k = \mathbf{u}_k + c^1_k\,\mathbf{y}_1 + ... + c^p_k\,\mathbf{y}_p \\ \mathbf{y}_i \cdot \mathbf{w}_k = 0 \quad \Rightarrow \quad c^i_k = -\mathbf{y}_i \cdot \mathbf{u}_k \end{array}$$

Since the vectors $\mathbf{u}_1,...,\mathbf{u}_k$ may be linearly dependent, the case $\mathbf{w}_k = \mathbf{0}$ may occur. If $\mathbf{w}_k \neq \mathbf{0}$, the basis vector $\mathbf{y}_{p+1} = \mathbf{w}_k / |\mathbf{w}_k|$ is determined. The basis $\mathbf{y}_1,...,\mathbf{y}_r$ is determined in m steps. Each vector $\mathbf{a} \in S^n$ may be expressed as a linear combination of these basis vectors :

$$\mathbf{a} = a^i\,\mathbf{u}_i = a^i\,(z^k_i\,\mathbf{y}_k) = v^k\,\mathbf{y}_k$$

Extension of a basis : Let **Y** be an orthonormal basis of a subspace S^n of the euclidean space $\mathbb{R}^n$. Then there is a special orthonormal basis **X** of the space $\mathbb{R}^n$ which contains **Y** as a subset and may be constructed from the canonical basis **E** of the space $\mathbb{R}^n$ with little effort.

Let the dimension of the basis **Y** be r. The basis vectors $\mathbf{y}_1,...,\mathbf{y}_r$ are copied from the basis **X**, that is $\mathbf{x}_i = \mathbf{y}_i$ for $i = 1,...,r$. The basis vectors $\mathbf{x}_{r+1},...,\mathbf{x}_n$ are determined in at most n steps using the orthogonalization procedure. In step m, an auxiliary vector $\mathbf{w}_m$ is constructed from the vector $\mathbf{e}_m$, using the basis vectors $\mathbf{x}_1,...,\mathbf{x}_p$ already determined and the scalar products $\mathbf{x}_i \cdot \mathbf{e}_m$. If $\mathbf{w}_m = \mathbf{0}$, the basis remains unchanged. Otherwise, the basis is extended by the vector $\mathbf{x}_{p+1} = \mathbf{w}_m / |\mathbf{w}_m|$. The procedure ends with the basis vector $\mathbf{x}_n$ after at most n steps.

$$\mathbf{w}_m = \mathbf{e}_m - (\mathbf{x}_1 \cdot \mathbf{e}_m)\mathbf{x}_1 - ... - (\mathbf{x}_p \cdot \mathbf{e}_m)\mathbf{x}_p$$

$$\mathbf{w}_m \neq \mathbf{0} \quad \Rightarrow \quad \mathbf{x}_{p+1} = \mathbf{w}_m / |\mathbf{w}_m|$$

Proof : Extension of an orthonormal basis

Before the first iteration, the r basis vectors $\mathbf{x}_1,...,\mathbf{x}_r$ which are copied from the basis **Y** of the subspace S^n are known. Before the m-th step, p basis vectors $\mathbf{x}_1,...,\mathbf{x}_p$ with $p < m + r$ are known. In the m-th step, the auxiliary vector $\mathbf{w}_m$ is constructed as a linear combination of the vectors $\mathbf{x}_1,...,\mathbf{x}_p$ and the vector $\mathbf{e}_m$ of the canonical basis **E**. The coefficients c^i_m are determined such that $\mathbf{w}_m$ is orthogonal to each of the vectors $\mathbf{x}_1,...,\mathbf{x}_p$.

$$\mathbf{w}_m = \mathbf{e}_m + c^1_m \mathbf{x}_1 + ... + c^p_m \mathbf{x}_p$$

$$\mathbf{x}_i \cdot \mathbf{w}_m = 0 \quad \Rightarrow \quad c^i_m = - \mathbf{x}_i \cdot \mathbf{e}_m$$

The vector $\mathbf{e}_m$ may be a linear combination of the vectors $\mathbf{x}_1,...,\mathbf{x}_p$, so that $\mathbf{w}_m = \mathbf{0}$ may hold. For $\mathbf{w}_m \neq \mathbf{0}$, the basis vector $\mathbf{x}_{p+1} = \mathbf{w}_m / |\mathbf{w}_m|$ is determined. After step m the space spanned by $\mathbf{x}_1,...,\mathbf{x}_{p+1}$ contains at least the basis vectors $\mathbf{e}_1,...,\mathbf{e}_m$, so that after step n the space spanned by **X** contains the basis **E**. Hence **X** is a basis of $\mathbb{R}^n$.

Example 1 : Construction of an orthonormal basis

Let the covariant basis $\mathbf{B}_*$ of Example 2 in Section 9.2.2 be given. Thus the basis vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ are known. The basis vectors $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are constructed iteratively. The basis $\mathbf{X}$ is orthonormal.

$$\mathbf{b}_1 = \begin{array}{|r|}\hline 1.0000 \\ \hline -1.0000 \\ \hline 1.0000 \\ \hline \end{array} \qquad \mathbf{b}_2 = \begin{array}{|r|}\hline -0.5000 \\ \hline 2.5000 \\ \hline -2.5000 \\ \hline \end{array} \qquad \mathbf{b}_3 = \begin{array}{|r|}\hline 0.5000 \\ \hline -1.5000 \\ \hline 2.5000 \\ \hline \end{array}$$

$\mathbf{w}_1 = \mathbf{b}_1$ $\qquad$ $\mathbf{w}_1 \cdot \mathbf{w}_1 = 3.0000$ $\qquad$ $|\mathbf{w}_1| = 1.7321$

$$\mathbf{x}_1 = \begin{array}{|r|}\hline 0.5774 \\ \hline -0.5774 \\ \hline 0.5774 \\ \hline \end{array}$$

$\mathbf{x}_1 \cdot \mathbf{b}_2 = -\ 3.1754$ $\qquad$ $\mathbf{w}_2 = \mathbf{b}_2 + 3.1754\ \mathbf{x}_1$ $\qquad$ $|\mathbf{w}_2| = 1.6330$

$$\mathbf{w}_2 = \begin{array}{|r|}\hline 1.3333 \\ \hline 0.6667 \\ \hline -0.6667 \\ \hline \end{array} \qquad \mathbf{x}_2 = \begin{array}{|r|}\hline 0.8165 \\ \hline 0.4082 \\ \hline -0.4082 \\ \hline \end{array}$$

$\mathbf{x}_1 \cdot \mathbf{b}_3 = \ \ 2.5980$ $\qquad$ $\mathbf{w}_3 = \mathbf{b}_3 - 2.5980\ \mathbf{x}_1 + 1.2246\ \mathbf{x}_2$

$\mathbf{x}_2 \cdot \mathbf{b}_3 = -\ 1.2246$ $\qquad$ $|\mathbf{w}_3| = 0.7071$

$$\mathbf{w}_3 = \begin{array}{|r|}\hline 0.0000 \\ \hline 0.5000 \\ \hline 0.5000 \\ \hline \end{array} \qquad \mathbf{x}_3 = \begin{array}{|r|}\hline 0.0000 \\ \hline 0.7071 \\ \hline 0.7071 \\ \hline \end{array}$$

$\mathbf{x}_1 \cdot \mathbf{x}_1 = 1.0000$ $\qquad$ $\mathbf{x}_1 \cdot \mathbf{x}_2 = 0.0000$

$\mathbf{x}_2 \cdot \mathbf{x}_2 = 1.0000$ $\qquad$ $\mathbf{x}_2 \cdot \mathbf{x}_3 = 0.0000$

$\mathbf{x}_3 \cdot \mathbf{x}_3 = 1.0000$ $\qquad$ $\mathbf{x}_3 \cdot \mathbf{x}_1 = 0.0000$

Example 2 : Extension of the basis of a subspace

Let the vectors $\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3$ which span a subspace S^4 of the euclidean space $\mathbb{R}^4$ be given : $S^4 = \text{span}(\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3)$. The basis $\mathbf{Y}$ of S^4 contains only two vectors $\mathbf{y}_1$ and $\mathbf{y}_2$, since $\mathbf{u}_3$ is a linear combination of $\mathbf{u}_1$ and $\mathbf{u}_2$ and hence also a linear combination of $\mathbf{y}_1$ and $\mathbf{y}_2$.

$$\mathbf{u}_1 = \begin{bmatrix} 1.0000 \\ -2.0000 \\ 1.0000 \\ -0.5000 \end{bmatrix} \qquad \mathbf{u}_2 = \begin{bmatrix} 3.0000 \\ 0.6000 \\ -1.0000 \\ 1.4000 \end{bmatrix} \qquad \mathbf{u}_3 = \begin{bmatrix} -0.5000 \\ -2.3000 \\ 1.5000 \\ -1.2000 \end{bmatrix}$$

$$\mathbf{w}_1 = \mathbf{u}_1 \qquad \mathbf{w}_1 \cdot \mathbf{w}_1 = 6.2500 \qquad |\mathbf{w}_1| = 2.5000$$

$$\mathbf{y}_1 = \begin{bmatrix} 0.4000 \\ -0.8000 \\ 0.4000 \\ -0.2000 \end{bmatrix}$$

$$\mathbf{y}_1 \cdot \mathbf{u}_2 = 0.0400 \qquad \mathbf{w}_2 = \mathbf{u}_2 - 0.0400\ \mathbf{y}_1 \qquad |\mathbf{w}_2| = 3.5098$$

$$\mathbf{w}_2 = \begin{bmatrix} 2.9840 \\ 0.6320 \\ -1.0160 \\ 1.4080 \end{bmatrix} \qquad \mathbf{y}_2 = \begin{bmatrix} 0.8502 \\ 0.1801 \\ -0.2895 \\ 0.4012 \end{bmatrix}$$

$$\begin{aligned} \mathbf{y}_1 \cdot \mathbf{u}_3 &= 2.4800 \\ \mathbf{y}_2 \cdot \mathbf{u}_3 &= -1.7550 \end{aligned} \qquad \mathbf{w}_3 = \mathbf{u}_3 - 2.4800\ \mathbf{y}_1 + 1.7550\ \mathbf{y}_2 = \mathbf{0}$$

The basis vectors $\mathbf{y}_1$, $\mathbf{y}_2$ of the subspace are copied into the orthonormal basis $\mathbf{X}$ for $\mathbb{R}^4$ with the designations $\mathbf{x}_1$, $\mathbf{x}_2$. The basis vectors $\mathbf{x}_3$ and $\mathbf{x}_4$ are constructed from the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the canonical basis.

$$\begin{aligned} \mathbf{e}_1 \cdot \mathbf{x}_1 &= 0.4000 \\ \mathbf{e}_1 \cdot \mathbf{x}_2 &= 0.8502 \end{aligned} \qquad \begin{aligned} \mathbf{w}_1' &= \mathbf{e}_1 - 0.4000\ \mathbf{x}_1 - 0.8502\ \mathbf{x}_2 \\ |\mathbf{w}_1'| &= 0.3423 \end{aligned}$$

$$\mathbf{w}_1' = \begin{bmatrix} 0.1172 \\ 0.1669 \\ 0.0861 \\ -0.2611 \end{bmatrix} \qquad \mathbf{x}_3 = \begin{bmatrix} 0.3424 \\ 0.4876 \\ 0.2515 \\ -0.7628 \end{bmatrix}$$

$$\begin{aligned} \mathbf{e}_2 \cdot \mathbf{x}_1 &= -0.8000 \\ \mathbf{e}_2 \cdot \mathbf{x}_2 &= 0.1801 \\ \mathbf{e}_2 \cdot \mathbf{x}_3 &= 0.4876 \end{aligned} \qquad \begin{aligned} \mathbf{w}_2' &= \mathbf{e}_2 + 0.8000\ \mathbf{x}_1 - 0.1801\ \mathbf{x}_2 - 0.4876\ \mathbf{x}_3 \\ |\mathbf{w}_2'| &= 0.2997 \end{aligned}$$

$$\mathbf{w}_2' = \begin{bmatrix} 0.0000 \\ 0.0898 \\ 0.2495 \\ 0.1397 \end{bmatrix} \qquad \mathbf{x}_4 = \begin{bmatrix} 0.0000 \\ 0.2996 \\ 0.8325 \\ 0.4661 \end{bmatrix}$$

9.2.6 TRANSFORMATION OF BASES

Introduction : The algebraic structure of a vector space is preserved under a linear mapping to another vector space (see Section 3.6). A linear mapping of a vector space is completely described by the mapping of a basis of the space, since every vector of the space may be represented as a linear combination of the basis vectors $\mathbf{b}^1,...,\mathbf{b}^n$ and the mapping is homomorphic.

$$f(\mathbf{u}) = f(u_1\mathbf{b}^1 + ... + u_n\mathbf{b}^n) = u_1 f(\mathbf{b}^1) + ... + u_n f(\mathbf{b}^n)$$

Affine mapping : A linear mapping between two complete euclidean vector spaces is said to be affine if the two spaces have the same dimension. A linear mapping between bases of such vector spaces is called an affine basis transformation. Affine transformations for covariant and contravariant bases as well as the corresponding transformations of the covariant and contravariant coordinates of vectors are treated in the following.

Transformation of a basis : Let $\mathbf{B}_*$ be a covariant basis of the euclidean space $\mathbb{R}^n$ with the basis vectors $\mathbf{b}_1,...,\mathbf{b}_n$. The basis $\mathbf{B}_*$ is linearly mapped to a covariant basis $\bar{\mathbf{B}}_*$ of $\mathbb{R}^n$ by representing each of the vectors $\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n$ of the basis $\bar{\mathbf{B}}_*$ as a linear combination of the basis vectors $\mathbf{b}_1,...,\mathbf{b}_n$. To describe the inverse transformation, each of the vectors $\mathbf{b}_1,...,\mathbf{b}_n$ of the basis $\mathbf{B}_*$ is represented as a linear combination of the basis vectors $\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n$.

$$\bar{\mathbf{b}}_k = a^1_{.k}\,\mathbf{b}_1 + ... + a^n_{.k}\,\mathbf{b}_n \qquad k = 1,...,n$$

$$\mathbf{b}_k = \bar{a}^1_{.k}\,\bar{\mathbf{b}}_1 + ... + \bar{a}^n_{.k}\,\bar{\mathbf{b}}_n$$

The notation $a^1_{.k}$ and $\bar{a}^1_{.k}$ for the transformation coefficients is compatible with the summation convention and allows the coefficients to be arranged in the transformation matrices $\mathbf{A}$ and $\bar{\mathbf{A}}$. The rule "row index before column index" applies as usual. The placeholder **.** on the level of the subscripts allows the indices to be arranged in columns behind the core symbol a.

$$\begin{array}{c|c}
 & \begin{array}{|c|c|c|c|} \hline a^1_{.1} & a^1_{.2} & \cdots & a^1_{.n} \\ \hline a^2_{.1} & a^2_{.2} & & a^2_{.n} \\ \hline \vdots & & \ddots & \vdots \\ \hline a^n_{.1} & a^n_{.2} & \cdots & a^n_{.n} \\ \hline \end{array} \\
\begin{array}{|c|c|c|c|} \hline \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \\ \hline \end{array} & \begin{array}{|c|c|c|c|} \hline \bar{\mathbf{b}}_1 & \bar{\mathbf{b}}_2 & \cdots & \bar{\mathbf{b}}_n \\ \hline \end{array}
\end{array}
\qquad
\begin{array}{c|c}
 & \begin{array}{|c|c|c|c|} \hline \bar{a}^1_{.1} & \bar{a}^1_{.2} & \cdots & \bar{a}^1_{.n} \\ \hline \bar{a}^2_{.1} & \bar{a}^2_{.2} & & \bar{a}^2_{.n} \\ \hline \vdots & & \ddots & \vdots \\ \hline \bar{a}^n_{.1} & \bar{a}^n_{.2} & \cdots & \bar{a}^n_{.n} \\ \hline \end{array} \\
\begin{array}{|c|c|c|c|} \hline \bar{\mathbf{b}}_1 & \bar{\mathbf{b}}_2 & \cdots & \bar{\mathbf{b}}_n \\ \hline \end{array} & \begin{array}{|c|c|c|c|} \hline \mathbf{b}_1 & \mathbf{b}_2 & \cdots & \mathbf{b}_n \\ \hline \end{array}
\end{array}$$

$$\bar{\mathbf{B}}_* = \mathbf{B}_*\,\mathbf{A} \qquad\qquad \mathbf{B}_* = \bar{\mathbf{B}}_*\,\bar{\mathbf{A}}$$

$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ implies det $\bar{\mathbf{B}}_*$ = det $\mathbf{B}_* \cdot$ det $\mathbf{A}$. Since the determinants of the basis matrices $\mathbf{B}_*$ and $\bar{\mathbf{B}}_*$ are non-zero, the determinant of the transformation matrix $\mathbf{A}$ must also be non-zero. Hence $\mathbf{A}$ has an inverse $\mathbf{A}^{-1}$. Multiplying the equation $\mathbf{B}_* = \bar{\mathbf{B}}_* \bar{\mathbf{A}}$ by $(\mathbf{B}^*)^T$ from the left and substituting $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ yields $\bar{\mathbf{A}} = \mathbf{A}^{-1}$:

$$\begin{aligned} \mathbf{I} &= (\mathbf{B}^*)^T \mathbf{B}_* = (\mathbf{B}^*)^T \bar{\mathbf{B}}_* \bar{\mathbf{A}} = (\mathbf{B}^*)^T (\mathbf{B}_* \mathbf{A}) \bar{\mathbf{A}} = \mathbf{A} \bar{\mathbf{A}} \\ \bar{\mathbf{A}} &= \mathbf{A}^{-1} \end{aligned}$$

Transformation rules for bases : The transformation rule for a covariant basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ is given by $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$. The transformation rule for the dual basis $\mathbf{B}^*$ is accordingly given by $\bar{\mathbf{B}}^* = \mathbf{B}^* \mathbf{C}$. The relationship between the transformation matrices $\mathbf{A}$ and $\mathbf{C}$ follows from the definition of the dual bases :

$$\begin{aligned} \mathbf{I} &= (\bar{\mathbf{B}}_*)^T \bar{\mathbf{B}}^* = (\mathbf{B}_* \mathbf{A})^T (\mathbf{B}^* \mathbf{C}) = \mathbf{A}^T \mathbf{C} \\ \mathbf{C} &= (\mathbf{A}^{-1})^T = \bar{\mathbf{A}}^T \end{aligned}$$

$\bar{\mathbf{B}}^* = \mathbf{B}^* \mathbf{C} = \mathbf{B}^* \bar{\mathbf{A}}^T$ implies $\mathbf{B}^* = \bar{\mathbf{B}}^* \mathbf{A}^T$. Hence the following transformation rules hold for dual bases :

$$\begin{array}{lllllllll} \bar{\mathbf{B}}_* & = & \mathbf{B}_* \mathbf{A} & \bar{\mathbf{b}}_m & = & \mathbf{b}_k \, a^k_{.m} & \bar{b}_{im} & = & b_{ik} \, a^k_{.m} \\ \bar{\mathbf{B}}^* & = & \mathbf{B}^* \bar{\mathbf{A}}^T & \bar{\mathbf{b}}^m & = & \mathbf{b}^k \, \bar{a}^m_{.k} & \bar{b}^{im} & = & b^{ik} \, \bar{a}^m_{.k} \\ \mathbf{B}_* & = & \bar{\mathbf{B}}_* \bar{\mathbf{A}} & \mathbf{b}_m & = & \bar{\mathbf{b}}_k \, \bar{a}^k_{.m} & b_{im} & = & \bar{b}_{ik} \, \bar{a}^k_{.m} \\ \mathbf{B}^* & = & \bar{\mathbf{B}}^* \mathbf{A}^T & \mathbf{b}^m & = & \bar{\mathbf{b}}^k \, a^m_{.k} & b^{im} & = & \bar{b}^{ik} \, \bar{a}^m_{.k} \\ \mathbf{I} & = & \mathbf{A} \, \bar{\mathbf{A}} & & & & \delta^m_i & = & a^m_{.k} \, \bar{a}^k_{.i} \end{array}$$

Rotation of a covariant basis : Let B and C be bases of the vector space $\mathbb{R}^n$. Let the covariant basis vectors be $(\mathbf{b}_1,...,\mathbf{b}_n)$ and $(\mathbf{c}_1,...,\mathbf{c}_n)$. The basis $\mathbf{C}_*$ is said to be a basis obtained by rotating the basis $\mathbf{B}_*$ if the scalar product of every pair of basis vectors in $\mathbf{B}_*$ is equal to the scalar product of the corresponding pair of basis vectors in $\mathbf{C}_*$. Thus the bases $\mathbf{B}_*$ and $\mathbf{C}_*$ lead to the same metric. The magnitudes of the basis vectors in $\mathbf{B}_*$ and $\mathbf{C}_*$ are pairwise equal. The angles between corresponding pairs of basis vectors in $\mathbf{B}_*$ and $\mathbf{C}_*$ are also equal.

$$\mathbf{b}_i \cdot \mathbf{b}_m = \mathbf{c}_i \cdot \mathbf{c}_m \qquad \text{i, m} = 1,...,n$$

The general law for the transformation between the bases B and C is $\mathbf{C}_* = \mathbf{B}_* \mathbf{A}$ with the transformation matrix $\mathbf{A}$. The equality $\mathbf{G}_* = \mathbf{B}_*^T \mathbf{B}_* = \mathbf{C}_*^T \mathbf{C}_*$ of the metrics is due to special properties of $\mathbf{A}$. Setting $\mathbf{C}_* = \mathbf{R}_0 \mathbf{B}_*$ with the rotation matrix $\mathbf{R}_0$ (which is premultiplied, in contrast to $\mathbf{A}$, so that every basis vector in $\mathbf{C}_*$ depends only on the basis vector with the same index in $\mathbf{B}_*$!), one obtains the following rotation conditions :

$$\begin{aligned}
c_{im} &= r_{i.}^{\;k}\, b_{km} && \Leftrightarrow && \mathbf{C}_* = \mathbf{R}_o \mathbf{B}_* \\
\mathbf{C}_*^T \mathbf{C}_* &= \mathbf{B}_*^T \mathbf{R}_o^T \mathbf{R}_o \mathbf{B}_* = \mathbf{B}_*^T \mathbf{B}_* && \Leftrightarrow && \mathbf{R}_o^T \mathbf{R}_o = \mathbf{I} \\
r_{i.}^{\;s}\, r_{m.}^{\;t}\, \delta^{im} &= \delta^{st}
\end{aligned}$$

Hence the rotation matrix $\mathbf{R}_0$ is orthonormal. The transformation matrix $\mathbf{A}$ is derived from the rotation matrix $\mathbf{R}_0$. In contrast to $\mathbf{R}_0$, it is not orthonormal.

$$\begin{aligned}
c_{im} &= b_{ik}\, a^{k}_{\;.m} && \Leftrightarrow && \mathbf{C}_* = \mathbf{B}_* \mathbf{A} \\
\mathbf{C}_* &= \mathbf{B}_* \mathbf{A} = \mathbf{R}_o \mathbf{B}_* && \Leftrightarrow && \mathbf{A} = (\mathbf{B}^*)^T \mathbf{R}_o \mathbf{B}_* \\
a^{i}_{\;.m} &= b^{si}\, r_{s.}^{\;t}\, b_{tm}
\end{aligned}$$

Rotation of a contravariant basis : The rotation conditions for the contravariant basis $\mathbf{B}^*$ dual to $\mathbf{B}_*$ are derived in analogy with the rotation conditions for $\mathbf{B}_*$. The rotated basis is designated by $\mathbf{C}^*$, the rotation matrix by $\mathbf{R}^o$.

$$\begin{aligned}
c^{im} &= r^{i}_{\;.k}\, b^{km} && \Leftrightarrow && \mathbf{C}^* = \mathbf{R}^o \mathbf{B}^* \\
r^{i}_{\;.s}\, r^{m}_{\;.t} \delta_{im} &= \delta^{st} && \Leftrightarrow && (\mathbf{R}^o)^T \mathbf{R}^o = \mathbf{I}
\end{aligned}$$

The general transformation law for the covariant basis is $\mathbf{C}^* = \mathbf{B}^* \bar{\mathbf{A}}^T$ with the transformation matrix $\bar{\mathbf{A}} = \mathbf{A}^{-1}$. The equation $\bar{\mathbf{A}}\,\mathbf{A} = \mathbf{I}$ implies the equation $(\mathbf{R}^o)^T \mathbf{R}^o = \mathbf{I}$ for the rotation matrix of the dual basis.

$$\begin{aligned}
c^{im} &= b^{ik}\, \bar{a}_{k.}^{\;m} && \Leftrightarrow && \mathbf{C}^* = \mathbf{B}^* \bar{\mathbf{A}}^T \\
\mathbf{C}^* &= \mathbf{B}^* \bar{\mathbf{A}}^T = \mathbf{R}^o \mathbf{B}^* && \Leftrightarrow && \bar{\mathbf{A}} = (\mathbf{B}^*)^T (\mathbf{R}^o)^T \mathbf{B}_* \\
\bar{\mathbf{A}}\,\mathbf{A} &= (\mathbf{B}^*)^T (\mathbf{R}^o)^T \mathbf{R}_o \mathbf{B}_* = \mathbf{I} && \Leftrightarrow && (\mathbf{C}^*)^T \mathbf{C}_* = (\mathbf{R}^o)^T \mathbf{R}_o = \mathbf{I}
\end{aligned}$$

Reflection of a basis : Let B and C be bases of a vector space $\mathbb{R}^n$. Let the covariant basis vectors be $(\mathbf{b}_1,...,\mathbf{b}_n)$ and $(\mathbf{c}_1,...,\mathbf{c}_n)$. The basis $\mathbf{C}_*$ is called a reflection of the basis $\mathbf{B}_*$ with respect to the basis vector $\mathbf{b}_k$ if the scalar product of every pair of basis vectors in $\mathbf{B}_*$ is equal to the scalar product of the corresponding pair of basis vectors in $\mathbf{C}_*$, except the scalar products which contain the vector $\mathbf{b}_k$ exactly once : The sign of these products is reversed.

$$\begin{aligned}
\mathbf{c}_i \cdot \mathbf{c}_i &= \mathbf{b}_i \cdot \mathbf{b}_i && i, m = 1,...,n \\
\mathbf{c}_i \cdot \mathbf{c}_m &= \mathbf{b}_i \cdot \mathbf{b}_m && i, m \ne k \\
\mathbf{c}_i \cdot \mathbf{c}_k &= -\,\mathbf{b}_i \cdot \mathbf{b}_k && i \ne k \\
\mathbf{c}_k \cdot \mathbf{c}_m &= -\,\mathbf{b}_k \cdot \mathbf{b}_m && m \ne k
\end{aligned}$$

In the metric of $\mathbf{C}_*$, the coefficients of the metric of $\mathbf{B}_*$ in row k and in column k are multiplied by –1. The diagonal element g_{kk} of the metric remains unchanged.

The diagonal sign matrix $\mathbf{V}_k$ with the diagonal element –1 in row k and the diagonal elements 1 in all other rows is introduced to represent the reflection condition.

$$\mathbf{C}_*^T \mathbf{C}_* = \mathbf{V}_k^T \mathbf{B}_*^T \mathbf{B}_* \mathbf{V}_k$$

$$\mathbf{V}_k = \begin{array}{|c|c|c|c|c|l} \hline 1 & & & & & \\ \hline & \ddots & & & & \\ \hline & & -1 & & & \leftarrow k \\ \hline & & & \ddots & & \\ \hline & & & & 1 & \\ \hline \end{array}$$

The general law for the transformation between the bases $\mathbf{B}_*$ and $\mathbf{C}_*$ is $\mathbf{C}_* = \mathbf{B}_* \mathbf{A}$ with the transformation matrix $\mathbf{A}$. The reflection condition is satisfied due to special properties of $\mathbf{A}$. Setting $\mathbf{C}_* = \mathbf{R}_o \mathbf{B}_* \mathbf{V}_k$, one shows as in the case of a rotation of a basis that $\mathbf{R}_o$ is an orthonormal rotation matrix.

$$\mathbf{C}_*^T \mathbf{C}_* = \mathbf{V}_k^T \mathbf{B}_*^T \mathbf{R}_o^T \mathbf{R}_o \mathbf{B}_* \mathbf{V}_k = \mathbf{V}_k^T \mathbf{B}_*^T \mathbf{B}_* \mathbf{V}_k \quad \Rightarrow \quad \mathbf{R}_o^T \mathbf{R}_o = \mathbf{I}$$

The transformation matrix $\mathbf{A}$ is derived from the rotation matrix $\mathbf{R}_o$ and the sign matrix $\mathbf{V}_k$. The transformation matrix for a reflection differs from the transformation matrix for a rotation in that the sign of the k-th column is reversed.

$$\mathbf{C}_* = \mathbf{B}_* \mathbf{A} = \mathbf{R}_o \mathbf{B}_* \mathbf{V}_k \quad \Rightarrow \quad \mathbf{A} = (\mathbf{B}^*)^T \mathbf{R}_o \mathbf{B}_* \mathbf{V}_k$$

Proper reflection of a basis : A reflection is called a proper reflection if the rotation matrix $\mathbf{R}_o$ is the unit matrix $\mathbf{I}$. A proper reflection reverses the direction of the basis vector $\mathbf{b}_k$ while leaving all other basis vectors unchanged.

$$\mathbf{R}_o = \mathbf{I} \quad \Rightarrow \quad \mathbf{c}_k = -\mathbf{b}_k \quad \wedge \quad \bigwedge_{i \neq k} \mathbf{c}_i = \mathbf{b}_i$$

A basis may be properly reflected at several basis vectors $\{\mathbf{b}_{k_1},\ldots,\mathbf{b}_{k_s}\}$. The rotation matrix of the combined reflection is the unit matrix $\mathbf{I}$. The sign matrix $\mathbf{V}_s$ of the combined reflection contains the diagonal coefficient –1 in rows $k_1,\ldots,k_s$.

$$\mathbf{C}_* = \mathbf{B}_* \mathbf{V}_s \quad \Rightarrow \quad \mathbf{A} = \mathbf{V}_s$$

Transformation of vector coordinates : Let the representation of a vector $\mathbf{u}$ of the euclidean space $\mathbb{R}^n$ in the contravariant basis $\mathbf{B}^*$ of $\mathbb{R}^n$ be $\mathbf{u}_*$ with the coordinates u_i. The basis $\mathbf{B}^*$ is transformed into the basis $\bar{\mathbf{B}}^*$ of $\mathbb{R}^n$ with a matrix $\mathbf{A}$, that is $\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^T$. Let the representation of the vector $\mathbf{u}$ in the transformed basis $\bar{\mathbf{B}}^*$ be $\bar{\mathbf{u}}_*$ with the coordinates $\bar{u}_i$. The transformation rule for the vector coordinates is obtained by substituting the identity $\mathbf{A}\bar{\mathbf{A}} = \mathbf{I}$ into the equation $\mathbf{u} = \mathbf{B}^* \mathbf{u}_*$:

$$\mathbf{u} = u_i \mathbf{b}^i = \mathbf{B}^* \mathbf{u}_*$$

$$\mathbf{u} = \mathbf{B}^* (\mathbf{A}\bar{\mathbf{A}})^T \mathbf{u}_* = (\mathbf{B}^* \bar{\mathbf{A}}^T)(\mathbf{A}^T \mathbf{u}_*) = \bar{\mathbf{B}}^* \bar{\mathbf{u}}_*$$

$$\mathbf{u} = \bar{u}_i \bar{\mathbf{b}}^i \quad \text{with} \quad \bar{\mathbf{u}}_* = \mathbf{A}^T \mathbf{u}_*$$

Transformation rules for vector coordinates : The following transformation rules hold for the coordinates of a vector **u** in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ and in the transformed bases $\bar{\mathbf{B}}_*$ and $\bar{\mathbf{B}}^*$ of the euclidean space :

$$\begin{array}{lllllllll}
\mathbf{u} & = & \mathbf{B}^*\mathbf{u}_* = \bar{\mathbf{B}}^*\bar{\mathbf{u}}_* & \bar{\mathbf{u}}_* & = & \mathbf{A}^T\mathbf{u}_* & \bar{u}_i & = & a^k_{.i}\ u_k \\
\mathbf{u} & = & \mathbf{B}_*\mathbf{u}^* = \bar{\mathbf{B}}_*\bar{\mathbf{u}}^* & \bar{\mathbf{u}}^* & = & \bar{\mathbf{A}}\ \mathbf{u}^* & \bar{u}^i & = & \bar{a}^i_{.k}\ u^k \\
\mathbf{u} & = & \bar{\mathbf{B}}^*\bar{\mathbf{u}}_* = \mathbf{B}^*\mathbf{u}_* & \mathbf{u}_* & = & \bar{\mathbf{A}}^T\bar{\mathbf{u}}_* & u_i & = & \bar{a}^k_{.i}\ \bar{u}_k \\
\mathbf{u} & = & \bar{\mathbf{B}}_*\bar{\mathbf{u}}^* = \mathbf{B}_*\mathbf{u}^* & \mathbf{u}^* & = & \mathbf{A}\ \bar{\mathbf{u}}^* & u^i & = & a^i_{.k}\ \bar{u}^k
\end{array}$$

Transformation of the metric : Let the metric of a covariant basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ be $\mathbf{G}_*$. The basis $\mathbf{B}_*$ is transformed into the basis $\bar{\mathbf{B}}_* = \mathbf{B}_*\mathbf{A}$ using the matrix **A**. Let the metric of the basis $\bar{\mathbf{B}}_*$ be $\bar{\mathbf{G}}_*$. Then the transformation rule for the covariant metric follows from the definition of the metric :

$$\begin{aligned}
\bar{\mathbf{G}}_* &= (\bar{\mathbf{B}}_*)^T\bar{\mathbf{B}}_* = (\mathbf{B}_*\mathbf{A})^T(\mathbf{B}_*\mathbf{A}) = \mathbf{A}^T(\mathbf{B}_*)^T\mathbf{B}_*\mathbf{A} \\
\bar{\mathbf{G}}_* &= \mathbf{A}^T\mathbf{G}_*\mathbf{A}
\end{aligned}$$

The transformation rule for a contravariant metric $\mathbf{G}^*$ follows by substituting the basis transformation $\bar{\mathbf{B}}^* = \mathbf{B}^*\bar{\mathbf{A}}^T$ into the definition of the contravariant metric $\bar{\mathbf{G}}^*$:

$$\begin{aligned}
\bar{\mathbf{G}}^* &= (\bar{\mathbf{B}}^*)^T\bar{\mathbf{B}}^* = (\mathbf{B}^*\bar{\mathbf{A}}^T)^T(\mathbf{B}^*\bar{\mathbf{A}}^T) = \bar{\mathbf{A}}(\mathbf{B}^*)^T\mathbf{B}^*\mathbf{A}^T \\
\bar{\mathbf{G}}^* &= \bar{\mathbf{A}}\ \mathbf{G}^*\ \bar{\mathbf{A}}^T
\end{aligned}$$

The coordinate forms of the transformations of the metric are as follows :

$$\begin{aligned}
\bar{g}_{st} &= a^i_{.s}\ a^k_{.t}\ g_{ik} \\
\bar{g}^{st} &= \bar{a}^s_{.i}\ \bar{a}^t_{.k}\ g^{ik}
\end{aligned}$$

Rotation of a vector : Let $\mathbf{B}_*$ and $\mathbf{B}^*$ be dual bases of the vector space $\mathbb{R}^n$. The basis $\mathbf{B}_*$ is rotated to $\mathbf{C}_*$ using the matrix $\mathbf{R}_o$. Let the basis dual to $\mathbf{C}_*$ be $\mathbf{C}^*$. Let the coordinates of a vector $\mathbf{u} \in \mathbb{R}^n$ be $\mathbf{u}_*$ in the basis $\mathbf{B}^*$ and $\bar{\mathbf{u}}_*$ in the basis $\mathbf{C}^*$.

$$\begin{aligned}
\mathbf{u} &= \mathbf{B}^*\mathbf{u}_* \\
\mathbf{C}_* &= \mathbf{R}_o\,\mathbf{B}_* = \mathbf{B}_*\mathbf{A} \quad \wedge \quad \mathbf{A} = (\mathbf{B}^*)^T\mathbf{R}_o\,\mathbf{B}_*
\end{aligned}$$

The coordinates of the vector **u** satisfy the general transformation rule $\bar{\mathbf{u}}_* = \mathbf{A}^T\mathbf{u}_*$. Substituting **A** shows that $\bar{\mathbf{u}}_*$ contains the coordinates of **u** in the basis $\mathbf{C}^*$.

$$\begin{aligned}
\bar{\mathbf{u}}_* &= \mathbf{A}^T\mathbf{u}_* = \mathbf{B}_*^T\,\mathbf{R}_o^T\,\mathbf{B}^*\,\mathbf{u}_* = (\mathbf{R}_o\,\mathbf{B}_*)^T\,\mathbf{u} = \mathbf{C}_*^T\,\mathbf{u} \\
\mathbf{u} &= \mathbf{C}^*\,\bar{\mathbf{u}}_*
\end{aligned}$$

Proper reflection of a vector : Let $\mathbf{C}_*$ be the proper reflection of a basis $\mathbf{B}_*$ of the space $\mathbb{R}^n$ with respect to the basis vector $\mathbf{b}_k$. Let the coordinates of a vector $\mathbf{u} \in \mathbb{R}^n$ in the dual basis $\mathbf{B}^*$ be $\mathbf{u}_*$. The covariant coordinates in the basis $\mathbf{C}^*$ are transformed using the general rule $\bar{\mathbf{u}}_* = \mathbf{A}^T \mathbf{u}_*$. Substituting $\mathbf{A} = \mathbf{V}_k$ for the proper reflection shows that $\bar{\mathbf{u}}_*$ contains the coordinates of $\mathbf{u}$ in the basis $\mathbf{C}^*$. The vector is reflected by reversing the sign of its k-th coordinate.

$$\begin{aligned} \bar{\mathbf{u}}_* &= \mathbf{A}^T \mathbf{u}_* \quad \wedge \quad \mathbf{A} = \mathbf{V}_k \\ \bar{\mathbf{u}}_* &= \mathbf{V}_k^T \mathbf{B}_*^T \mathbf{u} = \mathbf{C}_*^T \mathbf{u} \\ \mathbf{u} &= \mathbf{C}^* \bar{\mathbf{u}}_* \end{aligned}$$

Example 1 : Transformation of a basis

Let the covariant basis $\mathbf{B}_*$ and the contravariant basis $\mathbf{B}^*$ of Example 2 in Section 9.2.2 as well as the transformation matrix $\mathbf{A}$ and its inverse $\bar{\mathbf{A}} = \mathbf{A}^{-1}$ be given. The transformed bases $\bar{\mathbf{B}}_*$ and $\bar{\mathbf{B}}^*$ are calculated. Their product is the identity matrix $\mathbf{I}$.

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$$

0.8000	–0.5000	0.4000
–0.2000	1.2000	–0.6000
0.5000	–0.8000	1.0000

1.0000	–0.5000	0.5000
–1.0000	2.5000	–1.5000
1.0000	–2.5000	2.5000

1.1500	–1.5000	1.2000
–2.0500	4.7000	–3.4000
2.5500	–5.5000	4.4000

$$\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^T$$

1.6000	–0.2222	–0.9778
0.4000	1.3333	0.8667
–0.4000	0.8889	1.9111

1.2500	0.5000	0.0000
0.0000	1.0000	1.0000
–0.2500	0.5000	1.0000

2.2000	0.3889	–0.7889
0.0000	2.2222	2.7778
–0.6000	1.6111	2.5889

$$(\bar{\mathbf{B}}_*)^T \bar{\mathbf{B}}^* = \mathbf{I}$$

2.2000	0.3889	–0.7889
0.0000	2.2222	2.7778
–0.6000	1.6111	2.5889

1.1500	–2.0500	2.5500
–1.5000	4.7000	–5.5000
1.2000	–3.4000	4.4000

1.0000	0.0000	0.0000
0.0000	1.0000	0.0000
0.0000	0.0000	1.0000

Example 2 : Transformation of a vector

Let the vector **w** of Example 1 in Section 9.2.3 with the covariant representation $\mathbf{w}_*$ and the contravariant representation $\mathbf{w}^*$ in the bases $\mathbf{B}^*$ and $\mathbf{B}_*$ of the preceding example be given. The representations $\bar{\mathbf{w}}_*$ and $\bar{\mathbf{w}}^*$ of the vector **w** in the transformed bases $\bar{\mathbf{B}}^*$ and $\bar{\mathbf{B}}_*$ are calculated. A check confirms that $\bar{\mathbf{B}}^* \bar{\mathbf{w}}_* = \mathbf{w}$.

$$\bar{\mathbf{w}}_* = \mathbf{A}^T \mathbf{w}_*$$

–1.5000
4.7500
–4.2500

0.8000	–0.2000	0.5000
–0.5000	1.2000	–0.8000
0.4000	–0.6000	1.0000

–4.2750
9.8500
–7.7000

$$\bar{\mathbf{w}}^* = \bar{\mathbf{A}}\, \mathbf{w}^*$$

1.0000
0.0000
–1.0000

1.6000	0.4000	–0.4000
–0.2222	1.3333	0.8889
–0.9778	0.8667	1.9111

2.0000
–1.1111
–2.8889

$$\bar{\mathbf{B}}^* \bar{\mathbf{w}}_* = \mathbf{w}$$

–4.2750
9.8500
–7.7000

2.2000	0.3889	–0.7889
0.0000	2.2222	2.7778
–0.6000	1.6111	2.5889

0.5000
0.5000
–1.5000

Example 3 : Rotation of a basis in the space $\mathbb{R}^2$

Let the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ and the orthonormal matrix $\mathbf{R}$ be given. The dual bases $\mathbf{C}_* = \mathbf{R}\mathbf{B}_*$ and $\mathbf{C}^* = \mathbf{R}\mathbf{B}^*$ are determined by rotating B with $\mathbf{R}$.

$$\mathbf{B}_* = \frac{1}{2}\begin{bmatrix} 4 & 1 \\ 2 & 2 \end{bmatrix} \qquad \mathbf{B}^* = \frac{1}{3}\begin{bmatrix} 2 & -2 \\ -1 & 4 \end{bmatrix} \qquad \mathbf{R} = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$$

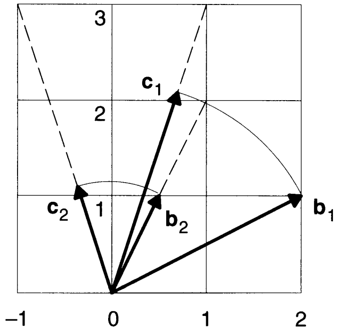
$$\mathbf{C}_* = \frac{1}{2\sqrt{2}}\begin{bmatrix} 2 & -1 \\ 6 & 3 \end{bmatrix} \qquad \mathbf{C}^* = \frac{1}{3\sqrt{2}}\begin{bmatrix} 3 & -6 \\ 1 & 2 \end{bmatrix} \qquad \mathbf{C}_*^T\,\mathbf{C}^* = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

The transformation matrix $\mathbf{D} = (\mathbf{B}^*)^T\mathbf{R}\,\mathbf{B}_*$ in $\mathbf{C}_* = \mathbf{B}_*\,\mathbf{D}$ is not orthonormal. The transformation law $\mathbf{C}_* = \mathbf{B}_*\,\mathbf{D}$ is satisfied.

$$\mathbf{D} = \frac{1}{6\sqrt{2}}\begin{bmatrix} -2 & -5 \\ 20 & 14 \end{bmatrix} \qquad \mathbf{D}^T\mathbf{D} = \frac{1}{72}\begin{bmatrix} 404 & 290 \\ 290 & 221 \end{bmatrix}$$

$$\mathbf{C}_* = \frac{1}{12\sqrt{2}}\begin{bmatrix} 4 & 1 \\ 2 & 2 \end{bmatrix} * \begin{bmatrix} -2 & -5 \\ 20 & 14 \end{bmatrix} = \frac{1}{2\,\sqrt{2}}\begin{bmatrix} 2 & -1 \\ 6 & 3 \end{bmatrix}$$

The images of the basis vectors show the rotation through the angle $\pi/4$:



9.2.7 ORIENTATION AND VOLUME

Determinants of bases : The basis matrix **B** of a basis of a euclidean space $\mathbb{R}^n$ is quadratic and has a determinant det **B**. Since the basis vectors are linearly independent, this determinant is non-zero. It may be positive or negative.

$$\text{general} \quad : \quad \det \mathbf{B} \neq 0$$

The absolute value of the determinant of an orthonormal basis **B** is 1, since the determinant of a matrix product is equal to the product of the determinants of the matrices and the determinant of a matrix is equal to the determinant of its transpose. The determinant of the canonical basis **E** is 1.

$$\begin{array}{lllll} \text{orthonormal} & : & \mathbf{B}^T\mathbf{B} = \mathbf{I} & \Rightarrow \quad \det \mathbf{B} \cdot \det \mathbf{B} = 1 & \Rightarrow \quad |\det \mathbf{B}| = 1 \\ \text{canonical} & : & \det \mathbf{E} = 1 & & \end{array}$$

The determinants of dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ are reciprocal values. The determinant of the metric **G** of a basis **B** is always positive.

$$\begin{array}{llllll} \text{dual} & : & (\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I} & \Rightarrow & \det \mathbf{B}_* \cdot \det \mathbf{B}^* = 1 & \\ \text{metric} & : & (\mathbf{B}_*)^T \mathbf{B}_* = \mathbf{G}_* & \Rightarrow & \det \mathbf{G}_* = (\det \mathbf{B}_*)^2 > 0 & \\ & & (\mathbf{B}^*)^T \mathbf{B}^* = \mathbf{G}^* & \Rightarrow & \det \mathbf{G}^* = (\det \mathbf{B}^*)^2 > 0 & \\ & & \det \mathbf{G}_* \cdot \det \mathbf{G}^* & = & (\det \mathbf{B}_* \cdot \det \mathbf{B}^*)^2 & = 1 \end{array}$$

Orientation of bases : The transformation of a basis **B** with a transformation matrix **A** is invertible. Hence the determinants of the transformation matrix **A** and of its inverse $\bar{\mathbf{A}}$ are non-zero.

$$\begin{array}{lll} \bar{\mathbf{B}} = \mathbf{B}\mathbf{A} \;\wedge\; \mathbf{B} = \bar{\mathbf{B}}\bar{\mathbf{A}} & \Rightarrow & \bar{\mathbf{B}} = \bar{\mathbf{B}}\bar{\mathbf{A}}\mathbf{A} \\ \det \bar{\mathbf{A}} \cdot \det \mathbf{A} = 1 & \Rightarrow & \det \mathbf{A},\ \det \bar{\mathbf{A}} \neq 0 \end{array}$$

The bases **B** and $\bar{\mathbf{B}}$ are said to have the same orientation if the determinants of the basis matrices have the same sign. With $\bar{\mathbf{B}} = \mathbf{B}\mathbf{A}$, the bases have the same orientation if the determinant of the transformation matrix **A** is positive :

$$\begin{array}{lll} \bar{\mathbf{B}} = \mathbf{B}\mathbf{A} & \Rightarrow & \det \bar{\mathbf{B}} = \det \mathbf{B} \cdot \det \mathbf{A} \\ \det \mathbf{A} > 0 & \Rightarrow & \det \mathbf{B} \cdot \det \bar{\mathbf{B}} > 0 \end{array}$$

The relation "identically oriented" is an equivalence relation in the set of bases of a space $\mathbb{R}^n$ which partitions this set into two classes. The one class contains bases whose determinant is positive. The determinants of the bases in the other class are negative. The space $\mathbb{R}^n$ is equipped with an orientation if one of the two classes is defined as positive (right-handed), the other as negative (left-handed). The class of bases with positive determinant is usually taken to be positively oriented (right-handed). This orientation of the space is assumed in the following. Since $\det \mathbf{B}_* \cdot \det \mathbf{B}^* = 1$, dual bases have the same orientation.

Volume of a basis : The volume of a body is defined in the geometry of the euclidean space $\mathbb{R}^3$. As a generalization of this concept, the volume of a basis $\mathbf{B}$ of the euclidean space $\mathbb{R}^n$ is defined in vector algebra. The determinant of the basis matrix $\mathbf{B}$ is called the volume of the basis and is designated by b. The sign of the volume corresponds to the orientation of the basis. Since bases with positive determinants are taken to be positively oriented, $b = \det \mathbf{B}$. The volumes of covariant and contravariant bases are different and are distinguished using the symbol $*$.

$$b = \det \mathbf{B} \qquad b_* = \det \mathbf{B}_* \qquad b^* = \det \mathbf{B}^*$$

According to this definition, the volume of the canonical basis of the euclidean space $\mathbb{R}^n$ is given by $\det \mathbf{E} = 1$. All orthonormal bases $\mathbf{B}$ have the volume 1 or -1, since $|\det \mathbf{B}| = 1$. The volumes of dual bases are reciprocal values. For a transformation of an arbitrary basis $\mathbf{B}$ with the matrix $\mathbf{A}$, the volume of the transformed basis $\bar{\mathbf{B}}$ is given by the product of the determinants of $\mathbf{B}$ and $\mathbf{A}$:

$$\text{dual} \quad : \quad (\mathbf{B}_*)^T \mathbf{B}^* = \mathbf{I} \quad \Rightarrow \quad b_* \cdot b^* = 1$$

$$\text{general} \quad : \quad \bar{\mathbf{B}} = \mathbf{BA} \quad \Rightarrow \quad \det \bar{\mathbf{B}} = \det \mathbf{B} \cdot \det \mathbf{A} \quad \Rightarrow \quad \bar{b} = b \det \mathbf{A}$$

Example 1 : Reversal of the orientation of a basis

Exactly one basis vector $\mathbf{e}_m$ in the canonical basis $\mathbf{E}$ is replaced by the inverse vector $-\mathbf{e}_m$. The resulting basis is designated by $\mathbf{E}'$; its determinant is given by $\det \mathbf{E}' = -1$.

The orientation of a general basis $\mathbf{B}$ is reversed by replacing exactly one basis vector $\mathbf{b}_m$ by the inverse vector $-\mathbf{b}_m$. The resulting basis is designated by $\mathbf{B}'$.

$$\mathbf{B}' = \mathbf{B}\mathbf{E}' \quad \Rightarrow \quad \det \mathbf{B}' = \det \mathbf{B} \cdot \det \mathbf{E}' = -\det \mathbf{B}$$

9.3 TENSOR ALGEBRA

9.3.1 INTRODUCTION

Tensors may be defined in different vector spaces and with different metrics. This section treats only the algebra of tensors which are defined as linear scalar mappings of vector m-tuples in euclidean real vector spaces. A different basis of the vector space may be freely chosen for every vector of the m-tuple. However, the coordinates of all vectors of the m-tuple are often referred to the same basis or to a pair of dual bases. This is assumed in the following unless explicitly stated otherwise.

A tensor is described by its coordinates. Each coordinate is the image of an m-tuple of basis vectors. The image of an arbitrary vector m-tuple is obtained by expressing each vector of the m-tuple as a linear combination of the associated basis vectors. Since the mapping is linear, the image of the vector m-tuple is a linear combination of the coordinates of the tensor. The tensor is represented using covariant, contravariant or mixed coordinates, according to the basis chosen.

Tensor operations allow new tensors to be constructed from given tensors. There are fixed rules for determining the coordinates of the new tensors from the coordinates of the given tensors. For example, the sum, the product, the contraction and the contracted product of tensors are also tensors. These tensor operations are often used to express relationships between physical quantities.

The values of the coordinates of a tensor may satisfy special conditions. This leads to special tensors : unit tensors, metric tensors, isotropic tensors as well as symmetric and antisymmetric tensors. The completely antisymmetric permutation tensor of a vector space is used to express determinants and vector products (for instance to determine area vectors). The coordinates of the permutation tensor depend only on the determinant of the basis of the vector space. The product of the covariant and the contravariant permutation tensor of a space is the unit tensor, whose coordinates are basis-independent. Since contracted products of permutation tensors often occur in physical problems, rules for contracting the unit tensor are derived.

Every tensor is a mapping of vector m-tuples. The coordinates of the tensor refer to a certain basis. If the basis for a vector of the m-tuple is changed, the coordinates of the tensor change. There are fixed rules for determining the coordinates of the tensor in the new basis if the values of the coordinates in the old basis and the transformation rules for the basis are given. Two cases of these rules are distinguished : transformations from a basis to its dual basis and transformations from a basis to an arbitrary basis of the space.

The number m of vectors in a tensor definition is called the rank of the tensor. Tensors of rank 1 are often called vectors. Their coordinates are represented as vectors. Tensors of rank 2 are called dyads. Their coordinates are represented as matrices. The transformation rules for tensors of ranks 1 and 2 may alternatively be represented either in coordinate notation or in matrix notation. Tensors of ranks 1 and 2 are combined to yield bilinear and quadratic forms, inner products and scalar products.

The values of the coordinates of dyads may satisfy certain conditions. This leads to special dyads : zero dyad, unit dyad, regular dyad, unitary dyad, symmetric and antisymmetric dyads. The decomposition of general dyads into the sum of a symmetric and an antisymmetric dyad, the determination of eigenvalues and eigenvectors for symmetric dyads and the decomposition of regular dyads into the product of a unitary and a symmetric dyad are treated.

9.3.2 TENSORS

Introduction : A basis of the euclidean vector space $\mathbb{R}^n$ is not unique. However, there are quantities in the space $\mathbb{R}^n$ whose value is uniquely determined and therefore independent of the arbitrarily chosen basis. Such quantities are essential for the understanding and mathematical description of physical phenomena. For example, the following quantities are independent of the choice of basis :

(1) The magnitude of a vector.
(2) The angle between two vectors.
(3) The volume of a parallelepiped.
(4) The magnitude of the component of the stress vector in a fixed direction on a surface with fixed surface normal.

The examples show that the quantities which are independent of the basis are described by vectors which are fixed relative to the canonical basis :

(1) The vector whose magnitude is determined.
(2) The two vectors which enclose the angle.
(3) The n independent vectors which define the edges of the parallelepiped.
(4) The normal vector of the surface and the direction vector of the stress component.

The examples also show that the value of the basis-independent quantity is a scalar. Changes in the vectors which describe the quantity lead to changes in the scalar :

(1) The value of the magnitude of the vector.
(2) The value of the angle.
(3) The value of the volume.
(4) The value of the magnitude of the stress component.

The examples show that a quantity is independent of the choice of basis if it is specified by vectors whose image is a scalar. A transformation of the basis in which the vector coordinates are specified does not affect the value of this scalar, since the vectors are fixed relative to the canonical basis.

Vector mappings : A mapping is called a vector mapping (vector function) if the domain of the mapping consists of m-tuples $(\mathbf{u}_1,...,\mathbf{u}_m)$ of vectors. The vector mapping is said to be scalar-valued (scalar) if the target consists of scalars (numbers). The vector mapping is said to be vector-valued (vectorial) if the target consists of vectors. The vector mapping is said to be real if the scalars and the coordinates of the vectors are real numbers.

scalar : $t : \mathbb{R}^n \times ... \times \mathbb{R}^n \rightarrow \mathbb{R}$ with $t(\mathbf{u}_1,...,\mathbf{u}_m) = w$
vectorial : $t : \mathbb{R}^n \times ... \times \mathbb{R}^n \rightarrow \mathbb{R}^s$ with $t(\mathbf{u}_1,...,\mathbf{u}_m) = \mathbf{w}$

Linear scalar vector mapping : A scalar vector mapping is said to be linear if it is structurally compatible (see Section 3.6). Thus if a vector $\mathbf{u}_i$ in the m-tuple is replaced by the vector sum $\mathbf{a}+\mathbf{b}$, the image of the m-tuple is equal to the sum of the images of the m-tuples with the vectors $\mathbf{a}$ and $\mathbf{b}$. If a vector $\mathbf{u}_i$ of the m-tuple is replaced by its s-fold multiple $s\mathbf{u}_i$, then the image of the m-tuple is the s-fold multiple of the image of the m-tuple with the vector $\mathbf{u}_i$.

$$\begin{aligned} t(\mathbf{u}_1,...,\mathbf{a}+\mathbf{b},...,\mathbf{u}_m) &= t(\mathbf{u}_1,...,\mathbf{a},...,\mathbf{u}_m) + t(\mathbf{u}_1,...,\mathbf{b},...,\mathbf{u}_m) \\ t(\mathbf{u}_1,...,s\mathbf{u}_i,...,\mathbf{u}_m) &= s\,t(\mathbf{u}_1,...,\mathbf{u}_i,...,\mathbf{u}_m) \end{aligned}$$

Tensor : Consider the m-fold cartesian product $\mathbb{R}^n \times ... \times \mathbb{R}^n$ of n-dimensional spaces $\mathbb{R}^n$. The integers $m,n \geq 0$ are arbitrary. A linear scalar mapping t of the m-tuples of vectors is called an n-dimensional tensor of rank m. The tensor is said to be real if the vector mapping t is real.

$$\begin{aligned} &t : \mathbb{R}^n \times ... \times \mathbb{R}^n \rightarrow \mathbb{R} \quad \text{with} \quad t(\mathbf{u}_1,...,\mathbf{u}_m) = w \\ &t(\mathbf{a}+\mathbf{b},\mathbf{c}) = t(\mathbf{a},\mathbf{c}) + t(\mathbf{b},\mathbf{c}) \\ &t(p\mathbf{a},\mathbf{c}) \quad = p\,t(\mathbf{a},\mathbf{c}) \end{aligned}$$

Example 1 : Tensors

(1) Every mapping $t : \mathbb{R} \rightarrow \mathbb{R}$ from scalars to scalars is a tensor of rank 0.

(2) Let the vector $\mathbf{f} \in \mathbb{R}^n$ be fixed. The mapping $t : \mathbb{R}^n \rightarrow \mathbb{R}$ with $t(\mathbf{u}) = \mathbf{f} \cdot \mathbf{u}$ is linear due to the properties of the scalar product in Section 9.2.1. It defines a tensor of rank 1 with target T :

$$\begin{aligned} &T = \{a \in \mathbb{R} \mid a = \mathbf{f} \cdot \mathbf{u} \ \wedge \ \mathbf{u} \in \mathbb{R}^n\} \\ &t(\mathbf{u}+s\mathbf{w}) = \mathbf{f} \cdot (\mathbf{u}+s\mathbf{w}) = \mathbf{f} \cdot \mathbf{u} + s\mathbf{f} \cdot \mathbf{w} = t(\mathbf{u}) + s\,t(\mathbf{w}) \end{aligned}$$

(3) The mapping $t : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ with $t(\mathbf{u},\mathbf{w}) = \mathbf{u} \cdot \mathbf{w}$ is linear due to the properties of the scalar product in Section 9.2.1. It defines a tensor of rank 2 with target T :

$$\begin{aligned} &T = \{a \in \mathbb{R} \mid a = \mathbf{u} \cdot \mathbf{w} \ \wedge \ \mathbf{u},\mathbf{w} \in \mathbb{R}^n\} \\ &t(\mathbf{a}+s\mathbf{b},\mathbf{w}) = (\mathbf{a}+s\mathbf{b}) \cdot \mathbf{w} = \mathbf{a} \cdot \mathbf{w} + s\,\mathbf{b} \cdot \mathbf{w} = t(\mathbf{a},\mathbf{w}) + s\,t(\mathbf{b},\mathbf{w}) \end{aligned}$$

(4) Any n linearly independent vectors $\mathbf{b}_1,...,\mathbf{b}_n$ form a basis $\mathbf{B}$ of the space $\mathbb{R}^n$. The volume of the basis is $b = \det \mathbf{B}$. The mapping $t : \mathbb{R}^n \times ... \times \mathbb{R}^n \rightarrow \mathbb{R}$ with $t(\mathbf{b}_1,..., \mathbf{b}_n) = \det \mathbf{B}$ is linear by virtue of the properties of determinants. It defines a tensor of rank n with target T :

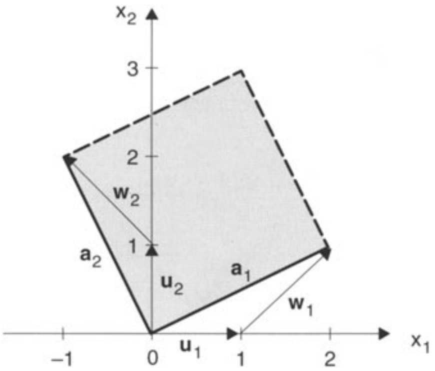
$$\begin{aligned} &T = \{a \in \mathbb{R} \mid a = \det \mathbf{B} \ \wedge \ \mathbf{b}_1,...,\mathbf{b}_n \in \mathbb{R}^n\} \\ &t(\mathbf{u}+s\mathbf{w}, \mathbf{b}_2,...,\mathbf{b}_n) = t(\mathbf{u}, \mathbf{b}_2,...,\mathbf{b}_n) + s\,t(\mathbf{w}, \mathbf{b}_2,..., \mathbf{b}_n) \end{aligned}$$

Example 2 : Linearity of the volume tensor of rank 2

The vectors $\mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^2$ define a surface. The linearity of the mapping $t : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ with $t(\mathbf{a}_1, \mathbf{a}_2) = \det[\mathbf{a}_1, \mathbf{a}_2]$ is shown for the components $\mathbf{u}_1, \mathbf{w}_1$ and $\mathbf{u}_2, \mathbf{w}_2$ of the vectors $\mathbf{a}_1, \mathbf{a}_2$.

$$t(\mathbf{u}_1 + \mathbf{w}_1, \mathbf{a}_2) = t(\mathbf{u}_1, \mathbf{a}_2) + t(\mathbf{w}_1, \mathbf{a}_2)$$

$$t(\mathbf{u}_1 + \mathbf{w}_1, \mathbf{u}_2 + \mathbf{w}_2) = t(\mathbf{u}_1, \mathbf{u}_2) + t(\mathbf{u}_1, \mathbf{w}_2) + t(\mathbf{w}_1, \mathbf{u}_2) + t(\mathbf{w}_1, \mathbf{w}_2)$$



$$\mathbf{a}_1 = \begin{pmatrix} 2 \\ 1 \end{pmatrix} \qquad \mathbf{a}_2 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

$$\mathbf{u}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \qquad \mathbf{u}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

$$\mathbf{w}_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \qquad \mathbf{w}_2 = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

$$t(\mathbf{a}_1, \mathbf{a}_2) = \det \begin{bmatrix} 2 & -1 \\ 1 & 2 \end{bmatrix} = 5$$

$$t(\mathbf{u}_1 + \mathbf{w}_1, \mathbf{a}_2) = \det \begin{bmatrix} 1 & -1 \\ 0 & 2 \end{bmatrix} + \det \begin{bmatrix} 1 & -1 \\ 1 & 2 \end{bmatrix} = 2 + 3 = 5$$

$$t(\mathbf{u}_1 + \mathbf{w}_1, \mathbf{u}_2 + \mathbf{w}_2) = \det \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} + \det \begin{bmatrix} 1 & -1 \\ 0 & 1 \end{bmatrix} +$$

$$\det \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} + \det \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = 1 + 1 + 1 + 2 = 5$$

Tensor coordinates : Let a tensor T in the euclidean space $\mathbb{R}^n$ be defined by the linear mapping $t(\mathbf{u},\mathbf{v},...,\mathbf{w})$ of an m-tuple of vectors. Every vector of the m-tuple $(\mathbf{u},...,\mathbf{w})$ may be represented as a linear combination of a basis $\mathbf{b}^1,...,\mathbf{b}^n$ of $\mathbb{R}^n$. By the linearity of the mapping t, the image of the m-tuple $(\mathbf{u},...,\mathbf{w})$ may be replaced by a linear combination of images of m-tuples of basis vectors :

$$\mathbf{u} = u_i\,\mathbf{b}^i \qquad \mathbf{v} = v_j\,\mathbf{b}^j \qquad \mathbf{w} = w_k\,\mathbf{b}^k$$

$$\begin{aligned} t(\mathbf{u},\mathbf{v},...,\mathbf{w}) &= t(u_i\,\mathbf{b}^i, \mathbf{v},...,\mathbf{w}) \\ &= u_i\,t(\mathbf{b}^i, \mathbf{v},...,\mathbf{w}) \\ &= u_i\,v_j\,t(\mathbf{b}^i, \mathbf{b}^j,...,\mathbf{w}) \\ &= u_i\,v_j...w_k\,t(\mathbf{b}^i, \mathbf{b}^j,...,\mathbf{b}^k) \end{aligned}$$

An image $t(\mathbf{b}^i, \mathbf{b}^j,...,\mathbf{b}^k)$ of basis vectors is called a coordinate of the tensor T in the basis $\mathbf{B}^* = \{\mathbf{b}^1,...,\mathbf{b}^n\}$. This coordinate is designated by $t^{i\,j...k}$. Since each of the m indices i,...,k can take the values 1,...,n, the tensor has n^m coordinates. The scalar value of the image $t(\mathbf{u},...,\mathbf{w})$ of an arbitrary m-tuple $(\mathbf{u},...,\mathbf{w})$ of vectors may be determined if the coordinates $u_i,..., w_k$ of the vectors and the coordinates $t^{i...k}$ of the tensor are known :

$$\begin{aligned} t^{i...k} &:= t(\mathbf{b}^i,...,\mathbf{b}^k) \\ t(\mathbf{u},...,\mathbf{w}) &= u_i\,...\,w_k\,t^{i...k} \end{aligned}$$

Types of tensor coordinates : Let a tensor T in the euclidean space $\mathbb{R}^n$ be defined by the linear mapping $t(\mathbf{u},...,\mathbf{w})$ of an m-tuple of vectors. Each vector $\mathbf{v}$ of the m-tuple may be associated with a different basis. The covariant coordinates $v_1,...,v_n$ of the vector $\mathbf{v}$ in the contravariant basis $\mathbf{b}^1,...,\mathbf{b}^n$ are arranged in a vector $\mathbf{v}_*$. Likewise, its contravariant coordinates $v^1,...,v^n$ in the covariant basis $\mathbf{b}_1,...,\mathbf{b}_n$ are arranged in a vector $\mathbf{v}^*$. Thus the vector $\mathbf{v}$ is represented as follows :

$$\mathbf{v} = v_i\,\mathbf{b}^i = \mathbf{B}^*\,\mathbf{v}_* \qquad\qquad \mathbf{v} = v^i\,\mathbf{b}_i = \mathbf{B}_*\,\mathbf{v}^*$$

The i-th vector of the m-tuple is associated with the i-th index column of the tensor coordinates. The position (subscript or superscript) of the index in the column corresponds to the position of the index of the basis vector. The index columns are arranged by using the symbol **.** (point) in free subscript positions. For example, if all vectors of a 4-tuple are described in the dual bases $\mathbf{b}_1,...,\mathbf{b}_n$ or $\mathbf{b}^1,...,\mathbf{b}^n$, the mapping $t(\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{z})$ may be described as follows :

$$\mathbf{u} = u_{s_1}\,\mathbf{b}^{s_1} \qquad\qquad \mathbf{w} = w_{s_3}\,\mathbf{b}^{s_3}$$

$$\mathbf{v} = v^{s_2}\,\mathbf{b}_{s_2} \qquad\qquad \mathbf{z} = z_{s_4}\,\mathbf{b}^{s_4}$$

$$\begin{aligned} t(\mathbf{b}^{s_1}, \mathbf{b}_{s_2}, \mathbf{b}^{s_3}, \mathbf{b}^{s_4}) &:= t^{s_1}_{\cdot}{}_{s_2}{}^{s_3}_{\cdot}{}^{s_4}_{\cdot} \\ t(\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{z}) &= u_{s_1}\,v^{s_2}\,w_{s_3}\,z_{s_4}\,t^{s_1}_{\cdot}{}_{s_2}{}^{s_3}_{\cdot}{}^{s_4}_{\cdot} \end{aligned}$$

$u_{s_1}, v^{s_2}, w_{s_3}, z_{s_4}$ coordinates of the vectors of the m-tuple

$t^{s_1}_{\cdot}{}_{s_2}{}^{s_3}_{\cdot}{}^{s_4}_{\cdot}$ coordinates of the tensor T

The coordinates of a tensor are said to be covariant if all indices of the coordinates are subscripts. In this case the tensor itself is often said to be covariant. The coordinates of a tensor are said to be contravariant if all indices of the coordinates are superscripts. Otherwise the coordinates of the tensor are said to be mixed.

$t_{s_1 \ldots s_m}$	covariant coordinates of the tensor T
$t^{s_1 \ldots s_m}$	contravariant coordinates of the tensor T
$t_{s_1 \,\cdot\, \ldots s_m}^{\, s_2}$	mixed coordinates of the tensor T

Designations : Let a tensor T in a euclidean space $\mathbb{R}^n$ be defined by the linear mapping $t(\mathbf{u},...,\mathbf{w})$ with $\mathbf{u},...,\mathbf{w} \in \mathbb{R}^n$. The following designations are used for this tensor :

T	target of the mapping t for all tuples $(\mathbf{u},...,\mathbf{w})$
$\mathbf{T}$	set of the coordinates of the tensor
$\mathbf{T}_*$	set of the covariant coordinates $t_{1 \ldots m}$ of the tensor
$\mathbf{T}^*$	set of the contravariant coordinates $t^{1 \ldots m}$ of the tensor

Representation of tensor coordinates : The coordinates of a tensor $t(\mathbf{u})$ of rank 1 are often represented as a vector $\mathbf{t}$ of dimension n. The representation $\mathbf{t}_*$ with the covariant coordinates and the representation $\mathbf{t}^*$ with the contravariant coordinates of the tensor must be distinguished.

$$\mathbf{t}_* = \begin{array}{|c|} \hline t_1 \\ \hline \vdots \\ \hline t_n \\ \hline \end{array} \qquad \mathbf{t}^* = \begin{array}{|c|} \hline t^1 \\ \hline \vdots \\ \hline t^n \\ \hline \end{array}$$

The coordinates of a tensor $T(\mathbf{u},\mathbf{w})$ of rank 2 are often represented as a quadratic matrix $\mathbf{T}$ of dimension n. The representation $\mathbf{T}_*$ with the covariant coordinates and the representation $\mathbf{T}^*$ with the contravariant coordinates of the tensor must be distinguished.

$$\mathbf{T}_* = \begin{array}{|c|c|c|c|} \hline t_{11} & t_{12} & \cdots & t_{1n} \\ \hline t_{21} & t_{22} & & t_{2n} \\ \hline \vdots & & \ddots & \vdots \\ \hline t_{n1} & t_{n2} & \cdots & t_{nn} \\ \hline \end{array} \qquad \mathbf{T}^* = \begin{array}{|c|c|c|c|} \hline t^{11} & t^{12} & \cdots & t^{1n} \\ \hline t^{21} & t^{22} & & t^{2n} \\ \hline \vdots & & \ddots & \vdots \\ \hline t^{n1} & t^{n2} & \cdots & t^{nn} \\ \hline \end{array}$$

Metric tensor : In a euclidean space $\mathbb{R}^n$ with the covariant basis $\mathbf{b}_1,\dots,\mathbf{b}_m$ and the contravariant basis $\mathbf{b}^1,\dots,\mathbf{b}^m$, the linear mapping $t(\mathbf{u},\mathbf{w}) = \mathbf{u} \cdot \mathbf{w}$ with $\mathbf{u},\mathbf{w} \in \mathbb{R}^n$ defines the metric tensor G. The coordinates of the metric tensor are the scalar products of the basis vectors. The metric tensor has the covariant coordinates g_{im} and the contravariant coordinates g^{im}.

$$g_{im} = \mathbf{b}_i \cdot \mathbf{b}_m \qquad i,m \in \{1,\dots,n\}$$

$$g^{im} = \mathbf{b}^i \cdot \mathbf{b}^m \qquad i,m \in \{1,\dots,n\}$$

The mixed coordinates of the metric tensor have either the value 0 or the value 1.

$$g_{i\,.}^{\;\;m} = \mathbf{b}_i \cdot \mathbf{b}^m = \delta_i^m \qquad i,m \in \{1,\dots,n\}$$

Symmetric tensors : A tensor T is said to be symmetric in the covariant indices i and m or symmetric in the contravariant indices r and s if interchanging these indices does not change the values of the coordinates of the tensor. In the following formulas, the symbol _ stands for subscripts or superscripts with fixed values.

$$t_{\_i\_\_m\_} = t_{\_m\_\_i\_}$$

$$t_{\_\,.\,\_\_\_\,.\,\_}^{\;\;r\;\;\;\;\;\;s} = t_{\_\,.\,\_\_\_\,.\,\_}^{\;\;s\;\;\;\;\;\;r}$$

Antisymmetric tensors : A tensor T is said to be antisymmetric in the covariant indices i and m or antisymmetric in the contravariant indices r and s if interchanging these indices changes the sign of the coordinates of the tensor but not their magnitude.

$$t_{\_i\_\_m\_} = -t_{\_m\_\_i\_}$$

$$t_{\_\,.\,\_\_\_\,.\,\_}^{\;\;r\;\;\;\;\;\;s} = -t_{\_\,.\,\_\_\_\,.\,\_}^{\;\;s\;\;\;\;\;\;r}$$

9.3.3 TRANSFORMATION OF TENSOR COORDINATES

Coordinates with dual indices : The coordinates of a tensor T may alternatively be specified as covariant, contravariant or mixed coordinates. If the i-th vector of the m-tuple in the mapping $t(\mathbf{u}_1,...,\mathbf{u}_m)$ is represented using covariant coordinates in the contravariant basis $\mathbf{b}^1,...,\mathbf{b}^n$, the tensor coordinates have the form $t_{--\cdot-}^{\quad i}$. If instead the i-th vector of the m-tuple is represented using contravariant coordinates in the covariant basis $\mathbf{b}_1,...,\mathbf{b}_n$, the tensor coordinates have the form t_{--i-}. The coordinates t_{--i-} and $t_{--\cdot-}^{\quad i}$ are called coordinates with dual indices. The subscript i is called a covariant index, the superscript i is called a contravariant index.

Rules for dual indices : Since the choice of basis does not change the scalar values of a tensor T, the coordinates of the tensor are transformed according to definite rules if the basis for the i-th vector of the m-tuple is replaced by the dual basis for the same vector. This transformation is referred to as lowering or raising of indices.

The dual basis vectors obey the transformation rules $\mathbf{b}_i = g_{ik}\,\mathbf{b}^k$ and $\mathbf{b}^i = g^{ik}\,\mathbf{b}_k$ with the coordinates g_{ik} and g^{ik} of the metric tensor G. The relationship between the coordinates of the tensor T with dual indices i is derived using the linearity of the mapping $t(\mathbf{u}_1,...,\mathbf{u}_m)$:

lowering : $t(...,\mathbf{b}_i,...) = t(...,g_{ik}\,\mathbf{b}^k,...) = g_{ik}\,t(...,\mathbf{b}^k,...)$

$$t_{--i-} = g_{ik}\,t_{--\cdot-}^{\quad k}$$

raising : $t(...,\mathbf{b}^i,...) = t(...,g^{ik}\,\mathbf{b}_k,...) = g^{ik}\,t(...,\mathbf{b}_k,...)$

$$t_{--\cdot-}^{\quad i} = g^{ik}\,t_{--k-}$$

general : $t_{-i\cdot-}^{\quad k} = g_{ir}\,g^{ks}\,t_{-\cdot s-}^{\;\, r}$

Transformation rules for tensor coordinates : Let a tensor T in the euclidean space $\mathbb{R}^n$ be defined by the linear mapping $t(\mathbf{u}_1,...,\mathbf{u}_m)$. The covariant basis vectors $\mathbf{b}_1,...,\mathbf{b}_n$ for the i-th vector of the m-tuple are transformed into the basis vectors $\overline{\mathbf{b}}_1,...,\overline{\mathbf{b}}_n$ with the matrix $\mathbf{A}$, that is $\overline{\mathbf{B}}_* = \mathbf{B}_*\mathbf{A}$. Let the coordinates of the tensor be t_{--i-} before the transformation and $\overline{t}_{--i-}$ afterwards. The transformation rule for the coordinates is derived from the transformation rule $\overline{\mathbf{b}}_i = \mathbf{b}_k\,a^k_{\cdot i}$ for the basis vectors using the linearity of the mapping $t(\mathbf{u}_1,...,\mathbf{u}_m)$.

covariant : $t(...,\overline{\mathbf{b}}_i,...) = t(...,a^k_{\cdot i}\,\mathbf{b}_k,...) = a^k_{\cdot i}\,t(...,\mathbf{b}_k,...)$

$$\overline{t}_{--i-} = a^k_{\cdot i}\,t_{--k-}$$

The transformation rule for a contravariant index of the tensor coordinates follows analogously from the transformation rule $\overline{\mathbf{b}}^{\,i} = \mathbf{b}^k\,\overline{a}^{\,i}_{\cdot k}$:

contravariant : $t(...,\overline{\mathbf{b}}^i,...) = t(...,\overline{a}^{\,i}_{\cdot k}\,\mathbf{b}^k,...) = \overline{a}^{\,i}_{\cdot k}\,t(...,\mathbf{b}^k,...)$

$$\overline{t}_{--\cdot-}^{\quad i} = \overline{a}^{\,i}_{\cdot k}\,t_{--\cdot-}^{\quad k}$$

Since the mapping $t(\mathbf{u}_1, ..., \mathbf{u}_m)$ is linear, the transformations for the various indices of the tensor coordinates may be combined. For example, if the i-th and the k-th vector of the m-tuple are described in dual bases which are transformed with the matrices $\mathbf{A}$ and $\bar{\mathbf{A}} = \mathbf{A}^{-1}$, the result is :

general : $\bar{t}_{-i-\,.\,-}^{k} = a^{r}_{.i}\,\bar{a}^{k}_{.s}\,t_{-r-\,.\,-}^{s}$

Tensor character of a quantity : Let the values of a quantity with m indices in a euclidean space $\mathbb{R}^n$ be known for all combinations of index values in the range 1,...,n. Let each index be associated with a basis of $\mathbb{R}^n$. These quantities with m indices are the coordinates of a tensor if under a change of basis the rules for raising and lowering indices and the transformation rules for tensor coordinates are satisfied for arbitrary indices.

Proof : Tensor character of an indexed quantity

In the preceding sections the rules for raising and lowering indices and the transformation rules are shown to be necessary properties of tensor coordinates. These rules are also sufficient conditions if they imply the existence of a linear mapping $t(\mathbf{u}_1, ..., \mathbf{u}_m)$ which maps every m-tuple of vectors to a basis-independent scalar and leads to the given coordinates.

Let a basis be stipulated for each of the m indices, so that the vectors of the m-tuple may be represented in the form $\mathbf{u} = u_i\,\mathbf{b}^i$ or $\mathbf{u} = u^k\,\mathbf{b}_k$. The following sum is defined for the indexed quantities and the coordinates of the vectors :

$$S \;:=\; ...\,u_i\,u^k ...\,t_{-\,.\,k\,-}^{i}$$

By hypothesis, the transformation rules hold for the coordinates of the indexed quantity and the vectors if the bases are transformed with the matrix $\mathbf{A}$:

$$\bar{u}_r = u_i\,a^{i}_{.r} \qquad \bar{u}^s = u^k\,\bar{a}^{s}_{.k}$$

$$\bar{t}_{-\,.\,s\,-}^{r} = ...\,\bar{a}^{r}_{.x}\,a^{y}_{.s}\,...\,t_{-\,.\,y\,-}^{x}$$

The value of the sum is now calculated with the transformed coordinates and is designated by $\bar{S}$. Substituting the transformation formulas leads to $\bar{S} = S$:

$$\begin{aligned} \bar{S} &= ...\,\bar{u}_r\,\bar{u}^s ...\,\bar{t}_{-\,.\,s\,-}^{r} \\ &= ...\,u_i\,u^k\,a^{i}_{.r}\,\bar{a}^{s}_{.k}\,\bar{a}^{r}_{.x}\,a^{y}_{.s}\,...\,t_{-\,.\,y\,-}^{x} \\ &= ...\,u_i\,u^k\,\delta^{i}_{x}\,\delta^{y}_{k}\,...\,t_{-\,.\,y\,-}^{x} \\ &= ...\,u_i\,u^k\,...\,t_{-\,.\,k\,-}^{i} \;=\; S \end{aligned}$$

Thus there is a linear mapping $t(\mathbf{u}_1, ..., \mathbf{u}_m)$ which leads to the given coordinates and assigns every m-tuple $(\mathbf{u}_1, ..., \mathbf{u}_m)$ of vectors a scalar sum S whose value is independent of the choice of the coordinate systems for the vectors $\mathbf{u}_i$. Hence the indexed quantities $t_{-\,.\,k\,-}^{i}$ are the coordinates of a tensor.

9.3.4 OPERATIONS ON TENSORS

Introduction : The question arises whether new tensors may be constructed by operations on given tensors. This is the case if the operation leads to quantities with tensor character (see Section 9.3.3). In the following the sum, the direct product, the contraction and the contracted product of tensors are shown to lead to new tensors. These operations are used in formulating physical problems.

Sums of tensors : Let the tensors X and Y in the euclidean space $\mathbb{R}^n$ be defined by the linear mappings $x(\mathbf{u}_1,...,\mathbf{u}_m)$ and $y(\mathbf{u}_1,...,\mathbf{u}_m)$. The sum $X+Y$ of these tensors is a tensor Z which is defined by a linear mapping $z(\mathbf{u}_1,...,\mathbf{u}_m)$. The value of the mapping $z(\mathbf{u}_1,...,\mathbf{u}_m)$ is the sum of the values of the mappings x(...) and y(...) for the same values of the m-tuple.

$$z(\mathbf{u}_1,...,\mathbf{u}_m) = x(\mathbf{u}_1,...,\mathbf{u}_m) + y(\mathbf{u}_1,...,\mathbf{u}_m)$$

The linearity of the mappings x(...) and y(...) implies the linearity of the mapping z(...). Each coordinate of Z is equal to the sum of the coordinates of X and Y with the same indices.

$$\begin{aligned}
x(...,\mathbf{u},\mathbf{w},...) &= x(...,u^i\mathbf{b}_i,\, w_k\mathbf{b}^k,...) = ...u^i\, w_k ... x_{\_i\,.\,\_}^{\;\;\;\;k} \\
y(...,\mathbf{u},\mathbf{w},...) &= y(...,u^i\mathbf{b}_i,\, w_k\mathbf{b}^k,...) = ...u^i\, w_k ... y_{\_i\,.\,\_}^{\;\;\;\;k} \\
z(...,\mathbf{u},\mathbf{w},...) &= z(...,u^i\mathbf{b}_i,\, w_k\mathbf{b}^k,...) = ...u^i\, w_k ... z_{\_i\,.\,\_}^{\;\;\;\;k} \\
z(...) = x(...) + y(...) &\quad\Rightarrow\quad z_{\_i\,.\,\_}^{\;\;\;\;k} = x_{\_i\,.\,\_}^{\;\;\;\;k} + y_{\_i\,.\,\_}^{\;\;\;\;k}
\end{aligned}$$

Products of tensors : Let the tensors X and Y in a euclidean space $\mathbb{R}^n$ be defined by the linear mappings $x(\mathbf{u}_1,...,\mathbf{u}_r)$ and $y(\mathbf{w}_1,...,\mathbf{w}_s)$. The product (tensor product, direct product) of these tensors is a tensor Z which is defined by a linear mapping $z(\mathbf{u}_1,...,\mathbf{u}_r,\mathbf{w}_1,...,\mathbf{w}_s)$. The value of the mapping z(...) is the product of the values of the mappings x(...) and y(...) for the same values of the vectors $\mathbf{u}_i$ or $\mathbf{w}_k$.

$$z(\mathbf{u}_1,...,\mathbf{u}_r,\mathbf{w}_1,...,\mathbf{w}_s) = x(\mathbf{u}_1,...,\mathbf{u}_r)\, y(\mathbf{w}_1,...,\mathbf{w}_s)$$

The linearity of the mapping z(...) follows from the linearity of the mappings x(...) and y(...). Since the mappings x, y and z are determined for the same values of the vectors $\mathbf{u}_i$ or $\mathbf{w}_k$, the coordinates of Z are equal to the product of the coordinates of X and Y with the same indices.

$$\begin{aligned}
x(...,\mathbf{u},...) &= x(...,u^i\mathbf{b}_i,...) &&= ...u^i ... x_{\_i\_} \\
y(...,\mathbf{w},...) &= y(..., w_k\mathbf{b}^k,...) &&= ...\, w_k ... y_{\_\,.\,\_}^{\;\;k} \\
z(...,\mathbf{u},...,\mathbf{w},...) &= z(...,u^i\mathbf{b}_i,...,\, w_k\mathbf{b}^k,...) &&= ...u^i ...\, w_k ... z_{\_i\,\_\,.\,\_}^{\;\;\;\;\;\;k} \\
z(...) = x(...)\,y(...) &\quad\Rightarrow\quad z_{\_i\_\,.\,\_}^{\;\;\;\;\;\;k} = x_{\_i\_}\; y_{\_\,.\,\_}^{\;\;k}
\end{aligned}$$

The product of two tensors is generally not commutative :

$$z_{ik} = x_i\, y_k \neq y_i\, x_k$$

Contracted product of tensors of rank 1 : A linear mapping is desired which assigns any two tensors X, Y of rank 1 a tensor Z of rank 0. Let the coordinates of X in a basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ be x_i, and let the coordinates of Y in the dual basis $\mathbf{B}^*$ be y^k. Then the most general form of a linear relationship between the coordinates of the tensors is $z = t^{i}_{.k}\, x_i\, y^k$ with arbitrary coefficients $t^{i}_{.k} \in \mathbb{R}$. This form is now required to be independent of the choice of the basis pair $\mathbf{B}^*, \mathbf{B}_*$. Let the coordinates of X, Y after a basis transformation using an arbitrary transformation matrix **A** be $\bar{x}_i$, $\bar{y}^k$, and let the coefficients of $\mathbf{A}$, $\mathbf{A}^{-1}$ be $a^{r}_{.i}$, $\bar{a}^{s}_{.k}$. The values of z before and after the basis transformation are equated :

$$\begin{aligned} t^{i}_{.k}\, \bar{x}_i\, \bar{y}^k &= t^{i}_{.k}\, x_r\, a^{r}_{.i}\, y^s\, \bar{a}^{k}_{.s} \\ &= t^{r}_{.s}\, x_i\, a^{i}_{.r}\, y^k\, \bar{a}^{s}_{.k} \\ &= (a^{i}_{.r}\, t^{r}_{.s}\, \bar{a}^{s}_{.k})\, x_i\, y^k \;=\; t^{i}_{.k}\, x_i\, y^k \end{aligned}$$

The following condition is obtained by comparing coefficients :

$$t^{i}_{.k} = a^{i}_{.r}\, t^{r}_{.s}\, \bar{a}^{s}_{.k}$$

The matrix form of this condition is $\mathbf{T} = \mathbf{A\,T\,A}^{-1}$. The condition is satisfied for arbitrary transformation matrices **A** if and only if **T** is a multiple of the unit matrix. Up to a constant factor, the value $t^{i}_{.k} = \delta^{i}_{k}$ is thus uniquely determined by the required basis independence. The rule $z = \delta^{i}_{k}\, x_i\, y^k = x_k\, y^k$, though expressed in terms of the basis-dependent coordinates of X and Y, defines a tensor Z which is independent of the choice of bases. This tensor is called the contracted product of X and Y.

$$\text{Z is the contracted product of X and Y} \quad :\Leftrightarrow \quad z = x_i\, y^i$$

Orthonormal tensors of rank 1 : A set M = {x, y, z, ...} of tensors of rank 1 is said to be orthonormal if the contracted product of each of these tensors with itself has the value 1 and the contracted product of an arbitrary pair of different tensors x,z has the value 0.

$$\begin{aligned} \mathbf{x} \cdot \mathbf{x} &= x_i\, x^i = 1 \\ \mathbf{x} \cdot \mathbf{z} &= x_i\, z^i = 0 \end{aligned}$$

Example 1 : Contracted product of tensors of rank 1

Let the tensors x(**u**) and y(**w**) be defined by the scalar products $\mathbf{u} \cdot \mathbf{f}$ and $\mathbf{w} \cdot \mathbf{g}$ with fixed vectors **f**, **g**. The contracted product of the tensor X with the tensor Y is the scalar product $\mathbf{f} \cdot \mathbf{g}$, and hence a scalar :

$$\begin{aligned} &x(\mathbf{u}) = \mathbf{u} \cdot \mathbf{f} && \text{with} && x_i = x(\mathbf{b}_i) = \mathbf{b}_i \cdot (f_k\, \mathbf{b}^k) = f_i \\ &y(\mathbf{w}) = \mathbf{w} \cdot \mathbf{g} && \text{with} && y^i = y(\mathbf{b}^i) = \mathbf{b}^i \cdot (g^k\, \mathbf{b}_k) = g^i \\ &z = x_i\, y^i = f_i\, g^i = \mathbf{f} \cdot \mathbf{g} \end{aligned}$$

The scalar $\mathbf{f} \cdot \mathbf{g}$ does not depend on the choice of bases for the vectors **u**, **w** and **f**, **g**. Hence the contracted product Z of X and Y is a tensor.

Contracted product of general tensors : A linear mapping is desired which maps a tensor X of rank p and a tensor Y of rank u to a tensor Z of rank (p + u – 2). Let the coordinates of X in a basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ be $x_{i_1 \dots i_p}$, and let the coordinates of Y in the dual basis $\mathbf{B}^*$ be $y^{k_1 \dots k_u}$. The product of the tensors is to be contracted in an arbitrary index of X and an arbitrary index of Y, for example in i_p and k_1. Then the most general form of a linear relationship between the coordinates of the tensors is

$$z_{i_1 \dots i_{p-1}}{}^{k_2 \dots k_u} = t^{i_p}_{\cdot k_1} \, x_{i_1 \dots i_p} \, y^{k_1 \dots k_u}$$

with arbitrary coefficients $t^{i_p}_{\cdot k_1} \in \mathbb{R}$. In analogy with the contracted product of tensors of rank 1, this form is now required to be independent of the choice of the basis pair $\mathbf{B}^*$, $\mathbf{B}_*$. If the basis transformation is restricted to the indices i_p and k_1, the following condition is obtained in analogy with the case of a contracted product of tensors of rank 1 :

$$t^{i_p}_{\cdot k_1} \, \bar{x}_{i_1 \dots i_p} \, \bar{y}^{k_1 \dots k_u} \quad = \quad (a^{i_p}_{\cdot r} \, t^{r}_{\cdot s} \, \bar{a}^{s}_{\cdot k_1}) \, x_{i_1 \dots i_p} \, y^{k_1 \dots k_u}$$

Comparison of coefficients again yields :

$$t^{i_p}_{\cdot k_1} = a^{i_p}_{\cdot r} \, t^{r}_{\cdot s} \, \bar{a}^{s}_{\cdot k_1}$$

$$\mathbf{T} = \mathbf{A\,T\,A}^{-1} \quad \Rightarrow \quad t^{r}_{\cdot s} = \delta^{r}_{s}$$

$$z_{i_1 \dots i_{p-1}}{}^{k_2 \dots k_u} = x_{i_1 \dots i_{p-1} s} \, y^{s \, k_2 \dots k_u}$$

This rule, though expressed in terms of the basis-dependent coordinates of X and Y, thus defines a tensor Z which is independent of the choice of bases. This tensor is called the contracted product of X and Y.

Example 2 : Contracted product of general tensors

The state of stress of a continuum at a point P in the space $\mathbb{R}^3$ is described by the stress tensor **S** with the coordinates s_{ik}. The orientation of a surface at the point P is described by the surface normal **n** with the coordinates n^k. The stress vector **t** on the surface is the contracted product of the tensors **S** and **n** :

$$t_i = s_{ik} \, n^k \qquad\qquad i, k = 1,\dots,3$$

The contracted product of these tensors may also be represented in matrix form :

$$\mathbf{t} = \mathbf{S}\,\mathbf{n} \qquad \begin{array}{|c|c|c|}\hline & & \\ \hline\end{array} \qquad \begin{array}{|c|}\hline n^1 \\ \hline n^2 \\ \hline n^3 \\ \hline\end{array}$$

$$\begin{array}{|c|c|c|}\hline s_{11} & s_{12} & s_{13} \\ \hline s_{21} & s_{22} & s_{23} \\ \hline s_{31} & s_{32} & s_{33} \\ \hline\end{array} \quad \begin{array}{|c|}\hline t_1 \\ \hline t_2 \\ \hline t_3 \\ \hline\end{array}$$

Contraction of a tensor : Let a tensor X be defined by the linear mapping $x(\mathbf{u}_1, ..., \mathbf{u}_m)$. The tensor may be contracted in two arbitrary indices, say i and k. The contraction Z is a linear mapping $z(\mathbf{u}_1, ..., \mathbf{u}_{i-1}, \mathbf{u}_{i+1}, ..., \mathbf{u}_{k-1}, \mathbf{u}_{k+1}, ..., \mathbf{u}_m)$.

It is convenient to contract the tensor X in a contravariant index and a covariant index. Let the coordinates of X be $x^{i_1 \cdots i_r}{}_{k_1 \ldots k_s}$. If for instance the tensor X is contracted in the indices i_r and k_1, the coordinates $z^{i_1 \cdots i_{r-1}}{}_{k_2 \ldots k_s}$ of the contraction Z are determined in analogy with the contracted product of tensors by summing the coordinates of X over the indices i_r and k_1 :

$$z^{i_1 \ldots i_{r-1}}{}_{k_2 \ldots k_s} := x^{i_1 \ldots i_{r-1} m}{}_{m\, k_2 \ldots k_s}$$

Outer product of tensors : In some of the literature the term "outer product" refers to the general product of tensors, while elsewhere it is used only for the cross product of tensors of rank 1. To prevent ambiguities, the term is not used here.

Inner product of tensors : A product contracted in the last index of the coordinates of a tensor U and the first index of the coordinates of a tensor W is called the inner product of the tensors U and W. The inner product of tensors is designated by the symbol • in analogy with the scalar product of vectors. For example :

$$\mathbf{A} = \mathbf{U} \cdot \mathbf{W} \quad \Leftrightarrow \quad a_{im} = u_{ik}\, w^{k}{}_{\cdot m}$$

Rules of calculation for tensor operations : Tensor operations are applied to the coordinates of the tensors. The rules of calculation for the tensor operations defined in the preceding paragraphs are compiled in the following (i, j, k = 1,...,n) :

sum	: $z_{-i}{}^{k}{}_{\cdot -}$	=	$x_{-i}{}^{k}{}_{\cdot -} + y_{-i}{}^{k}{}_{\cdot -}$
product	: $z_{-i-}{}^{k}{}_{\cdot -}$	=	$x_{-i-}\; y_{-}{}^{k}{}_{\cdot -}$
contracted product	: $z^{i}{}_{\cdot -}{}^{k}{}_{\cdot r-s}$	=	$x^{i}{}_{\cdot -}{}^{k}{}_{\cdot}{}^{j}{}_{\cdot}\; y_{jr-s}$
contraction	: $z_{-}{}^{i}{}_{\cdot --k}$	=	$x_{-}{}^{i}{}_{\cdot -}{}^{j}{}_{\cdot -j-k}$

9.3.5 ANTISYMMETRIC TENSORS

Introduction : Antisymmetric tensors have special properties which are useful in formulating physical problems. In particular, the completely antisymmetric permutation tensor is used to represent physical quantities such as areas, volumes and moments. The coordinates of the permutation tensor are derived in the following. The tensor is used to define the parallelepipedal product, the vector product and the cross product of tensors.

Completely antisymmetric tensor : A tensor T is said to be completely antisymmetric if it is antisymmetric in every pair of covariant indices i, k and in every pair of contravariant indices r, s :

$$t_{\_i\_k\_} = -t_{\_k\_i\_} \qquad \wedge \qquad t_{\_}{}^{r}{}_{\_}{}^{s}{}_{\_} = -t_{\_}{}^{s}{}_{\_}{}^{r}{}_{\_}$$

Permutation tensor : The rank of a completely antisymmetric tensor in the euclidean space $\mathbb{R}^n$ may be less than the dimension n of the space. The completely antisymmetric tensor of rank n in the space $\mathbb{R}^n$ is called a permutation tensor. The coordinates of the covariant permutation tensor for a covariant basis $\mathbf{b}_1, ..., \mathbf{b}_n$ are uniquely determined. They are designated by $\varepsilon_{i_1...i_n}$. The coordinates of the contravariant permutation tensor for a contravariant basis $\mathbf{b}^1, ..., \mathbf{b}^n$ are also uniquely determined and are designated by $\varepsilon^{i_1...i_n}$.

$$\varepsilon(\mathbf{b}_{i_1}, ..., \mathbf{b}_{i_n}) = \varepsilon_{i_1...i_n}$$
$$\varepsilon(\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_n}) = \varepsilon^{i_1...i_n}$$

Coordinates of the permutation tensor : The coordinates $\varepsilon_{i_1...i_n}$ of the permutation tensor of rank n are to be determined. The definition of the antisymmetry of the tensor implies that a coordinate of the tensor has the value 0 if at least two indices of the coordinate are equal. This is proved by interchanging two identical indices :

$$\varepsilon_{i_1..k..k..i_n} = -\varepsilon_{i_1..k..k..i_n} = 0$$

The choice of $\varepsilon := \varepsilon_{1...n}$ determines all other coordinates of the permutation tensor, since the definition of antisymmetry implies that two coordinates of the tensor have the same value if their indices are mapped to each other by an even permutation (see Section 7.7.5) and differ only in sign if their indices are mapped to each other by an odd permutation. In the following, <...> designates any permutation of the numbers 1 to n.

$$\text{sgn } \langle i_1, ..., i_n \rangle = 1 : \quad \varepsilon_{i_1...i_n} = \varepsilon$$
$$\text{sgn } \langle i_1, ..., i_n \rangle = -1 : \quad \varepsilon_{i_1...i_n} = -\varepsilon$$

Permutation tensor in the canonical basis : For the canonical basis **E**, the free parameter ε of the permutation tensor is set to 1. The coordinates of the permutation tensor in the canonical basis are designated by $e_{i_1 \ldots i_n}$.

$$\begin{aligned} &\varepsilon(\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_n}) = e_{i_1 \ldots i_n} = e^{i_1 \ldots i_n} = \varepsilon(\mathbf{e}^{i_1}, \ldots, \mathbf{e}^{i_n}) \\ &\text{sgn } <i_1, \ldots, i_n> = 1 \ : \quad e_{i_1 \ldots i_n} = 1 \\ &\text{sgn } <i_1, \ldots, i_n> = -1 \ : \quad e_{i_1 \ldots i_n} = -1 \end{aligned}$$

The value of the free parameter of the permutation tensor in an arbitrary basis $\mathbf{B}_*$ follows directly from the choice $\varepsilon = 1$ for the canonical basis. The transformation rules for tensor coordinates and the linearity of the mapping $\varepsilon(\ldots)$ imply :

$$\begin{aligned} \varepsilon(\mathbf{b}_{i_1}, \ldots, \mathbf{b}_{i_n}) &= \varepsilon(b_{k_1 i_1} \, \mathbf{e}^{k_1}, \ldots, b_{k_n i_n} \, \mathbf{e}^{k_n}) \\ &= b_{k_1 i_1} \ldots b_{k_n i_n} \varepsilon(\mathbf{e}^{k_1}, \ldots, \mathbf{e}^{k_n}) \end{aligned}$$

If $<i_1, \ldots, i_n>$ is an even permutation, then the left-hand side of this equation is equal to ε and by the definition of determinants the right-hand side is equal to det $\mathbf{B}_*$. If the permutation is odd, both sides of the equation acquire a minus sign. Thus for every covariant basis :

$$\begin{aligned} &\varepsilon = \det \mathbf{B}_* \\ &\varepsilon_{i_1 \ldots i_n} = \varepsilon \, e_{i_1 \ldots i_n} \end{aligned}$$

Since the covariant and the contravariant canonical basis coincide, analogous relationships hold for every contravariant basis $\mathbf{B}^*$:

$$\begin{aligned} &\varepsilon = \det \mathbf{B}^* \\ &\varepsilon^{i_1 \ldots i_n} = \varepsilon \, e^{i_1 \ldots i_n} \end{aligned}$$

Thus the permutation tensor of the space $\mathbb{R}^n$ has the following coordinates in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$:

$$\begin{array}{lcll} \text{equal indices } i_r = i_s & : & \varepsilon_{i_1 \ldots i_n} = 0 & \varepsilon^{i_1 \ldots i_n} = 0 \\ \text{sgn } <i_1, \ldots, i_n> = 1 & : & \varepsilon_{i_1 \ldots i_n} = \det \mathbf{B}_* & \varepsilon^{i_1 \ldots i_n} = \det \mathbf{B}^* \\ \text{sgn } <i_1, \ldots, i_n> = -1 & : & \varepsilon_{i_1 \ldots i_n} = -\det \mathbf{B}_* & \varepsilon^{i_1 \ldots i_n} = -\det \mathbf{B}^* \end{array}$$

Representation of the determinant : The permutation tensor in the canonical basis may be used to express the determinant of a matrix **A** in component notation as follows :

$$\det \mathbf{A} = a^{1k_1} \ldots a^{nk_n} e_{k_1 \ldots k_n} = a_{1k_1} \ldots a_{nk_n} e^{k_1 \ldots k_n}$$

a^{im}, a_{im} coefficients of the matrix **A** of dimension n

$e_{k_1 \ldots k_n} = e^{k_1 \ldots k_n}$ coordinates of the permutation tensor in the canonical basis

Example 1 : Calculation of a determinant

The determinant of a matrix **A** of dimension 3 is calculated using the permutation tensor of the space $\mathbb{R}^3$.

$$\mathbf{A} = \begin{array}{|c|c|c|} \hline 0.5 & 0.9 & 0.8 \\ \hline 0.4 & 0.2 & 0.6 \\ \hline 0.3 & 0.1 & 0.7 \\ \hline \end{array}$$

$$\begin{aligned} \det \mathbf{A} &= e^{ikm}\, a_{1i}\, a_{2k}\, a_{3m} \\ &= e^{123}\, a_{11}\, a_{22}\, a_{33} + e^{312}\, a_{13}\, a_{21}\, a_{32} + e^{231}\, a_{12}\, a_{23}\, a_{31} + \\ &\quad\; e^{213}\, a_{12}\, a_{21}\, a_{33} + e^{321}\, a_{13}\, a_{22}\, a_{31} + e^{132}\, a_{11}\, a_{23}\, a_{32} \\ &= 0.5 * 0.2 * 0.7 + 0.8 * 0.4 * 0.1 + 0.9 * 0.6 * 0.3 - \\ &\quad\; 0.9 * 0.4 * 0.7 - 0.8 * 0.2 * 0.3 - 0.5 * 0.6 * 0.1 \\ &= -0.066 \end{aligned}$$

Unit tensor : The tensor product of the covariant permutation tensor $\varepsilon_{i_1 \dots i_n}$ of the space $\mathbb{R}^n$ with the contravariant permutation tensor $\varepsilon^{m_1 \dots m_n}$ of the same space is called the unit tensor (δ-tensor) of the space $\mathbb{R}^n$. Its coordinates are designated by $d^{m_1 \dots m_n}_{i_1 \dots i_n}$. The unit tensor is independent of the choice of basis. Its coordinates are determined from the coordinates of the permutation tensor in the canonical basis. They can also be expressed as the determinant of a matrix of Kronecker symbols δ^m_i.

$$d^{m_1 \dots m_n}_{i_1 \dots i_n} = \varepsilon_{i_1 \dots i_n}\, \varepsilon^{m_1 \dots m_n} = e_{i_1 \dots i_n}\, e^{m_1 \dots m_n} \qquad i_s, m_s, s = 1,\dots,n$$

$$d^{m_1 \dots m_n}_{i_1 \dots i_n} = \det \begin{array}{|c|c|c|} \hline \delta^{m_1}_{i_1} & & \delta^{m_n}_{i_1} \\ \hline & \ddots & \\ \hline \delta^{m_1}_{i_n} & & \delta^{m_n}_{i_n} \\ \hline \end{array} = \delta^{m_{k_1}}_{i_1} \dots \delta^{m_{k_n}}_{i_n}\, e_{k_1 \dots k_n}$$

Thus the unit tensor of the space $\mathbb{R}^n$ has the following coordinates $d^{m_1 \dots m_n}_{i_1 \dots i_n}$:

two equal covariant indices	:	0
two equal contravariant indices	:	0
$\operatorname{sgn} <i_1, \dots, i_n> = \operatorname{sgn} <m_1, \dots, m_n>$	:	1
$\operatorname{sgn} <i_1, \dots, i_n> = -\operatorname{sgn} <m_1, \dots, m_n>$	:	–1

Proof : Coordinates of the unit tensor

(1) Since the determinants of dual bases are reciprocal, the unit tensor is independent of the choice of basis :

$$d^{m_1\ldots m_n}_{i_1\ldots i_n} \;=\; \varepsilon_{i_1\ldots i_n}\,\varepsilon^{m_1\ldots m_n} \;=\; \det \mathbf{B}_*\det \mathbf{B}^*\, e_{i_1\ldots i_n}\, e^{m_1\ldots m_n} \;=\; e_{i_1\ldots i_n}\, e^{m_1\ldots m_n}$$

(2) In the following, equation (a) is shown to hold for the product of the coordinates of the permutation tensor in the canonical basis :

$$e_{i_1\ldots i_n}\, e^{m_1\ldots m_n} \;=\; \delta^{m_{k_1}}_{i_1}\ldots\,\delta^{m_{k_n}}_{i_n}\; e_{k_1\ldots k_n} \qquad \text{(a)}$$

If at least two of the indices $i_1,\ldots,i_n$ have the same value, for example $i_1 = i_2$, then $e_{i_1\ldots i_n} = 0$ on the left-hand side and $\delta^{k_1}_{i_1}\,\delta^{k_2}_{i_2} \neq 0$ on the right-hand side only for $k_1 = k_2$, and hence $e_{k_1\ldots k_n} = 0$. Equation (a) is therefore satisfied.

If all the indices $i_1,\ldots,i_n$ are different and at least two of the indices $m_1,\ldots,m_n$ are equal, for example $m_1 = m_2$, then $e^{m_1\ldots m_n} = 0$ on the left-hand side. Also, $m_1 = m_2$ and $i_r \neq i_s$ implies that either $\delta^{m_1}_{i_r} = 1$ and $\delta^{m_2}_{i_s} = 0$ or $\delta^{m_2}_{i_s} = 1$ and $\delta^{m_1}_{i_r} = 0$, and hence $\delta^{m_1}_{i_r}\,\delta^{m_2}_{i_s} = 0$ on the right-hand side. Equation (a) is therefore satisfied.

If all indices $i_1,\ldots,i_n$ are different and all indices $m_1,\ldots,m_n$ are different, the left-hand side of equation (a) is non-zero :

$$e_{i_1\ldots i_n}\, e^{m_1\ldots m_n} \;=\; \operatorname{sgn} \langle i_1,\ldots,i_n\rangle \cdot \operatorname{sgn} \langle m_1,\ldots,m_n\rangle$$

A term on the right-hand side of (a) is non-zero if all Kronecker symbols have the value 1, that is if $m_{k_1} = i_1,\ldots,m_{k_n} = i_n$. This implies :

$$\begin{aligned}
\operatorname{sgn} \langle i_1,\ldots,i_n\rangle &= \operatorname{sgn} \langle m_{k_1},\ldots,m_{k_n}\rangle \\
&= \operatorname{sgn} \langle m_1,\ldots,m_n\rangle \cdot \operatorname{sgn} \langle k_1,\ldots,k_n\rangle \\
\operatorname{sgn} \langle i_1,\ldots,i_n\rangle &= \operatorname{sgn} \langle m_1,\ldots,m_n\rangle\; e_{k_1\ldots k_n}
\end{aligned}$$

If both sides of this equation are multiplied by $\operatorname{sgn} \langle m_1,\ldots,m_n\rangle$, then using $\operatorname{sgn} \langle m_1,\ldots,m_n\rangle \cdot \operatorname{sgn} \langle m_1,\ldots,m_n\rangle = 1$ the right-hand side of (a) becomes :

$$e_{k_1\ldots k_n} \;=\; \operatorname{sgn} \langle i_1,\ldots,i_n\rangle \cdot \operatorname{sgn} \langle m_1,\ldots,m_n\rangle$$

Equation (a) is therefore satisfied; it leads to a representation of the coordinates of the unit tensor as determinants :

$$d^{m_1\ldots m_n}_{i_1\ldots i_n} \;=\; \delta^{m_{k_1}}_{i_1}\ldots\,\delta^{m_{k_n}}_{i_n}\; e_{k_1\ldots k_n} \;=\; \det \begin{array}{|c|c|c|}\hline \delta^{m_1}_{i_1} & & \delta^{m_n}_{i_1} \\ \hline & \ddots & \\ \hline \delta^{m_1}_{i_n} & & \delta^{m_n}_{i_n} \\ \hline \end{array}$$

Example 2 : Unit tensor of the space $\mathbb{R}^3$

The coordinates $d^{m_1\,m_2\,m_3}_{i_1\,i_2\,i_3}$ of the unit tensor of the space $\mathbb{R}^3$ with i_s, $m_s = 1, 2, 3$ are alternatively determined as the product $\varepsilon_{i_1 i_2 i_3}\,\varepsilon^{m_1 m_2 m_3}$ of permutation tensors of $\mathbb{R}^3$ or as the determinant of a matrix of Kronecker symbols $\delta^{m_s}_{i_s}$. The calculation of the 36 non-zero coordinates using the determinant representation is shown in the following.

$$d^{123}_{123} = d^{312}_{312} = d^{231}_{231} = d^{213}_{213} = d^{321}_{321} = d^{132}_{132} = \det\begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1$$

$$d^{123}_{312} = d^{312}_{231} = d^{231}_{123} = d^{213}_{321} = d^{321}_{132} = d^{132}_{213} = \det\begin{vmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{vmatrix} = 1$$

$$d^{123}_{231} = d^{312}_{123} = d^{231}_{312} = d^{213}_{132} = d^{321}_{213} = d^{132}_{321} = \det\begin{vmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{vmatrix} = 1$$

$$d^{123}_{213} = d^{312}_{132} = d^{231}_{321} = d^{213}_{123} = d^{321}_{231} = d^{132}_{312} = \det\begin{vmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{vmatrix} = -1$$

$$d^{123}_{321} = d^{312}_{213} = d^{231}_{132} = d^{213}_{312} = d^{321}_{123} = d^{132}_{231} = \det\begin{vmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{vmatrix} = -1$$

$$d^{123}_{132} = d^{312}_{321} = d^{231}_{213} = d^{213}_{231} = d^{321}_{312} = d^{132}_{123} = \det\begin{vmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{vmatrix} = -1$$

Contraction of the unit tensor : The unit tensor of the space $\mathbb{R}^n$ is a tensor of rank 2n with n covariant indices and n contravariant indices in the range $\{1, 2, ..., n\}$. The unit tensor of the space $\mathbb{R}^n$ is contracted by summing over a covariant and a contravariant index according to the general rule. The result of the contraction is a tensor of rank $2(n-1)$. The contraction may be continued for further pairs of indices until the contracted tensor is a scalar. The coordinates of the unit tensor of the space $\mathbb{R}^n$ contracted in $r = n - s$ pairs of indices are determined using the following formulas for $i_j, m_j, t_j \in \{1, ..., n\}$ and $k_j \in \{1, ..., n - r\}$:

$$d^{m_1 \dots m_s\; t_1 \dots t_r}_{i_1 \dots i_s\; t_1 \dots t_r} = r!\;\delta^{m_{k_1}}_{i_1} \dots \delta^{m_{k_s}}_{i_s}\; e_{k_1 \dots k_s}$$

Thus contraction in all n pairs of indices yields $d^{t_1 \dots t_n}_{t_1 \dots t_n} = n!$.

Proof : Coordinates of the contracted unit tensor of $\mathbb{R}^n$

The coordinates of the unit tensor of the space $\mathbb{R}^n$ are represented in the form (a) derived above. A coordinate is non-zero only if the indices $i_1, \dots, i_n$ are pairwise different and the indices $m_1, \dots, m_n$ are pairwise different.

$$d^{m_1 \dots m_n}_{i_1 \dots i_n} = \delta^{m_{k_1}}_{i_1} \dots \delta^{m_{k_n}}_{i_n} \; e_{k_1 \dots k_n} \qquad \text{(a)}$$

$$i_j, k_j, m_j \in \{1, \dots, n\}$$

The tensor is contracted by summing over $r = n - s$ pairs of indices (i_{s+1}, m_{s+1}), ..., (i_n, m_n). The coordinates of the contracted tensor are given by :

$$d^{m_1 \dots m_s \, m_{s+1} \dots m_n}_{i_1 \dots i_s \, m_{s+1} \dots m_n} = \delta^{m_{k_1}}_{i_1} \dots \delta^{m_{k_s}}_{i_s} \, \delta^{m_{k_{s+1}}}_{m_{s+1}} \dots \delta^{m_{k_n}}_{m_n} e_{k_1 \dots k_s \, k_{s+1} \dots k_n} \qquad \text{(b)}$$

Only the non-zero coordinates of the δ-tensor are considered in the following, since only they affect the value of the sum on the right-hand side of the equation. For these coordinates the indices $m_1, \dots, m_n$ are pairwise different. The indices $k_{s+1}, \dots, k_n$ of the e-tensor have the fixed values $s+1, \dots, n$, since the value of the Kronecker symbols $\delta^{m_{k_{s+j}}}_{m_{s+j}}$ is 1 only for $k_{s+j} = s + j$. Thus the coordinates of the e-tensor are non-zero only for indices $k_1, \dots, k_s$ which lie in the range $\{1, \dots, s\}$ and are pairwise different. The coordinates $e_{k_1 \dots k_n}$ and $e_{k_1 \dots k_s}$ of the e-tensors are transformed into $e_{1 \dots n}$ and $e_{1 \dots s}$, respectively, by the same permutation of the indices $k_1, \dots, k_s$, since the indices $k_{s+1}, \dots, k_n$ already have the values $s+1, \dots, n$. Hence $e_{k_1 \dots k_n}$ may be replaced by $e_{k_1 \dots k_s}$ in (b) if the indices $k_1, \dots, k_s$ are restricted to the range $\{1, \dots, s\}$.

For fixed values of the indices $m_1, \dots, m_s$ each of the indices $m_{s+1}, \dots, m_n$ can take one of $r = n - s$ different values. Hence the group of indices $m_{s+1}, \dots, m_n$ has a total of $r!$ different valuations, so that the sum on the right-hand side of (b) consists of $r!$ equal terms :

$$d^{m_1 \dots m_s \, m_{s+1} \dots m_n}_{i_1 \dots i_s \, m_{s+1} \dots m_n} = r! \; \delta^{m_{k_1}}_{i_1} \dots \delta^{m_{k_s}}_{i_s} \, e_{k_1 \dots k_s}$$

$$k_j \in \{1, \dots, s\}$$

The indices $m_{s+1}, \dots, m_n$ are renamed to $t_1, \dots, t_r$. This yields the above formulas for the coordinates of the contracted unit tensor.

Example 3 : Contractions of the unit tensor of $\mathbb{R}^3$

The coordinates of the unit tensor of the space $\mathbb{R}^3$ were already determined in Example 2. The coordinates of the contractions of this tensor may be determined directly by summing the non-zero coordinates of the unit tensor :

$$
\begin{array}{lllllllllllllll}
d^{12m}_{12m} & = & d^{23m}_{23m} & = & d^{31m}_{31m} & = & d^{21m}_{21m} & = & d^{32m}_{32m} & = & d^{13m}_{13m} & = & 1 \\
d^{12m}_{21m} & = & d^{23m}_{32m} & = & d^{31m}_{13m} & = & d^{21m}_{12m} & = & d^{32m}_{23m} & = & d^{13m}_{31m} & = & -1 \\
d^{1km}_{1km} & = & d^{12m}_{12m} & + & d^{13m}_{13m} & = & 1+1 & = & 2 \\
d^{2km}_{2km} & = & d^{21m}_{21m} & + & d^{23m}_{23m} & = & 1+1 & = & 2 \\
d^{3km}_{3km} & = & d^{31m}_{31m} & + & d^{32m}_{32m} & = & 1+1 & = & 2 \\
d^{ikm}_{ikm} & = & d^{1km}_{1km} & + & d^{2km}_{2km} & + & d^{3km}_{3km} & = & 2+2+2 & = & 6
\end{array}
$$

The same results may be obtained using the formulas for the coordinates of the contracted unit tensors :

$$
d^{m_1\, m_2\, t_1}_{i_1\, i_2\, t_1} = 1!\ \delta^{m_{k_1}}_{i_1}\ \delta^{m_{k_2}}_{i_2}\ e_{k_1 k_2} = \det \begin{array}{|c|c|}\hline \delta^{m_1}_{i_1} & \delta^{m_2}_{i_1} \\ \hline \delta^{m_1}_{i_2} & \delta^{m_2}_{i_2} \\ \hline \end{array}
$$

$$
d^{m_1\, t_1\, t_2}_{i_1\, t_1\, t_2} = 2!\ \delta^{m_{k_1}}_{i_1}\ e_{k_1} = 2 \det \boxed{\delta^{m_1}_{i_1}}
$$

$$
d^{t_1\, t_2\, t_3}_{t_1\, t_2\, t_3} = 3! = 6
$$

Isotropic tensors : Let all indices of a tensor T be referred to the same basis. Then the tensor is said to be isotropic (spherical) if a rotation of the basis does not change the values of the coordinates of the tensor. The coordinates of an isotropic tensor are thus the same in the original basis and in the rotated basis. If the coordinates of the tensor T in an arbitrary covariant basis $\mathbf{B}_*$ of the space $\mathbb{R}^n$ are $t_{i_1 \ldots i_m}$, then its coordinates in the rotated basis $\bar{\mathbf{B}}_* = \mathbf{R}_o\, \mathbf{B}_*$ are $\bar{t}_{i_1 \ldots i_m} = t_{i_1 \ldots i_m}$.

$$
\begin{array}{lllll}
\bar{\mathbf{B}}_* = \mathbf{B}_*\, \mathbf{A} & = \mathbf{R}_o\, \mathbf{B}_* & \text{with} & (\mathbf{R}^o)^{\mathrm{T}}\, \mathbf{R}_o = \mathbf{I} \\
\bar{\mathbf{B}}^* = \mathbf{B}^*\, \bar{\mathbf{A}}^{\mathrm{T}} & = \mathbf{R}^o\, \mathbf{B}^* & & \bar{\mathbf{A}} = \mathbf{A}^{-1}
\end{array}
$$

$$
\bar{t}_{i_1 \ldots i_m} = t_{s_1 \ldots s_m}\ a^{s_1}_{\cdot\ i_1} \ldots a^{s_m}_{\cdot\ i_m} = t_{i_1 \ldots i_m}
$$

$\mathbf{R}_o$, $\mathbf{R}^o$ rotation matrices

$\mathbf{A}$ transformation matrix with the coefficients $a^{i}_{\cdot\, m}$

Isotropic tensors are constructed from the permutation tensor and the metric tensor as follows :

1) The coordinates of the permutation tensor in the canonical basis $\mathbf{E}$ of the space $\mathbb{R}^n$ are $e_{i_1 \dots i_n} = e^{i_1 \cdots i_n}$. The coordinates of the permutation tensor in an arbitrary basis $\mathbf{B}_*$ are $\varepsilon_{i_1 \dots i_n} = e_{i_1 \dots i_n} \det \mathbf{B}_*$. A rotation of the basis $\mathbf{B}_*$ with the rotation matrix $\mathbf{R}_0$ leaves the coordinates of the permutation tensor unchanged :

$$\det \overline{\mathbf{B}}_* = \det (\mathbf{R}_0 \mathbf{B}_*) = \det \mathbf{R}_0 \det \mathbf{B}_* = \det \mathbf{B}_* \quad \text{since} \quad \mathbf{R}_0^T \mathbf{R}_0 = \mathbf{I}$$

$$\bar{\varepsilon}_{i_1 \dots i_n} = e_{i_1 \dots i_n} \det \overline{\mathbf{B}}_* = e_{i_1 \dots i_n} \det \mathbf{B}_* = \varepsilon_{i_1 \dots i_n}$$

2) Since the covariant coordinates $\varepsilon_{i_1 \dots i_n}$ and the contravariant coordinates $\varepsilon^{i_1 \cdots i_n}$ of the permutation tensor of the space $\mathbb{R}^n$ in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ are invariant under a rotation of the bases, the unit tensor of the space $\mathbb{R}^n$ is also invariant under a rotation of the dual bases.

$$\bar{d}^{\,i_1 \cdots i_n}_{m_1 \dots m_n} = \bar{\varepsilon}^{i_1 \dots i_n} \bar{\varepsilon}_{m_1 \dots m_n} = \varepsilon^{i_1 \dots i_n} \varepsilon_{m_1 \dots m_n} = d^{\,i_1 \cdots i_n}_{m_1 \dots m_n}$$

3) The metric $\mathbf{G}_*$ of a basis $\mathbf{B}_*$ is invariant under rotations of this basis. This is proved by substituting the transformation matrix $\mathbf{A} = (\mathbf{B}^*)^T \mathbf{R}_0 \mathbf{B}_*$ into the general transformation rule $\overline{\mathbf{G}}_* = \mathbf{A}^T \mathbf{G}_* \mathbf{A}$:

$$\overline{\mathbf{G}}_* = \mathbf{B}_*^T \mathbf{R}_0^T \mathbf{B}^* \mathbf{G}_* (\mathbf{B}^*)^T \mathbf{R}_0 \mathbf{B}_* = \mathbf{B}_*^T \mathbf{R}_0^T \mathbf{R}_0 \mathbf{B}_* = \mathbf{B}_*^T \mathbf{B}_*$$

$$\overline{\mathbf{G}}_* = \mathbf{G}_*$$

4) The metric tensor scaled by a constant c is an isotropic tensor. Since the product and the sum of tensors are also tensors, scaled products of the metric tensor and sums of such terms are also isotropic tensors.

$$\overline{c\, g_{im}} = c g_{im}$$

$$\overline{c\, g_{im} g_{rs}} = c g_{im}\, g_{rs}$$

5) For isotropic tensors of rank 4, three products of the metric tensor which differ in their indices may be formed. Hence a general isotropic tensor of rank 4 has the following form :

$$t_{imrs} = c_1\, g_{im}\, g_{rs} + c_2\, g_{ir}\, g_{ms} + c_3\, g_{is}\, g_{mr}$$

For the completely symmetric, isotropic tensor of rank 4, the free parameters c_1, c_2 and c_3 are all equal :

$$t_{imrs} = c\,(g_{im}\, g_{rs} + g_{ir}\, g_{ms} + g_{is}\, g_{mr})$$

For the isotropic tensor of rank 4 symmetric in the indices i,m and in the indices r,s, the free parameters c_2 and c_3 are equal :

$$t_{imrs} = c_1\, g_{im}\, g_{rs} + c_2\,(g_{ir}\, g_{ms} + g_{is}\, g_{mr})$$

For the isotropic tensor of rank 4 antisymmetric in the indices i,m and in the indices r,s, c_1 is zero, while the free parameters c_2 and c_3 differ only in sign :

$$t_{imrs} = c\,(g_{ir}\,g_{ms} - g_{is}\,g_{mr})$$

6) Isotropic tensors of higher rank are constructed analogously.

Axial tensors : A tensor in the space $\mathbb{R}^n$ is said to be axial if the sign of its coordinates depends on the orientation of the space (see Section 9.2.7). The permutation tensor is axial : Changing the orientation of the space leads to a sign change in the determinant det **B** of the basis, and hence changes the sign of the coordinates.

Polar tensors : A tensor in the space $\mathbb{R}^n$ is said to be polar if the sign of its coordinates is independent of the orientation of the space (see Section 9.2.7). The unit tensor is polar : Changing the orientation of the space changes the sign of the coordinates of the covariant and the contravariant permutation tensor, but not the sign of their product.

Contracted product of the permutation tensor and the unit tensor : In order to express the permutation tensor in terms of the unit tensor, a Kronecker symbol is introduced for each index of the permutation tensor :

$$\begin{aligned} \varepsilon_{m_1 \ldots m_n} &= \varepsilon_{k_1 \ldots k_n}\, \delta^{k_1}_{m_1} \cdots \delta^{k_n}_{m_n} \\ &= \det \mathbf{B}_*\, e_{k_1 \ldots k_n}\, \delta^{k_1}_{m_1} \cdots \delta^{k_n}_{m_n} \\ &= \det \mathbf{B}_* \det \begin{vmatrix} \delta^{1}_{m_1} & & \delta^{n}_{m_1} \\ & \ddots & \\ \delta^{1}_{m_n} & & \delta^{n}_{m_n} \end{vmatrix} \end{aligned}$$

$$\varepsilon_{m_1 \ldots m_n} = \det \mathbf{B}_*\, d^{1 \ldots n}_{m_1 \ldots m_n}$$

Contracted product of permutation tensors and metric tensors : The (n–s)-fold contraction of a product of covariant permutation tensors with contravariant metric tensors is a tensor with the following coordinates :

$$\begin{aligned}
&\varepsilon_{i_1\ldots i_n}\ \varepsilon_{k_1\ldots k_n}\ g^{i_{s+1}k_{s+1}}\ \ldots\ g^{i_n k_n} &&= \\
&\varepsilon_{i_1\ldots i_n}\ \varepsilon^{m_1\ldots m_n}\ g_{k_1 m_1}\ \cdots\ g_{k_n m_n}\ g^{i_{s+1}k_{s+1}}\ \ldots\ g^{i_n k_n} &&= \\
&d^{m_1\ldots m_n}_{i_1\ldots i_n}\ g_{k_1 m_1}\ \cdots\ g_{k_s m_s}\ \delta^{i_{s+1}}_{m_{s+1}}\ldots\delta^{i_n}_{m_n} &&= \\
&d^{m_1\ldots m_s\, m_{s+1}\ldots m_n}_{i_1\ldots i_s\, m_{s+1}\ldots m_n}\ g_{k_1 m_1}\ \cdots\ g_{k_s m_s} &&= \\
&(n-s)!\ g_{k_1 m_1}\ \cdots\ g_{k_s m_s}\ d^{m_1\ldots m_s}_{i_1\ldots i_s} && i_r, k_r, m_r = 1,\ldots,n
\end{aligned}$$

The completely contracted product of the permutation tensors and the metric tensors coincides with the completely contracted product of the dual permutation tensors :

$$\varepsilon_{i_1\ldots i_n}\ \varepsilon_{k_1\ldots k_n}\ g^{i_1 k_1}\ \ldots\ g^{i_n k_n} \ =\ \varepsilon_{i_1\ldots i_n}\ \varepsilon^{i_1\ldots i_n}\ =\ n\,!$$

Parallelepipedal product : Let n tensors **a**,...,**z** of rank 1 in the euclidean space $\mathbb{R}^n$ be given. Let their coordinates be $a_i, \ldots, z_m$ with $i, m \in \{1,\ldots,n\}$ in the covariant basis $\mathbf{B}_*$ and $a^i, \ldots, z^m$ in the contravariant basis $\mathbf{B}^*$, so that their coordinates in the canonical basis are $\mathbf{a} = \mathbf{B}^*\,\mathbf{a}_* = \mathbf{B}_*\,\mathbf{a}^*$ to $\mathbf{z} = \mathbf{B}^*\,\mathbf{z}_* = \mathbf{B}_*\,\mathbf{z}^*$. The contracted product of the tensors **a**,...,**z** with the permutation tensor of $\mathbb{R}^n$ is called the parallelepipedal product of the tensors **a**,...,**z**. The parallelepipedal product is a scalar and is designated by s.

$$\begin{aligned}
s &= \varepsilon_{i\ldots m}\ a^i\ldots z^m &&= e_{i\ldots m}\ a^i\ldots z^m \det \mathbf{B}_* \\
s &= \varepsilon^{i\ldots m}\ a_i\ldots z_m &&= e^{i\ldots m}\ a_i\ldots z_m \det \mathbf{B}^*
\end{aligned}$$

Vector product : Let n – 1 tensors **a**,...,**y** of rank 1 in the euclidean space $\mathbb{R}^n$ be given. Let their coordinates be $a_i, \ldots, y_k$ with $i, k \in \{1,\ldots,n\}$ in the covariant basis $\mathbf{B}_*$ and $a^i, \ldots, y^k$ in the contravariant basis $\mathbf{B}^*$, so that their coordinates in the canonical basis are $\mathbf{a} = \mathbf{B}^*\,\mathbf{a}_* = \mathbf{B}_*\,\mathbf{a}^*$ to $\mathbf{y} = \mathbf{B}^*\,\mathbf{y}_* = \mathbf{B}_*\,\mathbf{y}^*$. The contracted product of these tensors with the permutation tensor of $\mathbb{R}^n$ is called the vector product of the tensors **a**,...,**y** in the space $\mathbb{R}^n$. The vector product is a tensor of rank 1. Let its coordinates be z_m in the basis $\mathbf{B}_*$ and z^m in the basis $\mathbf{B}^*$, so that its coordinates in the canonical basis are $\mathbf{z} = \mathbf{B}^*\,\mathbf{z}_* = \mathbf{B}_*\,\mathbf{z}^*$.

$$\begin{aligned}
\mathbf{z}\ &= \mathbf{a} \times \ldots \times \mathbf{y} \\
z_m &= \varepsilon_{i\ldots km}\ a^i\ldots y^k &&= e_{i\ldots km}\ a^i\ldots y^k \det \mathbf{B}_* \\
z^m &= \varepsilon^{i\ldots km}\ a_i\ldots y_k &&= e^{i\ldots km}\ a_i\ldots y_k \det \mathbf{B}^*
\end{aligned}$$

Orthogonality of the vector product : Let the coordinates of the n – 1 tensors $\mathbf{a},...,\mathbf{y}$ of rank 1 in a basis $\mathbf{B}_*$ of the euclidean space $\mathbb{R}^n$ be $a_i,...,y_k$. The vector product $\mathbf{z} = \mathbf{a} \times ... \times \mathbf{y}$ is orthogonal to each of the tensors $\mathbf{a},...,\mathbf{y}$. In fact, the scalar product of $\mathbf{z}$ with an arbitrary tensor $\mathbf{u} \in \{\mathbf{a},...,\mathbf{y}\}$ is proportional to the determinant of a matrix $\mathbf{A}$ with two identical columns $\mathbf{u}_*$, and is therefore zero :

$$\mathbf{u} \cdot \mathbf{z} = u_m z^m = b^* e^{i...km} a_i ... u_j ... y_k u_m = b^* \det \mathbf{A} = 0$$

Vector products of basis vectors : Let the vectors $\mathbf{b}_1, ..., \mathbf{b}_n$ and $\mathbf{b}^1, ..., \mathbf{b}^n$ form dual bases of the euclidean space $\mathbb{R}^n$. In the canonical basis $\mathbf{E}$, let the coordinates of $\mathbf{b}_r$ be designated by b_{i_r} or b_{m_r} and the coordinates of $\mathbf{b}^s$ by b^{i_s} or b^{m_s}. The vector products of n – 1 basis vectors are :

$$\mathbf{u} = \mathbf{b}^1 \times ... \times \mathbf{b}^{n-1} \qquad u_{i_n} = e_{i_1 ... i_n} b^{i_1} ... b^{i_{n-1}}$$

$$\mathbf{w} = \mathbf{b}_1 \times ... \times \mathbf{b}_{n-1} \qquad w^{i_n} = e^{i_1 ... i_n} b_{i_1} ... b_{i_{n-1}}$$

The vectors $\mathbf{u}$ and $\mathbf{w}$ are generally not the same. Their coordinates refer to the canonical basis. The scalar product of $\mathbf{u}$ with each of the vectors $\mathbf{b}^1, ..., \mathbf{b}^{n-1}$ is zero, since it is equal to the determinant of a matrix with two identical columns $\mathbf{b}^s$. Analogously, the scalar product of $\mathbf{w}$ with each of the vectors $\mathbf{b}_1, ..., \mathbf{b}_{n-1}$ is zero.

$$\mathbf{u} \cdot \mathbf{b}^s = e_{i_1 ... i_s ... i_{n-1} m_s} b^{i_1} ... b^{i_s} ... b^{i_{n-1}} b^{m_s} = 0$$

$$\mathbf{w} \cdot \mathbf{b}_r = e^{i_1 ... i_r ... i_{n-1} m_r} b_{i_1} ... b_{i_r} ... b_{i_{n-1}} b_{m_r} = 0$$

The scalar product of the vector $\mathbf{u}$ with the basis vector $\mathbf{b}^n$ is equal to the determinant b^* of the basis $\mathbf{B}^*$. Analogously, the scalar product of $\mathbf{w}$ and $\mathbf{b}_n$ is equal to the determinant b_* of the basis $\mathbf{B}_*$.

$$\mathbf{u} \cdot \mathbf{b}^n = e_{i_1 ... i_n} b^{i_1} ... b^{i_n} = \det \mathbf{B}^* = b^*$$

$$\mathbf{w} \cdot \mathbf{b}_n = e^{i_1 ... i_n} b_{i_1} ... b_{i_n} = \det \mathbf{B}_* = b_*$$

Thus the vector product of n – 1 basis vectors yields the scaled dual n-th basis vector :

$$\mathbf{b}^1 \times ... \times \mathbf{b}^{n-1} = b^* \mathbf{b}_n$$

$$\mathbf{b}_1 \times ... \times \mathbf{b}_{n-1} = b_* \mathbf{b}^n$$

Cross product : The vector product $\mathbf{v}$ of two vectors $\mathbf{u}, \mathbf{w} \in \mathbb{R}^3$ is designated by $\mathbf{u} \times \mathbf{w}$ and is called a cross product.

$$\mathbf{v} = \mathbf{u} \times \mathbf{w}$$

$$v_m = \varepsilon_{ikm} u^i w^k = b_* e_{ikm} u^i w^k$$

$$v^m = \varepsilon^{ikm} u_i w_k = b^* e^{ikm} u_i w_k$$

The cross product **v** is orthogonal to the vectors **u** and **w**. Changing the order of the vectors **u** and **w** in the cross product changes the sign of the coordinates of **v**.

$$\mathbf{v}\cdot\mathbf{u} \;=\; \mathbf{v}\cdot\mathbf{w} \;=\; 0$$

$$\varepsilon_{ikm}\, u^i\, w^k \;=\; -\varepsilon_{kim}\, u^i\, w^k \;=\; -\varepsilon_{ikm}\, w^i\, u^k$$

$$\mathbf{u}\times\mathbf{w} \;=\; -\,\mathbf{w}\times\mathbf{u}$$

The cross product of two basis vectors yields the scaled dual third basis vector (see vector products of basis vectors). The cross product of a basis vector with itself is zero.

$$\mathbf{b}_i\times\mathbf{b}_k = b_*\,\mathbf{b}^m \qquad \mathbf{b}_i\times\mathbf{b}_i = 0 \qquad i,k,m \in \{1,2,3\}$$

$$\mathbf{b}^i\times\mathbf{b}^k = b^*\,\mathbf{b}_m \qquad \mathbf{b}^i\times\mathbf{b}^i = 0 \qquad i \neq k \neq m$$

Example 4 : Calculation of a parallelepipedal product
Let the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of the space $\mathbb{R}^3$ as well as the coordinates of the vectors **u**,**v**,**w** in the canonical basis **E** and in the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ be given.

$$\mathbf{B}_* = \begin{array}{|c|c|c|}\hline 1.00 & -0.50 & 0.50 \\ \hline -1.00 & 2.50 & -1.50 \\ \hline 1.00 & -2.50 & 2.50 \\ \hline\end{array} \qquad \mathbf{B}^* = \begin{array}{|c|c|c|}\hline 1.25 & 0.50 & 0 \\ \hline 0 & 1.00 & 1.00 \\ \hline -0.25 & 0.50 & 1.00 \\ \hline\end{array}$$

$$\mathbf{u} = \begin{array}{|c|}\hline -1.00 \\ \hline 1.00 \\ \hline 1.00 \\ \hline\end{array} \qquad \mathbf{u}_* = (\mathbf{B}_*)^T\mathbf{u} = \begin{array}{|c|}\hline -1.00 \\ \hline 0.50 \\ \hline 0.50 \\ \hline\end{array} \qquad \mathbf{u}^* = (\mathbf{B}^*)^T\mathbf{u} = \begin{array}{|c|}\hline -1.50 \\ \hline 1.00 \\ \hline 2.00 \\ \hline\end{array}$$

$$\mathbf{v} = \begin{array}{|c|}\hline 2.00 \\ \hline -1.00 \\ \hline 1.00 \\ \hline\end{array} \qquad \mathbf{v}_* = (\mathbf{B}_*)^T\mathbf{v} = \begin{array}{|c|}\hline 4.00 \\ \hline -6.00 \\ \hline 5.00 \\ \hline\end{array} \qquad \mathbf{v}^* = (\mathbf{B}^*)^T\mathbf{v} = \begin{array}{|c|}\hline 2.25 \\ \hline 0.50 \\ \hline 0 \\ \hline\end{array}$$

$$\mathbf{w} = \begin{array}{|c|}\hline 1.00 \\ \hline 2.00 \\ \hline -2.00 \\ \hline\end{array} \qquad \mathbf{w}_* = (\mathbf{B}_*)^T\mathbf{w} = \begin{array}{|c|}\hline -3.00 \\ \hline 9.50 \\ \hline -7.50 \\ \hline\end{array} \qquad \mathbf{w}^* = (\mathbf{B}^*)^T\mathbf{w} = \begin{array}{|c|}\hline 1.75 \\ \hline 1.50 \\ \hline 0 \\ \hline\end{array}$$

The parallelepipedal product of the vectors **u**, **v**, **w** is alternatively calculated with the coordinates of these vectors in each of the bases **E**, $\mathbf{B}^*$ and $\mathbf{B}_*$. The determinants det $\mathbf{B}_* = 2.00$ and det $\mathbf{B}^* = 0.50$ are used. The value of the parallelepipedal product is independent of the choice of the basis in which the coordinates of the vectors are specified.

canonical basis :
$$
\begin{aligned}
s &= e_{ikm}\, \overset{*}{u}^i\, \overset{*}{v}^k\, \overset{*}{w}^m = e^{ikm}\, \overset{*}{u}_i\, \overset{*}{v}_k\, \overset{*}{w}_m \\
&= e^{123}(-1.00 * 1.00 * 2.00) + e^{213}(-1.00 * 2.00 * 2.00) + \\
&\quad e^{231}(\ 1.00 * 1.00 * 1.00) + e^{321}(-1.00 * 1.00 * 1.00) + \\
&\quad e^{312}(\ 1.00 * 2.00 * 2.00) + e^{132}(-1.00 * 1.00 * 2.00) \\
&= (-2.00 + 1.00 + 4.00) - (-4.00 - 1.00 - 2.00) \\
s &= 10.00
\end{aligned}
$$

basis $\mathbf{B}^*$:
$$
\begin{aligned}
s &= e^{ikm}\, u_i\, v_k\, w_m \det \mathbf{B}^* \\
&= [e^{123}(-1.00 * 6.00 * 7.50) + e^{213}(-0.50 * 4.00 * 7.50) + \\
&\quad e^{231}(-0.50 * 5.00 * 3.00) + e^{321}(\ 0.50 * 6.00 * 3.00) + \\
&\quad e^{312}(\ 0.50 * 4.00 * 9.50) + e^{132}(-1.00 * 5.00 * 9.50)] * 0.50 \\
&= [(-45.00 - 7.50 + 19.00) - (-15 + 9.00 - 47.50)] * 0.50 \\
s &= 10.00
\end{aligned}
$$

basis $\mathbf{B}_*$:
$$
\begin{aligned}
s &= e_{ikm}\, u^i\, v^k\, w^m \det \mathbf{B}_* \\
&= [e^{123}(-1.50 * 0.50 * \ 0) + e^{213}(\ 1.00 * 2.25 * \ 0) + \\
&\quad e^{231}(\ 1.00 * \ 0 * 1.75) + e^{321}(\ 2.00 * 0.50 * 1.75) + \\
&\quad e^{312}(\ 2.00 * 2.25 * 1.50) + e^{132}(-1.50 * \ 0 * 1.50)] * 2.00 \\
&= [(0 + 0 + 6.75) - (0 + 1.75 + 0)] * 2.00 \\
s &= 10.00
\end{aligned}
$$

Example 5 : Calculation of a vector product

The vector product **f** of the vectors **u** and **w** of the preceding Example 4 is determined in the bases **E**, $\mathbf{B}^*$ and $\mathbf{B}_*$ of the space $\mathbb{R}^3$.

canonical basis :
$$
\begin{aligned}
\overset{*}{f}_i &= e_{ikm}\, \overset{*}{u}^k\, \overset{*}{w}^m \\
\overset{*}{f}_1 &= e_{123}(-1.00 * 2.00) + e_{132}(\ 1.00 * 2.00) = -4.00 \\
\overset{*}{f}_2 &= e_{231}(\ 1.00 * 1.00) + e_{213}(\ 1.00 * 2.00) = -1.00 \\
\overset{*}{f}_3 &= e_{312}(-1.00 * 2.00) + e_{321}(\ 1.00 * 1.00) = -3.00
\end{aligned}
$$

basis $\mathbf{B}^*$:
$$
\begin{aligned}
f_i &= e_{ikm}\, u^k\, w^m \det \mathbf{B}_* \\
f_1 &= [e_{123}(\ 1.00 * 0\ ) + e_{132}(\ 2.00 * 1.50)] * 2.00 = -6.00 \\
f_2 &= [e_{231}(\ 2.00 * 1.75) + e_{213}(-1.50 * 0\ )] * 2.00 = \ 7.00 \\
f_3 &= [e_{312}(-1.50 * 1.50) + e_{321}(\ 1.00 * 1.75)] * 2.00 = -8.00
\end{aligned}
$$

basis $\mathbf{B}_*$: $f^i = e^{ikm} u_k w_m \det \mathbf{B}^*$

$$f^1 = [e^{123}(-0.50 * 7.50) \quad + \quad e^{132}(\ \ 0.50 * 9.50)] * 0.50 = -4.25$$

$$f^2 = [e^{231}(-0.50 * 3.00) \quad + \quad e^{213}(\ \ 1.00 * 7.50\,)] * 0.50 = -4.50$$

$$f^3 = [e^{312}(-1.00 * 9.50) \quad + \quad e^{321}(-0.50 * 3.00)] * 0.50 = -4.00$$

The equality $\mathbf{B}_* \mathbf{f}^* = \mathbf{B}^* \mathbf{f}_* = \mathbf{f}$ shows that the calculations in the three bases lead to the same vector. This vector is orthogonal to **u** and **w** :

canonical basis : $\overset{*}{f}_i \overset{*}{u}^i = 4.00 * 1.00 - 1.00 * 1.00 - 3.00 * 1.00 = 0$

$\overset{*}{f}_i \overset{*}{w}^i = -4.00 * 1.00 - 1.00 * 2.00 + 3.00 * 2.00 = 0$

basis $\mathbf{B}^*$: $f_i u^i = 6.00 * 1.50 + 7.00 * 1.00 - 8.00 * 2.00 = 0$

$f_i w^i = -6.00 * 1.75 + 7.00 * 1.50 - 8.00 * 0 = 0$

basis $\mathbf{B}_*$: $f^i u_i = 4.25 * 1.00 - 4.50 * 0.50 - 4.00 * 0.50 = 0$

$f^i w_i = 4.25 * 3.00 - 4.50 * 9.50 + 4.00 * 7.50 = 0$

9.3.6 TENSORS OF FIRST AND SECOND RANK

Introduction : Tensors of rank 1 (vectors) and tensors of rank 2 (dyads) are often used in formulating physical problems. The coordinates of these tensors are arranged in vectors and matrices, respectively. In the following, the general properties of tensors are specialized for vectors and dyads. Both coordinate notation and matrix notation are used. Contracted products of dyads and vectors (bilinear and quadratic forms, associated tensors, principal tensors) and contracted products of dyads (inner product, scalar product) are especially important.

Vector : A tensor of rank 1 is often called a vector in the literature. The coordinates of a tensor of rank 1 in the space $\mathbb{R}^n$ are arranged in a vector **t** of dimension n. The coordinates may be specified in the covariant form t_i or in the contravariant form t^i.

$$\mathbf{t}_* = \begin{array}{|c|} \hline t_1 \\ \hline t_2 \\ \hline \vdots \\ \hline t_n \\ \hline \end{array} \qquad \mathbf{t}^* = \begin{array}{|c|} \hline t^1 \\ \hline t^2 \\ \hline \vdots \\ \hline t^n \\ \hline \end{array}$$

Vectors have the following properties :

commutative :	$\mathbf{t} + \mathbf{u}$	$=$	$\mathbf{u} + \mathbf{t}$
associative :	$\mathbf{t} + (\mathbf{u} + \mathbf{w})$	$=$	$(\mathbf{t} + \mathbf{u}) + \mathbf{w}$
zero vector :	$\mathbf{t} + \mathbf{0}$	$=$	$\mathbf{t}$
inverse :	$\mathbf{t} + (-\mathbf{t})$	$=$	$\mathbf{0}$
distributive :	$(a + c)\mathbf{t}$	$=$	$a\,\mathbf{t} + c\,\mathbf{t}$
	$a(\mathbf{t} + \mathbf{u})$	$=$	$a\,\mathbf{t} + a\,\mathbf{u}$

Note : The definition of a tensor of rank 1 as a linear scalar vector mapping t(**u**) contains a vector **u** as an argument. The definition of a tensor does not require the vector **u** to be a tensor. In the mapping $t : \mathbb{R}^n \rightarrow \mathbb{R}$, **u** is only required to be a vector in the domain $\mathbb{R}^n$. By contrast, each coordinate t_i of the tensor in a basis $\mathbf{B}_*$ is the image $t(\mathbf{b}_i)$ of a basis vector $\mathbf{b}_i$, and therefore an element of the target $\mathbb{R}$. The images of the n basis vectors $\mathbf{b}_1, ..., \mathbf{b}_n$ are arranged in a vector **t**. The vectors **u** and **t** are similar in their schematic representation, but they have different meanings.

Transformation of the basis : The coordinates of a tensor of rank 1 may be images of the vectors of a covariant basis $\mathbf{B}_*$ or of the dual contravariant basis $\mathbf{B}^*$. Transforming the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ into the bases $\bar{\mathbf{B}}_*$ and $\bar{\mathbf{B}}^*$ with a transformation matrix $\mathbf{A}$ and its inverse $\bar{\mathbf{A}}$ generally changes the coordinates of the tensor.

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$$

$$\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^{\mathrm{T}} \quad \wedge \quad \bar{\mathbf{A}} = \mathbf{A}^{-1}$$

$$\mathbf{A} = \begin{array}{|c|c|c|c|} \hline a^{1}_{.1} & a^{1}_{.2} & \cdots & a^{1}_{.n} \\ \hline a^{2}_{.1} & a^{2}_{.2} & & a^{2}_{.n} \\ \hline \vdots & & \ddots & \vdots \\ \hline a^{n}_{.1} & a^{n}_{.2} & \cdots & a^{n}_{.n} \\ \hline \end{array}$$

Dual coordinates of a vector : Let the coordinates $\mathbf{t}_*$ and $\mathbf{t}^*$ of a tensor of rank 1 be referred to the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$. Let the coordinates of the metric tensor be g_{im} in the basis $\mathbf{B}_*$ and g^{im} in the basis $\mathbf{B}^*$. Then according to Section 9.3.3 the following relationships hold between the vector coordinates in the dual bases :

$$t_i = g_{is}\, t^s \quad \Leftrightarrow \quad \mathbf{t}_* = \mathbf{G}_* \mathbf{t}^*$$

$$t^i = g^{is}\, t_s \quad \Leftrightarrow \quad \mathbf{t}^* = \mathbf{G}^* \mathbf{t}_*$$

Transformation of vector coordinates : According to Section 9.3.3, the following relationships hold between the coordinates $\mathbf{t}_*$ and $\mathbf{t}^*$ of a tensor of rank 1 in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ and the coordinates $\bar{\mathbf{t}}_*$ and $\bar{\mathbf{t}}^*$ of the same tensor in the transformed bases $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^{\mathrm{T}}$:

$$\bar{t}_i = a^{s}_{.i}\, t_s \quad \Leftrightarrow \quad \bar{\mathbf{t}}_* = \mathbf{A}^{\mathrm{T}} \mathbf{t}_*$$

$$\bar{t}^i = \bar{a}^{i}_{.s}\, t^s \quad \Leftrightarrow \quad \bar{\mathbf{t}}^* = \bar{\mathbf{A}}\, \mathbf{t}^*$$

Scalar products of vectors : The contracted product of two tensors **t** and **w** of rank 1 is called the scalar product of the tensors and is designated by $\mathbf{t} \cdot \mathbf{u}$. The scalar product may be represented either using covariant and contravariant coordinates or using coordinates of the same type together with the metric tensor.

$$\mathbf{t} \cdot \mathbf{w} = t_i\, w^i = t^i\, w_i = g^{im}\, t_i\, w_m = g_{im}\, t^i\, w^m$$

Dyad : A tensor of rank 2 is called a dyad. The coordinates of a dyad in the space $\mathbb{R}^n$ are arranged in a quadratic matrix **T** of dimension n. The coordinates may be specified in the covariant form t_{im}, the contravariant form t^{im} or the mixed forms $t_{i.}^{\,m}$ and $t^{i}_{.m}$.

$$\mathbf{T}_* = \begin{array}{|c|c|c|c|}\hline t_{11} & t_{12} & & t_{1n} \\ \hline t_{21} & t_{22} & & t_{2n} \\ \hline & & \ddots & \\ \hline t_{n1} & t_{n2} & & t_{nn} \\ \hline \end{array} \qquad \mathbf{T}^* = \begin{array}{|c|c|c|c|}\hline t^{11} & t^{12} & & t^{1n} \\ \hline t^{21} & t^{22} & & t^{2n} \\ \hline & & \ddots & \\ \hline t^{n1} & t^{n2} & & t^{nn} \\ \hline \end{array}$$

$$\mathbf{T}_o = \begin{array}{|c|c|c|c|}\hline t_{1.}^{\;1} & t_{1.}^{\;2} & & t_{1.}^{\;n} \\ \hline t_{2.}^{\;1} & t_{2.}^{\;2} & & t_{2.}^{\;n} \\ \hline & & \ddots & \\ \hline t_{n.}^{\;1} & t_{n.}^{\;2} & & t_{n.}^{\;n} \\ \hline \end{array} \qquad \mathbf{T}^o = \begin{array}{|c|c|c|c|}\hline t^{1}_{\;.1} & t^{1}_{\;.2} & & t^{1}_{\;.n} \\ \hline t^{2}_{\;.1} & t^{2}_{\;.2} & & t^{2}_{\;.n} \\ \hline & & \ddots & \\ \hline t^{n}_{\;.1} & t^{n}_{\;.2} & & t^{n}_{\;.n} \\ \hline \end{array}$$

A dyad is a scalar linear mapping $T : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$. The two factors of the cartesian product $\mathbb{R}^n \times \mathbb{R}^n$ are usually referred either to the same basis $\mathbf{B}_*$ or to the same basis $\mathbf{B}^*$ or to a pair of dual bases $\mathbf{B}_*$, $\mathbf{B}^*$. In this case the coordinate $t_{im} = T(\mathbf{b}_i, \mathbf{b}_m)$ is referred to the basis $\mathbf{B}_*$. However, the two factors of the cartesian product may also be referred to different bases, for instance $\mathbf{B}_*$ and $\mathbf{C}_*$. In this case the two indices of the coordinate $t_{im} = T(\mathbf{b}_i, \mathbf{c}_m)$ refer to different bases. Dyads have the following properties :

commutative	:	$\mathbf{T} + \mathbf{U}$	$=$	$\mathbf{U} + \mathbf{T}$
associative	:	$\mathbf{T} + (\mathbf{U} + \mathbf{W})$	$=$	$(\mathbf{T} + \mathbf{U}) + \mathbf{W}$
identity element	:	$\mathbf{T} + \mathbf{0}$	$=$	$\mathbf{T}$
inverse element	:	$\mathbf{T} + (-\mathbf{T})$	$=$	$\mathbf{0}$
distributive	:	$(a + b)\mathbf{T}$	$=$	$a\mathbf{T} + b\mathbf{T}$
		$a(\mathbf{T} + \mathbf{U})$	$=$	$a\mathbf{T} + a\mathbf{U}$

Dual coordinates of a dyad : Let the coordinates of a dyad T be referred to the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$. Let the coordinates of the metric tensor be g_{im} in the basis $\mathbf{B}_*$ and g^{im} in the basis $\mathbf{B}^*$. Then according to Section 9.3.3 the following relationships hold between the coordinates of the dyad in the dual bases. The symmetry $\mathbf{G}_* = \mathbf{G}_*^{\mathrm{T}}$ and $\mathbf{G}^* = (\mathbf{G}^*)^{\mathrm{T}}$ of the metric is not used in the formulas.

$$\begin{array}{lclcrcl}
t_{im} & = & g_{ir}\, g_{ms}\, t^{rs} & \Leftrightarrow & \mathbf{T}_* & = & \mathbf{G}_*\, \mathbf{T}^*\, (\mathbf{G}_*)^{\mathrm{T}} \\
t^{im} & = & g^{ir}\, g^{ms}\, t_{rs} & \Leftrightarrow & \mathbf{T}^* & = & \mathbf{G}^*\, \mathbf{T}_*\, (\mathbf{G}^*)^{\mathrm{T}} \\
t_{i.}^{\;m} & = & g_{ir}\, g^{ms}\, t^{r}_{\;.s} & \Leftrightarrow & \mathbf{T}_o & = & \mathbf{G}_*\, \mathbf{T}^o\, (\mathbf{G}^*)^{\mathrm{T}} \\
t^{i}_{\;.m} & = & g^{ir}\, g_{ms}\, t_{r.}^{\;s} & \Leftrightarrow & \mathbf{T}^o & = & \mathbf{G}^*\, \mathbf{T}_o\, (\mathbf{G}_*)^{\mathrm{T}} \\
t_{im} & = & g_{ir}\, t^{r}_{\;.m} & \Leftrightarrow & \mathbf{T}_* & = & \mathbf{G}_*\, \mathbf{T}^o \\
t^{im} & = & g^{ir}\, t_{r.}^{\;m} & \Leftrightarrow & \mathbf{T}^* & = & \mathbf{G}^*\, \mathbf{T}_o
\end{array}$$

Transformation of dyad coordinates : According to Section 9.3.3 the following relationships hold between the coordinates of a dyad in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ and the coordinates of the same dyad in the transformed bases $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^{\mathrm T}$:

$$\begin{array}{lcl@{\qquad}c@{\qquad}lcl}
\bar t_{im} & = & a^{r}_{.i}\, a^{s}_{.m}\, t_{rs} & \Leftrightarrow & \bar{\mathbf{T}}_* & = & \mathbf{A}^{\mathrm T}\, \mathbf{T}_*\, \mathbf{A} \\
\bar t^{im} & = & \bar a^{i}_{.r}\, \bar a^{m}_{.s}\, t^{rs} & \Leftrightarrow & \bar{\mathbf{T}}^* & = & \bar{\mathbf{A}}\ \mathbf{T}^*\, \bar{\mathbf{A}}^{\mathrm T} \\
\bar t_{i.}^{m} & = & a^{r}_{.i}\, \bar a^{m}_{.s}\, t_{r.}^{s} & \Leftrightarrow & \bar{\mathbf{T}}_{o} & = & \mathbf{A}^{\mathrm T}\, \mathbf{T}_{o}\, \bar{\mathbf{A}}^{\mathrm T} \\
\bar t^{i}_{.m} & = & \bar a^{i}_{.r}\, a^{s}_{.m}\, t^{r}_{.s} & \Leftrightarrow & \bar{\mathbf{T}}^{o} & = & \bar{\mathbf{A}}\ \mathbf{T}^{o}\, \mathbf{A}
\end{array}$$

Products of tensors of rank 1 : Let X and Y be tensors of rank 1 in the metric space $\mathbb{R}^n$. Their coordinates are specified in the vectors **x** and **y**. The tensor product of tensors of rank 1 is a dyad $\mathbf{T} = \mathbf{xy}$. The tensor product is generally not commutative, that is $\mathbf{xy} \neq \mathbf{yx}$. The coordinates of the tensors X and Y may be referred to different bases of the space $\mathbb{R}^n$. For example, the covariant coordinates of Y may be referred to the basis $\mathbf{B}_*$ and the covariant coordinates of Y to the basis $\mathbf{C}_* \neq \mathbf{B}_*$.

covariant :	$t_{im} = x_i\, y_m$		$i, m = 1,...,n$
contravariant :	$t^{im} = x^i\, y^m$		
mixed :	$t_{i.}^{m} = x_i\, y^m$	$t^{i}_{.m} = x^i\, y_m$	

Inner products of dyads : The inner product of the dyads **T** and **U** is a dyad **W** and is designated by $\mathbf{T} \cdot \mathbf{U}$. According to the general definition of the inner product of tensors, the last index of the coordinates of T is contracted with the first index of the coordinates of U. This corresponds to the product of the matrices **T** and **U**. The inner product of two dyads is not commutative.

$$\mathbf{W} = \mathbf{T} \cdot \mathbf{U} \quad \Rightarrow \quad \begin{array}{lclcl}
w_{im} & = & t_{ik}\, u^{k}_{.m} & = & t_{i.}^{k}\ u_{km} \\
w^{im} & = & t^{ik}\, u_{k.}^{m} & = & t^{i}_{.k}\, u^{km} \\
w_{i.}^{m} & = & t_{ik}\, u^{km} & = & t_{i.}^{k}\ u_{k.}^{m} \\
w^{i}_{.m} & = & t^{ik}\, u_{km} & = & t^{i}_{.k}\ u^{k}_{.m}
\end{array}$$

The inner product of dyads has the following properties :

associative :	$(\mathbf{T} \cdot \mathbf{U}) \cdot \mathbf{R}$	$=$	$\mathbf{T} \cdot (\mathbf{U} \cdot \mathbf{R})$		
distributive :	$\mathbf{T} \cdot (\mathbf{U} + \mathbf{R})$	$=$	$\mathbf{T} \cdot \mathbf{U} + \mathbf{T} \cdot \mathbf{R}$		
	$(\mathbf{T} + \mathbf{U}) \cdot \mathbf{R}$	$=$	$\mathbf{T} \cdot \mathbf{R} + \mathbf{U} \cdot \mathbf{R}$		
scaled :	$s(\mathbf{T} \cdot \mathbf{U})$	$=$	$(s\mathbf{T}) \cdot \mathbf{U}$	$=$	$\mathbf{T} \cdot (s\mathbf{U})$

Scalar products of dyads : The completely contracted product of two dyads **T** and **U** is called the scalar product of the dyads. If the first index of the coordinates of the first dyad **T** is contracted with the first index of the coordinates of the second dyad **U** and the second index of **T** with the second index of **U**, the scalar product is designated by **T : U**. Otherwise the scalar product is designated by **T ·· U**. The two scalar products coincide if and only if one of the two dyads is symmetric.

$$s = \mathbf{T} : \mathbf{U} \quad :\Leftrightarrow \quad s = t_{im}\, u^{im}$$

$$s = \mathbf{T} \cdot\cdot\, \mathbf{U} \quad :\Leftrightarrow \quad s = t_{im}\, u^{mi}$$

The scalar product of dyads has the following properties :

$$\begin{array}{lll} \text{commutative :} & \mathbf{T} : \mathbf{U} = \mathbf{U} : \mathbf{T} \\ \text{distributive :} & \mathbf{T} : (\mathbf{U} + \mathbf{W}) = \mathbf{T} : \mathbf{U} + \mathbf{T} : \mathbf{W} \\ \text{scaled :} & a(\mathbf{T} : \mathbf{U}) = (a\mathbf{T}) : \mathbf{U} = \mathbf{T} : (a\mathbf{U}) \\ \text{positive :} & \mathbf{T} : \mathbf{T} > 0 \quad \text{for} \quad \mathbf{T} \neq \mathbf{0} \end{array}$$

Bilinear and quadratic forms : The contracted product of a dyad **T** with two tensors **x** and **y** of rank 1 is called a bilinear form. The bilinear form is a scalar c and is designated by $\mathbf{x} \cdot \mathbf{T} \cdot \mathbf{y}$. The order of the vectors in a bilinear form is generally relevant, that is $\mathbf{x} \cdot \mathbf{T} \cdot \mathbf{y} \neq \mathbf{y} \cdot \mathbf{T} \cdot \mathbf{x}$. For a symmetric dyad, however, $\mathbf{x} \cdot \mathbf{T} \cdot \mathbf{y} = \mathbf{y} \cdot \mathbf{T} \cdot \mathbf{x}$.

$$\begin{aligned} c = \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{y} &= t_{im}\, x^i\, y^m = t^{im}\, x_i\, y_m \\ &= t_{i\,.}^{\;m}\, x^i\, y_m = t^{\,i}_{\,.\,m}\, x_i\, y^m \end{aligned}$$

If the tensors **x** and **y** of a bilinear form are equal, the form is called a quadratic form (a quadric) and is designated by $q = \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x}$.

$$q = t^{im}\, x_i\, x_m = t_{im}\, x^i\, x^m = t_{i\,.}^{\;m}\, x^i\, x_m = t^{\,i}_{\,.\,m}\, x_i\, x^m$$

Associated tensors : The contracted product of a dyad **T** and a tensor **x** of rank 1 is called the tensor associated with T and x. If the right-hand index of the coordinates of T is contracted, the resulting product is called the right associated tensor and is designated by $\mathbf{T} \cdot \mathbf{x}$. If the left-hand index of the coordinates of T is contracted, the resulting product is called the left associated tensor and is designated by $\mathbf{x} \cdot \mathbf{T}$. The right and left associated tensors of T and x coincide if and only if the dyad is symmetric.

$$\begin{array}{lcll} \mathbf{y}_* = \mathbf{T}_o \cdot \mathbf{x}_* & : & y_i = t_{i\,.}^{\;m}\, x_m \\ \mathbf{y}^* = \mathbf{T}^o \cdot \mathbf{x}^* & : & y^i = t^{\,i}_{\,.\,m}\, x^m \\ \mathbf{z}_* = \mathbf{x}_* \cdot \mathbf{T}^o & : & z_i = t^{\,m}_{\,.\,i}\, x_m \\ \mathbf{z}^* = \mathbf{x}^* \cdot \mathbf{T}_o & : & z^i = t_{m\,.}^{\;i}\, x^m \end{array}$$

Principal tensors : Let a dyad T be real. If there is a real tensor u of rank 1 such that the tensor associated with T and u is proportional to u, then u is called a principal tensor of T. The vector **u** of the coordinates of u is called an eigenvector of the coordinate matrix **T**. The factor of proportionality p is called an eigenvalue of **T**. If the dyad T is not symmetric, the left and right principal tensors of the dyad are different. In the following (see Characteristic polynomial) the four forms of the eigenvalue problem are shown to lead to the same eigenvalues p.

$$\begin{array}{lllllllll}
\mathbf{T}_o \cdot \mathbf{u}_* & = & p\,\mathbf{u}_* & : & t_{i.}^{\;m}\, u_m & = & p\,u_i & = & p\,\delta_i^m\, u_m \\
\mathbf{T}^o \cdot \mathbf{u}^* & = & p\,\mathbf{u}^* & : & t^{i}_{\;.m}\, u^m & = & p\,u^i & = & p\,\delta^i_m\, u^m \\
\mathbf{w}_* \cdot \mathbf{T}^o & = & p\,\mathbf{w}_* & : & t^{i}_{\;.m}\, w_i & = & p\,w_m & = & p\,\delta^i_m\, w_i \\
\mathbf{w}^* \cdot \mathbf{T}_o & = & p\,\mathbf{w}^* & : & t_{i.}^{\;m}\, w^i & = & p\,w^m & = & p\,\delta_i^m\, w^i
\end{array}$$

The covariant and contravariant coordinates of the dyad T are obtained from the mixed coordinates by multiplication with the metric **G**. Multiplying the preceding equations with the metric leads to the following equations for the principal tensors :

$$\begin{array}{lllllll}
\mathbf{T}^* \cdot \mathbf{u}_* & = & p\,\mathbf{G}^* \cdot \mathbf{u}_* & : & t^{im}\, u_m & = & p\,g^{im}\, u_m \\
\mathbf{T}_* \cdot \mathbf{u}^* & = & p\,\mathbf{G}_* \cdot \mathbf{u}^* & : & t_{im}\, u^m & = & p\,g_{im}\, u^m \\
\mathbf{w}_* \cdot \mathbf{T}^* & = & p\,\mathbf{w}_* \cdot \mathbf{G}^* & : & t^{im}\, w_i & = & p\,g^{im}\, w_i \\
\mathbf{w}^* \cdot \mathbf{T}_* & = & p\,\mathbf{w}^* \cdot \mathbf{G}_* & : & t_{im}\, w^i & = & p\,g_{im}\, w^i
\end{array}$$

Characteristic polynomial : The equations for the eigenvalues and eigenvectors of a coordinate matrix **T** are homogeneous systems of equations with the variables p, **u** and **w**. These systems of equations have non-trivial solutions if the determinant of the coefficient matrix is zero.

$$\begin{array}{lllllll}
(t_{i.}^{\;m} - p\,\delta_i^m)\,u_m & = & 0 & \Rightarrow & \det(\mathbf{T}_o - p\,\mathbf{I}) & = & 0 \\
(t^{i}_{\;.m} - p\,\delta^i_m)\,u^m & = & 0 & \Rightarrow & \det(\mathbf{T}^o - p\,\mathbf{I}) & = & 0
\end{array}$$

Since the determinant of the metric **G** is not zero, the conditions involving the covariant and the contravariant coordinates of T lead to the same equations for the eigenvalues p :

$$\begin{array}{lllllll}
(t_{im} - p\,g_{im})\,u^m & = & 0 & \Rightarrow & g_{ik}(t^{k}_{\;.m} - p\,\delta^k_m)\,u^m & = & 0 \\
 & & & & \det \mathbf{G}_* \det(\mathbf{T}^o - p\,\mathbf{I}) & = & 0 \\
(t^{im} - p\,g^{im})\,u_m & = & 0 & \Rightarrow & g^{ik}(t_{k.}^{\;m} - p\,\delta_k^m)\,u_m & = & 0 \\
 & & & \Rightarrow & \det \mathbf{G}^* \det(\mathbf{T}_o - p\,\mathbf{I}) & = & 0
\end{array}$$

Due to the relationships between the mixed coordinates of a dyad, the two equations for the eigenvalues p coincide :

$$\begin{aligned}\det(\mathbf{T}^{\mathrm{o}} - p\mathbf{I}) &= \det(\mathbf{G}^{*}\mathbf{T}_{\mathrm{o}}\mathbf{G}_{*}^{\mathrm{T}} - p\,\mathbf{G}^{*}\mathbf{G}_{*}^{\mathrm{T}}) \\ &= \det(\mathbf{G}^{*})\det(\mathbf{T}_{\mathrm{o}} - p\mathbf{I})\det(\mathbf{G}_{*}^{\mathrm{T}}) \\ &= \det(\mathbf{T}_{\mathrm{o}} - p\mathbf{I})\end{aligned}$$

The equations for the left and right principal tensors of T also lead to the same eigenvalues. For example, for the mixed coordinates of T and $\mathbf{u} \cdot \mathbf{w} \neq 0$:

$$t_{i\,.}^{\;m}\, u_m = p\, u_i \qquad\qquad t_{i\,.}^{\;m}\, w^i = s\, w^m$$

$$t_{i\,.}^{\;m}\, w^i\, u_m = p\, u_i\, w^i = s\, w^m\, u_m \quad \Rightarrow \quad p = s$$

For a real dyad T the left-hand side of the equation, $\det(\mathbf{T}_{\mathrm{o}} - p\mathbf{I})$, is a polynomial of degree n with real coefficients. This polynomial is called the characteristic polynomial of the dyad T and is designated by C(p).

$$C(p) = \det(\mathbf{T}_{\mathrm{o}} - p\mathbf{I}) = (t_{1\,.}^{\;k_1} - p\,\delta_1^{k_1}) \ldots (t_{n\,.}^{\;k_n} - p\,\delta_n^{k_n})\, e_{k_1 \ldots k_n} = 0$$

Eigenvalues of a dyad : The solutions of the characteristic equation C(p) = 0 of a dyad T are called the eigenvalues of the dyad T. According to the fundamental theorem of algebra, every algebraic equation of degree n has exactly n roots (solutions) in the set $\mathbb{C}$ of complex numbers, where k-fold roots are counted k times. Thus there are eigenvalues of real dyads which are not real.

$$C(p_i) = 0 \quad \wedge \quad p_i \in \mathbb{C} \quad \wedge \quad i = 1,\ldots,n$$

Eigenvectors of a dyad : Substituting an eigenvalue p_i into the eigenvalue equation $\mathbf{T}_{\mathrm{o}} \cdot \mathbf{u}_{*} = p\,\mathbf{u}_{*}$ or $\mathbf{w}^{*} \cdot \mathbf{T}_{\mathrm{o}} = p\,\mathbf{w}^{*}$ yields a homogeneous linear system of equations with the solutions $\mathbf{u}_{*i}$ or $\mathbf{w}^{*i}$. These solutions are called the i-th right and left eigenvectors of the dyad T, respectively. The i-th left and right eigenvectors of a dyad are generally different. If the eigenvalue p_i is complex, the eigenvectors $\mathbf{u}_{*i}$ and $\mathbf{w}^{*i}$ are also complex. A pair (p,u) consisting of an eigenvalue p and the corresponding eigenvector u is called an eigenstate of the dyad.

$$\begin{aligned}\mathbf{T}_{\mathrm{o}} \cdot \mathbf{u}_{*i} &= p_i\, \mathbf{u}_{*i} & \qquad p_i &\in \mathbb{C} \\ \mathbf{w}^{*i} \cdot \mathbf{T}_{\mathrm{o}} &= p_i\, \mathbf{w}^{*i} & \qquad \mathbf{u}_{*i},\ \mathbf{w}^{*i} &\in \mathbb{C}^n \\ & & i &= 1,\ldots,n\end{aligned}$$

Example 1 : Eigenvalues of a dyad in the space $\mathbb{R}^2$

The characteristic polynomial of a dyad T in the real space $\mathbb{R}^2$ is quadratic. The sign of the discriminant $(I_1)^2 - 4I_2$ determines the character of the solution. For $(I_1)^2 - 4I_2 \geq 0$ the eigenvalues are real, otherwise the eigenvalues are complex.

determinant : $\det \begin{vmatrix} t_{1\,.}^{\,1} - p & t_{1\,.}^{\,2} \\ t_{2\,.}^{\,1} & t_{2\,.}^{\,2} - p \end{vmatrix} = 0$

equation : $p^2 - I_1\, p + I_2 = 0$

$I_1 = t_{1\,.}^{\,1} + t_{2\,.}^{\,2}$

$I_2 = t_{1\,.}^{\,1}\, t_{2\,.}^{\,2} - t_{1\,.}^{\,2}\; t_{2\,.}^{\,1}$

eigenvalues : $p_1 = (0.5\, I_1) + \sqrt{(0.5\, I_1)^2 - I_2}$

$p_2 = (0.5\, I_1) - \sqrt{(0.5\, I_1)^2 - I_2}$

Example 2 : Complex eigenstates of a dyad in the space $\mathbb{R}^2$

The real dyad T has the complex eigenstates $(p_1, \mathbf{u}_{*1})$ and $(p_1, \mathbf{u}_{*2})$. The eigenvectors $\mathbf{u}_{*1}$ and $\mathbf{u}_{*2}$ are conjugate to each other.

determinant : $\det \begin{vmatrix} 6 - p & 5 \\ -5 & -2 - p \end{vmatrix} = 0$

eigenstates : $p_1 = 2 + 3i$ $\qquad p_2 = 2 - 3i$

$\mathbf{u}_{*1} = \begin{bmatrix} -4 - 3i \\ 5 \end{bmatrix} \qquad \mathbf{u}_{*2} = \begin{bmatrix} -4 + 3i \\ 5 \end{bmatrix}$

Example 3 : Eigenstates of a dyad in the space $\mathbb{R}^3$

The characteristic polynomial of a dyad in the space $\mathbb{R}^3$ is cubic. The solutions of the characteristic equation may be real or complex. The eigenstates of a symmetric dyad S are determined in the following. All eigenvalues and eigenvectors of this dyad are real.

determinant : $\det \begin{vmatrix} 0.0791 - p & -0.5789 & -1.0226 \\ -0.5789 & 0.9789 - p & -1.6736 \\ -1.0226 & -1.6736 & 2.9421 - p \end{vmatrix} = 0$

equation : $p^3 - I_1\, p^2 + I_2\, p - I_3 = 0$

$$I_1 = t_{1.}^{\,1} + t_{2.}^{\,2} + t_{3.}^{\,3} = 4.0$$

$$I_2 = (t_{1.}^{\,1}\, t_{2.}^{\,2} - t_{1.}^{\,2}\, t_{2.}^{\,1}) + (t_{2.}^{\,2}\, t_{3.}^{\,3} - t_{3.}^{\,2}\, t_{2.}^{\,3}) + (t_{3.}^{\,3}\, t_{1.}^{\,1} - t_{1.}^{\,3}\, t_{3.}^{\,1}) = -1.0$$

$$I_3 = \det S = -4.0$$

eigenvalues : $a = \sqrt{(I_1)^2 - 3 I_2} = \sqrt{19}$

$$b = 2(I_1)^3 - 9\, I_1 I_2 + 27 I_3 = 56$$

$$\phi = \arccos \frac{b}{2a^3} = 1.2259$$

$$p_i = \frac{1}{3}\left(I_1 + 2a \cos \frac{\phi + 2\pi(i-1)}{3}\right)$$

$$p_1 = 4.0 \qquad p_2 = -1.0 \qquad p_3 = 1.0$$

The eigenvector $\mathbf{u}_1$ belonging to the eigenvalue $p_1 = 4.0$ is determined using the system of equations $\mathbf{T}_o \mathbf{u}_1 = p_1 \mathbf{u}_1$. The last coordinate of the auxiliary vector **w** is arbitrarily set to 1.0. The auxiliary vector **w** determined by solving the system of equations is then normalized such that its magnitude is 1.0. The other eigenvectors are calculated analogously.

$$\underbrace{\begin{bmatrix} -3.9209 & -0.5789 & -1.0226 \\ -0.5789 & -3.0211 & -1.6736 \\ -1.0226 & -1.6736 & -1.0579 \end{bmatrix}}_{\mathbf{T}_o - 4.0\,\mathbf{I}} * \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}}_{\mathbf{w}} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}}_{\mathbf{0}}$$

eigenvectors : $\underbrace{\begin{bmatrix} -0.1623 \\ -0.4542 \\ 0.8760 \end{bmatrix}}_{\mathbf{u}_1} \qquad \underbrace{\begin{bmatrix} 0.7071 \\ 0.5657 \\ 0.4243 \end{bmatrix}}_{\mathbf{u}_2} \qquad \underbrace{\begin{bmatrix} 0.6883 \\ -0.6883 \\ -0.2294 \end{bmatrix}}_{\mathbf{u}_3}$

9.3.7 PROPERTIES OF DYADS

Introduction : The values of the coordinates of a dyad may lead to special properties of the dyad. Unitary dyads, symmetric and antisymmetric dyads and the decomposition of a regular dyad into the inner product of a unitary and a symmetric dyad are especially important.

Zero dyad : The scalar vector mapping $\mathrm{N} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ with $\mathrm{N}(\mathbf{u},\mathbf{v}) = 0$ which takes the value zero for arbitrary 2-tuples is called the zero dyad (zero tensor of rank 2). All coordinates of the zero dyad are zero. Adding the zero dyad to an arbitrary dyad T leaves T unchanged.

$$\mathbf{N} \;=\; \begin{array}{|c|c|c|}\hline 0 & & 0 \\ \hline & \ddots & \\ \hline 0 & & 0 \\ \hline \end{array}$$

Unit dyad : The scalar vector mapping $\mathrm{I} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$ which takes the value δ_i^m (Kronecker symbol) for arbitrary 2-tuples of dual basis vectors $\mathbf{b}_i$, $\mathbf{b}^m$ is called the unit dyad. The diagonal coefficients of the coordinate matrix $\mathbf{I}$ of the unit dyad are 1, all other coefficients are 0. Contracted multiplication of a vector w or a dyad W with the unit dyad leaves w and W unchanged.

$$\mathbf{I} \;=\; \begin{array}{|c|c|c|}\hline 1 & & 0 \\ \hline & \ddots & \\ \hline 0 & & 1 \\ \hline \end{array}$$

$$\mathrm{I}(\mathbf{b}_i\,,\,\mathbf{b}^m) \;=\; \mathrm{I}(\mathbf{b}^m,\,\mathbf{b}_i\,) \;=\; \delta_i^m$$

$$t_i \;=\; \delta_i^k\, w_k \qquad \Rightarrow \qquad t_i \;=\; w_i$$

$$t_{im} \;=\; \delta_i^k\, w_{km} \qquad \Rightarrow \qquad t_{im} \;=\; w_{im}$$

Unitary dyads : A dyad T is said to be right-unitary if for arbitrary tensors **x** and **y** of rank 1 in the euclidean space $\mathbb{R}^n$ the scalar product $\mathbf{x} \cdot \mathbf{y}$ is equal to the scalar product $\mathbf{u} \cdot \mathbf{w}$ of the right associated tensors $\mathbf{u} = \mathbf{T} \cdot \mathbf{x}$ and $\mathbf{w} = \mathbf{T} \cdot \mathbf{y}$.

$$\text{T is right-unitary} \quad \Rightarrow \quad \mathbf{x} \cdot \mathbf{y} \;=\; \mathbf{u} \cdot \mathbf{w} \quad \text{with} \quad \mathbf{u} \;=\; \mathbf{T} \cdot \mathbf{x}, \;\; \mathbf{w} \;=\; \mathbf{T} \cdot \mathbf{y}$$

A dyad T is said to be left-unitary if for arbitrary tensors **x** and **y** of rank 1 in the euclidean space $\mathbb{R}^n$ the scalar product $\mathbf{x} \cdot \mathbf{y}$ is equal to the scalar product $\mathbf{u} \cdot \mathbf{w}$ of the left associated tensors $\mathbf{u} = \mathbf{x} \cdot \mathbf{T}$ and $\mathbf{w} = \mathbf{y} \cdot \mathbf{T}$.

$$\text{T is left-unitary} \quad \Rightarrow \quad \mathbf{x} \cdot \mathbf{y} \;=\; \mathbf{u} \cdot \mathbf{w} \quad \text{with} \quad \mathbf{u} \;=\; \mathbf{x} \cdot \mathbf{T}, \;\; \mathbf{w} \;=\; \mathbf{y} \cdot \mathbf{T}$$

A right-unitary dyad is also left-unitary; it is therefore called a unitary dyad. A dyad is unitary (orthonormal) if the product of the transpose of its coordinate matrix $\mathbf{T}_*$ with the matrix $\mathbf{T}^*$ of its dual coordinates is the unit matrix $\mathbf{I}$. The magnitudes of the tensors $\mathbf{x}$ and $\mathbf{u}$ are equal. The magnitudes of the tensors $\mathbf{y}$ and $\mathbf{w}$ are also equal. The angle θ enclosed by the tensors $\mathbf{x}, \mathbf{y}$ is equal to the angle enclosed by the tensors $\mathbf{u}, \mathbf{w}$.

$$\mathbf{T} \text{ is unitary} \quad \Rightarrow \quad \mathbf{T}_*^T \, \mathbf{T}^* = \mathbf{I}$$

$$|\mathbf{x}| = |\mathbf{u}| \quad \wedge \quad |\mathbf{y}| = |\mathbf{w}| \quad \wedge \quad \cos\theta = \frac{\mathbf{x} \cdot \mathbf{y}}{|\mathbf{x}|\,|\mathbf{y}|} = \frac{\mathbf{u} \cdot \mathbf{w}}{|\mathbf{u}|\,|\mathbf{w}|}$$

Proof : Properties of unitary dyads

The coordinates of the associated tensors $\mathbf{u} = \mathbf{T} \cdot \mathbf{x}$ and $\mathbf{w} = \mathbf{T} \cdot \mathbf{y}$ are $u_i = t_{ik} x^k$ and $w^i = t^{im} y_m$. The property $x_i y^i = u_i w^i$ of a right-unitary dyad leads to the following properties of the coordinates of the right-unitary dyad :

$$u_i w^i = t_{ik} t^{im} x^k y_m = x^k y_k \quad \Rightarrow \quad t_{ik} t^{im} = \delta_k^m$$

$$\mathbf{T}_*^T \, \mathbf{T}^* = (\mathbf{T}^*)^T \, \mathbf{T}_* = \mathbf{I}$$

The coordinates of the associated tensors $\mathbf{u} = \mathbf{x} \cdot \mathbf{T}$ and $\mathbf{w} = \mathbf{y} \cdot \mathbf{T}$ are $u_i = t_{ki} x^k$ and $w^i = t^{mi} y_m$. The property $x_i y^i = u_i w^i$ of a left-unitary dyad leads to the following properties of the coordinates of the left-unitary dyad :

$$u_i w^i = t_{ki} t^{mi} x^k y_m = x^k y_k \quad \Rightarrow \quad t_{ki} t^{mi} = \delta_k^m$$

$$\mathbf{T}_* \mathbf{T}^{*T} = \mathbf{T}^* \, \mathbf{T}_*^T = \mathbf{I}$$

The condition $\mathbf{T}_*^T \, \mathbf{T}^* = \mathbf{I}$ implies $\mathbf{T}_*^T = (\mathbf{T}^*)^{-1}$, and thus also $\mathbf{T}^* \; \mathbf{T}_*^T = \mathbf{T}^* \; (\mathbf{T}^*)^{-1} = \mathbf{I}$, so that a right-unitary dyad is also a left-unitary dyad. The magnitudes of the associated tensors are obtained as follows :

$$|u|^2 = \mathbf{u} \cdot \mathbf{u} = t_{ik} \; t^{im} x^k x_m = x^k x_k = \mathbf{x} \cdot \mathbf{x} = |x|^2$$

$$|w|^2 = \mathbf{w} \cdot \mathbf{w} = t_{ik} t^{im} \; y^k y_m = y^k y_k = \mathbf{y} \cdot \mathbf{y} = |y|^2$$

Hence $|\mathbf{x}| = |\mathbf{u}|$ and $|\mathbf{y}| = |\mathbf{w}|$. Since by definition $\mathbf{x} \cdot \mathbf{y} = \mathbf{u} \cdot \mathbf{w}$, the angles enclosed by $\mathbf{x}, \mathbf{y}$ and by $\mathbf{u}, \mathbf{w}$ are equal.

Regular, singular and inverse dyads : A dyad is said to be regular if the determinant of its coefficient matrix is not zero. A dyad is said to be singular if the determinant of its coefficient matrix is zero. The dyad U is called the inverse of the dyad T if the coordinate matrix $\mathbf{U}$ is the inverse of the coordinate matrix $\mathbf{T}$. Regular dyads have an inverse, singular dyads do not have an inverse.

$$
\begin{aligned}
&\text{T is regular} && :\Leftrightarrow && \det \mathbf{T} \neq 0 \\
&\text{T is singular} && :\Leftrightarrow && \det \mathbf{T} = 0 \\
&\text{inverse dyad} && : && \mathbf{U} = \mathbf{T}^{-1}
\end{aligned}
$$

Definite and semidefinite dyads : The definiteness of a dyad T in the space $\mathbb{R}^n$ is determined by the value of the quadric $\mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x}$ for arbitrary vectors $\mathbf{x} \neq \mathbf{0}$ of $\mathbb{R}^n$. The dyad is said to be definite if the value of the quadric is non-zero for all $\mathbf{x} \neq \mathbf{0}$. The dyad is said to be positive (negative) definite if the value of the quadric is positive (negative) for all $\mathbf{x} \neq \mathbf{0}$. The dyad is said to be positive (negative) semidefinite if the value of the quadric is positive (negative) or zero for all $\mathbf{x}$.

$$
\begin{aligned}
&\text{T is definite} && :\Leftrightarrow && \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x} \neq 0 \\
&\text{T is positive definite} && :\Leftrightarrow && \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x} > 0 \\
&\text{T is negative definite} && :\Leftrightarrow && \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x} < 0 \\
&\text{T is positive semidefinite} && :\Leftrightarrow && \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x} \geq 0 \\
&\text{T is negative semidefinite} && :\Leftrightarrow && \mathbf{x} \cdot \mathbf{T} \cdot \mathbf{x} \leq 0
\end{aligned}
$$

Symmetric dyads : A dyad is said to be symmetric if its coordinates satisfy one of the following equivalent conditions :

$$
\begin{aligned}
s_{im} &= s_{mi} & \Leftrightarrow && s^{im} &= s^{mi} & \Leftrightarrow && s_{i.}^{m} &= s_{.i}^{m} \\
\mathbf{S}_* &= \mathbf{S}_*^T & \Leftrightarrow && \mathbf{S}^* &= (\mathbf{S}^*)^T & \Leftrightarrow && \mathbf{S}_o &= (\mathbf{S}^o)^T
\end{aligned}
$$

Dyads with mixed coordinates satisfying $s_{i.}^{m} = s_{.i}^{m}$ are symmetric. By contrast, dyads with coordinates satisfying $s_{i.}^{m} = s_{m.}^{i}$ and $s_{.m}^{i} = s_{.i}^{m}$ are generally not symmetric.

$$
\begin{aligned}
s_{i.}^{m} &= s_{.i}^{m} & \Rightarrow && s_{im} &= g_{ir}\, s_{.m}^{r} = g_{ir}\, s_{m.}^{r} = s_{mi} \\
s_{.m}^{i} &= s_{.i}^{m} & \Rightarrow && s_{im} &= g_{ir}\, s_{.m}^{r} = \sum_{r} g_{ir}\, s_{.r}^{m} \neq s_{mi}
\end{aligned}
$$

Eigenstates of symmetric dyads : The n eigenvalues and the n eigenvectors of a symmetric dyad S in the euclidean space $\mathbb{R}^n$ are real. The left and right eigenvectors of a symmetric dyad are pairwise equal. In fact, the symmetry condition $s_{i.}^{m} = s_{.i}^{m}$ implies :

$$
\begin{aligned}
\mathbf{S}_o \cdot \mathbf{u}_* = p\mathbf{u}_* \quad & \Leftrightarrow \quad s_{i.}^{m}\, u_m = p\delta_i^m\, u_m \\
& \Leftrightarrow \quad s_{.i}^{m}\, u_m = p\delta_m^i\, u_m \quad \Leftrightarrow \quad \mathbf{u}_* \cdot \mathbf{S}^o = p\mathbf{u}_*
\end{aligned}
$$

Orthogonality of the eigenvectors : Let the eigenvalues p and r of the eigenstates (p,**u**) and (r,**w**) of a symmetric dyad S be different. Then the eigenvectors **u** and **w** are orthogonal, since the coordinates of the tensors in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ satisfy the following equations :

$$s^{i}_{.m}\, u^{m} = p u^{i}$$

$$s^{i}_{.m}\, w_{i} = r\, w_{m}$$

$$s^{i}_{.m}\, u^{m}\, w_{i} = (s^{i}_{.m}\, u^{m})\, w_{i} = p\, u^{i}\, w_{i}$$

$$s^{i}_{.m}\, u^{m}\, w_{i} = (s^{i}_{.m}\, w_{i}\,)\, u^{m} = r\, w_{m} u^{m} = r\, u^{i}\, w_{i}$$

$$(p-r)u^{i}\, w_{i} = 0 \quad \wedge \quad p \neq r \quad \Rightarrow \quad u^{i}\, w_{i} = 0$$

If the eigenvalues p and r of the eigenstates (p, **u**) and (r, **w**) are equal, then every linear combination of the eigenvectors **u** and **w** is an eigenvector for S. Thus orthogonal eigenvectors $\mathbf{u}+c\mathbf{w}$ and $\mathbf{u}-c\mathbf{w}$ with $c\,|\mathbf{w}| = |\mathbf{u}|$ may be chosen.

$$(\mathbf{u}+c\,\mathbf{w})\cdot(\mathbf{u}-c\,\mathbf{w}) = \mathbf{u}\cdot\mathbf{u} - c^2\,\mathbf{w}\cdot\mathbf{w} = 0$$

Scaling of the eigenvectors : If **x** is an eigenvector for a symmetric dyad S, then the scaled eigenvector $\mathbf{u} = c\,\mathbf{x}$ with $c \in \mathbb{R}$ and $c \neq 0$ is also an eigenvector of S. Eigenvectors are often scaled such that the vector **u** has magnitude 1.

$$u_i\, u^i = 1 \quad \Rightarrow \quad c^2 x_i\, x^i = 1 \quad \Rightarrow \quad c = \pm \sqrt{x_i\, x^i}$$

Eigenmatrix of a symmetric dyad : For every symmetric dyad S in the space $\mathbb{R}^n$ there are n orthonormal eigenvectors. The coordinates of these eigenvectors may alternatively be specified in the canonical basis **E** or in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$. The eigenvectors are accordingly designated as follows :

$\overset{*}{\mathbf{u}}_1, \ldots, \overset{*}{\mathbf{u}}_n$	coordinates in the canonical basis **E**
$\mathbf{u}_1, \ldots, \mathbf{u}_n$	covariant coordinates of the eigenvectors
$\mathbf{u}^1, \ldots, \mathbf{u}^n$	contravariant coordinates of the eigenvectors

The eigenvectors are arranged in columns in the eigenmatrices **U**, $\mathbf{U}_*$ and $\mathbf{U}^*$ of the dyad S. These eigenmatrices do not represent dyads.

$$\mathbf{U} = \left[\begin{array}{c|c|c|c} \overset{*}{\mathbf{u}}_1 & \overset{*}{\mathbf{u}}_2 & \ldots & \overset{*}{\mathbf{u}}_n \end{array}\right] \qquad \mathbf{U}_* = \left[\begin{array}{c|c|c|c} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_n \end{array}\right] \qquad \mathbf{U}^* = \left[\begin{array}{c|c|c|c} \mathbf{u}^1 & \mathbf{u}^2 & \ldots & \mathbf{u}^n \end{array}\right]$$

The eigenmatrix $\mathbf{U}$ is orthonormal. Its inverse coincides with its transpose. Since the left and right inverses of a quadratic matrix are equal, $\mathbf{U}\mathbf{U}^{T} = \mathbf{I}$ also holds.

$$\mathbf{U}^{T}\mathbf{U} = \mathbf{I} \quad \Rightarrow \quad \mathbf{U}^{-1} = \mathbf{U}^{T}$$

$$\mathbf{U}^{-1} = \mathbf{U}^{T} \quad \wedge \quad \mathbf{U}\mathbf{U}^{-1} = \mathbf{I} \quad \Rightarrow \quad \mathbf{U}\mathbf{U}^{T} = \mathbf{I}$$

The covariant and contravariant coordinates satisfy $\mathbf{u} = \mathbf{B}_{*}\mathbf{u}^{*} = \mathbf{B}^{*}\mathbf{u}_{*}$, and hence $\mathbf{U} = \mathbf{B}_{*}\mathbf{U}^{*} = \mathbf{B}^{*}\mathbf{U}_{*}$. From $(\mathbf{B}_{*})^{T}\mathbf{B}^{*} = (\mathbf{B}^{*})^{T}\mathbf{B}_{*} = \mathbf{I}$ it follows that $\mathbf{U}^{*} = (\mathbf{B}^{*})^{T}\mathbf{U}$ and $\mathbf{U}_{*} = (\mathbf{B}_{*})^{T}\mathbf{U}$, and hence :

$$\mathbf{U}_{*}(\mathbf{U}^{*})^{T} = (\mathbf{B}_{*})^{T}\mathbf{U}\mathbf{U}^{T}\mathbf{B}^{*} = (\mathbf{B}_{*})^{T}\mathbf{B}^{*} = \mathbf{I}$$

$$\mathbf{U}^{*}(\mathbf{U}_{*})^{T} = (\mathbf{B}^{*})^{T}\mathbf{U}\mathbf{U}^{T}\mathbf{B}_{*} = (\mathbf{B}^{*})^{T}\mathbf{B}_{*} = \mathbf{I}$$

In summary, the eigenmatrix of a symmetric dyad satisfies the following equations :

$$\mathbf{U}^{T}\mathbf{U} = (\mathbf{U}_{*})^{T}\mathbf{U}^{*} = (\mathbf{U}^{*})^{T}\mathbf{U}_{*} = \mathbf{I}$$

$$\mathbf{U}\mathbf{U}^{T} = \mathbf{U}_{*}(\mathbf{U}^{*})^{T} = \mathbf{U}^{*}(\mathbf{U}_{*})^{T} = \mathbf{I}$$

Principal basis of symmetric dyads : The matrices of the mixed coordinates $s_{i.}^{m}$ and $s^{i}_{.m}$ of a symmetric dyad S are designated by $\mathbf{S}_{o}$ and $\mathbf{S}^{o}$, respectively, the diagonal matrix of the eigenvalues of S is designated by $\mathbf{P} = \mathbf{P}_{o} = \mathbf{P}^{o}$. Then the n eigenstates of the dyad S may be combined into a system of equations :

$$s_{i.}^{k}\, u_{km} = u_{ik}\, p^{k}_{.m} \quad \Rightarrow \quad \mathbf{S}_{o}\mathbf{U}_{*} = \mathbf{U}_{*}\mathbf{P}$$

$$s^{i}_{.k}\, u^{km} = u^{ik}\, p_{k.}^{m} \quad \Rightarrow \quad \mathbf{S}^{o}\mathbf{U}^{*} = \mathbf{U}^{*}\mathbf{P}$$

$$\mathbf{P} = \begin{array}{|c|c|c|c|}\hline p_{1.}^{1} & 0 & & 0 \\ \hline 0 & p_{2.}^{2} & & 0 \\ \hline & & \ddots & \\ \hline 0 & 0 & & p_{n.}^{n} \\ \hline \end{array} = \begin{array}{|c|c|c|c|}\hline p^{1}_{.1} & 0 & & 0 \\ \hline 0 & p^{2}_{.2} & & 0 \\ \hline & & \ddots & \\ \hline 0 & 0 & & p^{n}_{.n} \\ \hline \end{array}$$

Since the eigenmatrices are orthonormal, the matrices $\mathbf{S}_{o}$ and $\mathbf{S}^{o}$ may be decomposed into products of the matrices $\mathbf{U}_{*}$, $\mathbf{U}^{*}$ and $\mathbf{P}$ by multiplying the equations from the right by $(\mathbf{U}^{*})^{T}$ and $(\mathbf{U}_{*})^{T}$, respectively :

$$\mathbf{S}_{o} = \mathbf{U}_{*}\mathbf{P}(\mathbf{U}^{*})^{T}$$

$$\mathbf{S}^{o} = \mathbf{U}^{*}\mathbf{P}(\mathbf{U}_{*})^{T}$$

If the equations are instead multiplied from the left by $(\mathbf{U}^{*})^{T}$ and $(\mathbf{U}_{*})^{T}$, respectively, the diagonal matrix $\mathbf{P}$ is seen to equal the products $(\mathbf{U}^{*})^{T}\mathbf{S}_{o}\mathbf{U}_{*}$ and $(\mathbf{U}_{*})^{T}\mathbf{S}^{o}\mathbf{U}^{*}$. These products correspond to the transformation of the coordinates of the dyad under a transformation of the dual bases into $\bar{\mathbf{B}}_{*} = \mathbf{B}_{*}\mathbf{U}^{*}$ and $\bar{\mathbf{B}}^{*} = \mathbf{B}^{*}\mathbf{U}_{*}$:

$$\begin{aligned}
\mathbf{P} &= (\mathbf{U}^*)^T \mathbf{S}_o \mathbf{U}_* = (\mathbf{U}_*)^T \mathbf{S}^o \mathbf{U}^* \\
\bar{\mathbf{B}}_* &= \mathbf{B}_* \, \mathbf{U}^* = \mathbf{U} \\
\bar{\mathbf{B}}^* &= \mathbf{B}^* \, \mathbf{U}_* = \mathbf{U}
\end{aligned}$$

The transformed dual bases $\bar{\mathbf{B}}_*$ and $\bar{\mathbf{B}}^*$ are therefore identical and orthonormal. The vectors $\mathbf{u}_1, ..., \mathbf{u}_n$ in the eigenmatrix **U** are called the principal basis of the symmetric dyad S. The matrix of the mixed coordinates of the dyad S in the basis **U** is the diagonal matrix of the eigenvalues of S.

Quadric of a symmetric dyad : The properties of the quadric of a symmetric dyad are conveniently studied in the principal basis. Since in the principal basis the coordinates of a symmetric dyad form a diagonal matrix with the eigenvalues as diagonal elements, the quadric has the following form :

$$\mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} = p_{ii} \, x^i \, x^i$$

p_{ii} diagonal coordinates in **P**

x_i coordinates of the tensor **x**

The eigenvalues p_{ii} of the dyad determine the definiteness of the dyad :

$$\begin{array}{llll}
\bigwedge\limits_i \; p_{ii} > 0 & \Rightarrow & \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} > 0 : & \text{positive definite} \\
\bigwedge\limits_i \; p_{ii} < 0 & \Rightarrow & \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} < 0 : & \text{negative definite} \\
\bigwedge\limits_i \; p_{ii} \geq 0 & \Rightarrow & \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} \geq 0 : & \text{positive semidefinite} \\
\bigwedge\limits_i \; p_{ii} \leq 0 & \Rightarrow & \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} \leq 0 : & \text{negative semidefinite}
\end{array}$$

The determinant of the dyad S in the principal basis is equal to the product of the diagonal elements of its coordinate matrix **P**. Hence the dyad is regular if its eigenvalues p_{ii} are non-zero. The inverse dyad $\mathbf{P}^{-1}$ has the diagonal elements $1 / p_{ii}$. It is therefore also symmetric and regular.

$$\begin{aligned}
\det \mathbf{P} &= p_{11} \, p_{22} \cdots p_{nn} \\
\det \mathbf{P}^{-1} &= 1 / (p_{11} \, p_{22} \cdots p_{nn})
\end{aligned}$$

A definite symmetric dyad is regular. The determinant of a positive definite dyad is positive, the determinant of a negative definite dyad has the sign $(-1)^n$. A semidefinite symmetric dyad is singular.

Antisymmetric dyads : A dyad A is said to be antisymmetric if its coordinates satisfy one of the following equivalent conditions :

$$\begin{array}{ccccc}
a_{im} = -a_{mi} & \Leftrightarrow & a^{im} = -a^{mi} & \Leftrightarrow & a_{i\,.}^{\;m} = -a^{m}_{\;.\,i} \\
\mathbf{A}_* = -(\mathbf{A}_*)^T & \Leftrightarrow & \mathbf{A}^* = -(\mathbf{A}^*)^T & \Leftrightarrow & \mathbf{A}_o = (\mathbf{A}^o)^T
\end{array}$$

Dyads whose mixed coordinates satisfy $a_{i\,.}^{\;m} = -a_{\;.\,i}^{m}$ are antisymmetric. By contrast, dyads with the properties $a_{i\,.}^{\;m} = -a_{m\,.}^{\;\;i}$ and $a_{\;.\,m}^{i} = -a_{\;.\,i}^{m}$ are generally not antisymmetric.

$$a_{i\,.}^{\;m} = -a_{\;.\,i}^{m} \quad \Rightarrow \quad a_{im} = g_{is}\, a_{\;.\,m}^{s} = -g_{is}\, a_{m\,.}^{\;\;s} = -a_{mi}$$

$$a_{\;.\,m}^{i} = -a_{\;.\,i}^{m} \quad \Rightarrow \quad a_{im} = g_{is}\, a_{\;.\,m}^{s} = -\sum_{s} g_{is}\, a_{\;.\,s}^{m} \neq -a_{mi}$$

The diagonal elements of the covariant coordinate matrix $\mathbf{A}_*$ and of the contravariant coordinate matrix $\mathbf{A}^*$ of an antisymmetric dyad are zero. By contrast, the diagonal elements of the mixed coordinate matrices $\mathbf{A}_{\mathrm{o}}$ and $\mathbf{A}^{\mathrm{o}}$ of an antisymmetric dyad are not zero. However, the sum of these diagonal elements is zero.

$$a_{ii} = -a_{ii} \quad \Rightarrow \quad a_{ii} = 0$$

$$a^{ii} = -a^{ii} \quad \Rightarrow \quad a^{ii} = 0$$

$$a_{i\,.}^{\;m} = -a_{\;.\,i}^{m} \quad \Rightarrow \quad a_{i\,.}^{\;i} = g_{ir}\, g^{is}\, a_{\;.\,s}^{r} = -g_{ir} g^{is}\, a_{s\,.}^{\;r}$$

$$= -\delta_{r}^{s}\; a_{s\,.}^{\;r} = -a_{s\,.}^{\;s} = 0$$

Quadric of an antisymmetric dyad : The quadric of an antisymmetric dyad **A** is zero for every tensor **x** of the space $\mathbb{R}^n$.

$$a^{im}\, x_i\, x_m = a^{mi}\, x_m\, x_i = -a^{im}\, x_i\, x_m \quad \Rightarrow \quad a^{im}\, x_i\, x_m = 0$$

$$\mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x} = 0$$

Symmetric and antisymmetric components of a dyad : Every dyad D in the euclidean space $\mathbb{R}^n$ is the sum of a symmetric dyad S and an antisymmetric dyad A. Since the quadric of the antisymmetric dyad A is identically zero, the definiteness of the dyad D is determined only by its symmetric component S.

$$d_{im} = s_{im} + a_{im} = s_{mi} - a_{mi}$$

$$s_{im} = 0.5\,(d_{im} + d_{mi})$$

$$a_{im} = 0.5\,(d_{im} - d_{mi})$$

$$\mathbf{x} \cdot \mathbf{A} \cdot \mathbf{x} = 0 \quad \Rightarrow \quad \mathbf{x} \cdot \mathbf{D} \cdot \mathbf{x} = \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x}$$

Associated tensors of an antisymmetric dyad : The associated tensors **A** • **x** and **x** • **A** of an antisymmetric dyad A differ only in sign. The tensor **x** is orthogonal to the tensors **A** • **x** and **x** • **A** associated with **A** and **x**.

$$\mathbf{A} \cdot \mathbf{x} = a_{im} x^m = -a_{mi} x^m = -\,\mathbf{x} \cdot \mathbf{A}$$

$$\mathbf{x} \cdot (\mathbf{A} \cdot \mathbf{x}) = (\mathbf{x} \cdot \mathbf{A}) \cdot \mathbf{x} = 0$$

Scalar product of a symmetric and an antisymmetric dyad : The scalar product of a symmetric dyad S and an antisymmetric dyad A in the euclidean space $\mathbb{R}^n$ is identically zero.

$$s_{im}\, a^{im} = -s_{mi} a^{mi} = -s_{im} a^{im} = 0$$

$$s_{mi}\, a^{im} = -s_{im} a^{mi} = -s_{mi} a^{im} = 0$$

$$\mathbf{S} : \mathbf{A} = \mathbf{S} \cdot\cdot\, \mathbf{A} = 0$$

Polar decomposition of a dyad : Every regular dyad D in the euclidean space $\mathbb{R}^n$ is the inner product of a unitary dyad C and a positive definite symmetric dyad S. The following relationships hold for the coordinates of the dyads and the coordinates of an arbitrary tensor $\mathbf{x} \neq \mathbf{0}$ of rank 1 :

$$\begin{array}{lll} \mathbf{D}_{o} = \mathbf{C}_{*}\, \mathbf{S}^{*} & : & d_{i.}^{m} = c_{ik}\, s^{km} \\ \mathbf{D}^{o} = \mathbf{C}^{*}\, \mathbf{S}_{*} & : & d^{i}_{.m} = c^{ik}\, s_{km} \\ \mathbf{C}_{*}^{T}\, \mathbf{C}^{*} = \mathbf{I} & : & c_{ik}\, c^{km} = \delta_{i}^{k} \\ \mathbf{S}_{*} = \mathbf{S}_{*}^{T} & : & s_{im} = s_{mi} \\ \mathbf{x} \cdot \mathbf{S} \cdot \mathbf{x} > 0 & : & s_{im}\, x^{i}\, x^{m} > 0 \end{array}$$

For a given dyad D, the coordinates of the dyads C and S are uniquely determined. They are calculated in the following steps :

1. Determine the dyad $\mathbf{T}^{o} = \mathbf{D}_{o}^{T}\, \mathbf{D}^{o}$.
2. Decompose the dyad $\mathbf{T}^{o} = \mathbf{U}^{*}\, \mathbf{P}\, \mathbf{U}_{*}^{T}$.
3. Determine the dyad $\mathbf{S}_{*} = \mathbf{U}_{*}\, \mathbf{P}^{0.5}\, (\mathbf{U}_{*})^{T}$.
4. Determine the dyad $\mathbf{S}_{*}^{-1} = \mathbf{U}^{*}\, \mathbf{P}^{-0.5}\, (\mathbf{U}^{*})^{T}$.
5. Determine the dyad $\mathbf{C}^{*} = \mathbf{D}^{o}\, \mathbf{S}_{*}^{-1}$.

Proof : Polar decomposition of a symmetric dyad

It is to be shown that the dyads S and C exist for every regular dyad D, since otherwise the specified calculational steps cannot be carried out.

1. The coordinates of the dyad T are $t^{i}_{.m} = d_{k.}^{\ i}\, d^{k}_{.m}$. The dyad T is symmetric, that is $t^{i}_{.m} = t_{m.}^{\ i}$:

$$\begin{aligned} t_{m.}^{\ i} = g_{mr}\, g^{is}\, t^{r}_{.s} &= g_{mr}\, g^{is}\, d_{k.}^{\ r}\, d^{k}_{.s} = d_{km}\, d^{ki} \\ &= g_{kr}\, g^{ks}\, d^{r}_{.m}\, d_{s.}^{\ i} = d^{r}_{.m}\, d_{r.}^{\ i} = t^{i}_{.m} \end{aligned}$$

2. The dyad T is positive definite since, with the tensor $\mathbf{x} \neq \mathbf{0}$ and with the tensor $\mathbf{w} = \mathbf{D} \cdot \mathbf{x} \neq \mathbf{0}$ associated with the regular dyad D, the quadric of T satisfies :

$$t^{i}_{.m}\, x_{i}\, x^{m} = d_{k.}^{\ i}\, d^{k}_{.m}\, x_{i}\, x^{m} = (d_{k.}^{\ i}\, x_{i})(d^{k}_{.m}\, x^{m}) = w_{k}\, w^{k} > 0$$

3. The positive eigenvalues of the symmetric positive definite dyad T are arranged in the diagonal matrix **P**. The covariant coordinates of the eigenvectors are arranged in columns in the matrix $\mathbf{U}_*$, the contravariant coordinates in the matrix $\mathbf{U}^*$.

$$t^{i}_{.m} = u^{ir}\, p_{r.}^{\;r}\, u_{mr}$$

4. Since the eigenvalues of T are positive, it is possible to form the diagonal matrix $\mathbf{P}^{0.5}$ whose diagonal elements are the positive square roots of the diagonal elements of **P**. The dyad $\mathbf{S}^{o} = \mathbf{U}^* \mathbf{P}^{0.5} \mathbf{U}_*^{T}$ is symmetric :

$$s^{i}_{.m} = u^{ik}\,(p_{k.}^{\;k})^{0.5}\, u_{mk} = u^{ik}\,(p^{k}_{.k})^{0.5}\, u_{mk} = s_{m.}^{\;i}$$

5. The covariant coordinate matrix $\mathbf{S}_*$ and the contravariant coordinate matrix $\mathbf{S}^*$ of the dyad S are determined using the rules for dual indices :

$$\mathbf{S}_* = \mathbf{G}_* \mathbf{S}^{o} = \mathbf{G}_* \mathbf{U}^* \mathbf{P}^{0.5} (\mathbf{U}_*)^{T} = \mathbf{U}_* \mathbf{P}^{0.5} (\mathbf{U}_*)^{T}$$

$$\mathbf{S}^* = \mathbf{G}^* \mathbf{S}_{o} = \mathbf{G}^* \mathbf{U}_* \mathbf{P}^{0.5} (\mathbf{U}^*)^{T} = \mathbf{U}^* \mathbf{P}^{0.5} (\mathbf{U}^*)^{T}$$

6. The symmetric dyad T satisfies $\mathbf{S}_{o}^{T} \mathbf{S}^{o} = \mathbf{T}^{o} = \mathbf{D}_{o}^{T} \mathbf{D}^{o}$:

$$\begin{aligned} s_{k.}^{\;i}\; s^{k}_{.m} &= u_{kr}\,(p_{r.}^{\;r})^{0.5}\, u^{ir} u^{ks}\,(p_{s.}^{\;s})^{0.5}\, u_{ms} \\ &= u^{ir}\,(p_{r.}^{\;r}\, p_{s.}^{\;s})^{0.5}\, u_{ms}\, \delta_{r}^{s} \\ &= u^{ir}\, p_{r.}^{\;r}\, u_{mr} = t^{i}_{.m} \end{aligned}$$

7. The inverses of the matrices $\mathbf{S}_{o}$ and $\mathbf{S}^{o}$ are determined using the equations $(\mathbf{U}_*)^{T}\mathbf{U}^* = \mathbf{I}$ and $(\mathbf{U}^*)^{T}\mathbf{U}_* = \mathbf{I}$:

$$(\mathbf{S}^{o})^{-1} = (\mathbf{U}^* \mathbf{P}^{0.5} (\mathbf{U}_*)^{T})^{-1} = \mathbf{U}^* \mathbf{P}^{-0.5} (\mathbf{U}_*)^{T}$$

$$(\mathbf{S}_{o})^{-1} = (\mathbf{U}_* \mathbf{P}^{0.5} (\mathbf{U}^*)^{T})^{-1} = \mathbf{U}_* \mathbf{P}^{-0.5} (\mathbf{U}^*)^{T}$$

8. The covariant coordinate matrix $\mathbf{S}_*^{-1}$ and the contravariant coordinate matrix $(\mathbf{S}^*)^{-1}$ of the dyad S^{-1} are determined using the rules for dual indices :

$$(\mathbf{S}_*)^{-1} = (\mathbf{G}_* \mathbf{S}^{o})^{-1} = \mathbf{U}^* \mathbf{P}^{-0.5} (\mathbf{U}_*)^{T} \mathbf{G}^* = \mathbf{U}^* \mathbf{P}^{-0.5} (\mathbf{U}^*)^{T}$$

$$(\mathbf{S}^*)^{-1} = (\mathbf{G}^* \mathbf{S}_{o})^{-1} = \mathbf{U}_* \mathbf{P}^{-0.5} (\mathbf{U}^*)^{T} \mathbf{G}_* = \mathbf{U}_* \mathbf{P}^{-0.5} (\mathbf{U}_*)^{T}$$

9. The dyad C is defined by the equation $\mathbf{D}^{o} = \mathbf{C}^* \mathbf{S}_*$. The equation $\mathbf{D}_{o} = \mathbf{C}_* \mathbf{S}^*$ is derived from this definition.

$$d^{i}_{.m} = c^{ik}\, s_{km}$$

$$\begin{aligned} d_{i.}^{\;m} &= g_{ir}\, g^{mt}\, d^{r}_{.t} = g_{ir}\, g^{mt}\, c^{rk}\, s_{kt} = c_{i.}^{\;k}\; s_{k.}^{\;m} \\ &= g^{kr}\, g_{kt}\, c_{ir}\, s^{tm} = \delta_{t}^{r}\, c_{ir}\, s^{tm} = c_{ir}\, s^{rm} \end{aligned}$$

10. The dyad C with $\mathbf{C}^* = \mathbf{D}^{o}\,\mathbf{S}_*^{-1}$ and $\mathbf{C}_* = \mathbf{D}_{o}\,(\mathbf{S}^*)^{-1}$ is unitary:

$$\begin{aligned}
\mathbf{C}_*^{T}\,\mathbf{C}^* &= ((\mathbf{S}^*)^{-1})^{T}\,\mathbf{D}_{o}^{T}\,\mathbf{D}^{o}\,(\mathbf{S}_*)^{-1} \\
&= ((\mathbf{S}_{o})^{-1}\mathbf{G}_*)^{T}\,\mathbf{S}_{o}^{T}\,\mathbf{S}^{o}\,(\mathbf{S}^{o})^{-1}\mathbf{G}^* \\
&= \mathbf{G}_*\,(\mathbf{S}_{o}\,\mathbf{S}_{o}^{-1})^{T}\,(\mathbf{S}^{o}(\mathbf{S}^{o})^{-1})\mathbf{G}^* \\
\mathbf{C}_*^{T}\,\mathbf{C}^* &= \mathbf{I}
\end{aligned}$$

It remains to be shown that the decomposition of the regular dyad D into the unitary dyad C and the positive definite symmetric dyad S is unique. Let an arbitrary positive definite symmetric dyad S with the coordinate matrix $\mathbf{S}_{o}$ be given. Let the eigenvalues w_i of $\mathbf{S}_{o}$ be pairwise different. They are arranged in increasing order in the diagonal matrix $\mathbf{W}$; the eigenvectors $\mathbf{z}_i$ are arranged in the same order in the eigenmatrix $\mathbf{Z}_*$. Then the matrix $\mathbf{S}_{o}$ may be uniquely decomposed into the products $\mathbf{Z}_*\mathbf{W}(\mathbf{Z}^*)^{T} = \mathbf{Z}^*\mathbf{W}\,\mathbf{Z}_*^{T}$.

The matrix $\mathbf{S}_{o}$ is suitable for a polar decomposition of the dyad D if the condition $\mathbf{T}_{o} = \mathbf{D}_{o}^{T}\,\mathbf{D}_{o} = \mathbf{S}_{o}^{T}\,\mathbf{S}_{o} = \mathbf{Z}^*\,\mathbf{W}^{T}\mathbf{W}\,\mathbf{Z}_*^{T}$ is satisfied. Let the eigenvalues p_i of $\mathbf{T}_{o}$ be pairwise different. They are arranged in increasing order in the diagonal matrix $\mathbf{P}$; the eigenvectors $\mathbf{u}_i$ are arranged in the same order in the eigenmatrix $\mathbf{U}_*$. Then the matrix $\mathbf{T}_{o}$ may be uniquely decomposed into the product $\mathbf{U}^*\mathbf{P}\,\mathbf{U}_*^{T}$. Since the matrices $\mathbf{S}_{o}$ and $\mathbf{T}_{o}$ are real, symmetric and positive definite, they have complete sets of real eigenstates with positive eigenvalues. Hence $\mathbf{S}_{o}$ is suitable for a polar decomposition of the dyad D only if $\sqrt{w_i} = p_i$ and $\mathbf{Z}^* = \mathbf{U}^*$. If $\mathbf{S}_{o}$ has multiple eigenvalues, the same result is obtained by considering eigenspaces instead of eigenvectors for the multiple eigenvalues.

Trace of a dyad : The sum of the mixed diagonal coordinates of a dyad T is called the trace of the dyad T and is designated by tr **T**. The trace is a scalar invariant of T : If the dual bases $\mathbf{B}_*, \mathbf{B}^*$ of $\mathbb{R}^n$ are transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_*\,\mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^*\bar{\mathbf{A}}^{T}$, tr **T** remains invariant.

$$\begin{aligned}
\mathrm{tr}\,\mathbf{T} &= t_{i\,.}^{\;i} = t^{\,i}_{.\,i} \\
\bar{t}_{i\,.}^{\;i} &= a^{\,r}_{.\,i}\,\bar{a}^{\,i}_{.\,s}\,t_{r\,.}^{\;s} = \delta^{r}_{s}\,t_{r\,.}^{\;s} = t_{s\,.}^{\;s} \\
t_{i\,.}^{\;i} &= g_{ir}\,g^{is}\,t^{\,r}_{.\,s} = \delta^{s}_{r}\,t^{\,r}_{.\,s} = t^{\,s}_{.\,s}
\end{aligned}$$

Spherical dyad : A dyad K in the space $\mathbb{R}^n$ is called a spherical dyad if its n mixed diagonal coordinates have the same value and its remaining mixed coordinates are zero.

$$k_{i\,.}^{\;m} = k^{\,i}_{.\,m} = k\,\delta^{i}_{m}$$

Deviatorial decomposition of a dyad : A dyad D is called a deviator if its trace is zero. Every dyad T in $\mathbb{R}^n$ may be decomposed into a deviator D and a spherical dyad K with the diagonal coordinates $\frac{1}{n}$ tr $\mathbf{T}$:

$$\text{tr}\,\mathbf{T} = t_{i.}^{\,i}$$

$$d_{i.}^{\,m} = t_{i.}^{\,m} - \frac{1}{n}\,\delta_i^m\,\text{tr}\,\mathbf{T}$$

Proof : Deviatorial decomposition of a dyad

$$d_{i.}^{\,i} = t_{i.}^{\,i} - \frac{1}{n}\,\delta_i^i\,\text{tr}\,\mathbf{T} = \text{tr}\,\mathbf{T} - \text{tr}\,\mathbf{T} = 0$$

Example 1 : Polar decomposition of a dyad

Let two dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ as well as the mixed coordinate matrix $\mathbf{D}_o$ of a dyad D in the euclidean space $\mathbb{R}^2$ be given. The polar decomposition of the dyad D is to be determined. For this purpose the coordinate matrix $\mathbf{D}^o$ of D and the coordinate matrix $\mathbf{T}^o = (\mathbf{D}_o)^T\mathbf{D}^o$ of the dyad T are calculated.

$$\mathbf{B}_* = \begin{array}{|c|c|}\hline 1.0000 & -1.0000 \\ \hline 0 & 1.0000 \\ \hline \end{array} \qquad \mathbf{G}_* = (\mathbf{B}_*)^T\,\mathbf{B}_* = \begin{array}{|c|c|}\hline 1.0000 & -1.0000 \\ \hline -1.0000 & 2.0000 \\ \hline \end{array}$$

$$\mathbf{B}^* = \begin{array}{|c|c|}\hline 1.0000 & 0 \\ \hline 1.0000 & 1.0000 \\ \hline \end{array} \qquad \mathbf{G}^* = (\mathbf{B}^*)^T\,\mathbf{B}^* = \begin{array}{|c|c|}\hline 2.0000 & 1.0000 \\ \hline 1.0000 & 1.0000 \\ \hline \end{array}$$

$$\mathbf{D}_o = \begin{array}{|c|c|}\hline 2.0000 & -0.5000 \\ \hline 0.5000 & 1.0000 \\ \hline \end{array} \qquad \mathbf{D}^o = \mathbf{G}^*\mathbf{D}_o\mathbf{G}_*^T = \begin{array}{|c|c|}\hline 4.5000 & -4.5000 \\ \hline 2.0000 & -1.5000 \\ \hline \end{array}$$

$$\mathbf{T}^o = \begin{array}{|c|c|}\hline 10.0000 & -9.7500 \\ \hline -0.2500 & 0.7500 \\ \hline \end{array}$$

The real symmetric dyad T has the real eigenvalues p_1 and p_2. One of the contravariant coordinates of each of the eigenvectors $\mathbf{u}_1^*$ and $\mathbf{u}_2^*$ is arbitrarily given the value 1.0. These eigenvectors are transformed to the canonical basis and to the covariant basis $\mathbf{B}_*$.

$$(10.00 - p)\,(0.75 - p) - 9.75 * 0.25 = 0 \quad \Rightarrow \quad \begin{aligned} p_1 &= 0.4936 \\ p_2 &= 10.2564 \end{aligned}$$

$$\mathbf{T}^{o}\mathbf{u}_i^{*} = p_i\mathbf{u}_i^{*} \quad \Rightarrow \quad \mathbf{u}_1^{*} = \begin{array}{|c|}\hline 1.0256 \\ \hline 1.0000 \\ \hline\end{array} \qquad \mathbf{u}_2^{*} = \begin{array}{|c|}\hline 1.0000 \\ \hline -0.0263 \\ \hline\end{array}$$

$$\mathbf{u}_i = \mathbf{B}_{*}\mathbf{u}_i^{*} \quad \Rightarrow \quad \mathbf{u}_1 = \begin{array}{|c|}\hline 0.0256 \\ \hline 1.0000 \\ \hline\end{array} \qquad \mathbf{u}_2 = \begin{array}{|c|}\hline 1.0263 \\ \hline -0.0263 \\ \hline\end{array}$$

$$\mathbf{u}_{*i} = \mathbf{G}_{*}\mathbf{u}_i^{*} \quad \Rightarrow \quad \mathbf{u}_{*1} = \begin{array}{|c|}\hline 0.0256 \\ \hline 0.9744 \\ \hline\end{array} \qquad \mathbf{u}_{*2} = \begin{array}{|c|}\hline 1.0263 \\ \hline -1.0526 \\ \hline\end{array}$$

The eigenvectors are scaled such that their magnitude is 1, and they are arranged in the coordinate matrices $\mathbf{U}^{*}$ and $\mathbf{U}_{*}$. Then the matrices $\mathbf{S}_{*}$, $\mathbf{S}_{*}^{-1}$, $\mathbf{C}^{*}$ and $\mathbf{C}_{*}$ are calculated.

$$\mathbf{P} = \begin{array}{|c|c|}\hline 0.4936 & 0 \\ \hline 0 & 10.2564 \\ \hline\end{array} \qquad \mathbf{U}^{*} = \begin{array}{|c|c|}\hline 1.0253 & 0.9740 \\ \hline 0.9997 & -0.0256 \\ \hline\end{array} \qquad \mathbf{U}_{*} = \begin{array}{|c|c|}\hline 0.0256 & 0.9997 \\ \hline 0.9741 & -1.0253 \\ \hline\end{array}$$

$$\mathbf{S}_{*} = \mathbf{U}_{*}\,\mathbf{P}^{0.5}\,(\mathbf{U}_{*})^{T} = \begin{array}{|c|c|}\hline 3.2011 & -3.2651 \\ \hline -3.2651 & 4.0333 \\ \hline\end{array}$$

$$\mathbf{S}_{*}^{-1} = \mathbf{U}^{*}\,\mathbf{P}^{-0.5}(\mathbf{U}^{*})^{T} = \begin{array}{|c|c|}\hline 1.7926 & 1.4512 \\ \hline 1.4512 & 1.4228 \\ \hline\end{array}$$

$$\mathbf{C}^{*} = \mathbf{D}^{o}\,\mathbf{S}_{*}^{-1} = \begin{array}{|c|c|}\hline 1.5365 & 0.1280 \\ \hline 1.4084 & 0.7682 \\ \hline\end{array}$$

$$\mathbf{C}_{*} = \mathbf{G}_{*}\,\mathbf{C}^{*}\,(\mathbf{G}_{*})^{T} = \begin{array}{|c|c|}\hline 0.7682 & -1.4084 \\ \hline -0.1280 & 1.5365 \\ \hline\end{array}$$

The dyad U is unitary, since $\mathbf{C}_{*}^{T}\,\mathbf{C}^{*} = \mathbf{I}$. Thus the given dyad D has been decomposed into the product $\mathbf{D}^{o} = \mathbf{C}^{*}\,\mathbf{S}_{*}$ of the unitary dyad C and the symmetric positive definite dyad S.

Example 2 : Deviator of a dyad

Let dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ and the contravariant coordinate matrix $\mathbf{T}^*$ of a dyad T in the euclidean space $\mathbb{R}^2$ be given. The deviator D of the dyad T is to be determined. For this purpose the mixed coordinate matrix $\mathbf{T}_o = \mathbf{G}_* \mathbf{T}^*$ and the trace of T are calculated.

$$\mathbf{B}_* = \begin{array}{|c|c|}\hline 1 & -1 \\ \hline 0 & 1 \\ \hline\end{array} \qquad \mathbf{B}^* = \begin{array}{|c|c|}\hline 1 & 0 \\ \hline 1 & 1 \\ \hline\end{array} \qquad \mathbf{T}^* = \begin{array}{|c|c|}\hline 4 & 1 \\ \hline 1 & 5 \\ \hline\end{array}$$

$$\mathbf{G}_* = \begin{array}{|c|c|}\hline 1 & -1 \\ \hline -1 & 2 \\ \hline\end{array} \qquad \mathbf{T}_o = \begin{array}{|c|c|}\hline 3 & -4 \\ \hline -2 & 9 \\ \hline\end{array} \qquad \mathrm{tr}\,\mathbf{T} = 3 + 9 = 12$$

The mixed coordinate matrix $\mathbf{D}_o$ of the deviator D of T is $\mathbf{D}_o = \mathbf{T}_o - \frac{1}{2}(\mathrm{tr}\,\mathbf{T})\,\mathbf{I}$. The coordinates $d_{i\,.}^{\,m}$ of $\mathbf{D}_o$ in the bases $\mathbf{B}_*$ and $\mathbf{B}^*$ are referred to the canonical basis $\mathbf{E}$ using the transformation $\mathbf{D} = \mathbf{B}^* \mathbf{D}_o (\mathbf{B}_*)^{\mathrm{T}}$:

$$\mathbf{D}_o = \begin{array}{|c|c|}\hline -3 & -4 \\ \hline -2 & 3 \\ \hline\end{array} \qquad \mathbf{D} = \begin{array}{|c|c|}\hline 1 & -4 \\ \hline -4 & -1 \\ \hline\end{array}$$

9.3.8 TENSOR MAPPINGS

Introduction : A physical quantity, for instance the state of stress at a point in a body, may take different values over time. Each of these states of stress corresponds to a linear vector function $T_k(\mathbf{u}_1, ..., \mathbf{u}_m)$, and thus to a tensor in the sense of the tensor definition. This tensor is completely determined by its coordinate tuple in a basis system. The set of coordinate tuples for the different values of the physical quantity forms the coordinate set of the physical quantity.

The coordinate set for all values of a physical quantity is often called a tensor, although by definition it is the function $T_k(\mathbf{u}_1, ..., \mathbf{u}_m)$ for each individual value which is a tensor. This abbreviating diction is also used in the following. The coordinate sets of one or several tensors may be mapped to the coordinate sets of other tensors. Such mappings are used in particular to formulate physical problems. For example, the strain tensor and the material tensor are mapped to the stress tensor.

The coordinates of a tensor $T_k(\mathbf{u}_1, ..., \mathbf{u}_m)$ generally depend on the choice of a basis system. Accordingly, the result of a tensor mapping generally depends on this choice. However, there are tensor mappings whose value is independent of the choice of basis. For example, the eigenvalues of a symmetric dyad are invariant under transformations of the basis. Such invariants contain essential information about physical quantities. It turns out that different invariants of a tensor are dependent, so that a basis of invariants can be chosen. This allows the invariants of the tensor to be expressed as functions of such a basis of invariants.

Basis system : Every value of a tensor T_k in the space $\mathbb{R}^n$ is the scalar value of a linear vector function $T_k(\mathbf{u}_1, ..., \mathbf{u}_m)$ for given vectors $\mathbf{u}_1, ..., \mathbf{u}_m \in \mathbb{R}^n$. The coordinates $t_k^{i_1 ... i_m}$ of the tensor are uniquely determined if each index i_s of the coordinates is associated with a basis B^s. The cartesian product $B^1 \times ... \times B^m$ is called the basis system of the tensor coordinates.

$$t_k^{i_1 ... i_m} = T_k(\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_m}) \qquad \wedge \qquad \mathbf{b}^{i_s} \in B^s$$

Coordinate tuple : An n-dimensional tensor T_k of rank m is completely defined by its n^m coordinates in a basis system $B^1 \times ... \times B^m$. These n^m coordinates are called the coordinate tuple of the tensor T_k for the basis system $B^1 \times ... \times B^m$. The coordinate tuple is designated by $\mathbf{T}_k$. If the coordinate tuple $\mathbf{T}_k$ for a basis system is known, the function $T_k(...)$ is uniquely determined. The value of the tensor T_k for arbitrary vectors $\mathbf{u}_1, ..., \mathbf{u}_m$ is a linear combination of the coordinate tuple.

$$\mathbf{T}_k = \{t_k^{i_1 ... i_m} = T_k(\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_m}) \mid (\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_m}) \in B^1 \times ... \times B^m\}$$

$$\mathbf{u}_{i_s} = c_{i_s} \mathbf{b}^{i_s}$$

$$\mathbf{T}_k(\mathbf{u}_{i_1}, ..., \mathbf{u}_{i_m}) = c_{i_1} ... c_{i_m} \, t_k^{i_1 ... i_m}$$

Coordinate set : The k-th state of a physical quantity corresponds to a tensor $T_k(...)$ with the coordinate tuple $\mathbf{T}_k$. The set of coordinate tuples of the physical quantity for different states is called the coordinate set of the physical quantity and is designated by **T**. Often **T** is called a tensor which represents the physical quantity.

$$\mathbf{T} = \{\mathbf{T}_k \mid k = 1,2,...\}$$

Tensor mapping : A mapping is called a tensor mapping (tensor function) if its domain is the cartesian product of the coordinate sets of tensors. The tensor mapping is said to be scalar-valued if the target is a set of scalars (numbers). The tensor mapping is said to be tensor-valued if the target is the cartesian product of the coordinate sets of tensors. The tensors of a cartesian product may be of different rank. The number of factors in the cartesian product may be different for the domain and the target.

$$f : \mathbf{A} \times ... \times \mathbf{C} \rightarrow \mathbf{T} \times ... \times \mathbf{W}$$

f	name of the tensor function
A,...,**C**	coordinate sets of the tensors of the domain
T,...,**W**	coordinate sets of the tensors of the target

Transformation of the basis system : A fixed physical quantity, for example the state of stress at a point in a body, may be described in different basis systems, for instance $B^1 \times ... \times B^m$ and $\bar{B}^1 \times ... \times \bar{B}^m$. The vector function $T_k(...)$ which defines this tensor is not changed by the choice of basis system. However, the coordinate tuples $\mathbf{T}_k$ and $\bar{\mathbf{T}}_k$ of the tensor for the two basis systems are different, since the function $T_k(...)$ is applied to different basis vectors .

$$\mathbf{T}_k = \{t_k^{i_1 ... i_m} = T_k(\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_m}) \mid (\mathbf{b}^{i_1}, ..., \mathbf{b}^{i_m}) \in B^1 \times ... \times B^m\}$$

$$\bar{\mathbf{T}}_k = \{\bar{t}_k^{i_1 ... i_m} = T_k(\bar{\mathbf{b}}^{i_1}, ..., \bar{\mathbf{b}}^{i_m}) \mid (\bar{\mathbf{b}}^{i_1}, ..., \bar{\mathbf{b}}^{i_m}) \in \bar{B}^1 \times ... \times \bar{B}^m\}$$

Scalar invariants of a tensor : Let a scalar tensor function $f : \mathbf{T} \rightarrow \mathbb{R}$ with $f(\mathbf{T}_k) = c$ be defined for a fixed physical quantity described by the tensor $T_k(...)$. Thus the value of the scalar c depends on all coordinates of the set $\mathbf{T}_k$. This is often expressed as $f(t_k^{i_1 ... i_m}) = c$. The domain of the mapping contains every coordinate tuple which may be obtained by transforming the basis system of the tensor. If transforming the basis system does not change the value c of the function, the function value c is called a (scalar) invariant of the tensor T_k. Thus the scalar tensor function has the same value c for every coordinate tuple $\mathbf{T}_k$ of its domain **T**.

$$f(\mathbf{T}_k) = f(\bar{\mathbf{T}}_k) = c \qquad \mathbf{T}_k, \bar{\mathbf{T}}_k \in \mathbf{T}$$

As a special case, let all indices of the coordinates $t_{i_1 \dots i_m}$ of a tensor T_k be referred to the same basis $\mathbf{B}_*$. The corresponding basis system is $\mathbf{B}_* \times \dots \times \mathbf{B}_*$ (m-fold). The basis is transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$; the corresponding basis system is $\bar{\mathbf{B}}_* \times \dots \times \bar{\mathbf{B}}_*$ (m-fold). Let the coefficients of the transformation matrix $\mathbf{A}$ be $a^{s}_{.\,i}$. Then the tensor function $f : \mathbf{T} \to \mathbb{R}$ is an invariant if its value is the same for the coordinate tuples $\mathbf{T}_k$ and $\bar{\mathbf{T}}_k$:

$$f(t_{i_1 \dots i_m}) = \bar{f}(t_{i_1 \dots i_m}) = f(a^{s_1}_{.\,i_1} \dots a^{s_m}_{.\,i_m} \; t_{s_1 \dots s_m})$$

Scalar invariants of several tensors : Let a scalar tensor function $f : \mathbf{T} \times \dots \times \mathbf{W} \to \mathbb{R}$ with $f(\mathbf{T}_j, \dots, \mathbf{W}_k) = c$ be defined for several fixed physical quantities described by the tensors $T_j(\dots), \dots, W_k(\dots)$. Each of the tensors has its own basis system. The domain of the mapping f contains every ordered tuple $(\mathbf{T}_j, \dots, \mathbf{W}_k)$ which may be obtained by transforming the basis systems of the tensors. The function value c is called a (scalar) invariant if transforming the coordinate tuples does not change the value c of the function.

$$f : \mathbf{T} \times \dots \times \mathbf{W} \to \mathbb{R}$$

$$f(\mathbf{T}_j, \dots, \mathbf{W}_k) = f(\bar{\mathbf{T}}_j, \dots, \bar{\mathbf{W}}_k) = c \qquad \mathbf{T}_j \in \mathbf{T}, \quad \mathbf{W}_k \in \mathbf{W}$$

Invariance of the eigenvalues of a symmetric dyad : Let the coordinates $t_{i.}^{m}$ of a symmetric dyad in the dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of the space $\mathbb{R}^n$ be given. This dyad has n real eigenvalues $p_1, \dots, p_n$ with corresponding eigenvectors $\mathbf{u}_1, \dots, \mathbf{u}_n$. An eigenstate $(p, \mathbf{u})$ satisfies :

$$(t_{i.}^{m} - p\, \delta_i^m) u_m = 0$$

The bases are transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^T$ using the transformation matrix $\mathbf{A}$. The coordinates $\bar{t}_{i.}^{m}$ of the dyad for the dual bases $\bar{\mathbf{B}}_*$ and $\bar{\mathbf{B}}^*$ and the coordinates $\bar{u}_m$ of the eigenvector $\mathbf{u}$ are transformed according to the general rules :

$$\bar{t}_{r.}^{s} = a^{i}_{.\,r}\, \bar{a}^{s}_{.\,m}\, t_{i.}^{m}$$

$$\bar{\delta}_r^s = a^{i}_{.\,r}\, \bar{a}^{s}_{.\,m}\, \delta_i^m$$

$$\bar{u}_s = a^{k}_{.\,s}\, u_k$$

An eigenstate $(\bar{p}, \bar{\mathbf{u}})$ of the tensor T in the transformed basis satisfies :

$$(\bar{t}_{r.}^{s} - \bar{p}\, \bar{\delta}_r^s)\, \bar{u}_s = 0$$

$$a^{i}_{.\,r}\, (t_{i.}^{m} - \bar{p}\, \delta_i^m)\, \bar{a}^{s}_{.\,m}\, a^{k}_{.\,s}\, u_k = 0$$

$$\bar{a}^{r}_{.\,j}\, a^{i}_{.\,r}\, (t_{i.}^{m} - \bar{p}\, \delta_i^m)\, \delta_m^k\, u_k = 0$$

$$(t_{i.}^{m} - \bar{p}\, \delta_i^m)\, u_m = 0$$

But $(t_{i.}^{\ m} - p\,\delta_i^m)\, u_m = (t_{i.}^{\ m} - \bar{p}\,\delta_i^m)\, u_m = 0$ implies $p = \bar{p}$. Hence the eigenvalues of the symmetric dyad $\mathbf{T}_o$ are invariant under transformations of the basis of the coordinates of the dyad. If the eigenvalues are now ordered algebraically, for example $p_1 \le p_2 \le \ldots \le p_n$, the following scalar invariants are obtained for the tensor T :

$$\begin{aligned} f_1(\mathbf{T}_o) &= p_1 \\ &\vdots \\ f_n(\mathbf{T}_o) &= p_n \end{aligned}$$

Principal invariants of a symmetric dyad : Let the coordinates of a symmetric dyad $\mathbf{T}_o$ in dual bases $\mathbf{B}_*$ of the space $\mathbb{R}^n$ be $t_{i.}^{\ m}$ with i, m = 1,...,n. The eigenvalue problem for this dyad is shown in matrix form :

$$\begin{bmatrix} t_{1.}^{\ 1} & t_{1.}^{\ 2} & & t_{1.}^{\ n} \\ t_{2.}^{\ 1} & t_{2.}^{\ 2} & & t_{2.}^{\ n} \\ & & \ddots & \\ t_{n.}^{\ 1} & t_{n.}^{\ 2} & & t_{n.}^{\ n} \end{bmatrix} * \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = p \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}$$

The condition $\det(\mathbf{T}_o - p\mathbf{I}) = 0$ for non-trivial solutions $\mathbf{u}$ leads to the characteristic equation of the dyad. Its coefficients are designated by $I_1, I_2, \ldots, I_n$:

$$(-p)^n + I_1\,(-p)^{n-1} + \ldots + I_{n-1}\,(-p)^1 + I_n = 0$$

For the real symmetric dyad $\mathbf{T}_o$, the solutions $p_1 \le p_2 \le \ldots \le p_n$ of the characteristic equation (the eigenvalues) and the corresponding eigenvectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ are real. In the principal basis, the coordinate matrix of the dyad T is a diagonal matrix with the eigenvalues as diagonal elements. In this representation the coefficients of the characteristic equation take an especially simple form.

$$\mathbf{T} = \begin{bmatrix} p_1 & & & \\ & p_2 & & \\ & & \ddots & \\ & & & p_n \end{bmatrix}$$

$$\begin{aligned} (p_1 - p)\,(p_2 - p) \ldots (p_n - p) &= 0 \\ I_1 &= p_1 + p_2 + \ldots + p_n \\ I_2 &= p_1(p_2 + p_3 + \ldots + p_n) + p_2(p_3 + p_4 + \ldots + p_n) + \ldots + p_{n-1}\,p_n \\ &\vdots \\ I_n &= p_1 p_2 \ldots p_n \end{aligned}$$

Since the solutions $p_1,...,p_n$ of the characteristic equation are scalar invariants, the coefficients $I_1,...,I_n$ of the characteristic equation are also invariants of $\mathbf{T}_0$. They are called the principal invariants of the dyad and may be regarded as functions of the eigenvalues.

Basic invariants of a dyad : Consider the r-fold contracted r-fold products of a dyad T with itself. The resulting scalars are called the basic invariants of the dyad and are designated by $S_0, S_1, S_2, \ldots$. Their invariance under a change of the coordinate basis is proved using the transformation formulas for the tensor coordinates.

$$
\begin{aligned}
S_0 &= \delta^i_i \\
S_1 &= t_{i.}^{\;i} \\
S_2 &= t_{i.}^{\;k}\, t_{k.}^{\;i} \\
S_3 &= t_{i.}^{\;k}\, t_{k.}^{\;m}\, t_{m.}^{\;i} \\
S_4 &= \ldots \\
\bar{S}_2 &= \bar{t}_{i.}^{\;k}\, \bar{t}_{k.}^{\;i} \quad = \quad a^{r}_{.i}\, \bar{a}^{k}_{.s}\, t_{r.}^{\;s}\, a^{x}_{.k}\, \bar{a}^{i}_{z.}\, t_{x.}^{\;z} \;=\; \delta^r_z\, \delta^x_s\, t_{r.}^{\;s}\, t_{x.}^{\;z} \\
\bar{S}_2 &= t_{r.}^{\;s}\, t_{s.}^{\;r} \quad = \quad S_2
\end{aligned}
$$

Like the principal invariants, the basic invariants of a symmetric dyad may be expressed as functions of the eigenvalues using the representation of the dyad in the principal basis :

$$
\begin{aligned}
S_0 &= 1 \;\; + 1 \;\; + \ldots + 1 \\
S_1 &= p_1 + p_2 + \ldots + p_n \\
&\;\vdots \\
S_r &= p_1^r + p_2^r + \ldots + p_n^r \\
&\;\vdots
\end{aligned}
$$

The question arises whether the basic invariants for arbitrary values of the index r are independent. This question is studied using the Cayley-Hamilton Theorem. The formulation of this theorem requires a definition of powers of a tensor.

Powers of a dyad : The tensor mapping $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with $f(\mathbf{u}) = \mathbf{v} = \mathbf{T} \cdot \mathbf{u}$ is defined for a tensor u of rank 1. The composition of this mapping with itself is $f \circ f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ with $f \circ f\,(\mathbf{u}) = f(\mathbf{v}) = \mathbf{w}$. The coordinates of **v** and **w** are determined as follows :

$$
\begin{aligned}
v_i &= t_{i.}^{\;m}\, u_m \\
w_i &= t_{i.}^{\;k}\, v_k \;=\; t_{i.}^{\;k}\, t_{k.}^{\;m}\, u_m
\end{aligned}
$$

The contracted product $t_{i.}^{\;k}\, t_{k.}^{\;m}$ of the dyad T with itself is called the square of the dyad and is designated by T^2. The coordinates of T^2 are designated by $t_{i.}^{\;m(2)}$. The r-th power of the dyad T is defined by induction as the contracted product $T^r = T \cdot T^{r-1}$.

$$t_{i.}^{\;m(2)} \;=\; t_{i.}^{\;k}\, t_{k.}^{\;m}$$

$$t_{i.}^{\;m(r)} \;=\; t_{i.}^{\;k}\, t_{k.}^{\;m(r-1)}$$

The square of a dyad can only conveniently be defined using the mixed coordinates of the tensor. Otherwise products of the form $t_{ik}\, t^{km}\, t_{ms} \ldots$ must be used, in which covariant and contravariant coordinates alternate. The 0-th power of the dyad T is defined to be a tensor T^o which leaves the coordinates of T unchanged upon contracted multiplication. Its coordinates are given by the Kronecker symbol δ_i^m :

$$t_{i.}^{\;m(0)} \;=\; \delta_i^m$$

$$t_{i.}^{\;m} \;=\; \delta_i^k\, t_{k.}^{\;m}$$

Like the coordinates of the dyad T itself, the coordinates of the r-th power of T may be arranged in a matrix. The matrix for the r-th power of the dyad T is designated by $\mathbf{T}_0^r$. Its coordinates are $t_{i.}^{\;m(r)}$ with $i, m = 1,\ldots,n$.

$$\mathbf{T}_0^r = \left[\begin{array}{|c|c|c|c|}\hline t_{1.}^{\;1(r)} & t_{1.}^{\;2(r)} & & t_{1.}^{\;n(r)} \\ \hline t_{2.}^{\;1(r)} & t_{2.}^{\;2(r)} & & t_{2.}^{\;n(r)} \\ \hline & & \ddots & \\ \hline t_{n.}^{\;1(r)} & t_{n.}^{\;2(r)} & & t_{n.}^{\;n(r)} \\ \hline\end{array}\right]$$

$$t_{i.}^{\;m(r)} = t_{i.}^{\;k_1}\, t_{k_1.}^{\;k_2} \ldots t_{k_{r-1}.}^{\;k_r}\, t_{k_r.}^{\;m}$$

Powers of symmetric dyads : A real symmetric dyad S in the space $\mathbb{R}^n$ has n real eigenstates $(p, \mathbf{u}_*)$, which form a principal basis whose directions are determined by the eigenmatrix $\mathbf{U}_*$ (see Section 9.3.7). The mixed coordinate matrix $\mathbf{S}_o$ is decomposed using the diagonal matrix $\mathbf{P}$ of the eigenvalues :

$$\mathbf{S}_o \;=\; \mathbf{U}_* \mathbf{P}\, (\mathbf{U}^*)^T$$

Let the eigenvectors be normalized to the magnitude 1. Then the diagonal matrix $\mathbf{P}$ is replaced by $\mathbf{P}^2$ in the decomposition of the square $\mathbf{S}_0^2$, while the eigenmatrix stays the same. Hence the eigenstates of $\mathbf{S}_0^2$ are $(p^2, \mathbf{u}_*)$. The eigenbases of $\mathbf{S}_o$ and $\mathbf{S}_0^2$ coincide, and the ratio of corresponding eigenvalues is $p : p^2$.

$$\mathbf{S}_0^2 \;=\; \mathbf{U}_* \mathbf{P}\, (\mathbf{U}^*)^T \mathbf{U}_* \mathbf{P} (\mathbf{U}^*)^T \;=\; \mathbf{U}_* \mathbf{P}^2 (\mathbf{U}^*)^T$$

It follows analogously that $\mathbf{S}_o^r$ has the real eigenstates $(p^r, \mathbf{u}_*)$. Multiplying the n-th power of $\mathbf{S}_o$ with the corresponding coefficient c_{n-r} of the characteristic equation $(-p)^n + c_1(-p)^{n-1} + \dots + c_n = 0$ of $\mathbf{S}_o$ yields :

$$(-\mathbf{S}_o)^n + c_1(-\mathbf{S}_o)^{n-1} + \dots + c_n\mathbf{I} = \mathbf{U}_*\{(-\mathbf{P})^n + c_1(-\mathbf{P})^{n-1} + \dots + c_n\mathbf{I}\}(\mathbf{U}^*)^T$$

The curly brackets {...} contain a diagonal matrix. Every diagonal coefficient of this matrix is zero, since the corresponding coefficient of p satisfies the characteristic equation $(-p)^n + c_1(-p)^{n-1} + \dots + c_n = 0$. Hence the symmetric dyad $\mathbf{S}_o$ satisfies its own characteristic equation. This is a special case of the Cayley-Hamilton Theorem.

$$(-\mathbf{S}_o)^n + c_1(-\mathbf{S}_o)^{n-1} + \dots + c_n\mathbf{I} = \mathbf{0}$$

Powers of a tensor : The powers of a dyad are a special case of the powers of a tensor. The m-fold contracted product of a tensor with itself is called the square of the tensor T and is designated by T^2. Let the m-fold covariant and m-fold contravariant coordinates of T be $t_{k_1 \dots k_m}^{i_1 \dots i_m}$. Then the coordinates of T^2 are designated by $t_{k_1 \dots k_m}^{i_1 \dots i_m (2)}$. The r-th power of the tensor T is defined by induction as the contracted product of T and T^{r-1}.

$$t_{k_1 \dots k_m}^{i_1 \dots i_m (2)} = t_{s_1 \dots s_m}^{i_1 \dots i_m} \; t_{k_1 \dots k_m}^{s_1 \dots s_m}$$

$$t_{k_1 \dots k_m}^{i_1 \dots i_m (r)} = t_{s_1 \dots s_m}^{i_1 \dots i_m} \; t_{k_1 \dots k_m}^{s_1 \dots s_m (r-1)}$$

The 0-th power of T is defined to be a tensor T^0 which leaves the coordinate matrix of T unchanged upon m-fold contracted multiplication. The coordinates of T^0 are products of Kronecker symbols δ_s^i :

$$t_{k_1 \dots k_m}^{i_1 \dots i_m} = t_{s_1 \dots s_m}^{i_1 \dots i_m (0)} \; t_{k_1 \dots k_m}^{s_1 \dots s_m}$$

$$t_{s_1 \dots s_m}^{i_1 \dots i_m (0)} = \delta_{s_1}^{i_1} \dots \delta_{s_m}^{i_m}$$

Cayley-Hamilton Theorem : If the powers of the eigenvalue p in the characteristic equation $C(p) = 0$ of a dyad T are replaced by corresponding powers of the coordinate matrix $\mathbf{T}_o$ of the dyad, the resulting equation $C(\mathbf{T}_0) = \mathbf{0}$ is also satisfied (short version : Every dyad satisfies its own characteristic equation).

$$\begin{aligned} C(p) &= (-p)^n + c_1(-p)^{n-1} + \dots + c_{n-1}(-p) + c_n = 0 \\ C(\mathbf{T}_o) &= (-\mathbf{T}_o)^n + c_1(-\mathbf{T}_o)^{n-1} + \dots + c_{n-1}(-\mathbf{T}_o) + c_n\mathbf{I} = \mathbf{0} \end{aligned}$$

Proof : Cayley-Hamilton Theorem

The real dyad T is not assumed to be symmetric. The condition for non-trivial eigenvectors of the dyad T is $\det(\mathbf{T}_o - p\,\mathbf{I}) = 0$. For values of p which are not eigenvalues, a matrix $\mathbf{A}$ is formed by scaling the matrix $(\mathbf{T}_o - p\,\mathbf{I})$ by $\det(\mathbf{T}_o - p\,\mathbf{I})$.

$$(\mathbf{T}_o - p\,\mathbf{I})\,\mathbf{A} \;=\; \mathbf{I}\,\det(\mathbf{T}_o - p\,\mathbf{I})$$

Since the factor $(\mathbf{T}_o - p\,\mathbf{I})$ is linear in p and $\det(\mathbf{T}_o - p\,\mathbf{I})$ is a polynomial of degree n in p, the matrix $\mathbf{A}$ must be a sum of products of the scalars $p^0, p^1, ..., p^{n-1}$ with constant matrices $\mathbf{A}_i$. The determinant $\det(\mathbf{T}_o - p\,\mathbf{I})$ is replaced by the characteristic polynomial C(p).

$$\mathbf{A} \;=\; (-p)^{n-1}\,\mathbf{A}_1 + ... + (-p)\mathbf{A}_{n-1} \;+\; \mathbf{A}_n$$

$$(\mathbf{T}_o - p\,\mathbf{I})\,\{(-p)^{n-1}\mathbf{A}_1 + ... + \mathbf{A}_n\} \;=\; \{(-p)^n + c_1(-p)^{n-1} + ... + c_n\}\,\mathbf{I}$$

Comparing the coefficients of the scalars $p^0, p^1, ..., p^n$ yields a relationship between the coefficients c_i of the characteristic polynomial and the matrices $\mathbf{A}_i$:

$$\begin{array}{lrcllll}
p^n & : & \mathbf{I} & = & \mathbf{A}_1 & & \text{factor :} & (-\mathbf{T}_o)^n \\
p^{n-1} & : & c_1\,\mathbf{I} & = & \mathbf{A}_2 & + \;\mathbf{T}_o\,\mathbf{A}_1 & & (-\mathbf{T}_o)^{n-1} \\
 & & & \vdots & & & & \\
p^1 & : & c_{n-1}\,\mathbf{I} & = & \mathbf{A}_n & + \;\mathbf{T}_o\,\mathbf{A}_{n-1} & & (-\mathbf{T}_o)^1 \\
p^0 & : & c_n\,\mathbf{I} & = & & \;\;\;\mathbf{T}_o\,\mathbf{A}_n & & (-\mathbf{T}_o)^0
\end{array}$$

Each of these equations is multiplied by the specified power of $\mathbf{T}_o$ from the left. The equations are added. The terms on the right-hand side of the resulting equation cancel. The result proves the Cayley-Hamilton Theorem.

$$(-\mathbf{T}_o)^n \;+\; c_1(-\mathbf{T}_o)^{n-1} + ... + \; c_{n-1}(-\mathbf{T}_o)^1 \;+\; c_n\,\mathbf{I} \;=\; \mathbf{0}$$

Independent powers of a dyad : The independence of the powers of the coordinate matrix $\mathbf{T}_o$ of a dyad is studied using the Cayley-Hamilton Theorem. The equation is solved for $\mathbf{T}_o^n$:

$$-\,(-\mathbf{T}_o)^n \;=\; c_1(-\mathbf{T}_o)^{n-1} \;+\; c_2(-\mathbf{T}_o)^{n-2} \;+ ... +\; c_{n-1}(-\mathbf{T}_o)^1 \;+\; c_n\mathbf{I}$$

This equation is multiplied by $-\mathbf{T}_o$. The term $(-\mathbf{T}_o)^n$ is replaced by the preceding linear combination of $\mathbf{T}_o^{n-1}, ..., \mathbf{I}$.

$$-\,(-\mathbf{T}_o)^{n+1} \;=\; c_1(-\mathbf{T}_o)^n \;+\; c_2(-\mathbf{T}_o)^{n-1} \;+ ... +\; c_{n-1}(-\mathbf{T}_o)^2 \;+\; c_n(-\mathbf{T}_o)^1$$

$$-\,(-\mathbf{T}_o)^{n+1} \;=\; (c_2 \,-\, c_1c_1)\,(-\mathbf{T}_o)^{n-1} \;+ ... +\; (c_n \,-\, c_1c_{n-1})\,(-\mathbf{T}_o)^1 \;-\; c_1c_n\mathbf{I}$$

Thus the powers $\mathbf{T}_o^n$ and $\mathbf{T}_o^{n+1}$ may be expressed in terms of the powers $\mathbf{T}_o^{n-1}$,..., $\mathbf{T}_o^0$. All powers of $\mathbf{T}_o$ may be determined recursively in a similar manner :

$$
\begin{aligned}
(-\mathbf{T}_o)^{n+i} &= c_1^{(i)}(-\mathbf{T}_o)^{n-1} + \ldots + c_{n-1}^{(i)}(-\mathbf{T}_o)^{1} + c_n^{(i)}\,\mathbf{I} \\
c_1^{(i)} &= c_1 c_1^{(i-1)} - c_2^{(i-1)} \\
c_2^{(i)} &= c_2 c_1^{(i-1)} - c_3^{(i-1)} \\
&\vdots \\
c_{n-1}^{(i)} &= c_{n-1} c_1^{(i-1)} - c_n^{(i-1)} \\
c_n^{(i)} &= c_n c_1^{(i-1)}
\end{aligned}
$$

$c_k^{(i)}$ coefficient c_k in the recursion formula for $\mathbf{T}_o^{n+i}$

c_k coefficient c_k in the recursion formula for $\mathbf{T}_o^{n}$

Independent basic invariants of a dyad : The independence of the basic invariants of a dyad is studied using the Cayley-Hamilton Theorem. For this purpose the theorem is expressed in coordinate form for the matrix elements (i,m) :

$$(-1)^n\, t_{i\,.}^{\,k_1}\, t_{k_1 .}^{\,k_2} \ldots t_{k_n .}^{\,m} + \ldots + (-1)^1\, t_{i\,.}^{\,m}\, c_{n-1} + (-1)^0\, \delta_i^m\, c_n = 0$$

This equation is set up for i = m = 1,...,n. The sum of the equations leads to a relationship among the basic invariants $S_n, \ldots, S_0$. This equation is solved for S_n.

$$
\begin{aligned}
(-1)^n S_n + (-1)^{n-1} S_{n-1} c_1 + \ldots + (-1)^1 S_1 c_{n-1} + (-1)^0 S_o c_n &= 0 \\
(-1)^{n-1} S_n = (-1)^{n-1} S_{n-1} c_1 + \ldots + (-1)^1 S_1 c_{n-1} + (-1)^0 S_o c_n
\end{aligned}
$$

Multiplying the coordinate form of the theorem with $t_{m\,.}^{\,r}$ and adding the resulting equations for m = 1,...,n now yields :

$$(-1)^n\, t_{i\,.}^{\,k_1} \ldots t_{k_n .}^{\,k_{n+1}}\, t_{k_{n+1} .}^{\,r} + \ldots + (-1)^1\, t_{i\,.}^{\,k_1}\, t_{k_1 .}^{\,r}\, c_{n-1} + (-1)^0\, t_{i\,.}^{\,r}\, c_n = 0$$

This equation is set up for i = r = 1,...,n. The sum of the equations leads to a relationship among the basic invariants $S_{n+1},\ldots,S_1$. This equation is solved for S_{n+1}. The invariant S_n is replaced by the linear combination of $S_o,\ldots,S_{n-1}$ obtained above.

$$
\begin{aligned}
(-1)^n S_{n+1} + (-1)^{n-1} S_n c_1 + \ldots + (-1)^1 S_2 c_{n-1} + (-1)^0 S_1 c_n &= 0 \\
(-1)^{n+1} S_{n+1} = (-1)^{n-1} S_{n-1}(c_1 c_1 - c_2) + \ldots + (-1)^1 S_1 (c_1 c_{n-1} - c_n) + c_1 c_n S_o
\end{aligned}
$$

Thus the basic invariants S_n and S_{n+1} may be expressed in terms of $S_o ,\ldots, S_{n-1}$. All basic invariants S_{n+i} may be determined recursively in a similar manner. The coefficients $c_k^{(i)}$ satisfy the formulas specified for powers of dyads.

$$(-1)^{n+i} S_{n+i} = (-1)^{n-1} S_{n-1}\, c_1^{(i)} + \ldots + (-1)^1 S_1\, c_{n-1}^{(i)} + (-1)^0 S_o\, c_n^{(i)}$$

Basis of invariants of a dyad : For a dyad T in the space $\mathbb{R}^n$, the n principal invariants (coefficients of the characteristic equation) were shown to depend on the n eigenvalues (solutions of the characteristic equation) :

$$\begin{aligned}
I_1 &= p_1 + p_2 + \dots + p_n \\
I_2 &= p_1(p_2 + p_3 + \dots + p_n) + p_2(p_3 + p_4 + \dots + p_n) + \dots + p_{n-1}\, p_n \\
I_3 &= p_1 p_2(p_3 + \dots + p_n) + p_2 p_3(p_4 + \dots + p_n) + \dots + p_{n-2}\, p_{n-1}\, p_n \\
&\vdots \\
I_n &= p_1 p_2 \dots p_{n-1}\, p_n
\end{aligned}$$

The basic invariants with index ≥ n were shown to depend on the basic invariants $S_o, \dots, S_{n-1}$. The basic invariants were expressed as functions of the n eigenvalues. The invariance of the trace is used :

$$\begin{aligned}
S_1 &= p_1 + p_2 + \dots + p_n &&= \operatorname{tr} \mathbf{T}_o \\
S_2 &= p_1^2 + p_2^2 + \dots + p_n^2 &&= \operatorname{tr} \mathbf{T}_o^2 \\
&\vdots \\
S_k &= p_1^k + p_2^k + \dots + p_n^k &&= \operatorname{tr} \mathbf{T}_o^k
\end{aligned}$$

It follows that the invariants of a dyad are not independent. A subset of n independent invariants of a dyad is called a basis of invariants for the dyad. For instance, the n eigenvalues or the n principal invariants or n independent basic invariants may be chosen as a basis of invariants. The basis of invariants may also be a mixture of a total of n independent eigenvalues, principal invariants and basic invariants.

Basis of invariants of a tensor : The concept of a basis of invariants is transferred from dyads to general tensors. Let the following scalar invariants be defined for a fixed physical quantity described by the tensor T(...) :

$$\begin{aligned}
f_1 &: \mathbf{T} \to \mathbb{R} && \text{with} && f_1(\mathbf{T}) = c_1 \\
f_2 &: \mathbf{T} \to \mathbb{R} && \text{with} && f_2(\mathbf{T}) = c_2 \\
&\vdots \\
f_s &: \mathbf{T} \to \mathbb{R} && \text{with} && f_s(\mathbf{T}) = c_s
\end{aligned}$$

T coordinate set of the tensor

If all of the invariants $c_1, \dots, c_s$ may be expressed as functions of a subset $c_1, \dots, c_r$ of the invariants, then there is a relationship between the functions $f_1, \dots, f_s$. The invariants $c_1, \dots, c_r$ are called a basis of invariants for the tensor T. The basis of invariants is generally not unique.

$$c_i = c_i\,(c_1, \dots, c_r) \qquad\qquad i = 1, \dots, s$$

Basis of invariants of several tensors : The concept of a basis of invariants is extended from scalar functions of one tensor to scalar functions of several tensors. Let several scalar invariants be defined for fixed physical quantities described by the tensors T(...),...,W(...) :

$$g_i \;:\quad \mathbf{T}\times\ldots\times\mathbf{W} \to \mathbb{R} \qquad \text{with} \quad g_i\,(\mathbf{T},\ldots,\mathbf{W}) \;=\; c_i$$

$\mathbf{T}\times\ldots\times\mathbf{W}$ coordinate set of the tensors

A subset $c_1,\ldots,c_r$ of the invariants is called a basis of invariants of the cartesian product $\mathbf{T}\times\ldots\times\mathbf{W}$ if each of the invariants $c_1,\ldots,c_s$ may be expressed as a function of the subset $c_1,\ldots,c_r$.

$$c_i \;=\; c_i\,(c_1,\ldots,c_r)$$

Example 1 : Invariants of a symmetric dyad in $\mathbb{R}^3$

Let the coordinates $\mathbf{T}_o \;=\; (\mathbf{T}^o)^T$ of a symmetric dyad in dual bases $\mathbf{B}_*$ and $\mathbf{B}^*$ of the space $\mathbb{R}^3$ be $t_{i.}^{m}$ with i, m = 1,...,3. The eigenvalue problem is solved using the condition det $(\mathbf{T}_o - p\mathbf{I}) = 0$ for non-trivial eigenvectors. The condition leads to the following characteristic equation :

$$\begin{array}{|c|c|c|}\hline t_{1.}^{1} & t_{1.}^{2} & t_{1.}^{3} \\ \hline t_{2.}^{1} & t_{2.}^{2} & t_{2.}^{3} \\ \hline t_{3.}^{1} & t_{3.}^{2} & t_{3.}^{3} \\ \hline \end{array} \;*\; \begin{array}{|c|}\hline u_1 \\ \hline u_2 \\ \hline u_3 \\ \hline \end{array} \;=\; p \;\begin{array}{|c|}\hline u_1 \\ \hline u_2 \\ \hline u_3 \\ \hline \end{array}$$

$$\begin{aligned} (-p)^3 + \mathrm{I}_1\,(-p)^2 + \;\mathrm{I}_2\,(-p) + \mathrm{I}_3 \;&=\; 0 \\ \mathrm{I}_1 \;&=\; t_{i.}^{i} \;=\; t_{1.}^{1} \;+\; t_{2.}^{2} + t_{3.}^{3} \\ \mathrm{I}_2 \;&=\; t_{1.}^{2}\; t_{2.}^{1} \;+\; t_{2.}^{3}\; t_{3.}^{2} \;+\; t_{3.}^{1}\; t_{1.}^{3} \\ \mathrm{I}_3 \;&=\; \det \mathbf{T}_o \;=\; e_{ikm}\, t_{1.}^{i}\;\, t_{2.}^{k}\;\, t_{3.}^{m} \end{aligned}$$

The eigenvalues p_1, p_2, p_3 of T are determined using the formulas in Example 3 of Section 9.3.6. Since T is symmetric, p_1, p_2, p_3 are real. The principal invariants are the following functions of the eigenvalues :

$$\begin{aligned} \mathrm{I}_1 \;&=\; p_1 + p_2 + p_3 \\ \mathrm{I}_2 \;&=\; p_1 p_2 + p_2 p_3 + p_3 p_1 \\ \mathrm{I}_3 \;&=\; p_1 p_2 p_3 \end{aligned}$$

The basic invariants of the dyad may alternatively be obtained from the coordinates $t_{i\,.}^{\,m}$ or the eigenvalues p_i :

$$\begin{array}{lcl}
S_0 & = & \delta_i^i = 3 \\
S_1 & = & p_1 + p_2 + p_3 \quad = \quad t_{i\,.}^{\,i} = t_{1\,.}^{\,1} + t_{2\,.}^{\,2} + t_{3\,.}^{\,3} \\
S_2 & = & p_1^2 + p_2^2 + p_3^2 \quad = \quad t_{i\,.}^{\,m}\, t_{m\,.}^{\,i} \\
 & & \qquad\qquad\qquad\;\; = \quad t_{1\,.}^{\,1}\, t_{1\,.}^{\,1} + t_{1\,.}^{\,2}\, t_{2\,.}^{\,1} + t_{1\,.}^{\,3}\, t_{3\,.}^{\,1} \; + \\
 & & \qquad\qquad\qquad\qquad\quad\; t_{2\,.}^{\,1}\, t_{1\,.}^{\,2} + t_{2\,.}^{\,2}\, t_{2\,.}^{\,2} + t_{2\,.}^{\,3}\, t_{3\,.}^{\,2} \; + \\
 & & \qquad\qquad\qquad\qquad\quad\; t_{3\,.}^{\,1}\, t_{1\,.}^{\,3} + t_{3\,.}^{\,2}\, t_{2\,.}^{\,3} + t_{3\,.}^{\,3}\, t_{3\,.}^{\,3} \\
S_3 & = & p_1^3 + p_2^3 + p_3^3 \quad = \quad t_{i\,.}^{\,k}\, t_{k\,.}^{\,m}\, t_{m\,.}^{\,i}
\end{array}$$

The following relationships between the principal invariants and the basic invariants of the dyad may be confirmed by substituting the eigenvalues :

$$\begin{array}{lcl}
I_1 & = & S_1 \\
I_2 & = & \frac{1}{2}(S_2 - S_1^2) \\
I_3 & = & \frac{1}{6}(2S_3 - 3S_1S_2 + S_1^3)
\end{array}$$

9.4 TENSOR ANALYSIS

9.4.1 INTRODUCTION

Physical events take place in point spaces. In order to describe a physical quantity in a point space, a tensor is defined for this quantity at every point of the space. This requires a vector space. The point space is therefore associated with a vector space which contains every vector which can be formed as a difference of two points. The tensor at an arbitrary point is defined as a linear scalar mapping of an m-tuple of vectors of the associated vector space. The set of tensors which describes a physical quantity at every point of a point space is called a tensor field.

A tensor is represented by its coordinates. These coordinates are images of m-tuples of basis vectors. Thus a basis of the associated vector space is required at every point of the space in order to describe a tensor field. These bases may be chosen globally for the entire vector space or locally at every point of the point space. The difference between global and local bases leads to different coordinate systems for the tensor field.

In order to define a global coordinate system, a point of the point space is chosen as the origin and a basis of the associated vector space is chosen as a global basis. The position vector of an arbitrary point of the space is a linear combination of the global basis vectors. The coefficients of this linear combination are called the global coordinates of the point. The partial derivatives of the position vector with respect to the global coordinates are the global basis vectors. Hence the same global basis is associated with all points of the space.

In order to define local coordinate systems, every point of a point space is associated with its own basis of the associated vector space. To this end, the n global coordinates of the position vectors are expressed as functions of n local coordinates. The local basis vectors at an arbitrary point of the space are obtained as the partial derivatives of the position vector with respect to the local coordinates. The basis vectors at different points are generally not parallel, so that the coordinate lines are curved. The local coordinates are therefore also called curvilinear coordinates. The global coordinates are also called rectilinear coordinates.

The coordinates of a tensor in local bases at different points cannot be compared, since the basis vectors are not parallel. Thus partial derivatives of tensor coordinates with respect to local coordinates are not a measure of the change of the tensor field from point to point. The concept of the covariant derivatives of the tensor field is introduced in order to describe the spatial change of the tensor field. For this purpose the local coordinate system at the considered point is considered as a global coordinate system for an infinitesimal neighborhood of the point. In the infinitesimal neighborhood of the point the local coordinates of the tensor field are transformed to this global coordinate system. The partial derivatives of the transformed tensor coordinates with respect to the local coordinates are a measure of the change of the tensor field in the neighborhood of the considered point and are called the covariant derivatives of the tensor field.

Integrals of tensor fields and their derivatives over lines, surfaces and volumes are defined in the euclidean space $\mathbb{R}^3$. Infinitesimal line elements, surface elements and volume elements in global and in local coordinates are considered for this purpose. Operations on tensor fields which are useful in the mathematical formulation of physical problems are defined using tensor integrals over the surface of a body, divided by the volume of the body. These operations lead to the gradient of a scalar field, the divergence of a vector field and the curl of a vector field.

9.4.2 POINT SPACES

Point space : An n-tuple $(x_1, \ldots, x_n) \in \mathbb{R} \times \ldots \times \mathbb{R}$ is called a real point and is designated by X. The scalars x_i are called the coordinates of the point X. If the coordinates of the point X are arranged in a vector, this vector is designated by **x**. The set of n-tuples $(x_1, \ldots, x_n) \in \mathbb{R} \times \ldots \times \mathbb{R}$ is called the n-dimensional real point space and is designated by $\mathbb{R}^n$.

$$\mathbf{x} = (x_1, \ldots, x_n) = \begin{array}{|c|} \hline x_1 \\ \hline \vdots \\ \hline x_n \\ \hline \end{array}$$

$$\mathbb{R}^n = \{ \mathbf{x} \mid \mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R} \times \ldots \times \mathbb{R} \}$$

Associated vector space : Every ordered pair (X,Y) of points in a point space $\mathbb{R}^n$ is associated with a vector whose elements $(u_1, \ldots, u_n)$ are the differences of the coordinates of the points Y and X. The vector is designated by $\vec{XY}$, its vector representation by **u**. The vectors associated with the pairs of points in $\mathbb{R}^n$ form the vector space associated with the point space $\mathbb{R}^n$, which is often designated by V^n.

$$\mathbf{u} = (u_1, \ldots, u_n) = \begin{array}{|c|} \hline u_1 \\ \hline \vdots \\ \hline u_n \\ \hline \end{array}$$

$$u_i = y_i - x_i$$

Coordinate system : The point $(0, \ldots, 0) \in \mathbb{R}^n$ is called the origin of the point space $\mathbb{R}^n$ and is designated by O. If an arbitrary basis **B** of the vector space associated with $\mathbb{R}^n$ is chosen, then (O, **B**) is called a coordinate system of the point space.

Position vector : The vector **x** associated with the origin O and an arbitrary point X of a point space $\mathbb{R}^n$ is called the position vector of the point X. If the special coordinate system (O, **E**) with the canonical basis $(\mathbf{e}^1, \ldots, \mathbf{e}^n)$ of the associated vector space is chosen, the coordinates x_i of the point X and the coordinates of the position vector $\vec{OX}$ coincide :

$$\mathbf{x} = x_i \, \mathbf{e}^i$$

If an arbitrary covariant basis $\mathbf{B}_*$ and the dual contravariant basis $\mathbf{B}^*$ are chosen, the vector $\vec{OX}$ may also be expressed as a linear combination of these basis vectors. The coefficients of the linear combinations are called the contravariant and covariant coordinates of the point X, respectively.

$$\mathbf{x} \;=\; \mathbf{B}_* \, \mathbf{x}^* \;=\; x^i \, \mathbf{b}_i$$

$$\mathbf{x} \;=\; \mathbf{B}^* \, \mathbf{x}_* \;=\; x_i \, \mathbf{b}^i$$

$$\mathbf{x}^* \;=\; (x^1,...,x^n) \qquad \text{contravariant coordinates of X}$$

$$\mathbf{x}_* \;=\; (x_1,...,x_n\,) \qquad \text{covariant coordinates of X}$$

Due to the properties of dual bases, the covariant and contravariant coordinates of X may be determined using the basis matrices :

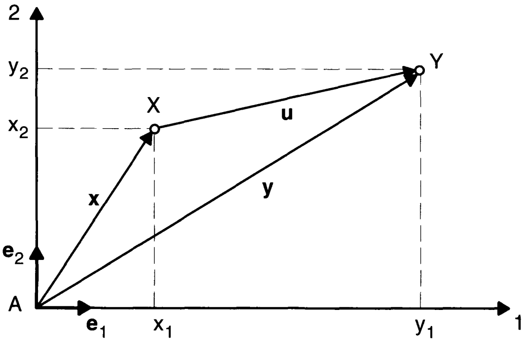
$$\mathbf{x}^* \;=\; (\mathbf{B}^*)^T \mathbf{x}$$

$$\mathbf{x}_* \;=\; (\mathbf{B}_*)^T \mathbf{x}$$

Cartesian coordinate system : A coordinate system (O, **B**) of a point space $\mathbb{R}^n$ is said to be cartesian if the basis **B** is orthonormal. A cartesian basis **B** is obtained by rotating the canonical basis **E** of the vector space associated with $\mathbb{R}^n$.

Example 1 : Geometric mapping of position vectors

In the two-dimensional space $\mathbb{R}^2$ every point X is specified by a pair (x_1,x_2) with $x_1,x_2 \in \mathbb{R}$. A vector space V^2 is associated with the point space $\mathbb{R}^2$ by associating any two points X and Y with a vector $\mathbf{u} = \overrightarrow{XY}$ with the elements (u_1,u_2), so that $u_i \;=\; y_i - x_i$. The origin O := (0,0) of the point space is mapped to a point A of the drawing plane. The canonical unit vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ of the canonical basis of V^2 are mapped to the horizontal and the vertical direction of the drawing plane, respectively. Then the position vectors **x** and **y** join the origin A with the points X and Y.



9.4.3 RECTILINEAR COORDINATES

Global basis : At every point of a point space a different basis of the associated vector space may be chosen. In the special case that a single basis **B** is chosen for the associated vector space and used at every point of the point space, this basis is said to be global (fixed). The basis vectors $\mathbf{b}_1,...,\mathbf{b}_n$ of a global basis are thus independent of the coordinates **x** of the point considered.

Global coordinate system : The origin O := (0,...,0) of the point space $\mathbb{R}^n$ and a global basis **B** of the associated vector space are said to form a global coordinate system (O, **B**) for the point space. In the following, the basis **B** is regarded as a covariant basis $(\mathbf{b}_1,...,\mathbf{b}_n)$, so that an arbitrary point X of the space $\mathbb{R}^n$ is specified by its contravariant coordinates $x^1,...,x^n$.

$$\mathbf{x} = x^i \, \mathbf{b}_i$$

x vector of the coordinates of X in the canonical basis of the vector space associated with $\mathbb{R}^n$

$\mathbf{b}_i$ basis vector whose coordinates are specified in the canonical basis of the vector space associated with $\mathbb{R}^n$

x^i contravariant coordinates of X in the basis $(\mathbf{b}_1,...,\mathbf{b}_n)$

Coordinate lines : In a point space $\mathbb{R}^n$ the subset M of points is formed which have the same coordinates $x^1,...,x^{m-1}, x^{m+1},...,x^n$. Thus only the values of the coordinate x^m of the points in M are different. This subset M of $\mathbb{R}^n$ is called a coordinate line for x^m.

The coordinates **x** and $\mathbf{x} + \Delta\mathbf{x}$ of neighboring points on a coordinate line for x^m differ only by the increment Δx^m of the coordinate x^m. Thus every coordinate line for x^m is a straight line parallel to the basis vector $\mathbf{b}_m$.

$$\Delta\mathbf{x} = (\mathbf{x} + \Delta\mathbf{x}) - \mathbf{x} = x^i \, \mathbf{b}_i + \Delta x^{(m)} \, \mathbf{b}_{(m)} - x^k \, \mathbf{b}_k$$

$$\Delta\mathbf{x} = \Delta x^{(m)} \, \mathbf{b}_{(m)}$$

The global basis vector $\mathbf{b}_m$ is constant in $\mathbb{R}^n$. Hence the coordinate line M is a straight line through a point **a** of M with the coordinate $x^m = 0$.

$$\mathbf{x} = \mathbf{a} + x^{(m)} \, \mathbf{b}_{(m)}$$

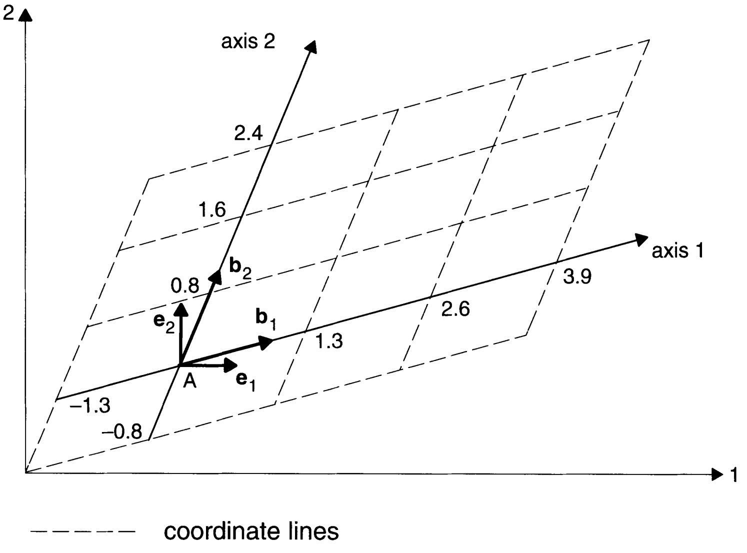
Global coordinate axes : The special coordinate line for x^m which contains the origin O of the coordinate system is called the coordinate axis x^m of the coordinate system (O, **B**). The axis x^m is the coordinate line for $\mathbf{a} = \mathbf{0}$. Since the global coordinate axes are coordinate lines and therefore rectilinear, the global coordinates $x^1,...,x^n$ are called rectilinear coordinates. The coordinate system has n axes :

$$\mathbf{x} = x^{(m)} \, \mathbf{b}_{(m)} \qquad m = 1,...,n$$

Rectilinear coordinate grid : A point $\mathbf{x} = x^i\ \mathbf{b}_i$ and increments $\Delta x^1,...,\Delta x^n$ of the global coordinates are chosen in a point space $\mathbb{R}^n$. The points $(x^1 + s_1\Delta x^1,...,$ $x^n + s_{(n)}\Delta x^{(n)})$ with $s_i \in \mathbb{Z}$ form a grid of points. Grid points with the same value for $n - 1$ of the coefficients $s_1,...,s_n$ lie on the same coordinate line. Every grid point lies at the intersection of n coordinate lines. The coordinate lines thus form a rectilinear coordinate grid.

Example 1 : Rectilinear coordinate grid in the point space $\mathbb{R}^2$

Let the origin (0,0) of the point space $\mathbb{R}^2$ be mapped to the point A of a drawing plane. The canonical basis $\mathbf{e}_1$, $\mathbf{e}_2$ of the vector space V^2 associated with $\mathbb{R}^2$ is mapped to orthogonal vectors of equal length in the drawing plane. A global basis $(\mathbf{b}_1, \mathbf{b}_2)$ is chosen for the vector space V^2. The following diagram shows the coordinate grid with the origin (0, 0) as a reference point and the increments 1.3, 0.8.



9.4.4 DERIVATIVES WITH RESPECT TO GLOBAL COORDINATES

Introduction : The description of physical states leads to tensors whose coordinates vary in the space considered. The concept of a tensor field is defined for such tensors. The variation of the tensor in the neighborhood of a point of the space is described by the partial derivatives of the tensor field with respect to the global coordinates.

Tensor field : Let the vector space V^n be associated with a point space $\mathbb{R}^n$. Let every point X of the point space be associated with a linear mapping $U(\mathbf{v}_1,...,\mathbf{v}_m)$ of a vector m-tuple. If the value of the mapping $U(\mathbf{v}_1,...,\mathbf{v}_m)$ for fixed values of the vectors $\mathbf{v}_1,...,\mathbf{v}_m \in V^n$ depends on the coordinates of the point X, the mapping U is called a tensor field. A tensor field may be scalar-valued, vector-valued or, in the general case, tensor-valued.

Covariant coordinates of a tensor field : A covariant global basis $\mathbf{B}_*$ with the vectors $\mathbf{b}_1,...,\mathbf{b}_n$ is chosen for the vector space V^n associated with a point space $\mathbb{R}^n$. The coordinates of the vectors $\mathbf{v}_i$ of a tensor field $U(\mathbf{v}_1,...\mathbf{v}_m)$ are referred to the basis $\mathbf{B}_*$. Then the images of the m-tuples $(\mathbf{b}_{i_1},...,\mathbf{b}_{i_m})$ of basis vectors are called the covariant coordinates of the tensor field.

$$u_{i_1...i_m} = U(\mathbf{b}_{i_1},...,\mathbf{b}_{i_m})$$

The contravariant coordinates of a point X of $\mathbb{R}^n$ in the covariant basis $\mathbf{B}_*$ are $x^1,..., x^n$. Although the basis vectors $\mathbf{b}_k$ are independent of the coordinates x^i of the point, the coordinates $u_{i_1...i_m}$ of the tensor field nevertheless depend on the coordinates of the point, since the mapping U (...) depends on $x^1,...,x^n$.

scalar tensor field : $u = u\ (x^1,...,x^n)$

vectorial tensor field : $u_i = u_i\ (x^1,...,x^n)$

dyadic tensor field : $u_{ik} = u_{ik}(x^1,...,x^n)$

Contravariant coordinates of a tensor field : A contravariant global basis $\mathbf{B}^*$ with the vectors $\mathbf{b}^1,..., \mathbf{b}^n$ is chosen for the vector space V^n associated with a point space $\mathbb{R}^n$. The coordinates of the vectors $\mathbf{v}_i$ of a tensor field $U(\mathbf{v}_1,...\mathbf{v}_m)$ are referred to the basis $\mathbf{B}^*$. Then the images of the m-tuples $(\mathbf{b}^{i_1},...,\mathbf{b}^{i_m})$ of basis vectors are called the contravariant coordinates of the tensor field.

$$u^{i_1...i_m} = U(\mathbf{b}^{i_1},...,\mathbf{b}^{i_m})$$

The covariant coordinates of a point X of $\mathbb{R}^n$ in the contravariant basis $\mathbf{B}^*$ are $x_1,...,x_n$. Although the basis vectors $\mathbf{b}^k$ are independent of the coordinates x_i of the point, the coordinates $u^{i_1...i_m}$ of the tensor field nevertheless depend on the coordinates of the point, since the mapping U(...) depends on $x_1,...,x_n$.

scalar tensor field	:	$u = u\ (x_1,...,x_n)$
vectorial tensor field	:	$u^i = u^i\ (x_1,...,x_n)$
dyadic tensor field	:	$u^{ik} = u^{ik}(x_1,...,x_n)$

Partial derivatives with respect to covariant coordinates : The value of a scalar tensor field $u(x_1,...,x_n)$ changes along a coordinate line. Let the values of the tensor field at the points $\mathbf{x}$ and $\mathbf{x} + \Delta x_{(m)} \mathbf{b}^{(m)}$ be u and $u + \Delta u$. The limit of the quotient $\Delta u / \Delta x_m$ is called the partial derivative of the tensor field u with respect to the covariant position coordinate x_m and is designated by $\partial u / \partial x_m$ or $u^{,m}$.

$$\frac{\partial u}{\partial x_m} := u^{,m} := \lim_{\Delta x_m \to 0} \frac{\Delta u}{\Delta x_m} \quad \text{with} \quad \Delta x_i = 0 \quad \text{for} \quad i \neq m$$

Partial derivatives with respect to contravariant coordinates : The value of a scalar tensor field $u(x^1,...,x^n)$ changes along a coordinate line. Let the values of the tensor field at the points $\mathbf{x}$ and $\mathbf{x} + \Delta x^{(m)} \mathbf{b}_{(m)}$ be u and $u + \Delta u$. The limit of the quotient $\Delta u / \Delta x^m$ is called the partial derivative of the tensor field u with respect to the contravariant position coordinate x^m and is designated by $\partial u / \partial x^m$ or $u_{,m}$.

$$\frac{\partial u}{\partial x^m} := u_{,m} := \lim_{\Delta x^m \to 0} \frac{\Delta u}{\Delta x^m} \quad \text{with} \quad \Delta x^i = 0 \quad \text{for} \quad i \neq m$$

Dual partial derivatives : Let the coordinates of the position vector of a point X of the point space $\mathbb{R}^n$ in the dual bases $\mathbf{B}^*$ and $\mathbf{B}_*$ of the associated vector space be $(x_1,...,x_n)$ and $(x^1,...,x^n)$, respectively. The indices of the coordinates are lowered and raised using the metric coefficients g_{ik} and g^{ik} of the bases.

$$x_i = g_{ik}\, x^k \qquad\qquad x^i = g^{ik}\, x_k$$

The metric coefficients of a global basis are constant in $\mathbb{R}^n$. The relationships between the partial derivatives of a scalar tensor function u with respect to the covariant and the contravariant coordinates of the position vector are determined using the chain rule.

$$\frac{\partial x_i}{\partial x^k} = g_{ik} \qquad\qquad \frac{\partial x^i}{\partial x_k} = g^{ik}$$

$$\frac{\partial u}{\partial x^k} = g_{ik} \frac{\partial u}{\partial x_i} \qquad\qquad \frac{\partial u}{\partial x_k} = g^{ik} \frac{\partial u}{\partial x^i}$$

Transformation of the partial derivatives : Let the coordinates of the position vector of a point X of the point space $\mathbb{R}^n$ in the dual bases $\mathbf{B}^*$ and $\mathbf{B}_*$ of the associated vector space be $(x_1,...,x_n)$ and $(x^1,...,x^n)$, respectively. The bases are transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_*\mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^*\bar{\mathbf{A}}^T$ using the coefficient matrix $\mathbf{A}$ and its inverse $\bar{\mathbf{A}}$. Then the coordinates of the position vector $\mathbf{x}$ are transformed according to the general transformation rules in Section 9.2.6 :

$$x_i = \bar{a}^k_{.i}\, \bar{x}_k \qquad\qquad x^i = a^i_{.k}\, \bar{x}^k$$

$\bar{x}_i$, $\bar{x}^i$ coordinates of X in the bases $\bar{\mathbf{B}}^*$, $\bar{\mathbf{B}}_*$

$\bar{a}^k_{.i}$ coordinates of the transformation matrix $\bar{\mathbf{A}}$

$a^i_{.k}$ coordinates of the transformation matrix $\mathbf{A}$

For global bases, the transformation matrices $\mathbf{A}$ and $\bar{\mathbf{A}}$ are constant in $\mathbb{R}^n$. The transformation rules for the partial derivatives of a scalar tensor function u are determined using the chain rule.

$$\frac{\partial x_i}{\partial \bar{x}_k} = \bar{a}^k_{.i} \qquad\qquad \frac{\partial x^i}{\partial \bar{x}^k} = a^i_{.k}$$

$$\frac{\partial u}{\partial \bar{x}_k} = \bar{a}^k_{.i}\, \frac{\partial u}{\partial x_i} \qquad\qquad \frac{\partial u}{\partial \bar{x}^k} = a^i_{.k}\, \frac{\partial u}{\partial x^i}$$

If the derivative $\partial u/\partial \bar{x}_k$ is designated by $\bar{u}^{,k}$ and the derivative $\partial u/\partial \bar{x}^k$ by $\bar{u}_{,k}$, the transformation rules justify the position of the indices in the comma form of the derivatives : The partial derivatives transform like the coordinates of the position vector.

$$\bar{x}^k = \bar{a}^k_{.i}\, x^i \qquad\qquad \bar{x}_k = a^i_{.k}\, x_i$$

$$\bar{u}^{,k} = \bar{a}^k_{.i}\, u^{,i} \qquad\qquad \bar{u}_{,k} = a^i_{.k}\, u_{,i}$$

Example 1 : Dual partial derivatives of a tensor field

Dual global coordinate systems $(O, \mathbf{B}_*)$ and $(O, \mathbf{B}^*)$ with the metrics $\mathbf{G}_* = \mathbf{B}_*^T \mathbf{B}_*$ and $\mathbf{G}^* = (\mathbf{B}^*)^T \mathbf{B}^*$ are chosen for the point space $\mathbb{R}^2$.

$$\mathbf{B}_* = \frac{1}{2} \begin{array}{|c|c|} \hline 2 & -2 \\ \hline 1 & 2 \\ \hline \end{array} \qquad \mathbf{B}^* = \frac{1}{3} \begin{array}{|c|c|} \hline 2 & -1 \\ \hline 2 & 2 \\ \hline \end{array}$$

$$\mathbf{G}_* = \frac{1}{4} \begin{array}{|c|c|} \hline 5 & -2 \\ \hline -2 & 8 \\ \hline \end{array} \qquad \mathbf{G}^* = \frac{1}{9} \begin{array}{|c|c|} \hline 8 & 2 \\ \hline 2 & 5 \\ \hline \end{array}$$

The relationship between the covariant coordinates (x_1, x_2) and the contravariant coordinates (x^1, x^2) of a point $X \in \mathbb{R}^2$ is established by the metric $\mathbf{G}_*$. Let the scalar tensor field $u = 2x_1 - 3x_2$ be defined in $\mathbb{R}^2$. Substituting x_1 and x_2 yields the contravariant form of the tensor field.

$$\begin{aligned} x_1 &= \tfrac{1}{4}(\ 5x^1 - 2x^2) \\ x_2 &= \tfrac{1}{4}(-2x^1 + 8x^2) \\ u &= 2x_1 - 3x_2 = 4x^1 - 7x^2 \end{aligned}$$

The partial derivatives of the tensor field are obtained by differentiating the functions $u(x_1, x_2)$ and $u(x^1, x^2)$. They satisfy the relationships between dual partial derivatives.

$$\frac{\partial u}{\partial x_1} = 2 \qquad \frac{\partial u}{\partial x_2} = -3$$

$$\frac{\partial u}{\partial x^1} = 4 \qquad \frac{\partial u}{\partial x^2} = -7$$

$$\frac{\partial u}{\partial x_1} = g^{11} \frac{\partial u}{\partial x^1} + g^{12} \frac{\partial u}{\partial x^2} = \frac{1}{9}(8 * 4 - 2 * 7) = 2$$

$$\frac{\partial u}{\partial x_2} = g^{21} \frac{\partial u}{\partial x^1} + g^{22} \frac{\partial u}{\partial x^2} = \frac{1}{9}(2 * 4 - 5 * 7) = -3$$

Example 2 : Transformed partial derivatives of a tensor field

The bases of the point space $\mathbb{R}^n$ in the preceding Example 2 are transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^* \bar{\mathbf{A}}^T$ using the matrix $\mathbf{A}$ and its inverse $\bar{\mathbf{A}}$.

$$\mathbf{A} = \begin{array}{|c|c|}\hline 1 & -1 \\ \hline 1 & 1 \\ \hline \end{array} \qquad \bar{\mathbf{A}} = \frac{1}{2} \begin{array}{|c|c|}\hline 1 & 1 \\ \hline -1 & 1 \\ \hline \end{array}$$

The relationship between the coordinates (x_1, x_2) of a point X in the basis $\mathbf{B}^*$ and the coordinates $(\bar{x}_1, \bar{x}_2)$ of the same point in the basis $\bar{\mathbf{B}}^*$ is determined using the inverse transformation matrix $\bar{\mathbf{A}}$. Let the tensor field $u = 2x_1 - 3x_2$ be defined in $\mathbb{R}^2$. Substituting x_1 and x_2 yields the tensor field $u(\bar{x}_1, \bar{x}_2)$.

$$\begin{aligned} x_1 &= \bar{a}^{1}_{.1}\,\bar{x}_1 + \bar{a}^{2}_{.1}\,\bar{x}_2 = 0.5\bar{x}_1 - 0.5\bar{x}_2 \\ x_2 &= \bar{a}^{1}_{.2}\,\bar{x}_1 + \bar{a}^{2}_{.2}\,\bar{x}_2 = 0.5\bar{x}_1 + 0.5\bar{x}_2 \\ u &= 2x_1 - 3x_2 = -0.5\bar{x}_1 - 2.5\bar{x}_2 \end{aligned}$$

The partial derivatives of the tensor field are obtained by differentiating the functions $u(x_1,x_2)$ and $u(\bar{x}_1, \bar{x}_2)$. They satisfy the relationships between transformed partial derivatives.

$$\frac{\partial u}{\partial x_1} = 2.0 \qquad \frac{\partial u}{\partial x_2} = -3.0$$

$$\frac{\partial u}{\partial \bar{x}_1} = -0.5 \qquad \frac{\partial u}{\partial \bar{x}_2} = -2.5$$

$$\frac{\partial u}{\partial \bar{x}_1} = \bar{a}^{1}_{.1} \frac{\partial u}{\partial x_1} + \bar{a}^{1}_{.2} \frac{\partial u}{\partial x_2} = 0.5 * 2 - 0.5 * 3 = -0.5$$

$$\frac{\partial u}{\partial \bar{x}_2} = \bar{a}^{2}_{.1} \frac{\partial u}{\partial x_1} + \bar{a}^{2}_{.2} \frac{\partial u}{\partial x_2} = -0.5 * 2 - 0.5 * 3 = -2.5$$

9.4.5 CURVILINEAR COORDINATES

Local basis : A global basis $\mathbf{B}_*$ for the associated vector space of a point space $\mathbb{R}^n$ is the same at every point X, and hence independent of the position vector $\mathbf{x}$. However, for certain problems (for example for the formulation of the behavior of cylindrical bodies) it may be convenient to choose a different basis $\bar{\mathbf{B}}_*$ for the vector space associated with $\mathbb{R}^n$ at different points in $\mathbb{R}^n$. This position-dependent basis is called a local basis. The local basis $\bar{\mathbf{B}}_*$ is specified by a transformation of the global basis $\mathbf{B}_*$. The transformation matrix $\mathbf{A}$ generally depends on position.

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A} \qquad \wedge \qquad \mathbf{A} = \mathbf{A}(\mathbf{x})$$

Local coordinate system : The position vector $\mathbf{x}$ of a point X of $\mathbb{R}^n$ and a local basis $\bar{\mathbf{B}}_*(\mathbf{x})$ are said to form a local coordinate system $(\mathbf{x}, \bar{\mathbf{B}}_*)$ of the point space. Every point of the point space has its own local coordinate system. This is not unique, since the local basis is chosen arbitrarily.

Specification of local bases : Generally the local basis at a point X of the space $\mathbb{R}^n$ may be chosen arbitrarily. A parametric specification of the local bases is often convenient. The global coordinates $(x^1,...,x^n)$ of the points of $\mathbb{R}^n$ in the basis $\mathbf{b}_1,...,\mathbf{b}_n$ are defined as functions of n independent parameters $y^1,...,y^n$. The functions $x^i\,(y^1,...,y^n)$ are generally non-linear. Hence the position vector $\mathbf{x}$ is a non-linear function of the parameters $y^1,...,y^n$.

$$\mathbf{x} = x^i\,\mathbf{b}_i$$
$$x^i = x^i\,(y^1,...,y^n)$$

The total (complete) differential d$\mathbf{x}$ of the position vector for arbitrary changes dy^i of the parameters $y^1,...,y^n$ is determined using the chain rule. The coefficient of the increment dy^m is chosen as the m-th basis vector at the point $\mathbf{x}$ and is designated by $\bar{\mathbf{b}}_m$.

$$d\mathbf{x} = \frac{\partial \mathbf{x}}{\partial x^i}\,\frac{\partial x^i}{\partial y^m}\,dy^m = \mathbf{b}_i\,\frac{\partial x^i}{\partial y^m}\,dy^m = \bar{\mathbf{b}}_m\,dy^m$$

$$\bar{\mathbf{b}}_m = \mathbf{b}_i\,\frac{\partial x^i}{\partial y^m}$$

The relationship between the global basis $\mathbf{b}_1,...,\mathbf{b}_n$ and a local basis $\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n$ is thus given by a transformation with a matrix $\mathbf{A}$ whose elements are the partial derivatives of the global coordinates x^i with respect to the parameters y^m. Since the functions $x^i(y^1,...,y^n)$ are non-linear, the partial derivatives $\partial x^i / \partial y^m$ are position-dependent. Thus the basis $\bar{\mathbf{B}}_*$ is local. The suitability of the local coordinate system depends on the choice of the functions $x^i\,(y^1,...,y^n)$.

$$\bar{\mathbf{B}}_* \quad = \quad \mathbf{B}_* \mathbf{A} \quad \text{with} \quad a^i_{.\,m} = \frac{\partial x^i}{\partial y^m}$$

$$\mathbf{A} \quad = \quad \begin{array}{|c|c|c|} \hline \frac{\partial x^1}{\partial y^1} & & \frac{\partial x^1}{\partial y^n} \\ \hline & \ddots & \\ \hline \frac{\partial x^n}{\partial y^1} & & \frac{\partial x^n}{\partial y^n} \\ \hline \end{array}$$

Functional determinant : If the functions $x^i(y^1,...,y^n)$ in the parametric specification of the local bases are chosen arbitrarily, a subset $\{x^{i_1},...,x^{i_m}\}$ of these functions may be dependent in a subset $S \subseteq \mathbb{R}^n$. Thus there is a function $f(x^{i_1},...,x^{i_m})$ which is zero at every point $\mathbf{x} \in S$. This implies that the increments $dx^{i_1},...,dx^{i_m}$ of the global coordinates at points $\mathbf{x} \in S$ are linearly dependent.

$$\bigwedge_{x \in S} \ (f(x^{i_1},...,x^{i_m}) \ = \ 0)$$

$$df \ = \ \frac{\partial f}{\partial x^{i_1}} \, dx^{i_1} \ + ... + \ \frac{\partial f}{\partial x^{i_m}} \, dx^{i_m} \ = \ 0 \qquad\qquad \mathbf{x} \in S$$

The increments dx^i of the global coordinates depend on the increments dy^k of the parameters. The total differentials dx^i are determined using the partial derivatives $\partial x^i / \partial y^k$. The partial derivatives are arranged in a matrix **J**, which is called the functional matrix (Jacobian matrix).

$$d\mathbf{x} \ = \ \mathbf{J} \, d\mathbf{y}$$

$$\begin{array}{|c|} \hline dx^1 \\ \hline \vdots \\ \hline dx^n \\ \hline \end{array} \ = \ \begin{array}{|c|c|c|} \hline \frac{\partial x^1}{\partial y^1} & & \frac{\partial x^1}{\partial y^n} \\ \hline & \ddots & \\ \hline \frac{\partial x^n}{\partial y^1} & & \frac{\partial x^n}{\partial y^n} \\ \hline \end{array} \ * \ \begin{array}{|c|} \hline dy^1 \\ \hline \vdots \\ \hline dy^n \\ \hline \end{array}$$

At every point **x** of the subset S the rows $i_1,...,i_m$ of the Jacobian matrix are linearly dependent. Hence the functional determinant det **J** is zero in the subset S.

Local coordinates : The global coordinates $x^1,...,x^n$ in the space $\mathbb{R}^n$ are independent. Their increments dx^i may be chosen arbitrarily. The functions $x^i(y^1,...,y^m)$ must therefore be chosen such that the functional determinant det **J** is non-zero. For such a choice of the functions $x^i(y^1,...,y^m)$, the functional matrix **J** is regular, so that every total differential d**x** of the position vector corresponds to a unique total differential d**y** of the parameters :

$$d\mathbf{y} \ = \ \mathbf{J}^{-1} \, d\mathbf{x}$$

If the condition det $\mathbf{J} \neq 0$ is satisfied, then the local basis $\bar{\mathbf{B}}_*$ is obtained from the global basis $\mathbf{B}_*$ using the transformation matrix $\mathbf{A} = \mathbf{J}$. The parameters $y^1,...,y^n$ are called the local coordinates of the point $(x^1,...,x^n)$ with $x^i = x^i(y^1,...,y^n)$.

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{J}$$

If the functional determinant det $\mathbf{J}$ is zero at a point $\mathbf{x} \in S$, then some of the local coordinates and the local basis vectors at the point $\mathbf{x}$ are not uniquely determined. Often this degenerate case cannot be avoided in subsets of $\mathbb{R}^n$ (see cylindrical and spherical coordinates in Examples 1 and 2 of this section).

Transformation of the local coordinates : Consider two local coordinate systems $x^i(y^1,...,y^n)$ and $x^i(z^1,...,z^n)$ in a point space $\mathbb{R}^n$. Let the specification $x^i(y^1,...,y^n)$ of the local coordinates $\mathbf{y}$ and the relationships $y^i(z^1,...,z^n)$ between the local coordinates $\mathbf{y}$ and $\mathbf{z}$ be given. Then the relationship between the increments dy^i and dz^k of the local coordinates is also known :

$$dy^i = \frac{\partial y^i}{\partial z^k} dz^k$$

At a point $\mathbf{x}$ of $\mathbb{R}^n$, an increment $d\mathbf{x}$ of the position vector may alternatively be expressed in the local basis $\mathbf{b}_1,...,\mathbf{b}_n$ defined by $\mathbf{y}$ or in the local basis $\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n$ defined by $\mathbf{z}$:

$$d\mathbf{x} = \mathbf{b}_i \, dy^i = \bar{\mathbf{b}}_k \, dz^k$$

$$\bar{\mathbf{b}}_k = \mathbf{b}_i \frac{\partial y^i}{\partial z^k}$$

The relationship between the local bases $\mathbf{B}_*$ for $\mathbf{y}$ and $\bar{\mathbf{B}}_*$ for $\mathbf{z}$ is expressed using a transformation matrix $\mathbf{A}$ whose coefficients are the partial derivatives of the local coordinates :

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$$

$$\mathbf{A} = \begin{array}{|c|c|c|} \hline \frac{\partial y^1}{\partial z^1} & & \frac{\partial y^1}{\partial z^n} \\ \hline & \ddots & \\ \hline \frac{\partial y^n}{\partial z^1} & & \frac{\partial y^n}{\partial z^n} \\ \hline \end{array}$$

Coordinate lines : Let the specification $x^i(y^1,...,y^n)$ of a local coordinate system in the point space $\mathbb{R}^n$ be given. Let the subset of points in $\mathbb{R}^n$ with the same coordinates $y^1,...,y^{m-1},y^{m+1},...,y^n$ be M. Thus only the values of the local coordinate y^m of the points in M are different. Then M is called a coordinate line for y^m in $\mathbb{R}^n$.

The local coordinates $\mathbf{y}$ and $\mathbf{y} + d\mathbf{y}$ of neighboring points $\mathbf{x}$ and $\mathbf{x} + d\mathbf{x}$ on a coordinate line for y^m differ only by the increment dy^m of the coordinate y^m. The vector $d\mathbf{x}$ is tangent to the coordinate line at the point $\mathbf{x}$.

$$
\begin{aligned}
d\mathbf{y} &= (0,...,0,dy^m,0,...,0)^T \\
d\mathbf{x} &= \mathbf{J}\ d\mathbf{y}
\end{aligned}
$$

The tangent vector $d\mathbf{x}$ is determined at different points $\mathbf{x}_1, \mathbf{x}_2,...$ of the coordinate line for the same increment $d\mathbf{y}$. Since the functional matrix $\mathbf{J}$ of the local coordinates $\mathbf{y}_1, \mathbf{y}_2, ...$ depends on the points considered, the tangent vectors $d\mathbf{x}_1, d\mathbf{x}_2, ...$ at these points are generally different. Thus the coordinate line is generally curved.

Local coordinate axis : Every point P in the space $\mathbb{R}^n$ is the origin of a local coordinate system. The point P lies on exactly n coordinate lines. If P has the local coordinates $(y^1,...,y^n)$, then the coordinates $y^1,...,y^{m-1},y^{m+1},...,y^n$ are constant on the m-th coordinate line. This coordinate line is called the m-th local coordinate axis at the point P. The m-th vector $\mathbf{b}_m$ of the local basis at the point P is tangent to the m-th local coordinate axis. Since the local axes are coordinate lines which are generally curved, the local coordinates $y^1,...,y^n$ are also called curvilinear coordinates.

Curved coordinate grid : A point with the local coordinates $(y^1,...,y^n)$ is chosen in a point space $\mathbb{R}^n$, together with increments $dy^1,...,dy^n$ of the local coordinates. The points $(y^1 + s_1 dy^1,...,y^n + s_n dy^n)$ with $s_i \in \mathbb{Z}$ form a grid. Grid points with the same value for $n - 1$ of the factors $s_1,...,s_n$ lie on a common coordinate line. Every grid point is at the intersection of n coordinate lines. Thus the coordinate lines form a curved coordinate grid.

Example 1 : Cylindrical coordinate system

Let the global basis of the point space $\mathbb{R}^3$ be the canonical basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Let the local coordinate system be cylindrical with the designations r, θ, z instead of y^1, y^2, y^3 :

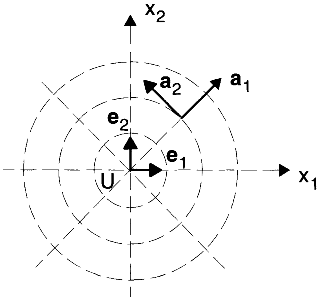
$$\begin{aligned} x^1 &= r\cos\theta & \qquad r &= y^1 : \text{ radial distance} \\ x^2 &= r\sin\theta & \theta &= y^2 : \text{ angle} \\ x^3 &= z & z &= y^3 : \text{ axial distance} \end{aligned}$$

The specification shows that the local coordinates of a point $\mathbf{x} \in \mathbb{R}^n$ are not unique : The local coordinates (r, θ, z), $(r, \theta + 2\pi s, z)$ and $(-r, \theta + (2s+1)\pi, z)$ with $s \in \mathbb{Z}$ are mapped to the same point $\mathbf{x} \in \mathbb{R}^n$. The functional matrix $\mathbf{J}$ is given by

$$\mathbf{J} = \begin{bmatrix} \cos\theta & -r\sin\theta & 0 \\ \sin\theta & r\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \end{bmatrix}$$

Since $\det \mathbf{J} = r$, the local coordinate system is well-defined for all points with $r > 0$. For the points $(0, 0, z)$ with $r = 0$, however, $\det \mathbf{J} = 0$. For these points the local coordinates r and z are uniquely determined, but the local coordinate θ and the basis vector $\mathbf{a}_2$ are not. The inverse of the Jacobian matrix is defined only for $r \neq 0$.

$$\mathbf{J}^{-1} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\frac{1}{r}\sin\theta & \frac{1}{r}\cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}$$



$(U\,;\ \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3)$	global coordinate system (x^1, x^2, x^3)
$(A\,;\ \mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$	local coordinate system (r, θ, z)
– – – –	coordinate lines

Example 2 : Spherical coordinate system

Let the global basis of the point space $\mathbb{R}^3$ be the canonical basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$. Let the local coordinate system be spherical with the designations r, θ, β instead of y^1, y^2, y^3 :

$$\begin{array}{lll} x^1 = r \cos\beta \cos\theta & r = y^1 : & \text{radial distance} \\ x^2 = r \cos\beta \sin\theta & \theta = y^2 : & \text{azimuthal angle} \\ x^3 = r \sin\beta & \beta = y^3 : & \text{polar angle} \end{array}$$

Due to the periodicity of the trigonometric functions the local coordinates of a point $\mathbf{x} \in \mathbb{R}^n$ are not unique. The functional matrix $\mathbf{J}$ is given by

$$\mathbf{J} = \begin{array}{|c|c|c|} \hline \cos\beta \cos\theta & -r \cos\beta \sin\theta & -r \sin\beta \cos\theta \\ \hline \cos\beta \sin\theta & r \cos\beta \cos\theta & -r \sin\beta \sin\theta \\ \hline \sin\beta & 0 & r \cos\beta \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline & & \\ \mathbf{a}_1 & \mathbf{a}_2 & \mathbf{a}_3 \\ & & \\ \hline \end{array}$$

Since $\det \mathbf{J} = r^2 \cos\beta$, the local coordinate system is well-defined for all points with $r \neq 0$ and $\cos\beta \neq 0$. For the points $(r, \theta, \frac{\pi}{2})$, however, $\det \mathbf{J} = 0$. For these points the local coordinates r and β are uniquely determined, but the local coordinate θ and the basis vector $\mathbf{a}_3$ are not. The basis vector $\mathbf{a}_2$ degenerates to $\mathbf{0}$. The determinant of $\mathbf{J}$ is also zero at the point $(0, \theta, \beta)$ with $r = 0$. For this point the local coordinates θ and β and the basis vector $\mathbf{a}_1$ are not uniquely determined. The basis vectors $\mathbf{a}_2$ and $\mathbf{a}_3$ degenerate to $\mathbf{0}$.

9.4.6 CHRISTOFFEL SYMBOLS

Introduction : Every point of a point space $\mathbb{R}^n$ with a local coordinate system $x^i(y^1,...,y^n)$ has its own local basis $\bar{\mathbf{B}}_*$. The local basis vectors $\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n$ are therefore functions of the local coordinates $y^1,...,y^n$. The partial derivative of a basis vector with respect to a local coordinate is a linear combination of the basis vectors. The coefficients of this linear combination are called Christoffel symbols.

Partial derivatives of a covariant basis : Let a global basis $\mathbf{B}_*$ and the specification $x^i(y^1,...,y^n)$ of a local coordinate system be given in a point space $\mathbb{R}^n$. Then the partial derivatives $\partial x^i / \partial y^m$ determine the local basis $\bar{\mathbf{B}}_*$. If the functional determinant is non-zero, the relationship between the local and the global basis vectors may be inverted :

$$\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{J} \quad : \quad \bar{\mathbf{b}}_i = \mathbf{b}_s \frac{\partial x^s}{\partial y^i}$$

$$\mathbf{B}_* = \bar{\mathbf{B}}_* \mathbf{J}^{-1} : \quad \mathbf{b}_s = \bar{\mathbf{b}}_m \frac{\partial y^m}{\partial x^s}$$

The global basis vectors $\mathbf{b}_i$ are independent of the local coordinates $y^1,...,y^n$. The partial derivative of a local basis vector $\bar{\mathbf{b}}_i$ with respect to the local coordinate y^k is therefore given by

$$\frac{\partial \bar{\mathbf{b}}_i}{\partial y^k} = \frac{\partial^2 x^s}{\partial y^i \, \partial y^k} \mathbf{b}_s = \frac{\partial^2 x^s}{\partial y^i \, \partial y^k} \cdot \frac{\partial y^m}{\partial x^s} \bar{\mathbf{b}}_m$$

The relationship between the local basis vectors and their partial derivatives with respect to the local coordinates is expressed using the Christoffel symbols. The expressions for the Christoffel symbols follow from the preceding equation. The Christoffel symbols are symmetric in the lower indices. They are not the coordinates of a tensor.

$$\frac{\partial \bar{\mathbf{b}}_i}{\partial y^k} = \Gamma^m_{ik} \bar{\mathbf{b}}_m$$

$$\Gamma^m_{ik} = \frac{\partial^2 x^s}{\partial y^i \, \partial y^k} \cdot \frac{\partial y^m}{\partial x^s}$$

Partial derivatives of a contravariant basis : The Christoffel symbols $\tilde{\Gamma}^i_{km}$ for the partial derivatives of the contravariant local basis vectors $\bar{\mathbf{b}}^i$ with respect to the local coordinates y^k are defined in analogy with the Christoffel symbols Γ^m_{ik} for the covariant basis :

$$\frac{\partial \bar{\mathbf{b}}^i}{\partial y^k} = \tilde{\Gamma}^i_{km} \bar{\mathbf{b}}^m$$

The relationship between the symbols $\underline{\Gamma}^m_{ik}$ and $\underline{\tilde{\Gamma}}^{\,i}_{km}$ is determined by partial differentiation of the orthonormality relation $\mathbf{b}_i \cdot \mathbf{b}^m = \delta^m_i$ of the dual bases with respect to the local coordinate y^k :

$$\begin{aligned}
\frac{\partial \bar{\mathbf{b}}_i}{\partial y^k} \cdot \bar{\mathbf{b}}^m \quad + \quad \bar{\mathbf{b}}_i \cdot \frac{\partial \bar{\mathbf{b}}^m}{\partial y^k} &= 0 \\
\Gamma^s_{ik}\, \bar{\mathbf{b}}_s \cdot \bar{\mathbf{b}}^m \quad + \quad \tilde{\Gamma}^m_{ks}\, \bar{\mathbf{b}}_i \cdot \bar{\mathbf{b}}^s &= 0 \\
\tilde{\Gamma}^m_{ki} &= -\Gamma^m_{ik}
\end{aligned}$$

Partial derivatives of the basis determinant : The determinant b_* of a local basis $\mathbf{B}_*$ in the point space $\mathbb{R}^n$ is expressed in terms of the permutation tensor $e^{i_1 \dots i_n}$ of the global canonical basis $\mathbf{E}$ and the coordinates b_{ki} of the local basis vectors $\mathbf{b}_i$:

$$b_* \quad = \quad e^{i_1 \dots i_n}\, b_{i_1 1} \dots b_{i_n n}$$

The partial derivative of the determinant b_* with respect to the local coordinate y^s is to be determined. The global permutation tensor $e^{i_1 \dots i_n}$ does not depend on the local coordinates. The partial derivatives of the coordinates b_{ki} of the local basis vectors are determined using the Christoffel symbols. The product rule yields :

$$\frac{\partial b_{ki}}{\partial y^s} \quad = \quad \Gamma^r_{is}\, b_{kr}$$

$$\frac{\partial b_*}{\partial y^s} \quad = \quad e^{i_1 \dots i_n} \,(\, \Gamma^r_{1s}\; b_{i_1 r}\, b_{i_2 2} \dots b_{i_n n} + \dots + \Gamma^r_{ns}\; b_{i_1 1} \dots b_{i_{n-1}\, n-1}\; b_{i_n r}\,)$$

For the first term on the right-hand side the sum over r is written out explicitly. For $r = 1$ the result is $\Gamma^1_{1s}\, b_*$. For $r \neq 1$ the result is proportional to the determinant of a matrix with two identical columns, and hence 0.

$$\begin{aligned}
e^{i_1 \dots i_n} \Gamma^r_{1s}\; b_{i_1 r}\, b_{i_2 2} \dots b_{i_n n} \quad &= \quad e^{i_1 \dots i_n} \Gamma^1_{1s}\; b_{i_1 1}\, b_{i_2 2} \dots b_{i_n n} \;+ \\
& e^{i_1 \dots i_n} \Gamma^2_{1s}\; b_{i_1 2}\, b_{i_2 2} \dots b_{i_n n} \;+ \dots \\
&= \quad \Gamma^1_{1s}\, b_*
\end{aligned}$$

The value $\Gamma^2_{2s}\, b_*$ is obtained analogously for the second term on the right-hand side of the equation for $\partial b_* / \partial y^s$. Altogether, the partial derivative of the basis determinant b_* with respect to the local coordinate y^s is given by the following sum :

$$\frac{\partial b_*}{\partial y^s} \quad = \quad \Gamma^i_{is}\, b_*$$

Partial derivatives of the determinant of a metric : The partial derivatives of the determinant of a metric are obtained by differentiating the relationship $g_* = (b_*)^2$ between the determinant b_* of a basis $\mathbf{B}_*$ and the determinant g_* of its metric $\mathbf{G}_* = \mathbf{B}_*^T\ \mathbf{B}_*$:

$$\frac{\partial g_*}{\partial y^s} = 2\ b_* \frac{\partial b_*}{\partial y^s} = 2\ g_*\ \Gamma^i_{is}$$

Determination of the Christoffel symbols from the metric : The knowledge of the metric $\mathbf{G}_*$ of a local coordinate system as a function of the local coordinates $y^1,...,y^n$ is sufficient for determining the Christoffel symbols of the coordinate system. For a metric coefficient $g_{ik} = \bar{\mathbf{b}}_i \cdot \bar{\mathbf{b}}_k$ one obtains :

$$\frac{\partial g_{ik}}{\partial y^s} = \frac{\partial \bar{\mathbf{b}}_i}{\partial y^s} \cdot \bar{\mathbf{b}}_k + \bar{\mathbf{b}}_i \cdot \frac{\partial \bar{\mathbf{b}}_k}{\partial y^s} = \Gamma^r_{is}\ \bar{\mathbf{b}}_r \cdot \bar{\mathbf{b}}_k + \Gamma^r_{ks}\ \bar{\mathbf{b}}_i \cdot \bar{\mathbf{b}}_r$$

Cyclic permutation of the indices yields three equations :

$$\frac{\partial g_{ik}}{\partial y^s} = g_{ik,s} = \Gamma^r_{is}\ g_{rk} + \Gamma^r_{ks}\ g_{ri}$$

$$\frac{\partial g_{ks}}{\partial y^i} = g_{ks,i} = \Gamma^r_{ki}\ g_{rs} + \Gamma^r_{si}\ g_{rk}$$

$$\frac{\partial g_{si}}{\partial y^k} = g_{si,k} = \Gamma^r_{sk}\ g_{ri} + \Gamma^r_{ik}\ g_{rs}$$

The first equation is subtracted from the sum of the second and the third equation. The result is multiplied by g^{ms}, the product is summed over s, and the symmetry of the Christoffel symbols in the lower indices is exploited.

$$g_{ks,i} + g_{si,k} - g_{ik,s} = 2\ \Gamma^r_{ik}\ g_{rs}$$

$$\Gamma^m_{ik} = \frac{1}{2}\ g^{ms}\ (g_{ks,i} + g_{is,k} - g_{ik,s})$$

Second partial derivatives of a covariant basis : Let a global basis $\mathbf{B}_*$ and the specification $x^i(y^1,...,y^n)$ of a local coordinate system with the basis $\bar{\mathbf{B}}_*$ be given in a point space $\mathbb{R}^n$. Since the global basis vectors $\mathbf{b}_s$ do not depend on the local coordinates $y^1,...,y^n$, the partial derivatives of the local basis vectors $\bar{\mathbf{b}}_i$ are formed as follows :

$$\frac{\partial \bar{\mathbf{b}}_i}{\partial y^k} = \frac{\partial^2 x^s}{\partial y^i\ \partial y^k}\ \mathbf{b}_s$$

$$\frac{\partial^2 \bar{\mathbf{b}}_i}{\partial y^k\ \partial y^m} = \frac{\partial^3 x^s}{\partial y^i\ \partial y^k\ \partial y^m}\ \mathbf{b}_s = \frac{\partial^3 x^s}{\partial y^i\ \partial y^k\ \partial y^m}\ \frac{\partial y^r}{\partial x^s}\ \bar{\mathbf{b}}_r$$

The symbols Λ^{r}_{ikm} are defined for the partial derivatives. Then the second partial derivatives of the basis vectors are given by

$$\frac{\partial^2 \overline{\mathbf{b}}_i}{\partial y^k \, \partial y^m} = \Lambda^{r}_{ikm} \, \overline{\mathbf{b}}_r$$

$$\Lambda^{r}_{ikm} := \frac{\partial^3 x^s}{\partial y^i \, \partial y^k \, \partial y^m} \, \frac{\partial y^r}{\partial x^s}$$

The symbol Λ^{r}_{ikm} is symmetric in the indices i, k, m.

Example 1 : Christoffel symbols for cylindrical coordinates

The Christoffel symbols for the cylindrical coordinates defined in Example 1 of Section 9.4.5 are determined using the general formulas :

$$\Gamma^{m}_{ik} = \frac{\partial^2 x^s}{\partial y^i \, \partial y^k} \cdot \frac{\partial y^m}{\partial x^s}$$

$$\begin{array}{|c|c|c|} \hline \frac{\partial^2 x^1}{\partial r^2} & \frac{\partial^2 x^1}{\partial r \partial \theta} & \frac{\partial^2 x^1}{\partial r \partial z} \\ \hline \frac{\partial^2 x^1}{\partial r \partial \theta} & \frac{\partial^2 x^1}{\partial \theta^2} & \frac{\partial^2 x^1}{\partial \theta \partial z} \\ \hline \frac{\partial^2 x^1}{\partial r \partial z} & \frac{\partial^2 x^1}{\partial \theta \partial z} & \frac{\partial^2 x^1}{\partial z^2} \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline 0 & -\sin\theta & 0 \\ \hline -\sin\theta & -r\cos\theta & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline \frac{\partial^2 x^2}{\partial r^2} & \frac{\partial^2 x^2}{\partial r \partial \theta} & \frac{\partial^2 x^2}{\partial r \partial z} \\ \hline \frac{\partial^2 x^2}{\partial r \partial \theta} & \frac{\partial^2 x^2}{\partial \theta^2} & \frac{\partial^2 x^2}{\partial \theta \partial z} \\ \hline \frac{\partial^2 x^2}{\partial r \partial z} & \frac{\partial^2 x^2}{\partial \theta \partial z} & \frac{\partial^2 x^2}{\partial z^2} \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline 0 & \cos\theta & 0 \\ \hline \cos\theta & -r\sin\theta & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline \frac{\partial^2 x^3}{\partial r^2} & \frac{\partial^2 x^3}{\partial r \partial \theta} & \frac{\partial^2 x^3}{\partial r \partial z} \\ \hline \frac{\partial^2 x^3}{\partial r \partial \theta} & \frac{\partial^2 x^3}{\partial \theta^2} & \frac{\partial^2 x^3}{\partial \theta \partial z} \\ \hline \frac{\partial^2 x^3}{\partial r \partial z} & \frac{\partial^2 x^3}{\partial \theta \partial z} & \frac{\partial^2 x^3}{\partial z^2} \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|} \hline \frac{\partial r}{\partial x^1} & \frac{\partial r}{\partial x^2} & \frac{\partial r}{\partial x^3} \\ \hline \frac{\partial \theta}{\partial x^1} & \frac{\partial \theta}{\partial x^2} & \frac{\partial \theta}{\partial x^3} \\ \hline \frac{\partial z}{\partial x^1} & \frac{\partial z}{\partial x^2} & \frac{\partial z}{\partial x^3} \\ \hline \end{array} = \begin{array}{|c|c|c|} \hline \cos\theta & \sin\theta & 0 \\ \hline -\frac{1}{r}\sin\theta & \frac{1}{r}\cos\theta & 0 \\ \hline 0 & 0 & 1 \\ \hline \end{array}$$

$$\Gamma^1_{ik} = \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline 0 & -r & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array} \qquad \Gamma^2_{ik} = \begin{array}{|c|c|c|}\hline 0 & \frac{1}{r} & 0 \\ \hline \frac{1}{r} & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array} \qquad \Gamma^3_{ik} = \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}$$

The determinant of the covariant basis is $b_* = r$. The partial derivatives of the basis determinant with respect to the local coordinates are determined first directly and then by summing the Christoffel symbols.

$$\begin{aligned} \frac{\partial b_*}{\partial r} &= 1 = (\Gamma^1_{11} + \Gamma^2_{21} + \Gamma^3_{31})\, r = (0 + \frac{1}{r} + 0)\, r \\ \frac{\partial b_*}{\partial \theta} &= 0 = (\Gamma^1_{12} + \Gamma^2_{22} + \Gamma^3_{32})\, r = (0 + 0 + 0)\, r \\ \frac{\partial b_*}{\partial z} &= 0 = (\Gamma^1_{13} + \Gamma^2_{23} + \Gamma^3_{33})\, r = (0 + 0 + 0)\, r \end{aligned}$$

Example 2 : Christoffel symbols for spherical coordinates

The Christoffel symbols for the spherical coordinates defined in Example 2 of Section 9.4.5 are determined using the general formulas :

$$\Gamma^m_{ik} = \frac{\partial^2 x^s}{\partial y^i \, \partial y^k} \cdot \frac{\partial y^m}{\partial x^s}$$

$$\begin{array}{|c|c|c|}\hline \frac{\partial^2 x^1}{\partial r^2} & \frac{\partial^2 x^1}{\partial r \partial \theta} & \frac{\partial^2 x^1}{\partial r \partial \beta} \\ \hline \frac{\partial^2 x^1}{\partial r \partial \theta} & \frac{\partial^2 x^1}{\partial \theta^2} & \frac{\partial^2 x^1}{\partial \theta \partial \beta} \\ \hline \frac{\partial^2 x^1}{\partial r \partial \beta} & \frac{\partial^2 x^1}{\partial \theta \partial \beta} & \frac{\partial^2 x^1}{\partial \beta^2} \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 0 & -\sin\theta \cos\beta & -\cos\theta \sin\beta \\ \hline -\sin\theta \cos\beta & -r\cos\theta \cos\beta & r\sin\theta \sin\beta \\ \hline -\cos\theta \sin\beta & r\sin\theta \sin\beta & -r\cos\theta \cos\beta \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|}\hline \frac{\partial^2 x^2}{\partial r^2} & \frac{\partial^2 x^2}{\partial r \partial \theta} & \frac{\partial^2 x^2}{\partial r \partial \beta} \\ \hline \frac{\partial^2 x^2}{\partial r \partial \theta} & \frac{\partial^2 x^2}{\partial \theta^2} & \frac{\partial^2 x^2}{\partial \theta \partial \beta} \\ \hline \frac{\partial^2 x^2}{\partial r \partial \beta} & \frac{\partial^2 x^2}{\partial \theta \partial \beta} & \frac{\partial^2 x^2}{\partial \beta^2} \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 0 & \cos\theta \cos\beta & -\sin\theta \sin\beta \\ \hline \cos\theta \cos\beta & -r\sin\theta \cos\beta & -r\cos\theta \sin\beta \\ \hline -\sin\theta \sin\beta & -r\cos\theta \sin\beta & -r\sin\theta \cos\beta \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|}\hline \frac{\partial^2 x^3}{\partial r^2} & \frac{\partial^2 x^3}{\partial r \partial \theta} & \frac{\partial^2 x^3}{\partial r \partial \beta} \\ \hline \frac{\partial^2 x^3}{\partial r \partial \theta} & \frac{\partial^2 x^3}{\partial \theta^2} & \frac{\partial^2 x^3}{\partial \theta \partial \beta} \\ \hline \frac{\partial^2 x^3}{\partial r \partial \beta} & \frac{\partial^2 x^3}{\partial \theta \partial \beta} & \frac{\partial^2 x^3}{\partial \beta^2} \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline 0 & 0 & \cos\beta \\ \hline 0 & 0 & 0 \\ \hline \cos\beta & 0 & -r\sin\beta \\ \hline \end{array}$$

$$\begin{array}{|c|c|c|}\hline \frac{\partial r}{\partial x^1} & \frac{\partial r}{\partial x^2} & \frac{\partial r}{\partial x^3} \\ \hline \frac{\partial \theta}{\partial x^1} & \frac{\partial \theta}{\partial x^2} & \frac{\partial \theta}{\partial x^3} \\ \hline \frac{\partial \beta}{\partial x^1} & \frac{\partial \beta}{\partial x^2} & \frac{\partial \beta}{\partial x^3} \\ \hline \end{array} = \begin{array}{|c|c|c|}\hline \cos\theta\ \cos\beta & \sin\theta\ \cos\beta & \sin\beta \\ \hline -\frac{1}{r}\sin\theta\ \sec\beta & \frac{1}{r}\cos\theta\sec\beta & 0 \\ \hline -\frac{1}{r}\cos\theta\ \sin\beta & -\frac{1}{r}\sin\theta\ \sin\beta & \frac{1}{r}\cos\beta \\ \hline \end{array}$$

$$\Gamma^1_{ik} = \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline 0 & -r\cos^2\beta & 0 \\ \hline 0 & 0 & -r \\ \hline \end{array}$$

$$\Gamma^2_{ik} = \begin{array}{|c|c|c|}\hline 0 & \frac{1}{r} & 0 \\ \hline \frac{1}{r} & 0 & -\tan\beta \\ \hline 0 & -\tan\beta & 0 \\ \hline \end{array}$$

$$\Gamma^3_{ik} = \begin{array}{|c|c|c|}\hline 0 & 0 & \frac{1}{r} \\ \hline 0 & \sin\beta\ \cos\beta & 0 \\ \hline \frac{1}{r} & 0 & 0 \\ \hline \end{array}$$

The determinant of the covariant basis is $b_* = r^2\cos\beta$. The partial derivatives of the basis determinant with respect to the local coordinates are determined first directly and then by summing the Christoffel symbols.

$$\frac{\partial b_*}{\partial r} = 2r\cos\beta = (\Gamma^1_{11} + \Gamma^2_{21} + \Gamma^3_{31})b_* = (0 + \tfrac{1}{r} + \tfrac{1}{r})\ r^2\cos\beta$$

$$\frac{\partial b_*}{\partial \theta} = 0 = (\Gamma^1_{12} + \Gamma^2_{22} + \Gamma^3_{32})b_* = (0 + 0 + 0)\ r^2\cos\beta$$

$$\frac{\partial b_*}{\partial \beta} = -r^2\sin\beta = (\Gamma^1_{13} + \Gamma^2_{23} + \Gamma^3_{33})b_* = (0 - \tan\beta + 0)\ r^2\cos\beta$$

9.4.7 DERIVATIVES WITH RESPECT TO LOCAL COORDINATES

Introduction : If the coordinates of a tensor field are referred to the local basis at every point of a point space $\mathbb{R}^n$, the coordinates of the tensor at different points of $\mathbb{R}^n$ cannot be compared. It is difficult to interpret partial derivatives of such coordinates. The concept of covariant derivatives of the local coordinates of a tensor field is therefore developed. These derivatives account for the effect of the variation of the basis on the coordinates of the tensor.

Tensor field of rank 1 : Let a tensor field U(**v**) of rank 1 be associated with the point space $\mathbb{R}^n$. Then every point of $\mathbb{R}^n$ is associated with its own linear mapping : Identical values of **v** generally lead to different values of U(**v**) at different points of $\mathbb{R}^n$. Thus the tensor U varies from point to point. Let **u** be the vector of the coordinates U($\mathbf{e}_i$) of the tensor at a point P in $\mathbb{R}^n$ in the global basis **E**. Generally **u** varies in the space $\mathbb{R}^n$.

Let the local coordinates of the point P be $(y^1,...,y^n)$. Let the vectors of the local basis $\mathbf{B}^*$ at the point P be $\mathbf{b}^i$. Then the contravariant coordinates of the tensor field U at the point P are the images $u^i = U(\mathbf{b}^i)$ of the local basis vectors. These coordinates are functions $u^i(y^1,...,y^n)$ of the local coordinates, since the tensor field U and the basis vectors $\mathbf{b}_i$ are functions of the local coordinates $y^1,...,y^n$:

$$\mathbf{u} = u^i\,\mathbf{b}_i$$

$\mathbf{u}$	vector of the tensor coordinates in the global basis
u^i	tensor coordinates in the local basis $\mathbf{B}^*$
$\mathbf{b}_i$	local basis vectors

Variation of the local basis : The position vector **x** of a point P in $\mathbb{R}^n$ is a function of the local coordinates $y^1,...,y^n$. The vectors $\mathbf{b}_i$ of the local basis $\mathbf{B}_*$ are the derivatives of the position vector **x** with respect to the local coordinates y^i. The contravariant basis $\mathbf{B}^*$ is obtained from the covariant basis $\mathbf{B}_*$. The variation of the local bases $\mathbf{B}_*$ and $\mathbf{B}^*$ with the local coordinates $y^1,...,y^n$ is known if the partial derivatives of the basis vectors $\mathbf{b}_i$ and $\mathbf{b}^i$ with respect to the local coordinates are known. These derivatives are determined using the Christoffel symbols :

$$\frac{\partial \mathbf{b}_i}{\partial y^k} = \Gamma^m_{ik}\,\mathbf{b}_m$$

$$\frac{\partial \mathbf{b}^i}{\partial y^k} = \Gamma^i_{km}\,\mathbf{b}^m$$

Γ^m_{ik}	Christoffel symbols

Covariant derivatives of a tensor field of rank 1 : The variation of a tensor field U with the local coordinates $y^1,...,y^n$ is known if the partial derivatives of the coordinate vector **u** with respect to the local coordinates are known. These derivatives are obtained using the product rule :

$$\frac{\partial \mathbf{u}}{\partial y^k} = \frac{\partial u^i}{\partial y^k}\,\mathbf{b}_i + u^i\,\frac{\partial \mathbf{b}_i}{\partial y^k} \qquad k=1,...,n$$

Replacing the derivatives of the basis vectors by the Christoffel symbols and the basis vectors themselves yields :

$$\frac{\partial \mathbf{u}}{\partial y^k} = \frac{\partial u^i}{\partial y^k}\,\mathbf{b}_i + u^i\,\Gamma^m_{ik}\,\mathbf{b}_m$$

By interchanging the indices i and m in the last term of the equation, the partial derivative of **u** may be expressed as follows :

$$\frac{\partial \mathbf{u}}{\partial y^k} = \left(\frac{\partial u^i}{\partial y^k} + \Gamma^i_{km}\,u^m\right)\mathbf{b}_i$$

The expression in parentheses is called the covariant derivative of u^i with respect to y^k and is designated by $u^i_{;k}$ (often also by $u^i|_k$). Using the designation $\mathbf{u}_{,k}$ introduced earlier for the partial derivative of **u** with respect to y_k, the partial derivatives of **u** may be written as follows :

$$\begin{aligned} \mathbf{u}_{,k} &= u^i_{;k}\,\mathbf{b}_i \qquad k=1,...,n \\ u^i_{;k} &= u^i_{,k} + \Gamma^i_{km}\,u^m \\ u^i_{;k} &= \text{covariant derivative of } u^i \text{ with respect to } y_k \end{aligned}$$

The change d**u** of the tensor field from the point **y** to the point **y** + d**y** is conveniently expressed by arranging the covariant derivatives of the tensor coordinates in a matrix $\mathbf{U}_{;y}$.

$$d\mathbf{u} = \mathbf{B}_*\,\mathbf{U}_{;y}\,d\mathbf{y}$$

$$\mathbf{U}_{;y} = \begin{array}{|c|c|c|} \hline u^1_{;1} & & u^1_{;n} \\ \hline & \ddots & \\ \hline u^n_{;1} & & u^n_{;n} \\ \hline \end{array}$$

Covariant derivatives of covariant tensor coordinates : The covariant derivatives of the contravariant coordinates of a tensor field of rank 1 are defined in the preceding section. The covariant derivatives of covariant coordinates $u_i(y^1,...,y^n)$ are determined analogously by considering the variation of the tensor field :

$$\mathbf{u} = u_i\,\mathbf{b}^i$$

$$\frac{\partial \mathbf{u}}{\partial y^k} = \frac{\partial u_i}{\partial y^k}\,\mathbf{b}^i + u_i\,\frac{\partial \mathbf{b}^i}{\partial y^k} = \frac{\partial u_i}{\partial y^k}\,\mathbf{b}^i - u_i\,\Gamma^i_{km}\,\mathbf{b}^m$$

The covariant derivative of the covariant coordinate u_i with respect to the local coordinate y^k is designated by $u_{i\,;\,k}$ (often also by $u_i\,|_k$). The partial derivatives of the coordinate vector **u** of the tensor field may now be expressed as follows :

$$\mathbf{u}_{,k} \;=\; u_{i\,;\,k}\,\mathbf{b}^i \qquad\qquad k=1,\dots,n$$

$$u_{i\,;\,k} \;=\; u_{i\,,\,k} - \Gamma^m_{ik}\,u_m$$

$u_{i\,;\,k}$ covariant derivative of u_i with respect to y^k

Transformation of the covariant derivatives : At a point P of a point space $\mathbb{R}^n$, let $\mathbf{B}_*$ be the basis for a local coordinate system $(y^1,\dots,y^n)$, and let $\bar{\mathbf{B}}_*$ be the basis for a local coordinate system $(z^1,\dots,z^n)$. The transformation matrix **A** with the coefficients $a^i_{.\,r}$ for the basis transformations $\bar{\mathbf{B}}_* = \mathbf{B}_*\mathbf{A}$ and $\bar{\mathbf{B}}^* = \mathbf{B}^*\bar{\mathbf{A}}^T$ is determined according to Section 9.4.5. Let the position vector of the point P be **x**, and let its local coordinates be $\mathbf{y} = (y^1,\dots,y^n)$ and $\mathbf{z} = (z^1,\dots,z^n)$. Let the position vector of a point Q near P be $\mathbf{x} + d\,\mathbf{x}$, and let its local coordinates be $\mathbf{y} + d\,\mathbf{y} = (y^1 + dy^1,\dots, y^n + dy^n)$ and $\mathbf{z} + d\,\mathbf{z} = (z^1 + dz^1,\dots,z^n + dz^n)$. Then the increment $d\mathbf{x}$ of the position vector is given by

$$d\mathbf{x} \;=\; \mathbf{B}_*\,d\mathbf{y} \;=\; \bar{\mathbf{B}}_*\,d\mathbf{z} \;=\; \mathbf{B}_*\mathbf{A}\,d\mathbf{z} \quad\Rightarrow\quad d\mathbf{y} \;=\; \mathbf{A}\,d\mathbf{z}$$

Let the coordinates of a tensor field of rank 1 be u^i and u_i for the coordinate system $(y^1,\dots,y^n)$ and $\bar{u}^i$, $\bar{u}_i$ for the coordinate system $(z^1,\dots,z^n)$. The covariant derivatives of the coordinates are arranged in the following matrices :

$\mathbf{U}^o_{;\,y}$ covariant derivatives $u^i_{;\,k}$ of u^i with respect to y^k

$\mathbf{U}_{*\,;\,y}$ covariant derivatives $u_{i\,;\,k}$ of u_i with respect to y^k

$\bar{\mathbf{U}}^o_{;\,z}$ covariant derivatives $\bar{u}^i_{;\,k}$ of $\bar{u}^i$ with respect to z^k

$\bar{\mathbf{U}}_{*\,;\,y}$ covariant derivatives $\bar{u}_{i\,;\,k}$ of $\bar{u}_i$ with respect to z^k

The increment $d\mathbf{u}$ of the tensor field from point P to point Q may be determined using any of these matrices :

$$\begin{aligned}
d\mathbf{u} &= u^i_{;\,k}\,\mathbf{b}_i\,dy^k &&= \mathbf{B}_*\,\mathbf{U}^o_{;\,y}\;d\mathbf{y}\\
d\mathbf{u} &= u_{i\,;\,k}\,\mathbf{b}^i\,dy^k &&= \mathbf{B}^*\,\mathbf{U}_{*\,;\,y}\,d\mathbf{y}\\
d\mathbf{u} &= \bar{u}^i_{;\,k}\,\bar{\mathbf{b}}_i\,dz^k &&= \bar{\mathbf{B}}_*\,\bar{\mathbf{U}}^o_{;\,z}\;d\mathbf{z}\\
d\mathbf{u} &= \bar{u}_{i\,;\,k}\,\mathbf{b}^i\,dz^k &&= \bar{\mathbf{B}}^*\,\bar{\mathbf{U}}_{*\,;\,z}\,d\mathbf{z}
\end{aligned}$$

Substituting the basis transformations and the relationship $d\mathbf{z} = \bar{\mathbf{A}}\,d\mathbf{y}$ leads to the following expressions for $d\mathbf{u}$:

$$\begin{aligned}
d\mathbf{u} &= \mathbf{B}_*\,\mathbf{U}^o_{;\,y}\;d\mathbf{y} &&= \mathbf{B}_*\,\mathbf{A}\,\bar{\mathbf{U}}^o_{;\,z}\;\bar{\mathbf{A}}\,d\mathbf{y}\\
d\mathbf{u} &= \mathbf{B}^*\,\mathbf{U}_{*\,;\,y}\,d\mathbf{y} &&= \mathbf{B}^*\,\bar{\mathbf{A}}^T\,\bar{\mathbf{U}}_{*\,;\,z}\;\bar{\mathbf{A}}\,d\mathbf{y}
\end{aligned}$$

These relationships hold for arbitrary increments d**y**. Hence the following transformation rules hold for the covariant derivatives of a tensor field of rank 1 :

$$\begin{array}{lllcl} \mathbf{U}^{o}_{;y} & = & \mathbf{A}\,\bar{\mathbf{U}}^{o}_{;z}\,\bar{\mathbf{A}} & \Leftrightarrow & u^{i}_{;k} = a^{i}_{.r}\,\bar{a}^{s}_{.k}\,\bar{u}^{r}_{;s} \\ \mathbf{U}_{*;y} & = & \bar{\mathbf{A}}^{T}\,\bar{\mathbf{U}}_{*;z}\,\bar{\mathbf{A}} & \Leftrightarrow & u_{;k} = \bar{a}^{r}_{.i}\,\bar{a}^{s}_{.k}\,\bar{u}_{r;s} \end{array}$$

The covariant derivatives transform like the coordinates of a dyad (see Section 9.3.6). Hence the covariant derivatives are the coordinates of a tensor.

Tensor field of rank 2 : Let the local coordinates of a point P in a point space $\mathbb{R}^n$ be $(y^1,...,y^n)$. Let the dual local bases at the point P be $\mathbf{B}_* = (\mathbf{b}_1,...,\mathbf{b}_n)$ and $\mathbf{B}^* = (\mathbf{b}^1,...,\mathbf{b}^n)$. If the coordinates of the basis vector $\mathbf{b}_i$ are designated by b_{ri}, the vector $\mathbf{e}_r$ of the canonical basis of the point space $\mathbb{R}^n$ may be expressed as follows :

$$\mathbf{B}^*\mathbf{B}_*^T = \mathbf{E} \quad \Rightarrow \quad \mathbf{e}_r = \mathbf{b}^i\, b_{ri}$$

Let the coordinates of a tensor field u of rank 2 be expressed as functions of the local coordinates. The coordinates $u(\mathbf{e}_r, \mathbf{e}_s)$ in the canonical basis **E** are arranged in a matrix **U**, the coordinates $u^{ik} = u(\mathbf{b}^i, \mathbf{b}^k)$ in the local basis are arranged in a matrix $\mathbf{U}^*$. The linearity of the vector mapping u leads to the following transformation rule for the tensor coordinates at the point P :

$$\begin{array}{lll} u(\mathbf{e}_r, \mathbf{e}_s) & = & u(\mathbf{b}^i\, b_{ri}, \mathbf{b}^k\, b_{sk}) \; = \; b_{ri}\, b_{sk}\, u^{ik} \\ \mathbf{U} & = & \mathbf{B}_*\mathbf{U}^*\mathbf{B}_*^T \; = \; u^{ik}\,\mathbf{b}_i\,\mathbf{b}_k^T \end{array}$$

Covariant derivatives of a tensor field of rank 2 : The partial derivatives of the coordinate matrix **U** of the tensor field with respect to the local coordinates are obtained using the product rule :

$$\begin{array}{lll} \dfrac{\partial \mathbf{U}}{\partial y^m} & = & \dfrac{\partial u^{ik}}{\partial y^m}\,\mathbf{b}_i\mathbf{b}_k^T \; + \; u^{ik}\,\dfrac{\partial \mathbf{b}_i}{\partial y^m}\,\mathbf{b}_k^T \; + \; u^{ik}\,\mathbf{b}_i\,\dfrac{\partial \mathbf{b}_k^T}{\partial y^m} \\ & = & u^{ik}_{,m}\,\mathbf{b}_i\mathbf{b}_k^T \; + \; u^{ik}\,\Gamma^{r}_{im}\,\mathbf{b}_r\,\mathbf{b}_k^T \; + \; u^{ik}\,\Gamma^{s}_{km}\,\mathbf{b}_i\,\mathbf{b}_s^T \end{array}$$

The covariant derivatives of the contravariant coordinates u^{ik} with respect to the local coordinates y^m are defined in analogy with the covariant derivatives of the contravariant coordinates of a tensor field of rank 1 and are designated by $u^{ik}_{;m}$:

$$\begin{array}{lll} \mathbf{U}_{,m} & = & u^{ik}_{;m}\,\mathbf{b}_i\,\mathbf{b}_k^T \\ u^{ik}_{;m} & = & u^{ik}_{,m} + \Gamma^{i}_{ms}\,u^{sk} + \Gamma^{k}_{ms}\,u^{is} \end{array}$$

$u^{ik}_{;m}$ covariant derivative of u^{ik} with respect to y^m

The covariant derivatives of the covariant coordinates and the mixed coordinates of a tensor field of rank 2 are obtained analogously :

$$\begin{array}{lll} u_{ik;m} & = & u_{ik,m} - \Gamma^{s}_{im}\,u_{sk} - \Gamma^{s}_{km}\,u_{is} \\ u^{i}_{.k;m} & = & u^{i}_{.k,m} + \Gamma^{i}_{ms}\,u^{s}_{.k} - \Gamma^{s}_{km}\,u^{i}_{.s} \\ u_{i.;m}^{\;\;k} & = & u_{i.,m}^{\;\;k} - \Gamma^{s}_{im}\,u_{s.}^{\;k} + \Gamma^{k}_{ms}\,u_{i.}^{\;s} \end{array}$$

Tensor field of rank 2 with mixed bases : Let the global coordinates of a point P in a point space $\mathbb{R}^n$ be $(x^1,...,x^n)$. Let a local coordinate system $x^i(y^1,...,y^n)$ with the dual local bases $\mathbf{B}_* = (\mathbf{b}_1,...,\mathbf{b}_n)$ and $\mathbf{B}^* = (\mathbf{b}^1,...,\mathbf{b}^n)$ and a local coordinate system $x^i(z^1,...,z^n)$ with the dual local bases $\bar{\mathbf{B}}_* = (\bar{\mathbf{b}}_1,...,\bar{\mathbf{b}}_n)$ and $\bar{\mathbf{B}}^* = (\bar{\mathbf{b}}^1,...,\bar{\mathbf{b}}^n)$ be defined in $\mathbb{R}^n$. Let the transformation matrix between the local bases be $\mathbf{A}$ with the coefficients $a^i_{.k}$:

$$d\mathbf{x} = \mathbf{b}^i \, dy^i = \bar{\mathbf{b}}_k \, dz^k$$

$$dy^i = a^i_{.k} \, dz^k$$

If the coordinates of $\mathbf{b}_i$ are designated by b_{ri} and those of $\bar{\mathbf{b}}_k$ by $\bar{b}_{sk}$, the vectors $\mathbf{e}_r$ and $\mathbf{e}_s$ of the canonical basis $\mathbf{E}$ of the point space $\mathbb{R}^n$ may be expressed as follows :

$$\mathbf{B}^* \mathbf{B}_*^T = \mathbf{E} \quad \Rightarrow \quad \mathbf{e}_r = \mathbf{b}^i \, b_{ri}$$

$$\bar{\mathbf{B}}^* \bar{\mathbf{B}}_*^T = \mathbf{E} \quad \Rightarrow \quad \mathbf{e}_s = \bar{\mathbf{b}}^k \, \bar{b}_{sk}$$

Let the coordinates of a tensor field u of rank 2 be expressed as functions of the local coordinates $z^1,...,z^n$. The coordinates $u(\mathbf{e}_r, \mathbf{e}_s)$ in the canonical basis $\mathbf{E}$ are arranged in a matrix $\mathbf{U}$. Let the first index of the coordinates $u^{ik} = u(\mathbf{b}^i, \bar{\mathbf{b}}^k)$ be referred to the basis $\mathbf{B}^*$, the second index to the basis $\bar{\mathbf{B}}^*$. The coordinates u^{ik} are arranged in the matrix $\mathbf{U}^*$. The linearity of the vector mapping u leads to the following transformation rule for the tensor coordinates at the point P :

$$u(\mathbf{e}_r, \mathbf{e}_s) = u(\mathbf{b}^i \, b_{ri} \, , \, \bar{\mathbf{b}}^k \, \bar{b}_{sk}) = b_{ri} \, b_{sk} \, u^{ik}$$

$$\mathbf{U} = \mathbf{B}_* \mathbf{U}^* \bar{\mathbf{B}}_*^T = u^{ik} \, \mathbf{b}_i \, \bar{\mathbf{b}}_k^T$$

Covariant derivatives of a tensor field with mixed bases : The partial derivatives of the coordinate matrix $\mathbf{U}$ of a tensor field of rank 2 with respect to the local coordinates are obtained using the product rule. The differentiation may be performed with respect to the local coordinates $y^1,...,y^n$ or with respect to the local coordinates $z^1,...,z^n$. The partial derivatives of the basis vectors $\bar{\mathbf{b}}_k$ with respect to z^m are expressed in terms of the Christoffel symbols $\bar{\Gamma}$ of the coordinates $z^1,...,z^n$. The partial derivatives of the basis vectors $\mathbf{b}_i$ with respect to the local coordinates y^m are expressed in terms of the Christoffel symbols Γ of the coordinates $y^1,...,y^n$.

$$\frac{\partial \mathbf{U}}{\partial z^m} = \frac{\partial u^{ik}}{\partial z^m} \mathbf{b}_i \, \bar{\mathbf{b}}_k^T + u^{ik} \frac{\partial \mathbf{b}_i}{\partial y^s} \frac{\partial y^s}{\partial z^m} \bar{\mathbf{b}}_k^T + u^{ik} \, \mathbf{b}_i \frac{\partial \bar{\mathbf{b}}_k^T}{\partial z^m}$$

$$= u^{ik}_{\ ,m} \, \mathbf{b}_i \, \bar{\mathbf{b}}_k^T + u^{ik} \, \Gamma^r_{is} \, a^s_{.m} \, \mathbf{b}_r \, \bar{\mathbf{b}}_k^T + u^{ik} \, \bar{\Gamma}^r_{km} \, \mathbf{b}_i \, \bar{\mathbf{b}}_r^T$$

The covariant derivatives of the contravariant coordinates u^{ik} of a tensor field of rank 2 with mixed bases with respect to the local coordinates z^m of the basis $\bar{\mathbf{B}}_*$ are defined in analogy with the case of tensor fields of rank 2 with identical bases for the two indices and are designated by $u^{ik}_{\ ;m}$.

$$\mathbf{U}_{,m} = u^{ik}_{;m}\, \mathbf{b}_i\, \mathbf{b}_k^T$$

$$u^{ik}_{;m} = u^{ik}_{,m} + \Gamma^i_{rs}\, a^s_{.m} u^{rk} + \bar{\Gamma}^k_{rm}\, u^{ir}$$

Γ^i_{rs} Christoffel symbols for the basis $(y^1,...,y^n)$

$\bar{\Gamma}^k_{rm}$ Christoffel symbols for the basis $(z^1,...,z^n)$

$a^s_{.m}$ coefficients of the transformation matrix $\mathbf{A}$ in $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$

Covariant derivatives of the metric : The coefficients of the local metric $\mathbf{G}_*$ are generally different at neighboring points P and Q. The partial derivatives of the metric coefficients are therefore non-zero. The covariant derivatives of the metric coefficients differ from the partial derivatives in that the effect of the change of basis from P to Q is accounted for (reference to the basis at the point P). Thus the change in the metric from P to Q is cancelled by the construction of the covariant derivatives : The covariant derivatives of the metric coefficients are zero.

The same result follows formally from the general formula for the covariant derivatives for tensor fields of rank 2 with $\mathbf{b}^i_{,m} = -\Gamma^i_{ms}\, \mathbf{b}^s$:

$$g^{ik}_{;m} = g^{ik}_{,m} + \Gamma^i_{ms}\, g^{sk} + \Gamma^k_{ms}\, g^{is}$$

$$g^{ik}_{,m} = \frac{\partial}{\partial y^m}(\mathbf{b}^i \cdot \mathbf{b}^k) = \frac{\partial \mathbf{b}^i}{\partial y^m} \cdot \mathbf{b}^k + \mathbf{b}^i \cdot \frac{\partial \mathbf{b}^k}{\partial y^m}$$

$$g^{ik}_{,m} = -\Gamma^i_{ms}\, g^{sk} - \Gamma^k_{ms}\, g^{is}$$

$$g^{ik}_{;m} = 0$$

Tensor density : The coordinates $u^i(y^1,...,y^n)$ of a tensor field of rank 1 may be written as products of the covariant basis determinant $b_* = \det \mathbf{B}_*$ with functions $w^i(y^1,...,y^n)$. Since b_* is the volume of the basis $\mathbf{B}_*$, the functions w^i are called the coordinates of a tensor density. Since the determinant b_* is not a scalar (tensor of rank 0), the tensor density is not a tensor. In formulating physical problems one often makes use of the property that the sum of the partial derivatives of the coordinates u^i of the tensor field may be expressed in terms of the tensor density as follows. The proof relies on the property $\frac{\partial b_*}{\partial y^i} = \Gamma^m_{mi}\, b_*$ proved in Section 9.4.6 :

$$u^i = b_*\, w^i$$

$$\frac{\partial u^i}{\partial y^i} = \frac{\partial b_*}{\partial y^i} w^i + b_* \frac{\partial w^i}{\partial y^i} = b_*(\Gamma^m_{mi}\, w^i + w^i_{,i})$$

Rules of calculation for covariant derivatives : The covariant derivatives of a tensor of rank r + s form a tensor of rank r + s + 1. The covariant derivatives with respect to a local coordinate y^m are formed in analogy with the derivatives for covariant and contravariant coordinates of a tensor field of rank 1.

$$\begin{aligned} u^{i_1\dots i_r}_{k_1\dots k_s;m} &= u^{i_1\dots i_r}_{k_1\dots k_s,m} + \Gamma^{i_1}_{tm}\, u^{t\, i_2\dots i_r}_{k_1\dots k_s} + \Gamma^{i_2}_{tm}\, u^{i_1\, t\, i_3\dots i_r}_{k_1\dots k_s} + \dots \\ &\quad - \Gamma^{t}_{k_1 m}\, u^{i_1\dots i_r}_{t k_2\dots k_s} - \Gamma^{t}_{k_2 m}\, u^{i_1\dots i_r}_{k_1\, t\, k_3\dots k_s} - \dots \end{aligned}$$

Sum rule : The covariant derivative of the sum of two tensors is the sum of the covariant derivatives of these tensors :

$$(u^{i_1\dots i_r}_{k_1\dots k_s} + w^{i_1\dots i_r}_{k_1\dots k_s})_{;m} = u^{i_1\dots i_r}_{k_1\dots k_s;m} + w^{i_1\dots i_r}_{k_1\dots k_s;m}$$

Product rule : The covariant derivative of the product of two tensors is determined as follows :

$$(u^{i}_{k}\, w^{r}_{s})_{;m} = u^{i}_{k;m}\, w^{r}_{s} + u^{i}_{k}\, w^{r}_{s;m}$$

Contraction rule : The contraction and the covariant derivative of a tensor commute :

$$u^{i_1\dots i_r}_{k_1\dots k_s} := w^{i_1\dots i_r\, t}_{k_1\dots k_s t} \qquad t = 1,\dots,n$$

$$u^{i_1\dots i_r}_{k_1\dots k_s;m} = w^{i_1\dots i_r\, 1}_{k_1\dots k_s 1;m} + \dots + w^{i_1\dots i_r\,(n)}_{k_1\dots k_s(n);m}$$

Covariant derivatives of the ε-tensor : The coordinates of the ε-tensor are first considered in the global canonical basis E. The Christoffel symbols for this basis are zero, since the metric is constant. The partial derivatives of the coordinates of the ε-tensor in the basis E are also zero. Hence the covariant derivatives of the ε-tensor for this choice of basis are zero. These derivatives are the coordinates of a tensor. The coordinates of this tensor in an arbitrary basis are obtained by a linear transformation, and are therefore also zero.

$$\varepsilon^{i_1\dots i_n}{}_{;m} = 0$$

$$\varepsilon_{i_1,\dots,i_n;m} = 0$$

Second covariant derivatives of a vector field : Let the local coordinates of two neighboring points P and Q in a point space $\mathbb{R}^n$ be given by $(y^1, ..., y^n)$ and $(y^1 + dy^1, ..., y^n + dy^n)$, respectively. Let the vectors of the local basis at the point P be $\mathbf{b}_1, ..., \mathbf{b}_n$. Let the coordinates u^i of a vector field be referred to the local basis at the considered point. The tensor $\mathbf{u}$(Q) at the point Q is determined using the series expansion at the point P :

$$\mathbf{u}(Q) \quad = \quad \mathbf{u}(P) + \frac{\partial \mathbf{u}}{\partial y^k} dy^k + \frac{1}{2} \frac{\partial^2 \mathbf{u}}{\partial y^k \partial y^m} dy^k dy^m + ...$$

The first derivatives of the vector field are expressed in terms of the Christoffel symbols Γ^r_{ik} :

$$\frac{\partial \mathbf{u}}{\partial y^k} \quad = \quad \frac{\partial u^i}{\partial y^k} \mathbf{b}_i + u^i \frac{\partial \mathbf{b}_i}{\partial y^k} \quad = \quad \frac{\partial u^i}{\partial y^k} \mathbf{b}_i + \Gamma^r_{ik} u^i \mathbf{b}_r$$

The second derivatives of the vector field are expressed in terms of the additional symbols Λ^r_{ikm} from Section 9.4.6 :

$$\begin{aligned} \frac{\partial^2 \mathbf{u}}{\partial y^k \partial y^m} \quad &= \quad \frac{\partial^2 u^i}{\partial y^k \partial y^m} \mathbf{b}_i + \frac{\partial u^i}{\partial y^k} \frac{\partial \mathbf{b}_i}{\partial y^m} + \frac{\partial u^i}{\partial y^m} \frac{\partial \mathbf{b}_i}{\partial y^k} + u^i \frac{\partial^2 \mathbf{b}_i}{\partial y^k \partial y^m} \\ &= \quad u^i_{,km} \mathbf{b}_i + \Gamma^r_{im} u^i_{,k} \mathbf{b}_r + \Gamma^r_{ik} u^i_{,m} \mathbf{b}_r + \Lambda^r_{ikm} u^i \mathbf{b}_r \end{aligned}$$

If the vector at the point Q is expressed in the covariant form

$$\mathbf{u}(Q) \quad = \quad \mathbf{u}(P) + u^i_{;k} dy^k \mathbf{b}_i + u^i_{;m} dy^m \mathbf{b}_i + \frac{1}{2} u^i_{;km} dy^k dy^m \mathbf{b}_i$$

then the covariant second derivatives of the coordinates of the vector field are obtained as

$$u^i_{;km} \quad = \quad u_{,km} + \Gamma^i_{sm} u^s_{,k} + \Gamma^i_{sk} \; u^s_{,m} + \Lambda^r_{skm} u^s$$

Since the symbols Λ^i_{skm} are symmetric in the indices k and m, the covariant second derivatives are also symmetric in the indices k and m.

Example 1 : Covariant derivatives with respect to cylindrical coordinates

A cylindrical coordinate system with the basis vectors $\mathbf{a}_i$ for the point space $\mathbb{R}^3$ is defined in Example 1 of Section 9.4.5. The Christoffel symbols for this basis are derived in Example 1 of Section 9.4.6. Let the contravariant coordinates of a tensor field $\mathbf{u}$ in $\mathbb{R}^3$ be given as follows :

$$\begin{array}{|c|}\hline u^1 \\ \hline u^2 \\ \hline u^3 \\ \hline \end{array} = \begin{array}{|c|}\hline r\cos\theta \\ \hline \sin\theta \\ \hline z \\ \hline \end{array} \qquad \begin{aligned} y^1 &= r \\ y^2 &= \theta \\ y^3 &= z \end{aligned}$$

The covariant derivatives of this field are given by :

$$\begin{aligned}
u^1_{;1} &= u^1_{,1} + \Gamma^1_{11}u^1 + \Gamma^1_{12}u^2 + \Gamma^1_{13}u^3 \\
&= \cos\theta + 0 + 0 + 0 = \cos\theta \\
u^1_{;2} &= u^1_{,2} + \Gamma^1_{21}u^1 + \Gamma^1_{22}u^2 + \Gamma^1_{23}u^3 \\
&= -r\sin\theta + 0 - r\sin\theta + 0 = -2r\sin\theta \\
u^2_{;1} &= u^2_{,1} + \Gamma^2_{11}u^1 + \Gamma^2_{12}u^2 + \Gamma^2_{13}u^3 \\
&= 0 + 0 + \frac{1}{r}\sin\theta + 0 = \frac{1}{r}\sin\theta \\
u^2_{;2} &= u^2_{,2} + \Gamma^2_{21}u^1 + \Gamma^2_{22}u^2 + \Gamma^2_{23}u^3 \\
&= \cos\theta + \frac{1}{r}(r\cos\theta) + 0 + 0 = 2\cos\theta \\
u^3_{;3} &= u^3_{,3} + \Gamma^3_{31}u^1 + \Gamma^3_{32}u^2 + \Gamma^3_{33}u^3 \\
&= 1 + 0 + 0 + 0 = 1
\end{aligned}$$

$$u^1_{;3} = u^2_{;3} = u^3_{;1} = u^3_{;2} = 0$$

9.4.8 TENSOR INTEGRALS

Introduction : Mathematical descriptions of physical states often contain integrals of tensor fields and their derivatives over lines, surfaces and volumes in the euclidean space $\mathbb{R}^3$. Infinitesimal line elements, surface elements and volume elements are defined to formulate such integrals. Scalar and vector integrals of scalar fields and of tensor fields of rank 1 are defined in terms of these elements. Integrals of general tensor fields may be defined analogously.

Tensor integrals are defined for global and for local coordinate systems. Which form is more convenient must be decided on the merits of the individual case. The global coordinate system simplifies the analytic integration, since the metric coefficients are constant, in contrast to the case of a local coordinate system. The integration range may, however, often be expressed more conveniently in local coordinates than in global coordinates.

Global coordinate systems : A vector space is associated with the point space $\mathbb{R}^3$ as described in Section 9.4.2. The global basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ of this vector space is the same at all points of $\mathbb{R}^3$. Let the global coordinates of a point P of $\mathbb{R}^3$ be (x^1, x^2, x^3). Then the position vector $\mathbf{x}$ of P is a linear combination of the basis vectors :

$$\mathbf{x} = x^1 \mathbf{b}_1 + x^2 \mathbf{b}_2 + x^3 \mathbf{b}_3 = x^i \mathbf{b}_i$$

The metric coefficients g_{im} of the global basis are the scalar products of the basis vectors, and are therefore the same at all points of $\mathbb{R}^3$. If the indices $<i, k, m>$ form an even permutation, the cross product of the basis vectors $\mathbf{b}_i$, $\mathbf{b}_k$ is equal to the dual basis vector $\mathbf{b}^m$ scaled by $b_* = \det \mathbf{B}_*$, and is therefore also the same at all points of $\mathbb{R}^3$.

$$\begin{aligned} \mathbf{b}_i \cdot \mathbf{b}_k &= g_{ik} && i, k, m \in \{1, 2, 3\} \\ \mathbf{b}_i \times \mathbf{b}_k &= b_* \, \mathbf{b}^m && i \neq k \neq m \\ \mathbf{b}_i \times \mathbf{b}_i &= \mathbf{0} && \operatorname{sgn} <i, k, m> = 1 \end{aligned}$$

In the special case that the canonical basis E with the orthonormal basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ and the metric coefficients (Kronecker symbols) δ_{im} is chosen as a global basis, these relationships take the following simple form :

$$\begin{aligned} \mathbf{e}_i \cdot \mathbf{e}_k &= \delta_{ik} && i, k, m \in \{1, 2, 3\} \\ \mathbf{e}_i \times \mathbf{e}_k &= \mathbf{e}^m = \mathbf{e}_m && i \neq k \neq m \\ \mathbf{e}_i \times \mathbf{e}_i &= \mathbf{0} && \operatorname{sgn} <i, k, m> = 1 \end{aligned}$$

Line : A subset of points of $\mathbb{R}^3$ whose global coordinates x^i are continuous functions $x^i(t)$ of a line parameter $t \in \mathbb{R}$ is called a line in the space $\mathbb{R}^3$ and is designated by L. A subset of the points of L whose parameters lie in a range $t_1 \leq t \leq t_2$ is called a line fragment and is designated by L_s. A subset of the points of L whose parameters lie in an infinitesimal range $t_1 \leq t \leq t_1 + dt$ is called a line element and is designated by dL.

$$\begin{aligned}
L &= \{\mathbf{x}(t) \mid \mathbf{x} = x^i(t)\,\mathbf{b}_i \quad \wedge \quad t \in \mathbb{R}\} \\
L_s &= \{\mathbf{x}(t) \mid \mathbf{x} = x^i(t)\,\mathbf{b}_i \quad \wedge \quad t_1 \leq t \leq t_2\} \\
dL &= \{\mathbf{x}(t) \mid \mathbf{x} = x^i(t)\,\mathbf{b}_i \quad \wedge \quad t_1 \leq t \leq t_1 + dt\}
\end{aligned}$$

In the vector space associated with $\mathbb{R}^3$, the endpoints $\mathbf{x}(t_1)$ and $\mathbf{x}(t_1 + dt)$ of a line element define an incremental vector, which is designated by $d\mathbf{x}$.

$$\begin{aligned}
d\mathbf{x} &= dx^i\,\mathbf{b}_i \\
dx^i &= x^i(t_1 + dt) - x^i(t_1)
\end{aligned}$$

The differentials dx^i of the global coordinates are expressed as functions of the differential dt of the line parameter. The derivatives of the global coordinates x^i with respect to the line parameter t are designated by $\dot{x}^i$. The length ds of the line element dL is defined to be the magnitude of the vector $d\mathbf{x}$.

$$\begin{aligned}
dx^i &= \frac{dx^i}{dt}\,dt = \dot{x}^i\,dt \qquad \text{with} \qquad \dot{x}^i = \frac{dx^i}{dt} \\
(ds)^2 &= d\mathbf{x} \cdot d\mathbf{x} = g_{im}\,dx^i\,dx^m \\
ds &= \sqrt{g_{im}\,\dot{x}^i\,\dot{x}^m}\;dt
\end{aligned}$$

Distance along a line : The position vector of a point on a line is a function $\mathbf{x}(t)$ of the line parameter t. The increment $d\mathbf{x}$ of the position vector of a continuous line is represented in the global basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ using the derivatives $\dot{x}^i$ of the global coordinates with respect to the line parameter :

$$\begin{aligned}
d\mathbf{x} &= \mathbf{c}\,dt \\
\mathbf{c}(t) &= c^i\,\mathbf{b}_i \qquad \text{with} \qquad c^i(t) = \frac{dx^i}{dt}
\end{aligned}$$

An arbitrary point A of the line with the line parameter t_0 is assigned the distance $s = 0$. Then for any point with the line parameter t_1 the distance s_1 from the point A may be determined :

$$\begin{aligned}
s_1 &= \int_0^{s_1} ds = \int_{t_0}^{t_1} \sqrt{\mathbf{c} \cdot \mathbf{c}}\;dt \\
s_1 &= \int_{t_0}^{t_1} \sqrt{g_{im}\,c^i\,c^m}\;dt
\end{aligned}$$

Points on the line may thus alternatively be specified by the line parameter t or by the distance s along the line. If the distance s is used, the increment d**x** of the position vector may be expressed in terms of the derivatives of the global coordinates with respect to the distance :

$$d\mathbf{x} = \mathbf{t}\, ds$$

$$\mathbf{t}(s) = w^i\, \mathbf{b}_i \qquad \text{with} \qquad w^i = \frac{dx^i}{ds}$$

Line integrals : Let the tensor field to be integrated be given as a function of the position vector **x**. Since the position vector can be expressed as a function of the distance s along the line or as a function of the line parameter t, the tensor field can also be specified as a function of s or t.

scalar field : $u = u(s)$ or $u = u(t)$

vector field : $\mathbf{u} = u^i(s)\, \mathbf{b}_i$ or $\mathbf{u} = u^i(t)\, \mathbf{b}_i$

The line integral is taken over a continuous line fragment L in the space $\mathbb{R}^3$. Like the tensor field, the incremental line element d**x** is expressed either as a function **t**(s) ds of the distance s or as a function **c**(t) dt of the line parameter t. The value of the line integral is a scalar h if the integrand is a scalar. Otherwise the result is a vector **h**.

scalar line integral of a scalar field

$$h = \int_{s_1}^{s_2} u(s)\, ds = \int_{t_1}^{t_2} u(t)\, \sqrt{\mathbf{c}\cdot\mathbf{c}}\; dt$$

vector line integral of a scalar field

$$\mathbf{h} = \int_{L} u\, d\mathbf{x} = \int_{s_1}^{s_2} u(s)\, \mathbf{t}\, ds = \int_{t_1}^{t_2} u(t)\, \mathbf{c}\, dt$$

$$\mathbf{h} = h^i\, \mathbf{b}_i$$

$$h^i = \int_{s_1}^{s_2} u(s)\, \frac{dx^i}{ds}\, ds = \int_{t_1}^{t_2} u(t)\, \frac{dx^i}{dt}\, dt$$

scalar line integral of a vector field

$$h = \int_{L} \mathbf{u}\cdot d\mathbf{x} = \int_{s_1}^{s_2} \mathbf{u}\cdot\mathbf{t}\, ds = \int_{t_1}^{t_2} \mathbf{u}\cdot\mathbf{c}\, dt$$

$$h = \int_{s_1}^{s_2} u^i(s)\, w^m(s)\, g_{im}\, ds = \int_{t_1}^{t_2} u^i(t)\, c^m(t)\, g_{im}\, dt$$

vector line integral of a vector field

$$\mathbf{h} = \int_{L} \mathbf{u} \times d\mathbf{x} = \int_{s_1}^{s_2} \mathbf{u} \times \mathbf{t}\, ds = \int_{t_1}^{t_2} \mathbf{u} \times \mathbf{c}\, dt$$

$$\mathbf{h} = h_i\, \mathbf{b}^i$$

$$h_i = \int_{s_1}^{s_2} \varepsilon_{ikm}\, u^k(s)\, w^m(s)\, ds = \int_{t_1}^{t_2} \varepsilon_{ikm}\, u^k\,(t)\, c^m(t)\, dt$$

Example 1 : Line integrals

A global coordinate system $(\mathbf{0}, \mathbf{E})$ with the orthonormal basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is chosen in the point space $\mathbb{R}^3$. The points $\mathbf{x} = x^i\, \mathbf{e}_i$ of a line in $\mathbb{R}^3$ are specified by a line parameter t :

$$\mathbf{x} = t^2\, \mathbf{e}_1 + \mathbf{e}_2 + t\, \mathbf{e}_3$$

$$\mathbf{c} = \frac{d\mathbf{x}}{dt} = 2t\,\mathbf{e}_1 + \mathbf{e}_3$$

The scalar and vector line integrals of the scalar field $u = t^3$ over the line fragment $0 \le t \le 1$ are calculated :

$$h = \int_0^1 u \sqrt{\mathbf{c} \cdot \mathbf{c}}\; dt = \int_0^1 \sqrt{4t^2 + 1}\; t^3\, dt = 0.4742$$

$$\mathbf{h} = \int_0^1 u\, \mathbf{c}\, dt = \int_0^1 t^3\, (2t\,\mathbf{e}_1 + \mathbf{e}_3)\, dt = 0.40\, \mathbf{e}_1 + 0.25\, \mathbf{e}_2$$

The scalar and vector line integrals of the vector field $\mathbf{u} = t^2\, \mathbf{e}_1 + t\,\mathbf{e}_2 + t^3\mathbf{e}_3$ over the line fragment $0 \le t \le 1$ are calculated :

$$h = \int_0^1 \mathbf{u} \cdot \mathbf{c}\, dt = \int_0^1 (t^2\, \mathbf{e}_1 + t\,\mathbf{e}_2 + t^3\mathbf{e}_3) \cdot (2t\,\mathbf{e}_1 + \mathbf{e}_3)\, dt$$

$$= \int_0^1 3\, t^3\, dt = 0.7500$$

$$\mathbf{h} = \int_0^1 \mathbf{u} \times \mathbf{c}\, dt = h^i\, \mathbf{e}_i$$

$$h^1 = \int_0^1 (u_2 c_3 - u_3 c_2)\, dt = \int_0^1 (t - 0) dt = 0.5000$$

$$h^2 = \int_0^1 (u_3 c_1 - u_1 c_3)\, dt = \int_0^1 (2t^4 - t^2) dt = 0.0667$$

$$h^3 = \int_0^1 (u_1 c_2 - u_2 c_1)\, dt = \int_0^1 (0 - 2t^2) dt = -0.6667$$

Surface : A subset of the points of $\mathbb{R}^3$ whose global coordinates are continuous functions $x^i(s,t)$ of two surface parameters $s, t \in \mathbb{R}$ is called a surface in the space $\mathbb{R}^3$ and is designated by F. A subset of the points of F whose parameter lie in the ranges $s_1 \le s \le s_2$ and $t_1 \le t \le t_2$ is called a surface fragment and is designated by F_s. A subset of the points of F whose parameters lie in infinitesimal ranges $s_1 \le s \le s_1 + ds$ and $t_1 \le t \le t_1 + dt$ is called a surface element and is designated by dF.

$$
\begin{aligned}
x^i &= x^i(s,t) \\
F &= \{\mathbf{x}(s,t) \mid \mathbf{x} = x^i\,\mathbf{b}_i \;\wedge\; s,t \in \mathbb{R}\} \\
F_s &= \{\mathbf{x}(s,t) \mid \mathbf{x} = x^i\,\mathbf{b}_i \;\wedge\; s_1 \le s \le s_2 \;\wedge\; t_1 \le t \le t_2\} \\
dF &= \{\mathbf{x}(s,t) \mid \mathbf{x} = x^i\,\mathbf{b}_i \;\wedge\; s_1 \le s \le s_1 + ds \;\wedge\; t_1 \le t \le t_1 + dt\}
\end{aligned}
$$

In the vector space associated with $\mathbb{R}^3$, the points $\mathbf{x}(s_1,t_1)$ and $\mathbf{x}(s_1 + ds, t_1)$ of a surface element dF define a vector d**u**, the points $\mathbf{x}(s_1,t_1)$ and $\mathbf{x}(s_1,t_1 + dt)$ define a vector d**w**. The vectors d**u** and d**w** are expressed as linear combinations of the global basis vectors $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$. The differentials du^i are expressed as functions of the parameter differential ds, the differentials dw^i are expressed as functions of the differential dt.

$$
\begin{aligned}
d\mathbf{u} &= du^i\,\mathbf{b}_i \qquad \text{with} \qquad du^i = x^i(s_1 + ds, t_1) - x^i(s_1, t_1) \\
d\mathbf{w} &= dw^i\,\mathbf{b}_i \qquad\qquad\qquad\; dw^i = x^i(s_1, t_1 + dt) - x^i(s_1, t_1) \\
du^i &= \frac{\partial x_i}{\partial s}\,ds \\
dw^i &= \frac{\partial x^i}{\partial t}\,dt
\end{aligned}
$$

The cross product d**u** × d**w** is called the area vector of the parallelogram spanned by the vectors d**u** and d**w** and is designated by d**a**. In order to determine the coordinates da_i of the area vector d**a**, the coordinate form of the vectors d**u** and d**w** is substituted into the cross product d**u** × d**w**. Then the cross product $\mathbf{b}_k \times \mathbf{b}_m$ is replaced by $\varepsilon_{ikm}\,\mathbf{b}^i$:

$$
\begin{aligned}
d\mathbf{a} &= d\mathbf{u} \times d\mathbf{w} = du^k\,dw^m\,\mathbf{b}_k \times \mathbf{b}_m = da_i\,\mathbf{b}^i \\
da_i &= \varepsilon_{ikm}\,\frac{\partial x^k}{\partial s}\,\frac{\partial x^m}{\partial t}\,ds\,dt = b_*\,e_{ikm}\,\frac{\partial x^k}{\partial s}\,\frac{\partial x^m}{\partial t}\,ds\,dt
\end{aligned}
$$

Surface area : The position vector of a point on a surface is a function $\mathbf{x}(s,t)$ of the surface parameters s and t. The incremental area vector d**a** of a surface element is represented in the global basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ using the derivatives of the global coordinates x^i with respect to the surface parameters :

$$
\begin{aligned}
d\mathbf{a} &= \mathbf{c}\,ds\,dt \\
\mathbf{c}(s,t) &= c_i\,\mathbf{b}^i \qquad \text{with} \qquad c_i(s,t) = b_*\,e_{ikm}\,\frac{\partial x^k}{\partial s}\,\frac{\partial x^m}{\partial t}
\end{aligned}
$$

The magnitude da of the incremental area vector d**a** is called the surface area of the surface element dF. The area a of the surface fragment F is obtained by integrating da.

$$\mathrm{da} = \sqrt{\mathrm{d}\mathbf{a} \cdot \mathrm{d}\mathbf{a}} = \sqrt{\mathbf{c} \cdot \mathbf{c}}\; \mathrm{ds\, dt}$$

$$a = \int_F \sqrt{c_i\, c_m\, g^{im}}\; \mathrm{ds\, dt}$$

Surface integrals : Let the tensor field to be integrated be given as a function of the position vector **x**. Since the position vector is known as a function of the surface parameter s and t, the tensor field can also be expressed as a function of s and t.

scalar field : $u = u\,(s,t)$

vector field : $\mathbf{u} = u_i(s,t)\; \mathbf{b}^i$

The surface integral is taken over a continuous surface fragment F in the space $\mathbb{R}^3$. The incremental area vector d**a** is expressed as a function **c** ds dt of the surface parameters with $\mathbf{c} = c_i\, \mathbf{b}^i$. The value of the surface integral is a scalar h if the integrand is scalar. Otherwise the result is a vector **h**.

scalar surface integral of a scalar field

$$h = \int_F u\, \mathrm{da} = \int_F u(s,t) \sqrt{\mathbf{c} \cdot \mathbf{c}}\; \mathrm{ds\, dt}$$

vector surface integral of a scalar field

$$\mathbf{h} = \int_F u\, \mathrm{d}\mathbf{a} = \int_F u\, \mathbf{c}\; \mathrm{ds\, dt}$$

$$\mathbf{h} = h_i\, \mathbf{b}^i \quad \text{with} \quad h_i = \int_F u(s,t)\; c_i(s,t)\; \mathrm{ds\, dt}$$

scalar surface integral of a vector field

$$h = \int_F \mathbf{u} \cdot \mathrm{d}\mathbf{a} = \int_F \mathbf{u} \cdot \mathbf{c}\; \mathrm{ds\, dt}$$

$$h = \int_F u_i\,(s,t)\; c_m(s,t)\; g^{im}\; \mathrm{ds\, dt}$$

vector surface integral of a vector field

$$h = \int_F \mathbf{u} \times \mathrm{d}\mathbf{a} = \int_F \mathbf{u} \times \mathbf{c}\; \mathrm{ds\, dt}$$

$$\mathbf{h} = h^i\, \mathbf{b}_i \quad \text{with} \quad h^i = \int_F \varepsilon^{ikm}\; u_k(s,t)\; c_m(s,t)\; \mathrm{ds\, dt}$$

Example 2 : Surface integrals

A global coordinate system $(\mathbf{0}, \mathbf{E})$ with the orthonormal basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is chosen in a point space $\mathbb{R}^3$. The points $\mathbf{x} = x^i\,\mathbf{e}_i$ of a cylindrical surface with radius r are specified using the surface parameters θ, z :

$$\begin{aligned} \mathbf{x} &= r\cos\theta\;\mathbf{e}_1 + r\sin\theta\;\mathbf{e}_2 + z\;\mathbf{e}_3 \qquad\qquad 0 \le \theta < 2\pi,\ 0 \le z \le a \\ \mathbf{c} &= e^{ikm}\,\frac{\partial x_k}{\partial \theta}\,\frac{\partial x_m}{\partial z}\,\mathbf{e}_i = r\cos\theta\;\mathbf{e}_1 + r\sin\theta\;\mathbf{e}_2 \end{aligned}$$

The scalar and vector surface integrals of the scalar field $u = r(\cos\theta + z\sin\theta)$ are calculated :

$$\begin{aligned} h &= \int_0^a\!\!\int_0^{2\pi} u\sqrt{\mathbf{c}\cdot\mathbf{c}}\;d\theta\,dz = \int_0^a\!\!\int_0^{2\pi} r^2\,(\cos\theta + z\sin\theta)\;d\theta\,dz = 0 \\ \mathbf{h} &= \int_0^a\!\!\int_0^{2\pi} u\,\mathbf{c}\;d\theta\,dz = \int_0^a\!\!\int_0^{2\pi} r^2\,(\cos\theta + z\sin\theta)(\cos\theta\;\mathbf{e}_1 + \sin\theta\;\mathbf{e}_2)\;d\theta\,dz \\ \mathbf{h} &= \pi r^2 a\,(\mathbf{e}_1 + 0.5\,a\;\mathbf{e}_2) \end{aligned}$$

The scalar and vector surface integrals of the vector field $\mathbf{u} = r\cos\theta\;\mathbf{e}_1 + r\sin\theta\;\mathbf{e}_2 - z\;\mathbf{e}_3$ are calculated :

$$\begin{aligned} h &= \int_0^a\!\!\int_0^{2\pi} \mathbf{u}\cdot\mathbf{c}\;d\theta\,dz &&= \int_0^a\!\!\int_0^{2\pi} r^2\,d\theta\,dz = 2\pi a r^2 \\ \mathbf{h} &= \int_0^a\!\!\int_0^{2\pi} \mathbf{u}\times\mathbf{c}\;d\theta\,dz &&= h^i\,\mathbf{e}_i \\ h^1 &= \int_0^a\!\!\int_0^{2\pi} (u_2 c_3 - u_3 c_2)\;d\theta\,dz &&= \int_0^a\!\!\int_0^{2\pi} (0 + zr\sin\theta)\;d\theta\,dz = 0 \\ h^2 &= \int_0^a\!\!\int_0^{2\pi} (u_3 c_1 - u_1 c_3)\;d\theta\,dz &&= \int_0^a\!\!\int_0^{2\pi} (-zr\cos\theta + 0)\;d\theta\,dz = 0 \\ h^3 &= \int_0^a\!\!\int_0^{2\pi} (u_1 c_2 - u_2 c_1)\;d\theta\,dz &&= \int_0^a\!\!\int_0^{2\pi} r^2\sin\theta\cos\theta\,(1-1)\;d\theta\,dz = 0 \end{aligned}$$

Volume : Let the global coordinates x^1, x^2, x^3 of the points of the space $\mathbb{R}^3$ with the global basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ be functions $x^i(y^1, y^2, y^3)$ of the local coordinates y^m. The set of points $\mathbf{x} \in \mathbb{R}^3$ is called the volume of the space $\mathbb{R}^3$ and is designated by V. A subset of the points of $\mathbb{R}^3$ whose local coordinates lie in given ranges $y^i_1 \le y^i \le y^i_2$ is called a volume fragment and is designated by V_s. A subset of points of $\mathbb{R}^3$ whose local coordinates lie in infinitesimal ranges $y^i_1 \le y^i \le y^i_1 + dy^i$ is called a volume element and is designated by dV.

$$\begin{aligned} V &= \{\mathbf{x} \mid \mathbf{x} = x^i\,(y^1, y^2, y^3)\,\mathbf{b}_i \quad \wedge \quad y^m \in \mathbb{R}\} \\ V_s &= \{\mathbf{x} \mid \mathbf{x} = x^i\,(y^1, y^2, y^3)\,\mathbf{b}_i \quad \wedge \quad y_1^m \le y^m \le y_2^m\} \\ dV &= \{\mathbf{x} \mid \mathbf{x} = x^i\,(y^1, y^2, y^3)\,\mathbf{b}_i \quad \wedge \quad y_1^m \le y^m \le y_1^m + dy^m\} \end{aligned}$$

Magnitude of the volume : Let a cuboidal volume element dV be given whose edges are parallel to the basis vectors $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ of the canonical basis of $\mathbb{R}^3$. Let the lengths of the edges be da^1, da^2, da^3. The determinant of the matrix whose columns contain the edge vectors of the cuboid is called the magnitude of the volume of the element and is designated by dv.

$$dV = \{\mathbf{x} \mid \mathbf{x} = \lambda^1\, da^1\, \mathbf{e}_1 + \lambda^2\, da^2\, \mathbf{e}_2 + \lambda^3\, da^3\, \mathbf{e}_3 \quad \wedge \quad 0 \le \lambda^i \le 1\}$$

$$dv = \det \left\{ \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 & \mathbf{e}_3 \end{bmatrix} * \begin{bmatrix} da^1 & & \\ & da^2 & \\ & & da^3 \end{bmatrix} \right\} = da^1\ da^2\ da^3$$

Let a global basis $\mathbf{B}_*$ of the space $\mathbb{R}^3$ and a parallelepipedal volume element dV whose edges are parallel to the basis vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ be given. Let the edge vectors be $\mathbf{b}_1 dx^1$, $\mathbf{b}_2 dx^2$, $\mathbf{b}_3 dx^3$. The magnitude dv of the volume of the parallelepipedon is the determinant of the matrix whose columns contain the edge vectors of the parallelepipedon. With $\det \mathbf{B}_* = b_*$ this is expressed as follows :

$$dV = \{\mathbf{x} \mid \mathbf{x} = \lambda^1\, dx^1\, \mathbf{b}_1 + \lambda^2\, dx^2\, \mathbf{b}_2 + \lambda^3\, dx^3\, \mathbf{b}_3 \quad \wedge \quad 0 \le \lambda^i \le 1\}$$

$$dv = \det \left\{ \begin{bmatrix} \mathbf{b}_1 & \mathbf{b}_2 & \mathbf{b}_3 \end{bmatrix} * \begin{bmatrix} dx^1 & & \\ & dx^2 & \\ & & dx^3 \end{bmatrix} \right\} = b_*\ dx^1\ dx^2\ dx^3$$

Let the coordinates of the points of the space $\mathbb{R}^3$ with the global basis $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ be functions $x^i(y^1, y^2, y^3)$ of the local coordinates y^1, y^2, y^3. Let the local basis vectors at a point $\mathbf{x}$ be $\overline{\mathbf{b}}_1, \overline{\mathbf{b}}_2, \overline{\mathbf{b}}_3$. Let a parallelepipedal volume element dV with the edge vectors $\overline{\mathbf{b}}_1 dy^1$, $\overline{\mathbf{b}}_2 dy^2$ and $\overline{\mathbf{b}}_3 dy^3$ be given. The magnitude dv of the volume is the determinant of the matrix whose columns contain the edge vectors of the parallelepipedon. With $\det \overline{\mathbf{B}}_* = \overline{b}_*$ this is expressed as follows :

$$\mathbf{x} = x^i(y^1, y^2, y^3)\,\mathbf{b}_i$$

$$d\mathbf{x} = \overline{\mathbf{b}}_i\, dy^i \qquad \text{with} \qquad \overline{\mathbf{b}}_i = \frac{\partial \mathbf{x}}{\partial y^i} = \frac{\partial x^m}{\partial y^i}\,\mathbf{b}_m$$

$$dV = \{\mathbf{x} \mid \mathbf{x} = \lambda^1\, dy^1\, \bar{\mathbf{b}}_1 + \lambda^2\, dy^2\, \bar{\mathbf{b}}_2 + \lambda^3\, dy^3\, \bar{\mathbf{b}}_3 \quad \wedge \quad 0 \le \lambda^i \le 1\}$$

$$dv = \det \begin{vmatrix} \bar{\mathbf{b}}_1 & \bar{\mathbf{b}}_2 & \bar{\mathbf{b}}_3 \end{vmatrix} * \begin{vmatrix} dy^1 & & \\ & dy^2 & \\ & & dy^3 \end{vmatrix} = \bar{b}_* \, dy^1\, dy^2\, dy^3$$

The determinant b_* of a global basis is independent of the coordinates x^1, x^2, x^3. By contrast, the determinant $\bar{b}_*$ of a local basis depends on the local coordinates y^1, y^2, y^3. The magnitude v of a volume fragment V may thus be determined using the following formulas :

$$v = \int_V da^1 da^2 da^3 = b_* \int_V dx^1 dx^2 dx^3 = \int_V \bar{b}_*\, dy^1 dy^2 dy^3$$

Volume integrals : Let a scalar field be given as a function $u(x^1, x^2, x^3)$ of the global coordinates or as a function $\bar{u}(y^1, y^2, y^3)$ of the local coordinates. Let the coordinates of a vector field with respect to the basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ be given as functions $u^i(x^1, x^2, x^3)$ of the global coordinates or as functions $\bar{u}^i(y^1, y^2, y^3)$ of the local coordinates.

scalar field : $u = u\,(x^1, x^2, x^3) \quad \vee \quad u = \bar{u}\,(y^1, y^2, y^3)$

vector field : $\mathbf{u} = u^i(x^1, x^2, x^3)\mathbf{b}_i \quad \vee \quad \mathbf{u} = \bar{u}^i(y^1, y^2, y^3)\mathbf{b}_i$

The volume integrals of the scalar field and the vector field over a volume fragment V of the space $\mathbb{R}^3$ are alternatively determined in global or in local coordinates.

volume integral of a scalar field

$$\begin{aligned} h &= \int_V u\, dv = \int_V u\,(x^1, x^2, x^3)\; b_*\, dx^1 dx^2 dx^3 \\ &= \int_V \bar{u}\,(y^1, y^2, y^3)\; \bar{b}_*\, dy^1 dy^2 dy^3 \end{aligned}$$

volume integral of a vector field

$$\mathbf{h} = \int_V \mathbf{u}\, dv = h^i\, \mathbf{b}_i$$

$$\begin{aligned} h^i &= \int_V u^i\, dv = \int_V u^i\,(x^1, x^2, x^3)\, b_*\, dx^1 dx^2 dx^3 \\ &= \int_V \bar{u}^i\,(y^1, y^2, y^3)\, \bar{b}_*\, dy^1 dy^2 dy^3 \end{aligned}$$

Example 3 : Volume of a sphere

The spherical coordinate system (r, θ, β) is defined in Example 2 of Section 9.4.5. The ranges $0 \le r \le a$, $0 \le \theta \le 2\pi$, $-0.5\pi \le \beta \le 0.5\pi$ define a spherical volume fragment with radius a. The functional matrix $\bar{\mathbf{B}}_*$ contains the coordinates of the local basis vectors $\bar{\mathbf{b}}_1, \bar{\mathbf{b}}_2, \bar{\mathbf{b}}_3$ at the point (r, θ, β) referred to the canonical basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. Let v be the volume of the sphere of radius a :

$$\bar{\mathbf{B}}_* = \begin{array}{|c|c|c|} \hline \cos\beta\ \cos\theta & -r\ \cos\beta \sin\theta & -r\ \sin\beta\ \cos\theta \\ \hline \cos\beta\ \sin\theta & r\ \cos\beta \cos\theta & -r\ \sin\beta\ \sin\theta \\ \hline \sin\beta & 0 & r\ \cos\beta \\ \hline \end{array}$$

$$\bar{b}_* = \det \bar{\mathbf{B}}_* = r^2 \cos\beta$$

$$v = \int_0^a \int_0^{2\pi} \int_{-0.5\pi}^{0.5\pi} r^2 \cos\beta \, d\beta \, d\theta \, dr = \frac{4}{3}\pi a^3$$

9.4.9 FIELD OPERATIONS

Introduction : In the mathematical formulation of physical problems one considers infinitesimal line elements, surface elements and volume elements whose size tends to zero. The limit of the ratio of the change of a tensor field between the endpoints of a line element to the length of the line element is called a directional derivative of the field. The limit of the ratio of an integral of a tensor field over the surface of a volume element to the magnitude of the volume of the element is called a volume derivative of the tensor field. The operation of taking such a limit is called a field operation.

Scalar fields and vector fields have different directional derivatives. Different surface integrals of scalar fields and vector fields lead to different volume derivatives : The vector integral of a scalar field leads to the gradient of the field, the scalar integral of a vector field leads to the divergence of the field, the vector integral of a vector field leads to the curl of the field, and the dyadic integral of a vector field leads to the vector gradient of the field. These field operations are defined in the following for local coordinate systems.

Coordinate system : The field operations are defined for the euclidean space $\mathbb{R}^3$ and local coordinates y^1, y^2, y^3. The coordinates x^1, x^2, x^3 of a point $\mathbf{x}$ in the global canonical basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ of $\mathbb{R}^3$ are given as functions $x^i(y^1, y^2, y^3)$ of the local coordinates. The local basis vectors $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ at the point $\mathbf{x}$ are determined by partial differentiation of the position vector :

$$
\begin{aligned}
\mathbf{x} &= x^1 \mathbf{e}_1 + x^2 \mathbf{e}_2 + x^3 \mathbf{e}_3 \\
x^i &= x^i(y^1, y^2, y^3) \\
\mathbf{b}_i &= \frac{\partial \mathbf{x}}{\partial y^i} = \frac{\partial x^1}{\partial y^i} \mathbf{e}_1 + \frac{\partial x^2}{\partial y^i} \mathbf{e}_2 + \frac{\partial x^3}{\partial y^i} \mathbf{e}_3
\end{aligned}
$$

Neighboring points of $\mathbb{R}^3$ are described by the differential $d\mathbf{x}$ of the position vector. The total differential of the position vector depends on the differentials dy^i of all local coordinates :

$$
d\mathbf{x} = \frac{\partial \mathbf{x}}{\partial y^i} dy^i = \mathbf{b}_1 dy^1 + \mathbf{b}_2 dy^2 + \mathbf{b}_3 dy^3
$$

Tensor field : In the definition of the field operations it is assumed that the scalar field is given as a function $u(y^1, y^2, y^3)$ and the coordinates of the vector field $\mathbf{u}$ are given as functions $u^i(y^1, y^2, y^3)$ of the local coordinates. The covariant coordinates u_i of the vector field are determined using the metric coefficients g_{im} of the basis $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$.

$$
\begin{aligned}
u &= u(y^1, y^2, y^3) \\
\mathbf{u} &= u^i \mathbf{b}_i \quad \text{with} \quad u^i = u^i(y^1, y^2, y^3) \\
\mathbf{u} &= u_i \mathbf{b}^i \quad \text{with} \quad u_i = g_{im} u^m
\end{aligned}
$$

Derivative of a scalar field with respect to a vector : Let a scalar field $u(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ and a vector $\mathbf{a}$ at a point $\mathbf{x}_o(y^1, y^2, y^3)$ of $\mathbb{R}^3$ be given. For an incremental parameter $ds \in \mathbb{R}$, the points $\mathbf{x}_o$ and $\mathbf{x}_o + \mathbf{a}\, ds$ are close to each other. The difference of the field values $u(\mathbf{x}_o + \mathbf{a}\, ds)$ and $u(\mathbf{x}_o)$ is divided by ds. The limit of this quotient for ds→0 is called the derivative of the scalar field with respect to the vector $\mathbf{a}$ at the point $\mathbf{x}_o$ and is designated by $\frac{\partial u}{\partial \mathbf{a}}$.

$$\frac{\partial u}{\partial \mathbf{a}} = \lim_{ds \to 0} \frac{u(\mathbf{x}_o + \mathbf{a}\, ds) - u(\mathbf{x}_o)}{ds}$$

The derivative of the position vector $\mathbf{x}(y^1, ..., y^n)$ with respect to the parameter s is obtained using the chain rule. The relationship between the coordinates a^i of the given vector $\mathbf{a}$ and the derivatives of the local coordinates y^i with respect to s follows from the condition $\frac{d\mathbf{x}}{ds} = \mathbf{a}$:

$$\frac{d\mathbf{x}}{ds} = \frac{\partial \mathbf{x}}{\partial y^i} \frac{dy^i}{ds} = \mathbf{b}_i \frac{dy^i}{ds}$$

$$\frac{d\mathbf{x}}{ds} = \mathbf{a} = \mathbf{b}_i\, a^i = \mathbf{b}_i \frac{dy^i}{ds}$$

$$a^i = \frac{dy^i}{ds}$$

The function value $u(\mathbf{x}_o + \mathbf{a}\, ds)$ is expressed as a Taylor series at the point $\mathbf{x}_o$ and substituted into the definition of the derivative. The terms containing second and higher derivatives of u vanish in the limit ds→0 since they contain a factor ds :

$$u(\mathbf{x}_o + \mathbf{a}\, ds) = u(\mathbf{x}_o) + \frac{\partial u}{\partial y^i} \frac{dy^i}{ds} ds + \frac{1}{2} \frac{\partial^2 u}{\partial y^i\, dy^m} \frac{dy^i}{ds} \frac{dy^m}{ds} (ds)^2 + \ldots$$

$$\frac{u(\mathbf{x}_o + \mathbf{a}\, ds) - u(\mathbf{x}_o)}{ds} = \frac{\partial u}{\partial y^i} a^i + \frac{1}{2} \frac{\partial^2 u}{\partial y^i\, dy^m} a^i a^m ds + \ldots$$

$$\frac{\partial u}{\partial \mathbf{a}} = \lim_{ds \to 0} \frac{u(\mathbf{x}_o + \mathbf{a}\, ds) - u(\mathbf{x}_o)}{ds} = \frac{\partial u}{\partial y^i} a^i$$

The partial derivatives of the field u with respect to the local coordinates y^i are arranged in a gradient vector $\mathbf{g}$. Then the derivative of the scalar field u with respect to the vector $\mathbf{a}$ is the scalar product of $\mathbf{g}$ and $\mathbf{a}$:

$$\mathbf{g} = \frac{\partial u}{\partial y^1} \mathbf{b}^1 + \frac{\partial u}{\partial y^2} \mathbf{b}^2 + \frac{\partial u}{\partial y^3} \mathbf{b}^3$$

$$\frac{\partial u}{\partial \mathbf{a}} = \mathbf{g} \cdot \mathbf{a}$$

Directional derivative of a scalar field : The derivative of a scalar field u with respect to a vector **a** depends on the magnitude of the vector **a**. However, if a unit vector **t** in the direction of **a** is considered, the derivative of the scalar field depends only on the direction of the vector **t**. The derivative of a scalar field u with respect to a unit vector $\mathbf{t} = t^i\, \mathbf{b}_i$ is called a directional derivative of the scalar field and is designated by $\frac{\partial u}{\partial \mathbf{t}}$.

$$\frac{\partial u}{\partial \mathbf{t}} = \lim_{ds \to 0} \frac{u(\mathbf{x}_o + \mathbf{t}\, ds) - u(\mathbf{x}_o)}{ds} = \frac{\partial u}{\partial y^i}\, t^i$$

$$\mathbf{t} \cdot \mathbf{t} = 1$$

The following relationship holds between the derivative of a scalar field with respect to a vector **a** and the directional derivative with respect to the unit vector **t** associated with **a** :

$$\mathbf{a} = |\mathbf{a}|\, \mathbf{t} \quad \wedge \quad \mathbf{t} \cdot \mathbf{t} = 1$$

$$\frac{\partial u}{\partial \mathbf{a}} = \mathbf{g} \cdot \mathbf{a} = |\mathbf{a}|\; \mathbf{g} \cdot \mathbf{t} = |\mathbf{a}|\; \frac{\partial u}{\partial \mathbf{t}}$$

Maximal value of the directional derivative : The question arises for which unit vector **n** the directional derivative of the scalar field is maximal. The scalar product $\mathbf{g} \cdot \mathbf{n}$ is maximal if the direction vector **n** is parallel to **g**, and hence **n** is a unit vector in the direction of **g**. If the magnitude of **g** is designated by g, the maximal directional derivative is equal to g.

$$\mathbf{g} = g\, \mathbf{n} \quad \text{with} \quad g = |\mathbf{g}|$$

$$\frac{\partial u}{\partial \mathbf{n}} = \mathbf{g} \cdot \mathbf{n} = g\, \mathbf{n} \cdot \mathbf{n} = g$$

The differential $du = \mathbf{g} \cdot d\mathbf{x}$ between neighboring points **x** and $\mathbf{x} + d\mathbf{x}$ is zero if the points lie on an isosurface u = const. Hence the vector **g** is normal to the isosurface. The directional derivative of the scalar field is maximal for the unit normal **n** of the isosurface u = const.

Directional derivative of a vector field : Let a vector field $\mathbf{u}(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ and a unit vector **t** at a point $\mathbf{x}_o(y^1, y^2, y^3)$ of $\mathbb{R}^3$ be given. The difference of the field values at the points $\mathbf{x}_o + \mathbf{t}\, ds$ and $\mathbf{x}_o$ with $ds \in \mathbb{R}$ is divided by ds. The limit of this quotient for ds→0 is called the directional derivative of the vector field **u** at the point $\mathbf{x}_o$ for the direction vector **t**. The directional derivative of a vector field is a vector and is designated by $\frac{\partial \mathbf{u}}{\partial \mathbf{t}}$.

$$\frac{\partial \mathbf{u}}{\partial \mathbf{t}} = \lim_{ds \to 0} \frac{\mathbf{u}(\mathbf{x}_o + \mathbf{t}\, ds) - \mathbf{u}(\mathbf{x}_o)}{ds}$$

The relationship between the coordinates t^i of the direction vector **t** and the derivatives of the local coordinates y^i with respect to the distance s is obtained using the chain rule :

$$\mathbf{t} \quad = \quad t^i\, \mathbf{b}_i$$

$$\frac{d\mathbf{x}}{ds} = \frac{\partial \mathbf{x}}{\partial y^i}\frac{dy^i}{ds} = \mathbf{b}_i \frac{dy^i}{ds} = \mathbf{b}_i\, t^i$$

$$\frac{dy^i}{ds} = t^i$$

On the line $\mathbf{x}_o + \mathbf{t}\,s$, the value of the vector field is a function $\mathbf{u}(s)$ of the distance s, since $\mathbf{x}_o$ and **t** are constant. The value $\mathbf{u}(\mathbf{x}_o + \mathbf{t}ds)$ at the point $\mathbf{x}_o + \mathbf{t}\,s$ is determined using a Taylor series for $\mathbf{u}(s)$ at the point $\mathbf{x}_o$ and is substituted into the definition of the directional derivative. The terms containing second and higher derivatives of **u** vanish in the limit ds→0 since they contain a factor ds.

$$\mathbf{u}(\mathbf{x}_o + \mathbf{t}\ ds) \quad = \quad \mathbf{u}(\mathbf{x}_o) + \frac{\partial \mathbf{u}}{\partial y^i}\frac{dy^i}{ds} ds \ + \ \frac{1}{2}\ \frac{\partial^2 \mathbf{u}}{\partial y^k\, \partial y^m}\ \frac{dy^k}{ds}\ \frac{dy^m}{ds}\ (ds)^2 \ + \ ...$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{t}} = \lim_{ds \to 0}\ \frac{\mathbf{u}(\mathbf{x}_o + \mathbf{t}\,ds) - \mathbf{u}(\mathbf{x}_o)}{ds} = \frac{\partial \mathbf{u}}{\partial y^i}\frac{dy^i}{ds}$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{t}} = \frac{\partial \mathbf{u}}{\partial y^i}\, t^i$$

The partial derivatives of the field are expressed in terms of the covariant derivatives $u_{i\,;\,m}$ of the field coordinates, which are arranged in a matrix $\mathbf{U}_{;\,y}$. The symbol ; y identifies the covariant derivatives contained in the matrix. Then the directional derivative of the vector field is the product of the basis $\mathbf{B}^*$, the matrix of derivatives $\mathbf{U}_{;\,y}$ and the direction vector **t** :

$$\frac{\partial \mathbf{u}}{\partial \mathbf{t}} = \mathbf{B}^*\, \mathbf{U}_{;\,y}\, \mathbf{t} \quad = \quad u_{i\,;\,m}\, t^m\, \mathbf{b}^i$$

Stationary values of the directional derivative : The question arises for which unit vectors **n** the magnitude of the directional derivative of the vector field is stationary. These unit vectors are the eigenvectors of the matrix $\mathbf{B}^*\, \mathbf{U}_{;\,y}$. If $\mathbf{U}_{;\,y}$ has the eigenstates $(p_k, \mathbf{n}_k)$, the magnitudes of the stationary directional derivatives are p_k.

$$\mathbf{B}^*\, \mathbf{U}_{;\,y}\, \mathbf{n} \ = \ p\, \mathbf{n} \qquad \Rightarrow \qquad u_{i\,;\,m}\, n^m\, \mathbf{b}^i \ = \ p\, \mathbf{n} \qquad \text{with} \qquad \mathbf{n} \cdot \mathbf{n} \ = \ 1$$

$$\left| \frac{\partial \mathbf{u}}{\partial \mathbf{n}} \right| = \sqrt{p_k^2\ \mathbf{n} \cdot \mathbf{n}} \ = \ p_k \qquad\qquad k = 1, 2, 3$$

Local volume element : Let $\mathbf{b}_1$, $\mathbf{b}_2$ and $\mathbf{b}_3$ be the local basis vectors at a point $\mathbf{a}(y^1, y^2, y^3)$ of the euclidean space $\mathbb{R}^3$. The vectors $\mathbf{b}_1 dy^1$, $\mathbf{b}_2 dy^2$ and $\mathbf{b}_3 dy^3$ span a parallelepipedal volume element dV. The surface of the parallelepipedon consists of six continuous surface elements, the faces dA_i of the parallelepipedon. The elements are defined by the following point sets :

$$\begin{array}{lclllll}
dV & = & \{ \mathbf{x} = \mathbf{a} + & s_1 dy^1 \, \mathbf{b}_1 & + \; s_2 dy^2 \, \mathbf{b}_2 & + \; s_3 dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \} \\
dA_1 & = & \{ \mathbf{x} = \mathbf{a} + & & + \; s_2 dy^2 \, \mathbf{b}_2 & + \; s_3 dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \} \\
dA_2 & = & \{ \mathbf{x} = \mathbf{a} + & s_1 dy^1 \, \mathbf{b}_1 & + & + \; s_3 dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \} \\
dA_3 & = & \{ \mathbf{x} = \mathbf{a} + & s_1 dy^1 \, \mathbf{b}_1 & + \; s_2 dy^2 \, \mathbf{b}_2 & + & \mid \; 0 \le s_i \le 1 \} \\
dA_4 & = & \{ \mathbf{x} = \mathbf{a} + & dy^1 \, \mathbf{b}_1 & + \; s_2 dy^2 \, \mathbf{b}_2 & + \; s_3 dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \} \\
dA_5 & = & \{ \mathbf{x} = \mathbf{a} + & s_1 dy^1 \, \mathbf{b}_1 & + \; dy^2 \, \mathbf{b}_2 & + \; s_3 dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \} \\
dA_6 & = & \{ \mathbf{x} = \mathbf{a} + & s_1 dy^1 \, \mathbf{b}_1 & + \; s_2 dy^2 \, \mathbf{b}_2 & + \; dy^3 \, \mathbf{b}_3 & \mid \; 0 \le s_i \le 1 \}
\end{array}$$

The outwardly directed area vectors $\mathbf{da}^i$ of the faces dA_i of the parallelepipedon, the surface areas da^i of the faces dA_i and the magnitude dv of the volume of the element dV are determined according to Section 9.4.8 and expressed in terms of the basis determinant $b_* = \det \mathbf{B}_*$ and the metric coefficients $g^{im} = \mathbf{b}^i \cdot \mathbf{b}^m$.

$$\begin{array}{llcccccc}
\text{faces 1 and 4} & : \; -\mathbf{da}^1 & = & \mathbf{da}^4 & = & (\mathbf{b}_2 dy^2) \times (\mathbf{b}_3 dy^3) & = & \mathbf{b}^1 b_* dy^2 \, dy^3 \\
\text{faces 2 and 5} & : \; -\mathbf{da}^2 & = & \mathbf{da}^5 & = & (\mathbf{b}_3 dy^3) \times (\mathbf{b}_1 dy^1) & = & \mathbf{b}^2 b_* dy^3 \, dy^1 \\
\text{faces 3 and 6} & : \; -\mathbf{da}^3 & = & \mathbf{da}^6 & = & (\mathbf{b}_1 dy^1) \times (\mathbf{b}_2 dy^2) & = & \mathbf{b}^3 b_* dy^1 \, dy^2
\end{array}$$

$$\begin{array}{llcccl}
\text{area } dA_1, dA_4 & : \; da^1 & = & da^4 & = & \sqrt{g^{11}} \, b_* dy^2 \, dy^3 \\
\text{area } dA_2, dA_5 & : \; da^2 & = & da^5 & = & \sqrt{g^{22}} \, b_* dy^3 \, dy^1 \\
\text{area } dA_3, dA_6 & : \; da^3 & = & da^6 & = & \sqrt{g^{33}} \, b_* dy^1 \, dy^2 \\
\text{magnitude } dV & : \; dv & = & b_* dy^1 \, dy^2 \, dy^3 & &
\end{array}$$

Gradient of a scalar field : Let a scalar field $u(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ be given. A point P of the space is enclosed in a piecewise continuous, closed surface F. The vector integral of the scalar field u over the surface F is divided by the magnitude v of the volume enclosed by F. If the limit of this quotient for v→0 exists, it is called the gradient of the scalar field at the point P and is designated by **grad** u.

$$\mathbf{grad}\, u \;=\; \lim_{v \to 0} \frac{1}{v} \int_F u \, \mathbf{da}$$

If the surface of a parallelepipedal volume element is chosen as the enclosing surface, the surface integral is equal to the sum of the surface integrals over the six faces of the parallelepipedon. On each face the direction of the surface normal is constant. The vector integral of the scalar field over the surface F then takes the following form :

$$\begin{aligned}\int_F u\,\mathbf{da} &= (u + \frac{\partial u}{\partial y^1}dy^1)\,\mathbf{da}^1 - u\,\mathbf{da}^1 + \\ &\quad (u + \frac{\partial u}{\partial y^2}dy^2)\,\mathbf{da}^2 - u\,\mathbf{da}^2 + \\ &\quad (u + \frac{\partial u}{\partial y^3}dy^3)\,\mathbf{da}^3 - u\,\mathbf{da}^3 \\ \int_F u\,\mathbf{da} &= \frac{\partial u}{\partial y^i}\,\mathbf{b}^i\, b_*\, dy^1\, dy^2\, dy^3\end{aligned}$$

Substituting this integral and the magnitude $dv = b_* dy^1\, dy^2\, dy^3$ of the volume into the definition of the gradient yields the differential form of the gradient :

$$\mathbf{grad}\; u = \frac{\partial u}{\partial y^i}\,\mathbf{b}^i = u_{,i}\,\mathbf{b}^i$$

$u_{,i}$ partial derivative of the scalar field $u(y^1, y^2, y^3)$ with respect to y^i

$\mathbf{b}^i$ local basis vector

The gradient **g** defined for the derivative of a scalar field u with respect to a vector and the gradient **grad** u defined here are identical.

Tensor character of the gradient : The local basis $\mathbf{B}_*$ is transformed into $\bar{\mathbf{B}}_* = \mathbf{B}_*\,\mathbf{A}$ using the matrix **A**. The coefficients of **A** are designated by $a^i_{.k}$, the coefficients of $\bar{\mathbf{A}} = \mathbf{A}^{-1}$ by $\bar{a}^k_{.m}$ and the local coordinates for the basis $\bar{\mathbf{B}}_*$ by $\bar{y}^i$. According to Sections 9.2.6 and 9.4.4, the following transformation rules hold for the basis vectors and the partial derivatives :

$$\bar{\mathbf{b}}^k = \bar{a}^k_{.m}\,\mathbf{b}^m$$

$$\frac{\partial u}{\partial \bar{y}^k} = a^i_{.k}\,\frac{\partial u}{\partial y^i}$$

Substituting these rules into the definition of the gradient shows that the gradient is a tensor.

$$\mathbf{grad}\; u = \frac{\partial u}{\partial \bar{y}^k}\,\bar{\mathbf{b}}^k = a^i_{.k}\,\frac{\partial u}{\partial y^i}\,\bar{a}^k_{.m}\,\mathbf{b}^m = \frac{\partial u}{\partial y^i}\,\mathbf{b}^i$$

Divergence of a vector field : Let a vector field **u** with the coordinates $u^i(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ be given. A point P of the space is enclosed in a piecewise continuous, closed surface F. The scalar integral of the vector field **u** over the surface F is divided by the magnitude v of the volume enclosed by F. If the limit of this quotient for v→0 exists, it is called the divergence of the vector field at the point P and is designated by div **u**. The divergence is a scalar and therefore a tensor.

$$\text{div}\,\mathbf{u} \quad = \quad \lim_{v \to 0} \frac{1}{v} \int_F \mathbf{u} \cdot \mathbf{da}$$

If the surface of a parallelepipedal volume element is chosen as the enclosing surface, the surface integral is equal to the sum of the surface integrals over the six faces of the parallelepipedon. On each face the direction of the surface normal is constant. Hence the scalar integral of the vector field over the surface F is given by

$$\begin{aligned} \int_F \mathbf{u} \cdot \mathbf{da} \quad &= \quad (u^i + u^i_{;1}\, dy^1)\, \mathbf{b}_i \cdot \mathbf{da}^1 \;-\; u^i \mathbf{b}_i \cdot \mathbf{da}^1 \;+ \\ &\quad\;\; (u^i + u^i_{;2}\, dy^2)\, \mathbf{b}_i \cdot \mathbf{da}^2 \;-\; u^i \mathbf{b}_i \cdot \mathbf{da}^2 \;+ \\ &\quad\;\; (u^i + u^i_{;3}\, dy^3)\, \mathbf{b}_i \cdot \mathbf{da}^3 \;-\; u^i \mathbf{b}_i \cdot \mathbf{da}^3 \\ \int_F \mathbf{u} \cdot \mathbf{da} \quad &= \quad u^i_{;m}\, \mathbf{b}_i \cdot \mathbf{b}^m\, b_* \, dy^1\, dy^2\, dy^3 \end{aligned}$$

Substituting this integral and the magnitude $dv = b_* dy^1\, dy^2\, dy^3$ of the volume into the definition of the divergence and using $\mathbf{b}_i \cdot \mathbf{b}^m = \delta_i^m$ yields the differential form of the divergence :

$$\text{div}\,\mathbf{u} \quad = \quad u^i_{;i} \quad = \quad u^1_{;1} + u^2_{;2} + u^3_{;3}$$

$u^i_{;m}$ covariant derivative of u^i with respect to y^m

Curl of a vector field : Let a vector field **u** with the coordinates $u_i(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ be given. A point P of the space is enclosed in a piecewise continuous, closed surface F. The vector integral of the vector field **u** over the surface F is divided by the magnitude v of the volume enclosed by F. If the limit of this quotient for v→0 exists, it is called the curl of the vector field at the point P and is designated by **rot u**.

$$\mathbf{rot\,u} \quad = \quad \lim_{v \to 0} \frac{1}{v} \int_F \mathbf{da} \times \mathbf{u}$$

The order of the factors **da** and **u** is chosen for historical reasons; it determines the sign of the curl. If the surface of a parallelepipedal volume element is chosen as the enclosing surface, the surface integral is equal to the sum of the surface integrals over the six faces of the parallelepipedon. On each face the direction of the vector field is considered to be constant. Then the vector integral of the vector field over the surface F is given by

$$\begin{aligned}\int_F \mathbf{da} \times \mathbf{u} &= (u_i + u_{i\,;\,1}\, dy^1)\, \mathbf{da}^1 \times \mathbf{b}^i \;-\; u_i\; \mathbf{da}^1 \times \mathbf{b}^i \;+\\ &\quad (u_i + u_{i\,;\,2}\, dy^2)\, \mathbf{da}^2 \times \mathbf{b}^i \;-\; u_i\, \mathbf{da}^2 \times \mathbf{b}^i \;+\\ &\quad (u_i + u_{i\,;\,3}\, dy^3)\, \mathbf{da}^3 \times \mathbf{b}^i \;-\; u_i\, \mathbf{da}^3 \times \mathbf{b}^i\\ \int_F \mathbf{da} \times \mathbf{u} &= u_{i\,;\,m}\; \mathbf{b}^m \times \mathbf{b}^i\; b_* \; dy^1\; dy^2\; dy^3\end{aligned}$$

Substituting this integral and the magnitude $dv = b_* dy^1 \; dy^2 \; dy^3$ of the volume into the definition of the curl yields the differential form of the curl :

$$\textbf{rot u} \;=\; u_{i\,;\,m}\; \mathbf{b}^m \times \mathbf{b}^i$$

$u_{i\,;\,m}$ covariant derivative of u_i with respect to y^m

The double sum is written out for $i, m = 1, 2, 3$. The differences $u_{i\,;\,k} - u_{k\,;\,i}$ of the covariant derivatives are replaced by the differences $u_{i\,,\,k} - u_{k\,,\,i}$ of the partial derivatives using the relationship between the covariant and partial derivatives and the Christoffel symbols Γ^m_{ik} from Section 9.4.7.

$$\textbf{rot u} \;=\; (u_{2;1} - u_{1;2})\mathbf{b}^1 \times \mathbf{b}^2 + (u_{3;2} - u_{2;3})\mathbf{b}^2 \times \mathbf{b}^3 + (u_{1;3} - u_{3;1})\mathbf{b}^3 \times \mathbf{b}^1$$

$$u_{i\,;\,k} - u_{k\,;\,i} \;=\; (u_{i,k} - \Gamma^m_{ik}\, u_m) - (u_{k,i} - \Gamma^m_{ki}\, u_m) \;=\; u_{i\,,k} - u_{k,i}$$

The cross products of the basis vectors are replaced by dual basis vectors and the basis determinant $b^* = \det \mathbf{B}^*$. For example, $\mathbf{b}^1 \times \mathbf{b}^2 \;=\; b^* \,\mathbf{b}_3$. The permutation tensor $\varepsilon^{ikm} = b^* e^{ikm}$ allows an alternative differential formulation of the curl :

$$\textbf{rot u} \;=\; \varepsilon^{ikm}\; u_{k,i}\; \mathbf{b}_m \;=\; \varepsilon^{ikm}\; u_{k\,;\,i}\; \mathbf{b}_m$$

$u_{k,i}$ partial derivative of u_k with respect to y^i

Tensor character of the curl : At a point P of a point space $\mathbb{R}^n$, let $\mathbf{B}_*$ be the basis for a local coordinate system $(y^1, ..., y^n)$, and let $\bar{\mathbf{B}}_*$ be the basis for a local coordinate system $(z^1, ..., z^n)$. Let the coordinates of a tensor field of rank 1 be u_i in the basis $\mathbf{B}_*$ and $\bar{u}_i$ in the basis $\bar{\mathbf{B}}_*$. Let the coefficients of the transformation matrix **A** in the coordinate transformation $\bar{\mathbf{B}}_* = \mathbf{B}_* \mathbf{A}$ be $a^i_{.\,k}$. According to Sections 9.2.6 and 9.4.7, the following transformation rules hold for the basis vectors and the covariant derivatives of the tensor coordinates :

$$\begin{aligned}\bar{\mathbf{b}}^i &= \bar{a}^{\,i}_{.\;k}\; \mathbf{b}^k\\ u_{i\,;\,k} &= \bar{a}^{\,r}_{.\,i}\; \bar{a}^{\,s}_{.\,k}\; \bar{u}_{r;\,s}\end{aligned}$$

Substituting these rules into the first differential form of the curl shows that the curl is a tensor :

$$\begin{aligned}\textbf{rot u} &= u_{i\,;\,k}\; \mathbf{b}^k \times \mathbf{b}^i \;=\; \bar{a}^{\,r}_{.\,i}\; \bar{a}^{\,s}_{.\,k}\; \bar{u}_{r;\,s}\;\; \mathbf{b}^i \times \mathbf{b}^k\\ \textbf{rot u} &= \bar{u}_{r;\,s}\; \bar{\mathbf{b}}^r \times \bar{\mathbf{b}}^s\end{aligned}$$

Gradient of a vector field : Let a vector field **u** with the coordinates $u^i(y^1, y^2, y^3)$ in the euclidean space $\mathbb{R}^3$ be given. A point P of the space is enclosed in a piecewise continuous, closed surface F. The integral of the dyad d**a u**^{T} over the surface F is divided by the magnitude v of the volume enclosed by F. The limit of this quotient for v→0 is called the gradient of the vector field (vector gradient) at the point P and is designated by **dya u**.

$$\mathbf{dya\,u} \;=\; \lim_{v\to 0} \frac{1}{v} \int_F d\mathbf{a}\,\mathbf{u}^T$$

If the surface of a parallelepipedal volume element is chosen as the enclosing surface, the surface integral is equal to the sum of the surface integrals over the six faces of the parallelepipedon. On each face the direction of the vector field is considered to be constant. The integral over the surface F in the expression for the gradient **dya u** of the vector field is then given by

$$\begin{aligned}\int_F d\mathbf{a}\,\mathbf{u}^T \;=\;& (u_i + u_{i\,;\,1}\,dy^1)\,d\mathbf{a}^1(\mathbf{b}^i)^T \;-\; u_i\;d\mathbf{a}^1\,(\mathbf{b}^i)^T \;+\\ & (u_i + u_{i\,;\,2}\,dy^2)\,d\mathbf{a}^2(\mathbf{b}^i)^T \;-\; u_i\;d\mathbf{a}^2\,(\mathbf{b}^i)^T \;+\\ & (u_i + u_{i\,;\,3}\,dy^3)\,d\mathbf{a}^3(\mathbf{b}^i)^T \;-\; u_i\;d\mathbf{a}^3\,(\mathbf{b}^i)^T\end{aligned}$$

$$\int_F d\mathbf{a}\,\mathbf{u}^T \;=\; u_{i\,;\,m}\,\mathbf{b}^m\,(\mathbf{b}^i)^T\,b_*\,dy^1\,dy^2\,dy^3$$

Substituting this integral and the magnitude $v = b_*\,dy^1\,dy^2\,dy^3$ of the volume into the definition of the vector gradient yields the differential form of the gradient :

$$\mathbf{dya\,u} \;=\; u_{i\,;\,m}\,\mathbf{b}^m\,(\mathbf{b}^i)^T$$

The coordinates of the gradient of a vector field are the covariant derivatives of the vector coordinates. Since the covariant derivatives are the coordinates of a tensor, the gradient of a vector field is a tensor of rank 2.

Summary of the field operations : The field operations for a scalar field $p(y^1, y^2, y^3)$ and a vector field $\mathbf{u}(y^1, y^2, y^3)$ are defined in the preceding sections. The differential forms of these field operations are compiled in the following.

gradient of a scalar field :

$$\mathbf{u} \;=\; \mathbf{grad}\,p \;=\; p_{,\,i}\,\mathbf{b}^i \qquad \text{with} \qquad p_{,\,i} \;=\; \frac{\partial p}{\partial y^i}$$

divergence of a vector field :

$$c \;=\; \text{div}\,\mathbf{u} \;=\; u^i_{;\,i} \qquad \text{with} \qquad u^i_{;\,m} \;=\; u^i_{,\,m} + \Gamma^i_{km}\,u^k$$

curl of a vector field :

$$\mathbf{r} \;=\; \mathbf{rot\,u} \;=\; r^i\,\mathbf{b}_i \qquad \text{with} \qquad r^i \;=\; \varepsilon^{ikm}\,u_{m\,,\,k} \;=\; \varepsilon^{ikm}\,u_{m\,;\,k}$$

gradient of a vector field :

$$\mathbf{D} \;=\; \mathbf{dya\,u} \;=\; u_{k\,;\,i}\,\mathbf{b}^i\,(\mathbf{b}^k)^T \qquad \text{with} \qquad u_{k\,;\,i} \;=\; u_{k\,,\,i} - \Gamma^m_{ki}\,u_m$$

**Example 1 : ** Field operations in cartesian coordinates

The cartesian coordinate system (x^1, x^2, x^3) leads to the orthonormal basis vectors $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$. Since the dual basis vectors $\mathbf{e}_i$ and $\mathbf{e}^i$ are equal, the dual coordinates $u^i(x^1, x^2, x^3)$ and $u_i(x^1, x^2, x^3)$ of a vector field $\mathbf{u}$ are also equal. The covariant derivatives of the vector field coincide with the partial derivatives, since the basis $\mathbf{E}$ is global.

$$\text{scalar field} : \quad p = p(x^1, x^2, x^3)$$

$$\text{vector field} : \quad \mathbf{u} = u^i(x^1, x^2, x^3)\,\mathbf{e}_i = u_i(x^1, x^2, x^3)\,\mathbf{e}^i$$

$$\text{gradient} : \quad \mathbf{grad}\,p = \frac{\partial p}{\partial x^1}\mathbf{e}^1 + \frac{\partial p}{\partial x^2}\mathbf{e}^2 + \frac{\partial p}{\partial x^3}\mathbf{e}^3$$

$$\text{divergence} : \quad \text{div}\,\mathbf{u} = \frac{\partial u^1}{\partial x^1} + \frac{\partial u^2}{\partial x^2} + \frac{\partial u^3}{\partial x^3} = u^i_{\ ,i}$$

$$\text{curl} : \quad \mathbf{rot\,u} = c^i\,\mathbf{e}_i \quad \text{with} \quad c^i = e^{ikm}\,u_{m,k}$$

$$\mathbf{rot\,u} = (u_{3,2} - u_{2,3})\,\mathbf{e}_1 + (u_{1,3} - u_{3,1})\,\mathbf{e}_2 + (u_{2,1} - u_{1,2})\,\mathbf{e}_3$$

$$\text{gradient} : \quad \mathbf{dya\,u} = u_{m,i}\,\mathbf{e}^i(\mathbf{e}^m)^T$$

**Example 2 : ** Field operations in cylindrical coordinates

The cylindrical coordinate system (r, θ, z) and its Christoffel symbols are defined in Examples 1 of Sections 9.4.5 and 9.4.6. The covariant basis vectors $\mathbf{b}_i$ are the partial derivatives of the position vector $\mathbf{x}$ with respect to the local coordinates r, θ, z. The contravariant basis vectors $\mathbf{b}^m$ are determined using the condition $\mathbf{b}_i \cdot \mathbf{b}^m = \delta_i^m$.

$$\text{scalar field} : \quad p = p(r, \theta, z)$$

$$\text{vector field} : \quad \mathbf{u} = u^i(r, \theta, z)\,\mathbf{b}_i$$

$$\text{basis} : \quad \mathbf{b}_1 = \begin{bmatrix} \cos\theta \\ \sin\theta \\ 0 \end{bmatrix} \qquad \mathbf{b}_2 = \begin{bmatrix} -r\sin\theta \\ r\cos\theta \\ 0 \end{bmatrix} \qquad \mathbf{b}_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

$$b_* = \det \mathbf{B}_* = r$$

$$\mathbf{b}^1 = \mathbf{b}_1 \qquad \mathbf{b}^2 = \frac{1}{r^2}\mathbf{b}_2 \qquad \mathbf{b}^3 = \mathbf{b}_3$$

$$\text{metric} : \quad \mathbf{G}_* = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r^2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

coordinates : $u_i = g_{ik}\, u^k$

$$u_1 = u^1 \qquad u_2 = (r)^2\, u^2 \qquad u_3 = u^3$$

gradient : $\mathbf{grad}\; p = \frac{\partial p}{\partial r}\,\mathbf{b}^1 + \frac{\partial p}{\partial \theta}\,\mathbf{b}^2 + \frac{\partial p}{\partial z}\,\mathbf{b}^3$

$$= \frac{\partial p}{\partial r}\,\mathbf{b}_1 + \frac{1}{r^2}\frac{\partial p}{\partial \theta}\,\mathbf{b}_2 + \frac{\partial p}{\partial z}\,\mathbf{b}_3$$

divergence : $\operatorname{div} \mathbf{u} = u^1_{;1} + u^2_{;2} + u^3_{;3}$

$$u^1_{;1} = \frac{\partial u^1}{\partial r} + \Gamma^1_{1m}\, u^m = \frac{\partial u^1}{\partial r}$$

$$u^2_{;2} = \frac{\partial u^2}{\partial \theta} + \Gamma^2_{2m}\, u^m = \frac{\partial u^2}{\partial \theta} + \frac{u^1}{r}$$

$$u^3_{;3} = \frac{\partial u^3}{\partial z} + \Gamma^3_{3m}\, u^m = \frac{\partial u^3}{\partial z}$$

curl : $\mathbf{rot}\; \mathbf{u} = b^*\, e^{ikm}\, u_{m,k}\, \mathbf{b}_i = c^i\, \mathbf{b}_i$

$$c^1 = \frac{1}{r}\left(\frac{\partial u_3}{\partial \theta} - \frac{\partial u_2}{\partial z}\right) = \frac{1}{r}\frac{\partial u^3}{\partial \theta} - r\,\frac{\partial u^2}{\partial z}$$

$$c^2 = \frac{1}{r}\left(\frac{\partial u_1}{\partial z} - \frac{\partial u_3}{\partial r}\right) = \frac{1}{r}\frac{\partial u^1}{\partial z} - \frac{1}{r}\frac{\partial u^3}{\partial r}$$

$$c^3 = \frac{1}{r}\left(\frac{\partial u_2}{\partial r} - \frac{\partial u_1}{\partial \theta}\right) = r\,\frac{\partial u^2}{\partial r} - \frac{1}{r}\frac{\partial u^1}{\partial \theta} + 2\,u^2$$

vector gradient : $u_{i\,;\,k} = u_{i,k} - \Gamma^m_{ik}\, u_m$

$$\mathbf{U}_{;\,y} = \begin{array}{|c|c|c|}\hline u_{1,1} & u_{1,2} - \frac{1}{r}\; u_2 & u_{1,3} \\ \hline u_{2,1} - \frac{1}{r}\, u_2 & u_{2,2} + r\,u_2 & u_{2,3} \\ \hline u_{3,1} & u_{3,2} & u_{3,3} \\ \hline \end{array}$$

$$= \begin{array}{|c|c|c|}\hline u^1_{,1} & u^1_{,2} - r u^2 & u^1_{,3} \\ \hline (r)^2 u^2_{,1} + r\,u^2 & (r)^2 u^2_{,2} + r\,u^1 & (r)^2 u^2_{,3} \\ \hline u^3_{,1} & u^3_{,2} & u^3_{,3} \\ \hline \end{array}$$

Example 3 : Field operations in spherical coordinates

The spherical coordinate system (r, θ, β) and its Christoffel symbols are defined in Examples 2 of Sections 9.4.5 and 9.4.6. The covariant basis vectors $\mathbf{b}_i$ are the partial derivatives of the position vector $\mathbf{x}$ with respect to the local coordinates r, θ, β. The contravariant basis vectors $\mathbf{b}^m$ are obtained using the condition $\mathbf{b}_i \cdot \mathbf{b}^m = \delta_i^m$.

scalar field : $p = p(r, \theta, \beta)$

vector field : $\mathbf{u} = u^i(r, \theta, \beta)\,\mathbf{b}_i$

basis :
$$\mathbf{b}_1 = \begin{array}{|c|}\hline \cos\beta\ \cos\theta \\ \hline \cos\beta\ \sin\theta \\ \hline \sin\beta \\ \hline \end{array} \qquad \mathbf{b}_2 = \begin{array}{|c|}\hline -r\cos\beta\ \sin\theta \\ \hline r\cos\beta\ \cos\theta \\ \hline 0 \\ \hline \end{array} \qquad \mathbf{b}_3 = \begin{array}{|c|}\hline -r\sin\beta\ \cos\theta \\ \hline -r\sin\beta\ \sin\theta \\ \hline r\cos\beta \\ \hline \end{array}$$

$$\mathbf{b}^1 = \mathbf{b}_1 \qquad \mathbf{b}^2 = \frac{1}{r^2\cos^2\beta}\mathbf{b}_2 \qquad \mathbf{b}^3 = \frac{1}{r^2}\mathbf{b}_3$$

$$b_* = \det \mathbf{B}_* = r^2\cos\beta \qquad b^* = \det \mathbf{B}^* = \frac{1}{r^2}\sec\beta$$

metric :
$$\mathbf{G}_* = \begin{array}{|c|c|c|}\hline 1 & 0 & 0 \\ \hline 0 & r^2\cos^2\beta & 0 \\ \hline 0 & 0 & r^2 \\ \hline \end{array}$$

coordinates : $u_i = g_{ik}\,u^k$

$$u_1 = u^1 \qquad u_2 = (r)^2\cos^2\beta\,u^2 \qquad u_3 = (r)^2 u^3$$

gradient :
$$\mathbf{grad}\,p = \frac{\partial p}{\partial r}\mathbf{b}^1 + \frac{\partial p}{\partial \theta}\mathbf{b}^2 + \frac{\partial p}{\partial \beta}\mathbf{b}^3$$
$$= \frac{\partial p}{\partial r}\mathbf{b}_1 + \frac{1}{r^2}\sec^2\beta\,\frac{\partial p}{\partial \theta}\mathbf{b}_2 + \frac{1}{r^2}\frac{\partial p}{\partial \beta}\mathbf{b}_3$$

divergence : $\operatorname{div}\mathbf{u} = {u^1}_{;1} + {u^2}_{;2} + {u^3}_{;3}$

$${u^1}_{;1} = \frac{\partial u^1}{\partial r} + \Gamma^1_{1m}\,u^m = \frac{\partial u^1}{\partial r}$$

$${u^2}_{;2} = \frac{\partial u^2}{\partial \theta} + \Gamma^2_{2m}\,u^m = \frac{\partial u^2}{\partial \theta} + \frac{u^1}{r} - u^3\tan\beta$$

$${u^3}_{;3} = \frac{\partial u^3}{\partial \beta} + \Gamma^3_{3m}\,u^m = \frac{\partial u^3}{\partial \beta} + \frac{u^1}{r}$$

curl : $\mathbf{rot\,u} = b^{*}\, e^{ikm}\, u_{m,k}\, \mathbf{b}_i = c^i\, \mathbf{b}_i$

$$c^1 = \frac{1}{r^2} \sec\beta \left(\frac{\partial u_3}{\partial\theta} - \frac{\partial u_2}{\partial\beta}\right) = \sec\beta \frac{\partial u^3}{\partial\theta} - \cos\beta \frac{\partial u^2}{\partial\beta} + 2 \sin\beta\, u^2$$

$$c^2 = \frac{1}{r^2} \sec\beta \left(\frac{\partial u_1}{\partial\beta} - \frac{\partial u_3}{\partial r}\right) = \sec\beta \left(\frac{1}{r^2} \frac{\partial u^1}{\partial\beta} - \frac{\partial u^3}{\partial r} - \frac{2}{r}\, u^3\right)$$

$$c^3 = \frac{1}{r^2} \sec\beta \left(\frac{\partial u_2}{\partial r} - \frac{\partial u_1}{\partial\theta}\right) = \cos\beta \frac{\partial u^2}{\partial r} - \frac{1}{r^2} \sec\beta \frac{\partial u^1}{\partial\theta} + \frac{2}{r} \cos\beta\, u^2$$

vector gradient : $u_{i\,;\,k} = u_{i,k} - \Gamma^m_{ik}\, u_m$

$$\mathbf{U}_{;\,y} = \begin{array}{|c|c|c|} \hline u_{1,1} & u_{1,2} - \frac{1}{r}\, u_2 & u_{1,3} - \frac{1}{r}\, u_3 \\ \hline u_{2,1} - \frac{1}{r}\, u_2 & u_{2,2} + r\cos^2\beta\, u_1 - \sin\beta\, \cos\beta\, u_3 & u_{2,3} + \tan\beta\, u_2 \\ \hline u_{3,1} - \frac{1}{r}\, u_3 & u_{3,2} + \tan\beta\, u_2 & u_{3,3} + r\, u_1 \\ \hline \end{array}$$

9.4.10 NABLA CALCULUS

Introduction : The representation of field operations in local coordinate systems leads to complicated expressions which impede the understanding of the mathematical formulation of physical problems. A simplified notation which facilitates understanding is therefore introduced by defining the nabla operator and the Laplace operator. The rules of calculation for these operators are called the nabla calculus. In actual calculations the operators are replaced by their forms for the coordinate system being used.

Nabla operator : The representation of spatial field operations is simplified by introducing the nabla operator. The operator is the sum of three terms ; each term is the product of a local basis vector $\mathbf{b}^i$ and a partial derivative operator $\partial/\partial y^i$:

$$\nabla := \mathbf{b}^1 \frac{\partial}{\partial y^1} + \mathbf{b}^2 \frac{\partial}{\partial y^2} + \mathbf{b}^3 \frac{\partial}{\partial y^3}$$

A knowledge of the basis $\mathbf{B}^*$ is sufficient for constructing the nabla operator. Note that the derivatives of the position vector $\mathbf{x}$ with respect to the local coordinates y^1, y^2, y^3 used in the nabla operator do not yield the basis vectors $\mathbf{b}^i$ of the nabla operator, but rather the dual basis vectors $\mathbf{b}_m$. Using the nabla operator, the field operations may be expressed as products :

gradient of a scalar field : $\nabla p = \textbf{grad}\ p$

divergence of a vector field : $\nabla \cdot \mathbf{u} = \text{div}\ \mathbf{u}$

curl of a vector field : $\nabla \times \mathbf{u} = \textbf{rot u}$

gradient of a vector field : $\nabla \mathbf{u}^T = \textbf{dya u}$

Proof : Action of the nabla operator

(1) The gradient of the scalar field $p(y^1, y^2, y^3)$ is the tensor product of the nabla operator with the scalar field p :

$$\nabla p = \frac{\partial p}{\partial y^1}\mathbf{b}^1 + \frac{\partial p}{\partial y^2}\mathbf{b}^2 + \frac{\partial p}{\partial y^3}\mathbf{b}^3 = \textbf{grad}\ p$$

(2) The divergence of the vector field $\mathbf{u}(y^1, y^2, y^3)$ is the scalar product of the nabla operator with the vector field $\mathbf{u}$. The partial derivatives of the local basis vectors are expressed in terms of the covariant derivatives. With $\mathbf{u}_{,k} = u^i_{;k}\,\mathbf{b}_i$ this yields :

$$\begin{aligned}
\nabla \cdot \mathbf{u} &= (\mathbf{b}^1 \frac{\partial}{\partial y^1} + \mathbf{b}^2 \frac{\partial}{\partial y^2} + \mathbf{b}^3 \frac{\partial}{\partial y^3}) \cdot \mathbf{u} \\
&= \mathbf{b}^1 \cdot \frac{\partial \mathbf{u}}{\partial y^1} + \mathbf{b}^2 \cdot \frac{\partial \mathbf{u}}{\partial y^2} + \mathbf{b}^3 \frac{\partial \mathbf{u}}{\partial y^3} \\
&= u^i_{;1}\,\mathbf{b}^1 \cdot \mathbf{b}_i + u^i_{;2}\,\mathbf{b}^2 \cdot \mathbf{b}_i + u^i_{;3}\,\mathbf{b}^3 \cdot \mathbf{b}_i
\end{aligned}$$

$$\nabla \cdot \mathbf{u} = u^1_{;1} + u^2_{;2} + u^3_{;3} = \text{div}\ \mathbf{u}$$

(3) The curl of the vector field $\mathbf{u}(y^1, y^2, y^3)$ is the cross product of the nabla operator and the vector field **u**. The cross product is expressed in terms of the permutation tensor ε^{ikm} of the basis $\mathbf{B}^*$.

$$\nabla \times \mathbf{u} = \varepsilon^{ikm}\, u_{k\,,\,i}\, \mathbf{b}_m$$

$$\nabla \times \mathbf{u} = \textbf{rot u}$$

(4) The gradient of the vector field $\mathbf{u}(y^1, y^2, y^3)$ is the tensor product of the nabla operator and the vector field **u**. The partial derivatives of the local basis vectors are taken into account by using covariant derivatives of the coordinates of the vector field.

$$\begin{aligned}
\nabla \mathbf{u}^T &= (\mathbf{b}^1 \frac{\partial}{\partial y^1} + \mathbf{b}^2 \frac{\partial}{\partial y^2} + \mathbf{b}^3 \frac{\partial}{\partial y^3})\,(u_1 \mathbf{b}^1 + u_2 \mathbf{b}^2 + u_3 \mathbf{b}^3)^T \\
&= \mathbf{b}^i\,(u_{1\,;\,i}\, \mathbf{b}^1 + u_{2\,;\,i}\, \mathbf{b}^2 + u_{3\,;\,i}\, \mathbf{b}^3)^T = u_{m\,;\,i}\, \mathbf{b}^i\, (\mathbf{b}^m)^T \\
\nabla \mathbf{u}^T &= \textbf{dya u}
\end{aligned}$$

Rules of calculation for differential operators : The rules are written for scalar fields p and t and vector fields **u** and **w**. The formulations using the operators **grad**, div, **rot**, and **dya** are equivalent to the formulations containing the nabla operator.

(R1) Sums of tensor fields

$$\begin{aligned}
\nabla(p+t) &= \nabla p + \nabla t &&\Leftrightarrow && \textbf{grad}\,(p+t) = \textbf{grad}\,p + \textbf{grad}\,t \\
\nabla \cdot (\mathbf{u}+\mathbf{w}) &= \nabla \cdot \mathbf{u} + \nabla \cdot \mathbf{w} &&\Leftrightarrow && \text{div}\,(\mathbf{u}+\mathbf{w}) = \text{div}\,\mathbf{u} + \text{div}\,\mathbf{w} \\
\nabla \times (\mathbf{u}+\mathbf{w}) &= \nabla \times \mathbf{u} + \nabla \times \mathbf{w} &&\Leftrightarrow && \textbf{rot}\,(\mathbf{u}+\mathbf{w}) = \textbf{rot u} + \textbf{rot w} \\
\nabla(\mathbf{u}+\mathbf{w})^T &= \nabla \mathbf{u}^T + \nabla \mathbf{w}^T &&\Leftrightarrow && \textbf{dya}\,(\mathbf{u}+\mathbf{w}) = \textbf{dya u} + \textbf{dya w}
\end{aligned}$$

(R2) Products of tensor fields

$$\begin{aligned}
\nabla(pt) &= p\nabla t + t\,\Delta p &&\Leftrightarrow && \text{grad}\,(pt) = p\,\text{grad}\,t + t\,\text{grad}\,p \\
\nabla \cdot (p\mathbf{u}) &= \nabla p \cdot \mathbf{u} + p\nabla \cdot \mathbf{u} &&\Leftrightarrow && \text{div}\,(p\mathbf{u}) = \textbf{grad}\,p \cdot \mathbf{u} + p\,\text{div}\,\mathbf{u} \\
\nabla \times (p\mathbf{u}) &= \nabla p \times \mathbf{u} + p\nabla \times \mathbf{u} &&\Leftrightarrow && \textbf{rot}\,(p\mathbf{u}) = \textbf{grad}\,p \times \mathbf{u} + p\,\textbf{rot u} \\
\nabla(p\mathbf{u})^T &= \nabla p\ \mathbf{u}^T + p\nabla \mathbf{u}^T &&\Leftrightarrow && \textbf{dya}\,(p\mathbf{u}) = \textbf{grad}\,p\mathbf{u}^T + p\,\textbf{dya u}
\end{aligned}$$

(R3) Products of vector fields

$$\begin{aligned}
\textbf{grad}\,(\mathbf{u} \cdot \mathbf{w}) &= (\textbf{dya u})\,\mathbf{w} + (\textbf{dya w})\mathbf{u} \\
\text{div}\,(\mathbf{u} \times \mathbf{w}) &= \mathbf{w} \cdot \textbf{rot u} - \mathbf{u} \cdot \textbf{rot w} \\
\textbf{rot}\,(\mathbf{u} \times \mathbf{w}) &= (\textbf{dya u})^T \mathbf{w} - (\textbf{dya w})^T \mathbf{u} + \mathbf{u}\,\text{div}\,\mathbf{w} - \mathbf{w}\,\text{div}\,\mathbf{u}
\end{aligned}$$

These rules of calculation may be conveniently expressed using covariant derivatives of the coordinates of the field :

$$\begin{aligned}
\textbf{grad}\,(\mathbf{u} \cdot \mathbf{w}) &= (u_{k\,;\,i}\, w^k + u^k\, w_{k\,;\,i})\mathbf{b}^i \\
\text{div}\,(\mathbf{u} \times \mathbf{w}) &= \varepsilon^{ikm}\,(u_{k\,;\,i}\, w_m + u_k\, w_{m\,;\,i}) \\
\textbf{rot}\,(\mathbf{u} \times \mathbf{w}) &= (u^i\, w^k - u^k\, w^i)_{;\,k}\, \mathbf{b}_i
\end{aligned}$$

Note : The nabla operator ∇ is written as a vector in these expressions. However, one must always keep in mind which quantity the derivatives are acting on. Expressions with nabla operators are generally not commutative.

Proof : Rules of calculation for differential operators

(R1) Sums of tensor fields

$$\begin{aligned}
\nabla(p+t) &= (\mathbf{b}^i \frac{\partial}{\partial y^i})(p+t) = \mathbf{b}^i \frac{\partial p}{\partial y^i} + \mathbf{b}^i \frac{\partial t}{\partial y^i} \\
&= \nabla p + \nabla t \\
\nabla \cdot (\mathbf{u}+\mathbf{w}) &= (\mathbf{b}^i \frac{\partial}{\partial y^i}) \cdot (u_k \mathbf{b}^k + w_m \mathbf{b}^m) = u_{k;i}\, \mathbf{b}^i \cdot \mathbf{b}^k + w_{m;i}\, \mathbf{b}^i \cdot \mathbf{b}^m \\
&= \nabla \cdot \mathbf{u} + \nabla \cdot \mathbf{w} \\
\nabla \times (\mathbf{u}+\mathbf{w}) &= (\mathbf{b}^i \frac{\partial}{\partial y^i}) \times (u_k \mathbf{b}^k + w_m \mathbf{b}^m) = u_{k;i}\, \mathbf{b}^i \times \mathbf{b}^k + w_{m;i}\, \mathbf{b}^i \times \mathbf{b}^m \\
&= \nabla \times \mathbf{u} + \nabla \times \mathbf{w} \\
\nabla(\mathbf{u}+\mathbf{w})^T &= (\mathbf{b}^i \frac{\partial}{\partial y^i})(u_k \mathbf{b}^k + w_m \mathbf{b}^m)^T = u_{k;i}\, \mathbf{b}^i (\mathbf{b}^k)^T + w_{m;i}\, \mathbf{b}^i (\mathbf{b}^m)^T \\
&= \nabla \mathbf{u}^T + \nabla \mathbf{w}^T
\end{aligned}$$

(R2) Products of tensor fields

$$\begin{aligned}
\nabla(pt) &= \mathbf{b}^i \frac{\partial (pt)}{\partial y^i} = p\, \mathbf{b}^i \frac{\partial t}{\partial y^i} + t\, \mathbf{b}^i \frac{\partial p}{\partial y^i} \\
&= p \nabla t + t \nabla p \\
\nabla \cdot (p\mathbf{u}) &= (\mathbf{b}^i \frac{\partial}{\partial y^i}) \cdot (p\, u^k\, \mathbf{b}_k) = p\, u^i_{;i} + u^i p_{,i} \\
&= p(\nabla \cdot \mathbf{u}) + \nabla p \cdot \mathbf{u} \\
\nabla \times (p\mathbf{u}) &= (\mathbf{b}^i \frac{\partial}{\partial y^i}) \times (p\, u_k\, \mathbf{b}^k) = \varepsilon^{ikm} (p\, u_k)_{;i}\, \mathbf{b}_m \\
&= p(\nabla \times \mathbf{u}) + \nabla \mathbf{p} \times \mathbf{u} \\
\nabla(p\mathbf{u})^T &= (\mathbf{b}^i \frac{\partial}{\partial y^i})\, (p\, u_k\, \mathbf{b}^k)^T = p\, u_{k;i}\, \mathbf{b}^i (\mathbf{b}^k)^T + p_{,i}\, \mathbf{b}^i\, u_k\, (\mathbf{b}^k)^T \\
&= p \nabla \mathbf{u}^T + \nabla p\, \mathbf{u}^T
\end{aligned}$$

(R3) Products of vector fields

– $$\nabla(\mathbf{u}\cdot\mathbf{w}) = (\mathbf{b}^i \frac{\partial}{\partial y^i})(u_k\, w^k) = u_{k;i}\, w^k\, \mathbf{b}^i + u^k\, w_{k;i}\, \mathbf{b}^i$$

$$(\mathbf{dya\ u})\,\mathbf{w} = u_{k;i}\,\mathbf{b}^i\,(\mathbf{b}^k)^T\cdot(w^s\,\mathbf{b}_s) = u_{k;i}\,w^k\,\mathbf{b}^i$$

$$(\mathbf{dya\ w})\,\mathbf{u} = w_{k;i}\,\mathbf{b}^i\,(\mathbf{b}^k)^T\cdot(u^s\,\mathbf{b}_s) = w_{k;i}\,u^k\,\mathbf{b}^i$$

$$\mathbf{grad}\,(\mathbf{u}\cdot\mathbf{w}) = \nabla(\mathbf{u}\cdot\mathbf{w}) = (\mathbf{dya\ u})\,\mathbf{w} + (\mathbf{dya\ w})\,\mathbf{u}$$

– $$\begin{aligned}\nabla(\mathbf{u}\times\mathbf{w}) &= (\mathbf{b}^i \frac{\partial}{\partial y^i})\cdot(\varepsilon^{skm}\,u_k\,w_m\,\mathbf{b}_s) = (\varepsilon^{ikm}\,u_k\,w_m)_{;i} \\ &= \varepsilon^{ikm}\,(u_{k;i}\,w_m + u_k\,w_{m;i})\end{aligned}$$

$$\text{div}\ (\mathbf{u}\times\mathbf{w}) = \mathbf{w}\cdot\mathbf{rot\ u} + \mathbf{u}\cdot\mathbf{rot\ w}$$

– $$\begin{aligned}\nabla\times(\mathbf{u}\times\mathbf{w}) &= (\mathbf{b}^k \frac{\partial}{\partial y^k})\times(\varepsilon_{rsm}\,u^r\,w^s\,\mathbf{b}^m) = \varepsilon^{ikm}(\varepsilon_{rsm}\,u^r\,w^s)_{;k}\,\mathbf{b}_i \\ &= (\delta^i_r\,\delta^k_s - \delta^i_s\,\delta^k_r)\,(u^r\,w^s)_{;k}\,\mathbf{b}_i = (u^i w^k)_{;k}\,\mathbf{b}_i - (u^k w^i)_{;k}\,\mathbf{b}_i\end{aligned}$$

$$(\mathbf{dya\ u})^T\,\mathbf{w} + \mathbf{u}\ \text{div}\ \mathbf{w} = (u^i_{;k}\,w^k + u^i\,w^k_{;k})\,\mathbf{b}_i = (u^i\,w^k)_{;k}\,\mathbf{b}_i$$

$$(\mathbf{dya\ w})^T\,\mathbf{u} + \mathbf{w}\ \text{div}\ \mathbf{u} = (w^i_{;k}\,u^k + w^i\,u^k_{;k})\,\mathbf{b}_i = (u^k\,w^i)_{;k}\,\mathbf{b}_i$$

$$\mathbf{rot}\ (\mathbf{u}\times\mathbf{w}) = (\mathbf{dya\ u})^T\,\mathbf{w} - (\mathbf{dya\ w})^T\,\mathbf{u} + \mathbf{u}\ \text{div}\ \mathbf{w} - \mathbf{w}\ \text{div}\ \mathbf{u}$$

Laplace operator : Repeated application of the nabla operator leads to different compositions of the differential operators already defined. However, the scalar product $\nabla\cdot\nabla$ of the nabla operator with itself cannot be expressed as a composition of two of the operators **grad**, div and **rot**. The Laplace operator $\Delta = \nabla\cdot\nabla$ is therefore introduced. This symbolic operator may be applied to scalar and vector fields.

$$\begin{aligned}
\nabla\cdot(\nabla p) &= \text{div}\ \mathbf{grad}\ p &&= \Delta p \\
\nabla\times(\nabla p) &= \mathbf{rot\ grad}\ p &&= \mathbf{0} \\
(\nabla\cdot\nabla)\mathbf{u} &= \Delta\mathbf{u} &&= \mathbf{grad}\ \text{div}\ \mathbf{u} - \mathbf{rot\ rot\ u} \\
\nabla(\nabla\cdot\mathbf{u}) &= \mathbf{grad}\ \text{div}\ \mathbf{u} &&= \Delta\mathbf{u} + \mathbf{rot\ rot\ u} \\
\nabla\cdot(\nabla\times\mathbf{u}) &= \text{div}\ \mathbf{rot\ u} &&= 0 \\
\nabla\times(\nabla\times\mathbf{u}) &= \mathbf{rot\ rot\ u} &&= \mathbf{grad}\ \text{div}\ \mathbf{u} - \Delta\mathbf{u}
\end{aligned}$$

The Laplace operator Δ is applied to a vector $\mathbf{u} = u_m\,\mathbf{b}^m$ by applying it to each coordinate u_m of the vector :

$$\Delta p = p_{;ik}\,g^{ik} \quad \text{with} \quad p_{;ik} = p_{,ik} - \Gamma^m_{ik}\,p_{,m}$$

$$\Delta\mathbf{u} = \Delta u_m\,\mathbf{b}^m \quad \text{with} \quad \Delta u_m = u_{m;ik}\,g^{ik}$$

g^{ik} metric coefficients of the basis $\mathbf{B}^*$

Proof : Compositions of the nabla operator

- **rot grad** p $= \varepsilon^{ikm}\, p_{,ik}\, \mathbf{b}_m = \mathbf{0}$

  example m = 1 : $(\varepsilon^{231}\, p_{,32} + \varepsilon^{321}\, p_{,23})\mathbf{b}_1 = (p_{,32} - p_{,23})\mathbf{b}_1 = \mathbf{0}$

- div **rot u** $= \text{div}\,(\varepsilon^{ikm}\, u_{k;i}\, \mathbf{b}_m) = \varepsilon^{ikm}\, u_{k;im} = 0$

  example k = 1 : $(\varepsilon^{312}\, u_{1;32} + \varepsilon^{213}\, u_{1;23}) = u_{1;32} - u_{1;23} = 0$

- Applying the Laplace operator to a scalar field p yields the contracted product of the covariant second derivatives $p_{;ik}$ of the scalar with the metric coefficients g^{ik}:

$$\Delta p = (\mathbf{b}^k \frac{\partial}{\partial y^k}) \cdot (\mathbf{b}^i \frac{\partial p}{\partial y^i}) = p_{;ik}\, \mathbf{b}^i \cdot \mathbf{b}^k$$

$$\Delta p = p_{;ik}\, g^{ik}$$

- The remaining three formulas contain $\Delta\mathbf{u}$ and differ only in the arrangement of the terms on the two sides of the equation. It suffices to prove on of these formulas.

$$\begin{aligned}
\textbf{rot u} &= \varepsilon^{ikm}\, u_{k;i}\, \mathbf{b}_m = g^{ir}\, g^{ks}\, g^{mt}\, \varepsilon_{rst}\, u_{k;i}\, \mathbf{b}_m \\
&= \varepsilon_{rst}\, g^{ir}\, u^s_{;i}\, \mathbf{b}^t \\
\textbf{rot rot u} &= \varepsilon^{jkm}\, w_{k;j}\, \mathbf{b}_m \\
&= \varepsilon^{mjk}\, \varepsilon_{rsk}\, g^{ir}\, u^s_{;ij}\, \mathbf{b}_m \\
&= (\delta^m_r\, \delta^j_s - \delta^m_s\, \delta^j_r)\, g^{ir}\, u^s_{;ij}\, \mathbf{b}_m \\
&= g^{im}\, u^s_{;is}\, \mathbf{b}_m - g^{ir}\, u^m_{;ir}\, \mathbf{b}_m \\
&= u^s_{;is}\, \mathbf{b}^i - g^{ir}\, u_{m;ir}\, \mathbf{b}^m \\
\textbf{rot rot u} &= \textbf{grad}\ \text{div}\ \mathbf{u} - \Delta\mathbf{u}
\end{aligned}$$

Example 1 : Form of the nabla operator in various coordinate systems

The unit vectors $\mathbf{e}^1$, $\mathbf{e}^2$, $\mathbf{e}^3$ are chosen as a basis for the cartesian coordinate system x^1, x^2, x^3. For the cylindrical coordinate system r, θ, z and the spherical coordinate system r, θ, β, the covariant basis vectors $\mathbf{b}_1$, $\mathbf{b}_2$, $\mathbf{b}_3$ are obtained by taking the partial derivative $\partial\mathbf{x}/\partial y^i$ of the position vector; the contravariant basis vectors $\mathbf{b}^1$, $\mathbf{b}^2$, $\mathbf{b}^3$ are determined by the duality condition $\mathbf{b}_i \cdot \mathbf{b}^m = \delta^m_i$.

general : $\nabla p = \mathbf{b}^1 \dfrac{\partial p}{\partial y^1} + \mathbf{b}^2 \dfrac{\partial p}{\partial y^2} + \mathbf{b}^3 \dfrac{\partial p}{\partial y^3}$

cartesian : $\mathbf{b}^1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \qquad \mathbf{b}^2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \qquad \mathbf{b}^3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$$\nabla p = \mathbf{b}^1 \frac{\partial p}{\partial x^1} + \mathbf{b}^2 \frac{\partial p}{\partial x^2} + \mathbf{b}^3 \frac{\partial p}{\partial x^3}$$

cylindrical :

$$\begin{array}{lcllcll} x^1 & = & r\cos\theta & r & = & y^1 : & \text{radial distance} \\ x^2 & = & r\sin\theta & \theta & = & y^2 : & \text{angle} \\ x^3 & = & z & z & = & y^3 : & \text{axial distance} \end{array}$$

$$\mathbf{b}^1 = \begin{array}{|c|}\hline \cos\theta \\ \hline \sin\theta \\ \hline 0 \\ \hline \end{array} \qquad \mathbf{b}^2 = \begin{array}{|c|}\hline -\frac{1}{r}\sin\theta \\ \hline \frac{1}{r}\cos\theta \\ \hline 0 \\ \hline \end{array} \qquad \mathbf{b}^3 = \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline 1 \\ \hline \end{array}$$

$$\nabla p = \mathbf{b}^1 \frac{\partial p}{\partial r} + \mathbf{b}^2 \frac{\partial p}{\partial \theta} + \mathbf{b}^3 \frac{\partial p}{\partial z}$$

spherical :

$$\begin{array}{lcllcll} x^1 & = & r\cos\beta\cos\theta & r & = & y^1 : & \text{radial distance} \\ x^2 & = & r\cos\beta\sin\theta & \theta & = & y^2 : & \text{azimuthal angle} \\ x^3 & = & r\sin\beta & \beta & = & y^3 : & \text{polar angle} \end{array}$$

$$\mathbf{b}^1 = \begin{array}{|c|}\hline \cos\beta\cos\theta \\ \hline \cos\beta\sin\theta \\ \hline \sin\beta \\ \hline \end{array} \qquad \mathbf{b}^2 = \begin{array}{|c|}\hline -\frac{1}{r}\sec\beta\,\sin\theta \\ \hline \frac{1}{r}\sec\beta\cos\theta \\ \hline 0 \\ \hline \end{array} \qquad \mathbf{b}^3 = \begin{array}{|c|}\hline -\frac{1}{r}\sin\beta\cos\theta \\ \hline -\frac{1}{r}\sin\beta\sin\theta \\ \hline \frac{1}{r}\cos\beta \\ \hline \end{array}$$

$$\nabla p = \mathbf{b}^1 \frac{\partial p}{\partial r} + \mathbf{b}^2 \frac{\partial p}{\partial \theta} + \mathbf{b}^3 \frac{\partial p}{\partial \beta}$$

Example 2 : Form of the Laplace operator in various coordinate systems

The Laplace operator is obtained using the general formula. The metric coefficients g^{ik} are determined with the basis vectors $\mathbf{b}^i$ of the preceding example. The Christoffel symbols in the covariant derivatives are taken from Examples 1 and 2 of Section 9.4.6.

general : $\Delta p = p_{;\,ik}\, g^{ik}$

cartesian : $$\Delta p = \frac{\partial^2 p}{\partial x_1^2} + \frac{\partial^2 p}{\partial x_2^2} + \frac{\partial^2 p}{\partial x_3^2}$$

cylindrical : $g^{11} = 1 \quad g^{22} = \frac{1}{r^2} \quad g^{33} = 1 \qquad$ otherwise $g^{ik} = 0$

$\Gamma^1_{22} = -r \qquad$ otherwise $\Gamma^m_{ii} = 0$

$$\Delta p = \frac{\partial^2 p}{\partial r^2} + \frac{1}{r^2}\frac{\partial^2 p}{\partial \theta^2} + \frac{1}{r}\frac{\partial p}{\partial r}$$

spherical : $g^{11} = 1 \quad g^{22} = \frac{1}{r^2}\sec^2\beta \quad g^{33} = \frac{1}{r^2} \qquad$ otherwise $g^{ik} = 0$

$\Gamma^1_{22} = -r\cos^2\beta \quad \Gamma^1_{33} = -r \quad \Gamma^3_{22} = \sin\beta\cos\beta \qquad$ otherwise $\Gamma^m_{ii} = 0$

$$\Delta p = \frac{\partial^2 p}{\partial r^2} + \frac{1}{(r\cos\beta)^2}\frac{\partial^2 p}{\partial \theta^2} + \frac{1}{r^2}\frac{\partial^2 p}{\partial \beta^2} + \frac{2}{r}\frac{\partial p}{\partial r} - \frac{\tan\beta}{r^2}\frac{\partial p}{\partial \beta}$$

9.4.11 SPECIAL VECTOR FIELDS

Introduction : A vector field is said to be special if the values of its coordinates have certain properties. The coordinates of a general vector field do not possess these properties. Conservative, irrotational, source-free and solenoidal vector fields are defined in the following. The problem of finding such fields leads to the differential equations of field theory, in particular to the differential equations of Laplace and Poisson.

Potential of a vector field : Assume that at every point of the space $\mathbb{R}^3$ with the local coordinates y^1, y^2, y^3 the gradient of a scalar field $p(\mathbf{y})$ is equal to the vector field $\mathbf{u}(\mathbf{y})$. Then the scalar field p is called the potential of the vector field **u**. The potential is only determined up to an additive constant c.

$$\mathbf{u} = u_i \, \mathbf{b}^i = \mathbf{grad}\, p$$

$$u_i = \frac{\partial p}{\partial y^i}$$

Conservative vector field : A vector field **u** is said to be conservative if the value of the scalar line integral of the vector field depends only on the endpoints $\mathbf{x}_1 = \mathbf{x}(\mathbf{y}_1)$ and $\mathbf{x}_2 = \mathbf{x}(\mathbf{y}_2)$ of the line. Every closed line integral of a conservative vector field is therefore zero.

$$\mathbf{u} \text{ is conservative} \quad \Leftrightarrow \quad \oint \mathbf{u} \cdot d\mathbf{x} = 0$$

A vector field **u** for which a potential p can be specified is conservative. The value of the line integral of the vector field **u** between given endpoints $\mathbf{x}_1$ and $\mathbf{x}_2$ depends only on the values $p_1 = p(\mathbf{x}_1)$ and $p_2 = p(\mathbf{x}_2)$ of the potential at the endpoints :

$$\int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{u} \cdot d\mathbf{x} = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{grad}\, p \cdot d\mathbf{x} = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \left(\frac{\partial p}{\partial y^i} \mathbf{b}^i\right) \cdot (\mathbf{b}_m \, dy^m)$$

$$\int_{\mathbf{x}_1}^{\mathbf{x}_2} \mathbf{u} \cdot d\mathbf{x} = \int_{\mathbf{x}_1}^{\mathbf{x}_2} \frac{\partial p}{\partial y^i} dy^i = \int_{p_1}^{p_2} dp = p_2 - p_1$$

Irrotational vector field : A vector field $\mathbf{u}(y^1, y^2, y^3)$ is said to be irrotational if its curl is zero at every point of $\mathbb{R}^3$. The vector gradient of an irrotational field is symmetric.

$$\mathbf{u} \text{ is irrotational} \quad :\Leftrightarrow \quad \mathbf{rot}\, \mathbf{u} = \mathbf{0}$$

$$\varepsilon^{ikm} u_{m\,;\,k} \mathbf{b}_i = \mathbf{0} \quad \Rightarrow \quad u_{k\,;\,m} = u_{m\,;\,k}$$

A conservative vector field **u** is irrotational. To prove this, consider the curl of the gradient of the potential p of the vector field :

$$\mathbf{u} = \mathbf{grad}\, p \quad \wedge \quad \mathbf{rot}\, \mathbf{grad}\, p = \mathbf{0} \quad \Rightarrow \quad \mathbf{rot}\, \mathbf{u} = \mathbf{0}$$

Source-free vector field : A vector field $\mathbf{u}(y^1, y^2, y^3)$ is said to be source-free if its divergence is zero at every point of $\mathbb{R}^3$. The trace of the vector gradient of a source-free vector field $\mathbf{u}$ is therefore zero. However, the vector gradient of a source-free field is not necessarily antisymmetric.

$$\mathbf{u} \text{ is source-free} \quad :\Leftrightarrow \quad \text{div } \mathbf{u} = 0$$

$$\text{tr } (\mathbf{dya\ u}) = u^1_{;1} + u^2_{;2} + u^3_{;3} = 0$$

Solenoidal vector field : A vector field $\mathbf{u}(y^1, y^2, y^3)$ is said to be solenoidal if its vector gradient is antisymmetric. A solenoidal field is source-free.

$$\mathbf{u} \text{ is solenoidal} \quad :\Leftrightarrow \quad u_{i\,;\,k} = -u_{k\,;\,i}$$

$$\text{div } \mathbf{u} = g^{ik}\, u_{k\,;\,i} = -g^{ik}\, u_{i\,;\,k} = -g^{ik}\, u_{k\,;\,i} = 0$$

The increment $d\mathbf{u}$ of a solenoidal field is orthogonal to the curl $\mathbf{r} = \mathbf{rot\ u}$ of the field and to the increment $d\mathbf{x}$ of the position vector :

$$d\mathbf{x} = dy^i\, \mathbf{b}_i$$

$$\mathbf{r} = r^i\, \mathbf{b}_i \qquad \text{with} \quad r^i = \varepsilon^{ikm}\, u_{m\,;\,k}$$

$$d\mathbf{u} = du_i\, \mathbf{b}^i \qquad \text{with} \quad du_i = u_{i\,;\,k}\, dy^k \quad \text{and} \quad u_{m\,;\,k} = \frac{1}{2}\, \varepsilon_{ikm}\, r^i$$

$$\mathbf{r} \cdot d\mathbf{u} = d\mathbf{x} \cdot d\mathbf{u} = 0$$

Proof : Properties of a solenoidal vector field

$$r^1 = \varepsilon^{1km}\, u_{m\,;\,k} = (u_{3\,;\,2} - u_{2\,;\,3})\, b^*$$

$$r^2 = \varepsilon^{2km}\, u_{m\,;\,k} = (u_{1\,;\,3} - u_{3\,;\,1})\, b^*$$

$$r^3 = \varepsilon^{3km}\, u_{m\,;\,k} = (u_{2\,;\,1} - u_{1\,;\,2})\, b^*$$

The contracted product of the permutation tensor ε_{ikm} and the curl r^i is exemplarily calculated for $k = 1$, $m = 2$ and for $k = 2$, $m = 1$:

$$\varepsilon_{ikm}\ r^i = e_{1km}\,(u_{3\,;\,2}) - u_{2\,;\,3} + e_{2km}\,(u_{1\,;\,3} - u_{3\,;\,1}) + e_{3km}\,(u_{2\,;\,1} - u_{1\,;\,2})$$

$$\varepsilon_{i12}\ r^i = e_{312}\,(u_{2\,;\,1} - u_{1\,;\,2}) = 2\, u_{2\,;\,1}$$

$$\varepsilon_{i21}\ r^i = e_{321}\,(u_{2\,;\,1} - u_{1\,;\,2}) = 2\, u_{1\,;\,2}$$

The relationships $\varepsilon_{i23}\ r^i = 2\, u_{3\,;\,2}$ and $\varepsilon_{i31}\ r^i = 2\, u_{1\,;\,3}$ are obtained in analogy with the relationship $\varepsilon_{i12}\, r^i = 2\, u_{2\,;\,1}$. The covariant derivatives of the coordinates of the solenoidal field may therefore be expressed as a contracted product of the permutation tensor and the curl of the field :

$$u_{m\,;\,k} = \frac{1}{2}\, \varepsilon_{ikm}\, r^i$$

The orthogonality of the vector d**u** to the vectors **r** and d**x** follows by substituting the expressions for the increment d**u** :

$$\begin{aligned} \mathbf{r}\cdot d\mathbf{u} &= r^i\, u_{k\,;\,i}\, dy^k &&= \frac{1}{2}\,\varepsilon_{ikm}\, r^i\, r^m\, dy^k &&= 0 \\ d\mathbf{x}\cdot d\mathbf{u} &= u_{k\,;\,i}\, dy^i\, dy^k &&= \frac{1}{2}\,\varepsilon_{ikm}\, r^m\, dy^i\, dy^k &&= 0 \end{aligned}$$

Laplace's differential equation : A scalar field p(**x**) whose gradient is a source-free and irrotational vector field is to be determined. The gradient of every continuous scalar field is irrotational. However, the gradient of the scalar field is source-free only if the scalar field satisfies the Laplace equation $\Delta p = 0$.

$$\begin{aligned} \mathbf{u} &= \mathbf{grad}\; p \\ \mathbf{rot}\;\mathbf{u} &= \varepsilon^{ikm}\,\frac{\partial^2 p}{\partial y^k\,\partial y^m}\,\mathbf{b}_i &&= \mathbf{0} \\ \text{div}\;\mathbf{u} &= \text{div}\;\mathbf{grad}\; p &&= \Delta p \\ \text{div}\;\mathbf{u} &= 0 \quad \Rightarrow \quad \Delta p &&= 0 \end{aligned}$$

Harmonic functions : A function is said to be harmonic if it satisfies the Laplace equation. The solution of the Laplace equation on a given domain V is determined by the values prescribed on the boundary ∂V of the domain (boundary conditions). The following cases are distinguished according to the type of boundary conditions imposed :

Dirichlet problem : The scalar field p(**x**) takes prescribed values $p_0(\mathbf{x})$ on the boundary ∂V.

$$\begin{aligned} \Delta p &= 0 && \text{for} \quad \mathbf{x} \in V \\ p &= p_0 && \text{for} \quad \mathbf{x} \in \partial V \end{aligned}$$

Neumann problem : The derivative of the scalar field p(**x**) in the direction of the unit normal of the boundary ∂V takes prescribed values $t_0(\mathbf{x})$.

$$\begin{aligned} \Delta p &= 0 && \text{for} \quad \mathbf{x} \in V \\ \frac{\partial p}{\partial n} &= t_0 && \text{for} \quad \mathbf{x} \in \partial V \end{aligned}$$

General problem : A linear combination of the scalar field p(**x**) and its derivative in the direction of the unit normal of the boundary ∂V with given coefficients a(**x**) and b(**x**) takes prescribed values $c_0(\mathbf{x})$.

$$\begin{aligned} \Delta p &= 0 && && \text{for} \quad \mathbf{x} \in V \\ ap + b\,\frac{\partial p}{\partial n} &= c_0 && \wedge \quad a^2 + b^2 \neq 0 && \text{for} \quad \mathbf{x} \in \partial V \end{aligned}$$

Poisson's differential equation : A scalar field p($\mathbf{x}$) whose gradient has a given divergence (source density) q($\mathbf{x}$) at every point $\mathbf{x}$ of a domain V is to be determined. This scalar field satisfies the Poisson equation. Due to the term q($\mathbf{x}$) the equation is inhomogeneous, in contrast to the Laplace equation.

$$\text{div}\,\mathbf{u} \;=\; \text{div}\,\mathbf{grad}\,\text{p} \;=\; \Delta\text{p} \;=\; \text{q}(\mathbf{x})$$

The solution of the Poisson equation for a given domain V and a given source density q($\mathbf{x}$) is determined by the values prescribed on the boundary ∂V of the domain (boundary conditions).

Vector potential : A vector field $\mathbf{u}$ which has a given curl (vorticity density) $\mathbf{w}(\mathbf{x})$ at every point $\mathbf{x}$ of a domain V is to be determined. Since the curl of the field $\mathbf{u}$ must satisfy the equation div $\mathbf{rot\,u} = 0$, the function $\mathbf{w}(\mathbf{x})$ cannot be chosen arbitrarily : It must satisfy the condition div $\mathbf{w} = 0$.

$$\mathbf{rot\,u} \;=\; \mathbf{w} \quad \wedge \quad \text{div}\,\mathbf{rot\,u} \;=\; 0 \qquad \Rightarrow \qquad \text{div}\,\mathbf{w} \;=\; 0$$

If the ansatz $\mathbf{u} = \mathbf{rot\,a}$ with div $\mathbf{a} = 0$ is chosen for the field $\mathbf{u}$, the vector field $\mathbf{a}$ must satisfy a Poisson equation. In analogy with the scalar potential p, which also satisfies a Poisson equation, the field $\mathbf{a}(\mathbf{x})$ is called the vector potential of the field $\mathbf{u}(\mathbf{x})$.

$$\Delta\mathbf{a} \;=\; \mathbf{grad}\,\text{div}\,\mathbf{a} \;-\; \mathbf{rot\,rot\,a}$$

$$\mathbf{u} \;=\; \mathbf{rot\,a} \quad \wedge \quad \text{div}\,\mathbf{a} \;=\; 0 \qquad \Rightarrow \qquad \Delta\mathbf{a} \;=\; -\mathbf{rot\,u} \;=\; -\mathbf{w}$$

It follows from $\Delta\mathbf{a} = -\mathbf{w}$ that the inhomogeneity $-\mathbf{w}$ is the source density of the field $\mathbf{a}$. The field $\mathbf{u}$ and the vorticity density $\mathbf{w}$ are obtained by differentiating the vector potential $\mathbf{a}$.

9.4.12 INTEGRAL THEOREMS

Introduction : It is often convenient to exploit the relationships between volume integrals, surface integrals and line integrals of tensor fields in the mathematical formulation of physical problems. The theorems of Gauss, Stokes and Green describe these relationships.

In order to formulate Gauss' theorem, a control volume with a closed surface is first defined in the space $\mathbb{R}^3$. The integral of the divergence of a vector field over the control volume is equal to the scalar integral of the vector field over the surface of the control volume.

In order to formulate Stokes' theorem, a control surface with a closed boundary is first defined in the space $\mathbb{R}^3$. A definite relationship between the surface normal and the orientation of the boundary is established. The scalar integral of the curl of a vector field over the control surface is equal to the scalar integral of the vector field over the oriented boundary.

Green's theorems follow from Gauss' theorem if the vector field is replaced by the product of a scalar field and the gradient of another scalar field. Green's theorems are particularly useful for representing potential fields.

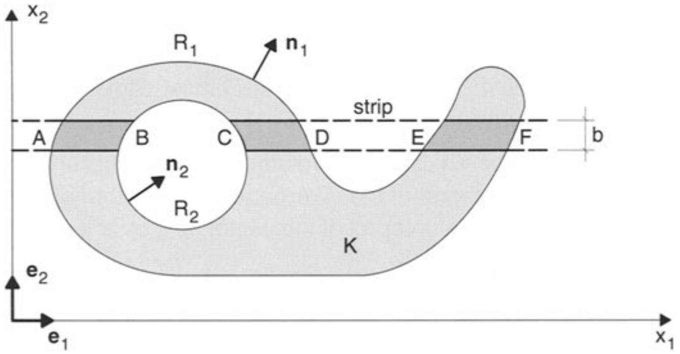
Control volume : Let a subspace V of the euclidean space $\mathbb{R}^3$ be connected. Let the boundary F of the subspace consist of one or more surface fragments which are piecewise continuous. Let the global coordinates x^1, x^2, x^3 of the points of the subspace V in the canonical basis $\mathbf{e}_1$, $\mathbf{e}_2$, $\mathbf{e}_3$ of the space $\mathbb{R}^3$ be bounded. Then the subspace V is called a control volume in the space $\mathbb{R}^3$.

Surface normal : The direction of the unit normal **n** of a surface F in the euclidean space $\mathbb{R}^3$ is not uniquely determined. If **n** is a normal of F at the point **x**, then $-\mathbf{n}$ is also a normal of F at the point **x**. The formulation of Gauss' theorem requires a distinction between the normals **n** and $-\mathbf{n}$.

The boundary of the control volume V consists of piecewise continuous surface fragments $F_1, F_2, \ldots$. At the inner points of a surface fragment F_i, the direction of the normal **n** is chosen such that for a sufficiently small magnitude $da > 0$ none of the points $\mathbf{x} + \mathbf{n}\, da$ belong to the volume V. The normal **n** is therefore directed outwards. On the boundary of F_i, the normal **n** is the limit of the normals of the neighboring points in F_i.

Note : The parametric representation of surfaces in Section 9.4.8 fixes the direction of the surface normals **n**. This is due to the fact that the surface parameters (s, t) of the point coordinates $x^i(s, t)$ form an ordered pair and the area vector is defined using the cross product $d\mathbf{a} = d\mathbf{u} \times d\mathbf{w}$, where $d\mathbf{u}$ contains the derivatives of the position coordinates x^i with respect to s and $d\mathbf{w}$ contains the derivatives with respect to t.

Example 1 : Control domain in the space $\mathbb{R}^2$



Consider the control domain K in the euclidean plane $\mathbb{R}^2$. The boundary R of the control domain consists of the closed curves R_1 and R_2. The unit normals $\mathbf{n}_1$ and $\mathbf{n}_2$ of the boundary curves are directed outwards. The space $\mathbb{R}^2$ is divided into strips of width b with the axis $\mathbf{e}_1$. The intersection of one of these strips with the control domain consists of the three substrips AB, CD and EF. An integral over K is replaced by the sum of the integrals over such substrips.

Integral of a partial derivative : Let a continuous scalar field p($\mathbf{x}$) be given in a control volume V. Assume that the partial derivatives $p_{,m}$ of the scalar field exist and are continuous. The integral I of a certain partial derivative $p_{,i}$ over the control volume is to be determined :

$$I = \int_V \frac{\partial p}{\partial x^i} \, dv \qquad\qquad i = 1, 2, 3$$

The space $\mathbb{R}^3$ is partitioned into infinitesimal prisms whose axes are parallel to the basis vector $\mathbf{e}_i$. The intersection of one of these prisms with the control volume generally consists of several subprisms. First the contribution I_e of a subprism V_e to the desired integral I is determined.

$$I_e = \int_{V_e} \frac{\partial p}{\partial x^{(i)}} \, ds \, dx^{(i)} = p \, ds \, \Big|_{A_1}^{A_2}$$

ds area of the cross section of V_e with normal $\mathbf{e}_i$

A_1, A_2 end surfaces of the subprism V_e

The end surfaces A_1 and A_2 of the subprism V_e are parts of the surface F of the control volume. Let $d\mathbf{a}_1$ and $d\mathbf{a}_2$ be the outer area vectors of the end surfaces. The component of $d\mathbf{a}_2$ in the direction of the prism axis $\mathbf{e}_i$ is the area vector $\mathbf{e}_i$ ds of the cross section ; the corresponding component of $d\mathbf{a}_1$ is the inverse vector $-\mathbf{e}_i$ ds.

$$ds = -\mathbf{e}_i \cdot d\mathbf{a}_1 = \mathbf{e}_i \cdot d\mathbf{a}_2$$

$$I_e = p_2\, \mathbf{e}_i \cdot d\mathbf{a}_2 + p_1 \mathbf{e}_i \cdot d\mathbf{a}_1$$

p_1, p_2 values of p on the infinitesimal end surfaces A_1, A_2

The contributions I_e of the subprisms which belong to the control volume are added. The union of the end surfaces A_i of these subprisms is the surface F of the control volume. In the limit da → 0 the volume integral I becomes an integral over the closed surface F.

$$I = \sum_e (p_1\, \mathbf{e}_i \cdot d\mathbf{a}_1 + p_2\, \mathbf{e}_i \cdot d\mathbf{a}_2)$$

$$\int_V p_{,i}\, dv = \int_F p\, \mathbf{e}_i \cdot d\mathbf{a}$$

The area vector d**a** is expressed as a product **n** da of the unit normal $\mathbf{n} = n_k\, \mathbf{e}^k$ and the surface area da. Substitution yields the coordinate form of the transformation rule :

$$\int_V p_{,i}\, dv = \int_F p\, n_i\, da$$

n_i coordinate of the outer unit normal **n** of the surface F

Gauss' theorem : Let a vector field **u**(**x**) be given in a control volume V. Assume that its coordinates u^i are continuous and that the partial derivatives $u^i_{,m}$ of the coordinates exist and are continuous. Then the integral of the divergence of **u** over the volume V is equal to the scalar surface integral of **u** over the surface F of the control volume.

$$\int_V \operatorname{div} \mathbf{u}\, dv = \int_F \mathbf{u} \cdot d\mathbf{a}$$

The area vector d**a** may be replaced by the product of the unit normal **n** and the surface area da. This yields the coordinate form of Gauss' theorem in local coordinates :

$$\int_V u^i_{;i}\, dv = \int_F u^i\, n_i\, da$$

$u^i_{;i}$ covariant derivative of the coordinate u^i with respect to y^i

n_i coordinates of the outer unit normal **n** of the surface F

Proof : Gauss' theorem

The proof is carried out in global cartesian coordinates using partial derivatives. The tensors div **u** and **u** • **da** may then be expressed in arbitrary local coordinate systems, so that the theorem also holds for local coordinates.

$$\begin{aligned}
\int\limits_V \left(\frac{\partial u^1}{\partial x^1} + \frac{\partial u^2}{\partial x^2} + \frac{\partial u^3}{\partial x^3}\right) dv &= \int\limits_V \frac{\partial u^1}{\partial x^1} dv + \int\limits_V \frac{\partial u^2}{\partial x^2} dv + \int\limits_V \frac{\partial u^3}{\partial x^3} dv \\
&= \int\limits_F u^1 n_1 \, da + \int\limits_F u^2 n_2 \, da + \int\limits_F u^3 n_3 \, da \\
\int\limits_V \text{div}\, \mathbf{u}\, dv &= \int\limits_F \mathbf{u} \cdot \mathbf{n}\, da
\end{aligned}$$

Corollaries to Gauss' theorem : The following formulas can be derived from Gauss' theorem for a scalar field p, a vector field **u** and a dyadic field **T**. The control volume is designated by V, its surface by F and the outer normal of the surface by **n**. Let the magnitude of the volume element be dv, and let the surface area of the surface element be da.

$$\begin{aligned}
&(1) & \int\limits_V \mathbf{grad}\, p \, dv &= \int\limits_F p \, \mathbf{da} \\
&(2) & \int\limits_V \mathbf{rot}\, \mathbf{u} \, dv &= \int\limits_F \mathbf{da} \times \mathbf{u} \\
&(3) & \int\limits_V p_1 \, \mathbf{grad}\, p_2 \, dv &= -\int\limits_V p_2 \, \mathbf{grad}\, p_1 \, dv + \int\limits_F p_1 \, p_2 \, \mathbf{da} \\
&(4) & \int\limits_V p \, \text{div}\, \mathbf{u} \, dv &= -\int\limits_V \mathbf{grad}\, p \cdot \mathbf{u} \, dv + \int\limits_F p \, \mathbf{u} \cdot \mathbf{da} \\
&(5) & \int\limits_V u_m \, t^{im}{}_{;\,i} \, dv &= -\int\limits_V u_{m;\,i} \, t^{im} \, dv + \int\limits_F u_m \, t^{im} \, n_i \, da
\end{aligned}$$

The theorems (3) to (5) are often used to integrate by parts in the formulation of physical problems.

Proof : Corollaries to Gauss' theorem

(1) For the scalar field p and a constant vector **c** one obtains :

$$\int\limits_V \text{div}\,(p\,\mathbf{c})\, dv = \int\limits_F p \, \mathbf{c} \cdot \mathbf{da}$$

$$\text{div}\,(p\,\mathbf{c}) = \mathbf{grad}\, p \cdot \mathbf{c} + p \, \text{div}\, \mathbf{c} = \mathbf{grad}\, p \cdot \mathbf{c}$$

$$\int\limits_V \mathbf{grad}\, p \cdot \mathbf{c} \, dv = \int\limits_F p \, \mathbf{c} \cdot \mathbf{da}$$

This relationship must hold for arbitrary constant vectors **c**. This implies that equation (1) holds.

(2) For the vector field **u** and a constant vector **c** one obtains :

$$\int_V \operatorname{div}(\mathbf{u}\times\mathbf{c})\,dv = \int_F (\mathbf{u}\times\mathbf{c})\cdot\mathbf{da}$$

$$\operatorname{div}(\mathbf{u}\times\mathbf{c}) = (\varepsilon^{ikm}\,u_k\,c_m)_{;\,i} = \varepsilon^{ikm}\,u_{k\,;\,i}\,c_m = \mathbf{rot\;u\cdot c}$$

$$\int_V \mathbf{rot\;u\cdot c}\,dv = \int_F \mathbf{da}\times\mathbf{u\cdot c}$$

This relationship must hold for arbitrary constant vectors **c**. This implies that equation (2) holds.

(3) For the scalar fields p_1 and p_2 one obtains using (1) :

$$\int_V \mathbf{grad}\,(p_1 p_2)\,dv = \int_F p_1 p_2\,\mathbf{da}$$

$$\mathbf{grad}\,(p_1 p_2) = p_1\,\mathbf{grad}\,p_2 + p_2\,\mathbf{grad}\,p_1$$

$$\int_V p_1\,\mathbf{grad}\,p_2\,dv = -\int_V p_2\,\mathbf{grad}\,p_1\;dv + \int_F p_1 p_2\,\mathbf{da}$$

(4) For the scalar field p and the vector field **u** one obtains :

$$\int_V \operatorname{div}(p\mathbf{u})\,dv = \int_F p\mathbf{u}\cdot\mathbf{da}$$

$$\operatorname{div}(p\mathbf{u}) = \mathbf{grad}\,p\cdot\mathbf{u} + p\operatorname{div}\mathbf{u}$$

$$\int_V p\operatorname{div}\mathbf{u}\,dv = -\int_V \mathbf{grad}\,p\cdot\mathbf{u}\,dv + \int_F p\,\mathbf{u}\cdot\mathbf{da}$$

(5) For the vector field **u** and the dyadic field **T** one obtains :

$$\int_V \operatorname{div}(\mathbf{T}\cdot\mathbf{u})\,dv = \int_F \mathbf{da}\cdot\mathbf{T}\cdot\mathbf{u}$$

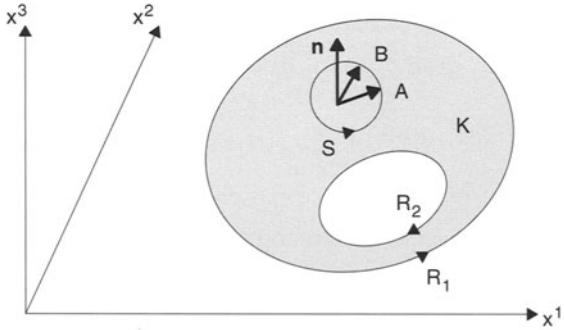
$$(t^{im}\,u_m)_{;\,i} = t^{im}\,u_{m\,;\,i} + t^{im}{}_{;\,i}\,u_m$$

$$\int_V t^{im}{}_{;\,i}\,u_m\,dv = -\int_V t^{im}\,u_{m;\,i}\,dv + \int_F t^{im}\,u_m\,n_i\,da$$

Control surface : Let a surface K in the euclidean space $\mathbb{R}^3$ be connected. Let the boundary R of the surface consist of one or more lines which are piecewise continuous. Let the global coordinates x^1, x^2, x^3 of the points of the surface K be bounded. Then the surface K is called a control surface in the space $\mathbb{R}^3$.

Orientation : The direction of the unit normal **n**(**x**) of a surface K in the euclidean space $\mathbb{R}^3$ is not uniquely determined. If **n** is a normal of K at the point **x**, then –**n** is also a normal of K at the point **x**. The formulation of Stokes' theorem requires a definite relationship between the normal **n** and the direction of the boundary R of the surface K to be established.

Consider a surface element S in the neighborhood of a point **x** of the surface K which contains **x** as an inner point. Let **a** and **b** be the vectors from **x** to two neighboring points A and B on the boundary of S. Assume that the vectors **a** and **b** are not parallel. Then the cross product $\mathbf{a} \times \mathbf{b}$ defines a direction of the surface normal **n** which is uniquely associated with the direction $A \rightarrow B$ of the boundary of S. The direction $A \rightarrow B$ is called the orientation of the boundary associated with **n**. If the orientation is changed from $A \rightarrow B$ to $B \rightarrow A$, the new cross product $\mathbf{b} \times \mathbf{a} = - \mathbf{a} \times \mathbf{b}$ leads to the complementary surface normal – **n**.



Stokes' theorem : Let a vector field **u**(**x**) be given on a control surface K. Assume that its coordinates u^i are continuous, and that the partial derivatives $u^i_{,m}$ of the coordinates exist and are continuous. Then the scalar integral of the curl of **u** over the control surface K is equal to the scalar integral of **u** over the boundary R of the control surface.

$$\int_K \textbf{rot u} \cdot d\mathbf{a} \qquad = \qquad \int_R \mathbf{u} \cdot d\mathbf{x}$$

The curl **rot u** is expressed using the permutation tensor ε^{ikm}, and the area vector d**a** is expressed as the product of the surface normal **n** and the surface area da. This yields the coordinate form of Stokes' theorem in local coordinates :

$$\int_K \varepsilon^{ikm} \, u_{m\,,\,k} \, n_i \, da \quad = \quad \int_R u_i \, dx^i$$

On each piece of the boundary R, the integration is carried out in the direction which corresponds to the relationship between the surface normal **n** and the orientation.

Proof : Stokes' theorem

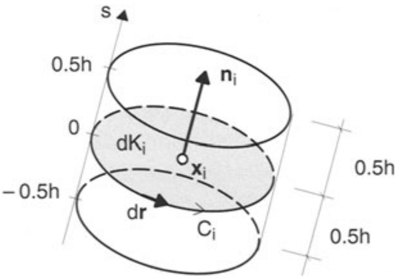
The following shell volume V is associated with the given control surface K :

$$V \;:=\; \{\mathbf{w} \mid \mathbf{w} = \mathbf{x} + s\,\mathbf{n} \quad \wedge \quad \mathbf{x} \in K \quad \wedge \quad -0.5h \le s \le 0.5h\}$$

$\mathbf{n}(\mathbf{x})$ unit normal of the control surface K at the point $\mathbf{x}$

h shell height

The control surface K is divided into infinitesimal surface elements dK_i with surface area da_i, reference point $\mathbf{x}_i \in dK_i$, boundary curve C_i and surface normal $\mathbf{n}_i$ at the point $\mathbf{x}_i$. A volume element dV_i is associated with the surface element dK_i :



$$dV_i := \{\mathbf{w} \mid \mathbf{w} = \mathbf{x} + s\,\mathbf{n}_i \quad \wedge \quad \mathbf{x} \in K_i \quad \wedge \quad -0.5h \le s \le 0.5h\}$$

The curl **rot u** of a tensor field **u** at the point $\mathbf{x}_i$ is defined as a limit for the infinitesimal volume dV_i with the magnitude dv_i :

$$\mathbf{rot\;u}(\mathbf{x}_i) \;=\; \lim_{dv_i \to 0} \frac{1}{dv_i} \int_{dS_i} d\mathbf{s} \times \mathbf{u}$$

$d\mathbf{s}$ area vector for an element of the surface dS_i of dV_i

The surface normal $\mathbf{n}_i$ is considered to be constant in the surface element dK_i. It is independent of s. In the limit h → 0 the magnitude of the infinitesimal volume is given by $dv_i = h\,da_i$. With an approximation error ε_i, the preceding equation implies :

$$\mathbf{rot\;u}(\mathbf{x}_i) \cdot \mathbf{n}_i \;=\; \lim_{h \to 0} \frac{1}{h\,da_i} \int_{dS_i} \mathbf{n}_i \cdot d\mathbf{s} \times \mathbf{u} \;+\; \varepsilon_i$$

The area vector of the surface fragment $s = 0.5\,h$ of dS_i is $d\mathbf{s} = \mathbf{n}_i\,da_i$. The area vector of the surface fragment $s = -0.5h$ has the opposite sign and the same magnitude. If **u** is considered to be approximately constant in dV_i, the contributions of the two surface fragments to the integral on the right-hand side of the equation cancel. On the remaining surface of dV_i one obtains :

$$\mathbf{n}_i \cdot \mathbf{ds} \times \mathbf{u} \;=\; \mathbf{u} \cdot \mathbf{n}_i \times \mathbf{ds} \;=\; h\,\mathbf{u} \cdot \mathbf{dr}$$

$\mathbf{dr}$ incremental tangent vector of the boundary curve C_i

The area vector of dK_i is designated by $d\mathbf{a}_i := \mathbf{n}_i da_i$. The expression $h\,\mathbf{u} \cdot \mathbf{dr}$ is substituted into the preceding equation :

$$\textbf{rot } \mathbf{u}(\mathbf{x}_i) \cdot d\mathbf{a}_i \;=\; \int_{C_i} \mathbf{u} \cdot \mathbf{dr} \;+\; \varepsilon_i \, da_i$$

The control surface K is divided into elements dK_i. The scalars $\textbf{rot } \mathbf{u}(\mathbf{x}_i) \cdot d\mathbf{a}_i$ are summed over the elements with $i = 1,...,k$:

$$\sum_i \textbf{rot } \mathbf{u}(\mathbf{x}_i) \cdot d\mathbf{a}_i \;=\; \sum_i \int_{C_i} \mathbf{u}(\mathbf{x}_i) \cdot \mathbf{dr} \;+\; \sum_i \varepsilon_i \, d\mathbf{a}_i$$

If the greatest distance between two points in dK_i tends to zero, the number k of elements increases without bound. The sum on the left-hand side becomes the integral $\int_K \textbf{rot } \mathbf{u} \cdot d\mathbf{a}$. In the summation over the boundary curves C_i on the right-hand side, the contributions of neighboring elements on the common boundary cancel since they have opposite orientations. The contributions on the boundary curves R of K are not cancelled, since every point of these boundary curves belongs to only one element dK_i.

For a certain subdivision of K, let $|\varepsilon_i| < \varepsilon$ for $i = 1,...,k$ with $\varepsilon > 0$. Then $\sum_i \varepsilon_i \, da_i \leq \varepsilon \sum_i da_i$. With decreasing size of the elements dK_i (see above), ε tends to zero, and thus $\varepsilon \sum_i da_i$ also tends to zero, since $\sum_i da_i$ is the surface area of the surface K. This yields Stokes' theorem :

$$\int_K \textbf{rot } \mathbf{u} \cdot d\mathbf{a} \;=\; \int_R \mathbf{u} \cdot \mathbf{dr}$$

Integral formula for $\mathbb{R}^2$: In the plane $\mathbb{R}^2$, the theorems of Gauss and Stokes take the following forms :

$$\int_K \left(\frac{\partial u_1}{\partial x^1} + \frac{\partial u_2}{\partial x^2} \right) dx^1 \, dx^2 \;=\; \int_R (u_1 n^1 + u_2 n^2) \, ds$$

$$\int_K \left(\frac{\partial u_2}{\partial x^1} - \frac{\partial u_1}{\partial x^2} \right) dx^1 \, dx^2 \;=\; \int_R (u_1 dx^1 + u_2 dx^2)$$

These formulas differ only in notation. To prove this, replace u_1 by $-u_2$ and u_2 by u_1 in the lower equation. Then substituting the relationships $dx^1 = -n^2 ds$ and $dx^2 = n^1 ds$ on the right-hand side transforms the lower formula into the upper one.

Green's theorems : Let continuous scalar fields p_1 and p_2 be given in a control volume V. Then substituting p_1 **grad** p_2 for the vector field **u** in Gauss' theorem yields :

$$\int_V \operatorname{div}(p_1\,\mathbf{grad}\,p_2)\,dv = \int_F p_1\,\mathbf{grad}\,p_2 \cdot \mathbf{da}$$

Green's first theorem is obtained from this formula by transforming the field operations and introducing the Laplace operator Δ from Section 9.4.10 :

$$\operatorname{div}(p_1\,\mathbf{grad}\,p_2) = \mathbf{grad}\,p_1 \cdot \mathbf{grad}\,p_2 + p_1 \operatorname{div}\mathbf{grad}\,p_2$$

$$\int_V (\mathbf{grad}\,p_1 \cdot \mathbf{grad}\,p_2 + p_1\,\Delta p_2)\,dv = \int_F p_1\,\mathbf{grad}\,p_2 \cdot \mathbf{da}$$

Green's second theorem is obtained from the first theorem by writing the theorem once for p_1 **grad** p_2 and once for p_2 **grad** p_1 :

$$\int_V (p_1\,\Delta p_2 - p_2 \Delta\,p_1)\,dv = \int_F (p_1\,\mathbf{grad}\,p_2 - p_2\mathbf{grad}\,p_1) \cdot \mathbf{da}$$

Green's third theorem is obtained from the first theorem by choosing $p_1 = 1$ and $p_2 = p$:

$$\int_V \Delta p\,dv = \int_F \mathbf{grad}\,p \cdot \mathbf{da}$$

Example 1 : Gauss' theorem

The integral of the divergence of a given vector field **u** over the volume of a cuboid $0 \le x \le a,\ 0 \le y \le b,\ 0 \le z \le c$ in the cartesian coordinate system (x, y, z) with the global basis $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ is determined.

$$\mathbf{u} = xyz\,\mathbf{e}_1 + (x^2+y^2)\,\mathbf{e}_2 + z^3\,\mathbf{e}_3$$

$$\begin{aligned}
\int_V \operatorname{div}\mathbf{u}\,dv &= \int_0^c\int_0^b\int_0^a (yz+2y+3z^2)\,dx\,dy\,dz \\
&= \int_0^c\int_0^b a\,(yz+2y+3z^2)\,dy\,dz \\
&= \int_0^c a\,(\tfrac{1}{2}\,b^2z+b^2+3z^2 b)\,dz \\
&= a\,(\tfrac{1}{4}\,b^2c^2+b^2c+bc^3)
\end{aligned}$$

By Gauss' theorem this integral is equal to the scalar integral of **u** over the surface of the cuboid :

$$\begin{array}{llllll}
x=0: & \mathbf{n}=-\mathbf{e}_1 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^b\int_0^c -xyz\,dz\,dy & = & 0 \\
x=a: & \mathbf{n}=\mathbf{e}_1 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^b\int_0^c ayz\,dz\,dy & = & \tfrac{1}{4}\,a\,b^2c^2 \\
y=0: & \mathbf{n}=-\mathbf{e}_2 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^a\int_0^c -x^2\,dz\,dx & = & -\tfrac{1}{3}\,a^3c \\
y=b: & \mathbf{n}=\mathbf{e}_2 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^a\int_0^c (x^2+b^2)dz\,dx & = & \tfrac{1}{3}\,a^3c+b^2ac \\
z=0: & \mathbf{n}=-\mathbf{e}_3 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^a\int_0^b -z^3\,dy\,dx & = & 0 \\
z=c: & \mathbf{n}=\mathbf{e}_3 & : & \int \mathbf{u}\cdot d\mathbf{a} = \int_0^a\int_0^b c^3\,dy\,dx & = & c^3ab
\end{array}$$

$$\int_F \mathbf{u}\cdot d\mathbf{a} = \tfrac{1}{4}\,a\,b^2c^2 + b^2ac + c^3ab = a\,(\tfrac{1}{4}\,b^2c^2+b^2c+bc^3)$$

Example 2 : Properties of an ellipse

Let the lengths of the axes of an ellipse in the space $\mathbb{R}^2$ be 2a and 2b. The surface area F, the static moment S and the moment of inertia I with respect to axis 1 of the ellipse are determined using the integral formula for $\mathbb{R}^2$. The parametric form of the boundary of the ellipse and the integral formula are :

$$x = a\cos\theta \qquad dx = -a\sin\theta\, d\theta \qquad 0 \le \theta \le 2\pi$$

$$y = b\sin\theta \qquad dy = b\cos\theta\, d\theta$$

$$\int_K \left(\frac{\partial u}{\partial x} + \frac{\partial w}{\partial y}\right) dx\, dy = \int_R (u\, dy - w\, dx)$$

The choice u = x and w = y yields the surface area F :

$$\int_K (1+1)\, da = \int_R (ab\cos^2\theta + ab\sin^2\theta)\, d\theta = 2F$$

$$F = \frac{1}{2} ab \int_0^{2\pi} d\theta = \pi ab$$

The choice u = xy and w = y^2 yields the static moment S :

$$\int_K (y+2y)\, da = \int_R (ab^2\sin\theta\cos^2\theta + ab^2\sin^3\theta)\, d\theta = 3S$$

$$S = \frac{1}{3} ab^2 \int_0^{2\pi} \sin\theta\, d\theta = 0$$

The choice u = xy^2 and w = y^3 yields the moment of inertia I :

$$\int_K (y^2 + 3y^2)\, da = \int_R (ab^3\sin^2\theta\cos^2\theta + ab^3\sin^4\theta)\, d\theta = 4I$$

$$I = \frac{1}{4} ab^3 \int_0^{2\pi} \sin^2\theta\, d\theta = \frac{\pi}{4} ab^3$$

10 STOCHASTICS

10.1 INTRODUCTION

An experiment is a process which can be repeated an arbitrary number of times under identical conditions. Every experiment yields a result. If an experiment always yields the same result, the experiment is deterministic. If there are different possible results due to random influences, the experiment is stochastic. Stochastics is a field of mathematics which deals with the results of stochastic experiments.

The simplest stochastic experiments are found in games of chance, whose results are called random events. Typical examples are drawing a lot with the possible events ”blank”, ”prize”, ”first prize” or throwing two dice with the possible throws 2 to 12. The set of all possible events is the event space of the experiment. Every possible event is assigned a probability with which it occurs when the experiment is performed. The algebra of sets for the events and the axioms for probabilities lead to the rules of the calculus of probabilities. The fundamentals of the calculus of probabilities for random events and its applications are treated in Section 10.2.

If a quantity in a stochastic experiment takes real values, then this quantity is mathematically treated as a random variable. Typical examples are the daily number of vehicles over a bridge or the daily precipitation at a location. The axioms and rules of probabilities for random events are transferred to random variables and lead to the definition of the probability distribution for a random variable. A random variable is characterized by values determined from its probability distribution. Typical characteristic values are the mean and the variance, which measures the quadratic deviation from the mean. Discrete and continuous random variables are distinguished. The theoretical foundations for random variables and the most important distributions for discrete and continuous random variables as well as their applications are treated in Section 10.3.

If several quantities which take real values are considered in a stochastic experiment, these quantities are mathematically treated as a random vector with several random variables. Typical examples are the daily numbers of vehicles which turn left, drive straight on or turn right at a junction, or the daily high water marks at different locations along a coast. The theoretical foundations for random variables are extended to random vectors. The random variables of a random vector may be dependent. The degree of linear dependence is described by correlation coefficients. The correlation of the random variables is an essential property of random vectors. The theoretical foundations for random vectors and examples of their application are treated in Section 10.4.

P.J Pahl et al., *Mathematical Foundations of Computational Engineering*

If a time-dependent quantity is considered in a stochastic experiment, this quantity is mathematically treated as a random function of time. Typical examples are the number of vehicles in a car park or the temperature at a location over the course of a day. Time-dependent stochastic processes are also called random processes. The theory of random processes is very extensive and is therefore treated for selected classes of random processes only. Markov processes in discrete and continuous time form an important class of random processes. They are based on a mathematical formulation of time-dependent stochastic processes as initial value problems. The long-term behavior of these processes and the resulting probability distributions are particularly important. Stationary processes in discrete and continuous time form another important class. They possess the same probability distribution at all points in time. The random quantities at different points in time are dependent to a certain degree. This dependence is described by the correlation function. The fundamentals of Markov processes together with their applications in the theory of queues and the fundamentals of stationary processes are treated in Section 10.5.

10.2 RANDOM EVENTS

10.2.1 INTRODUCTION

The calculus of probabilities deals with the rules for calculating probabilities for random events and is based on set theory. It includes elementary combinatorics, the algebra of events and the axioms and calculational rules for probabilities. The fundamentals of the calculus of probabilities for random events are treated in the following sections. Their application is illustrated using simple examples.

10.2.2 ELEMENTARY COMBINATORICS

Introduction : Combinatorics is a part of set theory. It is restricted to finite sets and deals with the different possibilities of choosing elements from a given set and arranging them in a tuple. The different procedures used for choice and arrangement lead to the definition of permutations, variations and combinations.

Set and tuple : Let a finite set of elements be given. Elements are selected from the set and arranged in a tuple in the order in which they are chosen. An element may be selected several times, so that a tuple may contain several identical elements. If the arrangement of the elements in the tuple is relevant, the tuple is said to be ordered. If the arrangement is irrelevant, the tuple is said to be unordered. To distinguish ordered and unordered tuples, the elements of the tuples are enclosed by round brackets and square brackets, respectively.

set	:	{ a, b, c, d, ... }
ordered tuple	:	(a, c, d, a, ...)
unordered tuple	:	[a, c, d, a, ...]

A tuple with k elements is called a k-tuple. Two ordered k-tuples are equal if they contain the same elements in the same arrangement. Two unordered k-tuples are equal if they contain the same elements in an arbitrary arrangement.

Permutation : Let a set with n elements be given. An ordered n-tuple which contains each element of the set exactly once is called a permutation without repetition. The number of different permutations without repetition is :

$$p(n) = 1 \cdot 2 \cdot \ldots \cdot (n-1) \cdot n = n\,!$$

An ordered tuple which contains each element e_j of the set exactly m_j times is called a permutation with repetition. The number of different permutations with repetition is :

$$\overset{*}{p}(n \mid m_1, m_2, \ldots, m_n) = \frac{(m_1 + m_2 + \ldots + m_n)!}{m_1! \cdot m_2! \cdot \ldots \cdot m_n!}$$

Variation : Let a set with n elements be given. An ordered k-tuple which does not contain any element of the set more than once is called a variation without repetition of class k. The number of different variations without repetition is :

$$v(n,k) = \frac{n!}{(n-k)!}$$

An ordered k-tuple which may contain elements of the set more than once is called a variation with repetition of class k. The number of different variations with repetition is :

$$\overset{*}{v}(n,k) = n^k$$

Combination : Let a set with n elements be given. An unordered k-tuple which does not contain any element of the set more than once is called a combination without repetition of class k. The number of different combinations without repetition is :

$$c(n,k) = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

An unordered k-tuple which may contain elements of the set more than once is called a combination with repetition of class k. The number of different combinations with repetition is :

$$\overset{*}{c}(n,k) = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!}$$

Example 1 : Combinatorics for a character set

Let a set { a, b, c } with three characters of the alphabet be given. The permutations, variations and combinations with and without repetition for this character set are formed.

Permutations without repetition

tuples : (a, b, c) (a, c, b) (b, a, c) (b, c, a) (c, a, b) (c, b, a)

number : $p(3) = 3! = 1 \cdot 2 \cdot 3 = 6$

Permutations with 2-fold occurrence of a and 1-fold occurrence of b, c

tuples : (a, a, b, c) (a, a, c, b) (a, b, a, c) (a, c, a, b) (a, b, c, a) (a, c, b, a)
(b, a, a, c) (b, a, c, a) (b, c, a, a) (c, a, a, b) (c, a, b, a) (c, b, a, a)

number : $\overset{*}{p}(3|2,1,1) = (2+1+1)! \;/\; (2! \cdot 1! \cdot 1!) = 24/2 = 12$

Variations without repetition of class $k=2$

tuples : (a, b) (a, c) (b, a) (b, c) (c, a) (c, b)

number : $v(3,2) = 3!/(3-2)! = 3!/1! = 6$

Variations with repetition of class $k=2$

tuples : (a, a) (a, b) (a, c) (b, a) (b, b) (b, c) (c, a) (c, b) (c, c)

number : $\overset{*}{v}(3,2) = 3^2 = 9$

Combinations without repetition of class $k = 2$

tuples : [a, b] [a, c] [b, c]

number : $c(3,2) = \binom{3}{2} = \dfrac{3 \cdot 2}{1 \cdot 2} = 3$

Combinations with repetition of class $k=2$

tuples : [a, a] [a, b] [a, c] [b, b] [b, c] [c, c]

number : $\overset{*}{c}(3,2) = \binom{3+2-1}{2} = \binom{4}{2} = \dfrac{4 \cdot 3}{1 \cdot 2} = 6$

Example 2 : Game of dice

Two dice are thrown in a game of chance. Each die yields a number from 1 to 6. Let the two dice be distinguished by different colors. The possible throws are variations with repetition. Their number is :

$$n = 6 \qquad k = 2 \qquad \overset{*}{v}(6,2) = 6^2 = 36$$

Let the two dice be indistinguishable. The possible throws are combinations with repetition. Their number is :

$$n = 6 \qquad k = 2 \qquad \overset{*}{c}(6,2) = \binom{6+2-1}{2} = \binom{7}{2} = \frac{7 \cdot 6}{1 \cdot 2} = 21$$

10.2.3 ALGEBRA OF EVENTS

Introduction : A stochastic experiment is a process in which several results are possible and the result which occurs is not predictable. The possible results are called elementary events and form the event space of the experiment. An event is a set of elementary events. Thus the treatment of events is reduced to the treatment of sets. In analogy with the algebra of sets, the resulting algebra is called the algebra of events.

Event space : Every possible result of an experiment is called an elementary event. Elementary events are mutually exclusive. The set of all elementary events is called the event space of the experiment.

$S := \{e_1, e_2, \ldots, e_n\}$

e_j elementary event

Event : A subset $A \subseteq S$ of the event space is called an event and is designated by A. The event occurs if and only if an elementary event contained in A occurs. The impossible event $\emptyset$ is an empty set of elementary events and therefore never occurs. The certain event S is identical with the event space and therefore always occurs.

event	:	$\emptyset \subseteq A \subseteq S$
impossible event	:	$\emptyset$
certain event	:	S

This definition reduces events to sets. The algebra of sets for events is called the algebra of events. The essential definitions of the algebra of events are compiled in the following.

Operations : The event "A or B" occurs if and only if A occurs or B occurs. It contains all elementary events which are contained in A or B. The event "A and B" occurs if and only if A and B both occur. It contains all elementary events which are contained in both A and B. These operations on events correspond to the union $\cup$ and the intersection $\cap$ in set theory.

A or B	:	$A \cup B$
A and B	:	$A \cap B$

Complementary event : The event $\bar{A}$ complementary to A occurs if and only if the event A does not occur. It contains all elementary events of the event space S which are not contained in A. The complement $\bar{A}$ corresponds to the difference of S and A in set theory.

complementary event	:	$\bar{A} = S - A$
rules	:	$A \cup \bar{A} = S \qquad A \cap \bar{A} = \emptyset$

Subevent : The event A is called a subevent of B if the event B occurs whenever the event A occurs. All elementary events of A are contained in B. This definition corresponds to the definition of a subset in set theory.

subevent : $A \subseteq B$

rules : $A \cap B = A$ $\qquad$ $A \cup B = B$

Incompatible events : Two events A and B are said to be incompatible if they cannot both occur. The intersection $A \cap B$ is the impossible event $\emptyset$.

incompatibility: $A \cap B = \emptyset$

Partition : A set of events A_j is called a partition of the event space S if the events are pairwise incompatible and their union is the event space S.

$$S = A_1 \cup A_2 \cup \ldots \cup A_n = \bigcup_{j=1}^{n} A_j \qquad A_j \cap A_k = \emptyset \quad \text{for} \quad j \neq k$$

An event B is partitioned into incompatible subevents $B_j \subseteq B$ by forming the intersections $(B \cap A_j)$.

$$B = B_1 \cup B_2 \cup \ldots \cup B_n = \bigcup_{j=1}^{n} B_j \qquad B_j = B \cap A_j$$

$$B_j \cap B_k = \emptyset \quad \text{for} \quad j \neq k$$

Example : Die events

Let a die with the numbers 1 to 6 be given. The result of a throw of the die is a number. Every possible number is an elementary event. The event space S contains the numbers 1 to 6. Let an event A be defined as the throw of an even number. The complementary event $\overline{A}$ is the throw of an odd number. The throw of an even or odd number is the certain event, which occurs in every throw. The throw of an even and odd number is the impossible event, which cannot occur. The throw of an even number and the throw of an odd number are incompatible events.

event space for the throw	$S = \{1, 2, 3, 4, 5, 6\}$
throw of an even number	$A = \{2, 4, 6\}$
throw of an odd number	$\overline{A} = \{1, 3, 5\}$
certain event	$A \cup \overline{A} = S$
impossible event	$A \cap \overline{A} = \emptyset$

10.2.4 PROBABILITY

Introduction : A stochastic experiment is performed several times. The number of cases in which a certain event occurs is determined and divided by the number of experiments performed. This leads to the relative frequency for the occurrence of the event. With increasing numbers, the relative frequency for an event becomes increasingly stable. This property suggests that every event may be assigned a probability as the ”limit” of the relative frequency. Since the existence of such a limit cannot be proved, axioms of probability theory are constructed in analogy with the definition of and the rules for relative frequencies. The fundamentals of the calculus of probabilities and their application are treated in the following.

Relative frequency : An experiment is carried out n times. A certain event A occurs m times as a result of the experiments. The quotient m / n is called the relative frequency for the occurrence of A and is designated by $H_n(A)$.

$$H_n(A) := \frac{m}{n} \qquad\qquad n > 0$$

Since $0 \leq m \leq n$ and $n > 0$, the relative frequency $H_n(A)$ is a rational number between 0 and 1.

$$0 \leq H_n(A) \leq 1$$

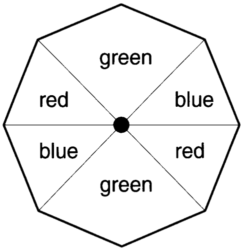
If the relative frequencies for two incompatible events A and B of an experiment are $H_n(A)$ and $H_n(B)$, the relative frequency $H_n(A \cup B)$ for the event $(A \cup B)$ is equal to the sum of the relative frequencies for A and B.

$$H_n(A \cup B) = H_n(A) + H_n(B) \quad \Leftarrow \quad A \cap B = \emptyset$$

The values of the relative frequency generally vary with the number n of performed experiments. It is observed that the fluctuations in the relative frequencies decrease with increasing n. This observation leads to the hypothesis that the relative frequencies converge to a limit with increasing n. However, this hypothesis cannot be proved mathematically.

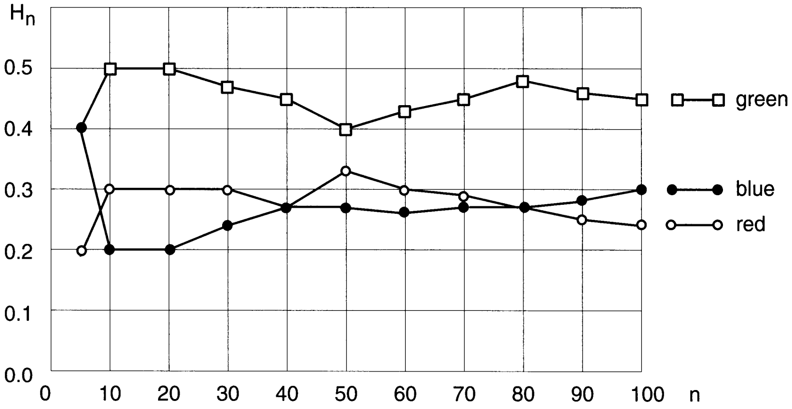
Example 1 : Top experiment

Let a regular octagonal top be divided into color sectors as illustrated. Each experiment consists in spinning the top and writing down the color sector on which it comes to rest. The experiment is repeated 100 times. The number m for the occurrence of the colors red, green and blue is shown in a table as a function of the number n of experiments. The corresponding relative frequency is shown diagrammatically.



top with color sectors

n	5	10	20	30	40	50	60	70	80	90	100
red m	1	3	6	9	11	16	18	20	21	22	24
green m	2	5	10	14	18	20	26	31	38	42	46
blue m	2	2	4	7	11	14	16	19	21	26	30



The relative frequencies for the occurrence of the colors red, green and blue after 100 experiments are 0.24, 0.46 and 0.30, respectively. The theoretically expected limits for the relative frequencies are 0.25, 0.50 and 0.25.

Axioms : Since it is not possible from a mathematical point of view to define probability as a limit of relative frequencies, probability is introduced as a quantity which satisfies certain axioms.

(A1) Every event A is assigned a non-negative real number P(A) as its probability.

$$P(A) \geq 0$$

(A2) The certain event S has the probability 1.

$$P(S) = 1$$

(A3) If two incompatible events A and B have the probabilities P(A) and P(B), the probability P(A ∪ B) for the event A ∪ B is equal to the sum of the probabilities of A and B.

$$P(A \cup B) \;=\; P(A) + P(B) \quad \Leftarrow \quad A \cap B = \emptyset$$

Rules : The following rules of calculation for probabilities follow from the axioms of probability theory :

(R1) The impossible event ∅ has the probability 0. The certain event S has the probability 1. An arbitrary event A has a probability between 0 and 1.

$$P(\emptyset) \le P(A) \le P(S) \qquad P(\emptyset) = 0 \qquad P(S) = 1$$

(R2) The event A and the complementary event $\bar{A}$ are incompatible. The event $A \cup \bar{A}$ is the certain event S. Accordingly, the sum of the probabilities of A and $\bar{A}$ is exactly 1.

$$P(A) + P(\bar{A}) \;=\; 1$$

(R3) The sum of the probabilities of A and B is equal to the sum of the probabilities of A ∪ B and A ∩ B. This rule is proved by considering the incompatible elementary events contained in A and B. If A and B are incompatible, the rule reduces to axiom (A3).

$$P(A) + P(B) \;=\; P(A \cup B) + P(A \cap B)$$

Conditional probability : The probability that an event A occurs given that the event B occurs is called a conditional probability; it is designated by P(A I B). The conditional probability is defined as follows :

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)} \qquad P(B) \ne 0$$

The definition of the conditional probability satisfies the first two axioms of probability theory. If the events A and B are incompatible, the conditional probability is 0. If A is a subevent of B, the conditional probability is P(A) / P(B) with P(A) ≤ P(B). If B is a subevent of A, the conditional probability is 1. The definition of the conditional probability implies the product rule of the calculus of probabilities, which also holds for P(B) = 0 :

(R4) $P(A \cap B) = P(A \mid B) \cdot P(B)$

Stochastic independence : Two events A and B are said to be stochastically independent if the probability for the occurrence of the one event is independent of the probability for the occurrence of the other event, that is :

$$\begin{aligned} P(A \mid B) &= P(A) \\ P(B \mid A) &= P(B) \\ P(A \cap B) &= P(A) \cdot P(B) \end{aligned}$$

Total probability theorem : Let the event space S be partitioned into the incompatible events A_j :

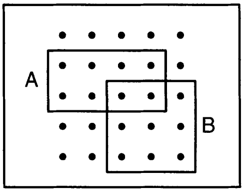
$$S = \bigcup_{j=1}^{n} A_j \qquad P(S) = \sum_{j=1}^{n} P(A_j) = 1$$

An event B is partitioned into the incompatible subevents $B_j = B \cap A_j$. The probability $P(B_j)$ is expressed in terms of the conditional probability $P(B \mid A_j)$. This yields the following formula for the total probability for B :

$$B = \bigcup_{j=1}^{n} (B \cap A_j) \qquad P(B) = \sum_{j=1}^{n} P(B \cap A_j) = \sum_{j=1}^{n} P(B \mid A_j) \cdot P(A_j)$$

Example 2 : Events with equal probabilities

In the illustrated event space with 25 elementary events, the events A and B are represented as point sets.



event space	:	25 elementary events
event A	:	8 elementary events
event B	:	9 elementary events
event $A \cap B$	:	2 elementary events

Let the probability of the elementary events be equal. Since the certain event contains all 25 elementary events and has the probability 1, the probability for each elementary event is 1/25. This leads to the following probabilities for the events A, B and $A \cap B$.

$$\begin{aligned} P(A) &= 8/25 = 0.32 \\ P(B) &= 9/25 = 0.36 \\ P(A \cap B) &= 2/25 = 0.08 \end{aligned}$$

The probabilities for the events $\bar{A}$ and $A \cup B$ are calculated according to rules (R2) and (R3). The results are readily verified using the illustrated event space.

$$\begin{aligned} P(\bar{A}) &= 1 - P(A) = 1 - 0.32 = 0.68 \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) = 0.32 + 0.36 - 0.08 = 0.60 \end{aligned}$$

The probability that event A occurs given that event B occurs is calculated as follows according to the definition of the conditional probability :

$$P(A \mid B) = P(A \cap B) / P(B) = 0.08 / 0.36 = 0.222$$

Example 3 : Throwing a die

Let a die with the numbers 1 to 6 be given. If the die is thrown, each number occurs with the same probability 1 / 6. The probability that at least one "6" is thrown in the course of three experiments is to be determined. To solve this problem formally, the following events are defined :

A_n	throw of "6" in experiment n	$P(A_n) = 1/6$
$\bar{A}_n$	throw other than "6" in experiment n	$P(\bar{A}_n) = 5/6$
A	throw of "6" in at most 3 experiments	$P(A)$
$\bar{A}$	throw other than "6" in 3 experiments	$P(\bar{A}) = 1 - P(A)$

The event A occurs if and only if either A_1 occurs or $\bar{A}_1$ and A_2 occur or $\bar{A}_1$ and $\bar{A}_2$ and A_3 occur. Since the events in the die experiment are incompatible and stochastically independent, the probability P(A) is calculated according to the sum and product rules :

$$A = A_1 \cup (\bar{A}_1 \cap A_2) \cup (\bar{A}_1 \cap \bar{A}_2 \cap A_3)$$

$$P(A) = P(A_1) + P(\bar{A}_1) \cdot P(A_2) + P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(A_3)$$

$$P(A) = \frac{1}{6} + \frac{5}{6} \cdot \frac{1}{6} + \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{1}{6} = \frac{91}{216}$$

The problem may be solved more easily by first determining the probability for the event $\bar{A}$ and then determining the probability for the event A. The event $\bar{A}$ occurs if and only if $\bar{A}_1$ and $\bar{A}_2$ and $\bar{A}_3$ occur. This yields :

$$\bar{A} = \bar{A}_1 \cap \bar{A}_2 \cap \bar{A}_3$$

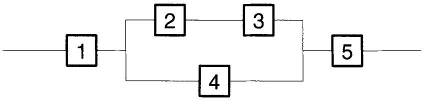
$$P(\bar{A}) = P(\bar{A}_1) \cdot P(\bar{A}_2) \cdot P(\bar{A}_3) = \frac{5}{6} \cdot \frac{5}{6} \cdot \frac{5}{6} = \frac{125}{216}$$

$$P(A) = 1 - P(\bar{A}) = 1 - \frac{125}{216} = \frac{91}{216}$$

10.2.5 RELIABILITY

Introduction : In technical applications, the reliability of systems is determined using the calculus of probabilities. Various systems are considered, such as traffic systems, supply systems, communication systems, computer systems or structural systems. Generally speaking, a system consists of elements which are arranged in a certain way. In analogy with electric circuits, serial, parallel and mixed arrangements of elements in a system are distinguished. The elements of a system fail with a certain probability. Depending on the arrangement of the elements and the failure probabilities of the elements, the system fails with a certain probability. The failure probability of the system is determined under simplifying assumptions according to the rules of the calculus of probabilities. The probability for non-failure is the reliability of the system.

Reliability of systems : In analogy with electric circuits, a system consists of elements arranged in series or in parallel. The system is represented in a block scheme. The following diagram shows a block scheme for a simple system :



block scheme of a system

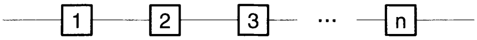
The failure of an element j is a random event; it is designated by A_j. The complementary event $\bar{A}_j$ is the non-failure of the element j. Accordingly, the random event A is introduced for the failure of the system, and the complementary event $\bar{A}$ is introduced for the non-failure of the system. The probability of failure is called the failure probability. The probability of non-failure is called reliability.

A_j	failure of the j-th element	$P(A_j) = p_j$
$\bar{A}_j$	non-failure of the j-th element	$P(\bar{A}_j) = q_j$
A	failure of the system	$P(A) = p$
$\bar{A}$	non-failure of the system	$P(\bar{A}) = q$

$$p + q = 1 \qquad p_j + q_j = 1$$

The reliability of the system critically depends on the arrangement of the elements in the system. Systems with serial, parallel and mixed arrangements of elements are treated in the following.

Serial system : The elements of a serial system are arranged in sequence. The system fails if at least one of the elements fails. If the elements are stochastically independent with respect to failure, then the probability for non-failure of the system is calculated from the probabilities of non-failure of the elements according to the product rule.



$$\overline{A} = \overline{A}_1 \cap \overline{A}_2 \cap \overline{A}_3 \cap \ldots \cap \overline{A}_n = \bigcap_{j=1}^{n} \overline{A}_j$$

$$q = q_1 * q_2 * q_3 * \ldots * q_n = \prod_{j=1}^{n} q_j$$

$$p = 1 - q$$

The reliability q of the serial system decreases with increasing number n of elements and tends to 0 for n → ∞. The failure probability p accordingly increases with increasing number n of elements and tends to 1 for n → ∞.

If the elements are stochastically dependent with respect to failure, then the probability for non-failure of the system is calculated from the conditional probabilities of non-failure of the elements according to the product rule :

$$P(\overline{A}) = P(\overline{A}_1) \cdot P(\overline{A}_2 \mid \overline{A}_1) \cdot P(\overline{A}_3 \mid \overline{A}_1 \cap \overline{A}_2) \cdot \ldots \cdot P(\overline{A}_n \mid \overline{A}_1 \cap \ldots \cap \overline{A}_{n-1})$$

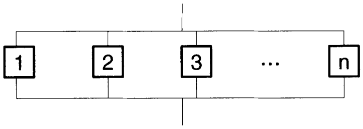
This formula is also valid if $\overline{A}_1$ is interchanged with an arbitrary $\overline{A}_j$. The conditional probabilities are less than or equal to 1. This implies the following bounds :

$$q \le q_j \qquad q \le \min_j \{q_j\}$$

$$p \ge p_j \qquad p \ge \max_j \{p_j\}$$

The reliability q of the serial system is less than or equal to the minimal reliability of an element. Accordingly, the failure probability p is greater or equal to the maximal failure probability of an element.

Parallel system : A parallel system exhibits a parallel arrangement of elements. It fails if all elements fail. If the elements are stochastically independent with respect to failure, then the probability for the failure of the system is calculated from the probabilities for the failure of the elements according to the product rule.



$$A = A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n = \bigcap_{j=1}^{n} A_j$$

$$p = p_1 * p_2 * p_3 * \ldots * p_n = \prod_{j=1}^{n} p_j$$

$$q = 1 - p$$

The failure probability p of the parallel system decreases with increasing number n of elements and tends to 0 for $n \rightarrow \infty$. The reliability q accordingly increases with increasing number n of elements and tends to 1 for $n \rightarrow \infty$.

If the elements are stochastically dependent with respect to failure, then the probability for the failure of the system is calculated from the conditional probabilities for the failure of the elements according to the product rule :

$$P(A) \;=\; P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot \ldots \cdot P(A_n \mid A_1 \cap \ldots \cap A_{n-1})$$

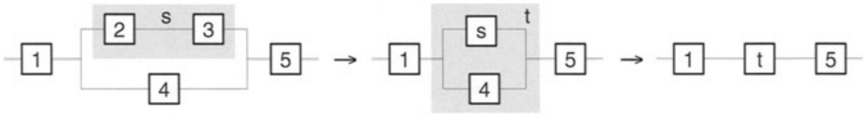
This formula is also valid if A_1 is interchanged with an arbitrary A_j. The conditional probabilities are less than or equal to 1. This implies the following bounds :

$$p \le p_j \qquad\qquad p \le \min_j \{p_j\}$$

$$q \ge q_j \qquad\qquad q \ge \max_j \{q_j\}$$

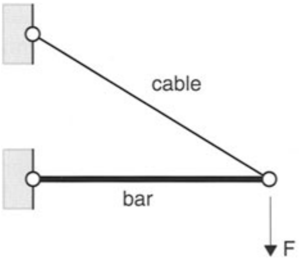
The failure probability p of the parallel system is less than or equal to the minimal failure probability of an element. Accordingly, the reliability q is greater or equal to the maximal reliability of an element. In contrast to a serial system, the stochastic independence of the elements with respect to failure is often not satisfied for a parallel system, since the failure of some elements places a higher load on the remaining elements, and thus to a higher failure probability.

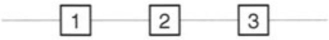
Mixed system : A mixed system exhibits an arrangement of elements which is partially serial and partially parallel. The system is recursively decomposed into components which possess either a serial or a parallel arrangement of elements or components. Thus the calculation of a mixed system is reduced to the calculation of its serial and parallel components. The decomposition of a simple mixed system into components is illustrated schematically.



Example 1 : Serial system

The following diagram shows a simple static system consisting of a bar and a cable to which a force F is applied. The static system does not fail if the bar and the cable do not fail due to the load. The bar does not fail if its strength is not exceeded and the bar does not buckle. The static system corresponds to a serial system.





A_1 : cable failure

A_2 : bar failure : strength exceeded

A_3 : bar failure : bar buckles

If the system is designed such that for a certain load F each elementary failure state A_j occurs with the same failure probability p_0, then the following reliability q is obtained for the static system :

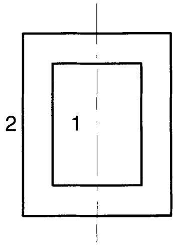
$$q = q_0^3 = (1 - p_0)^3$$

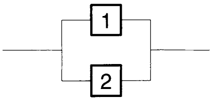
The failure probability p of the static system is the probability complementary to the reliability q. If the failure probability p_0 is very close to 0, terms of higher order may be neglected, and one obtains :

$$p = 1 - q = 1 - (1 - p_0)^3 = 3p_0 - 3p_0^2 + p_0^3 \approx 3p_0$$

Example 2 : Parallel system

The following diagram shows a rotationally symmetric container consisting of an inner container 1 with volume V_1 and an outer container 2 with volume V_2. The inner container is filled with gas and is subjected to a gas pressure g_1. If the inner container fails, the gas expands and applies the gas pressure $g_2 = g_1 V_1 / V_2$ to the outer container. The container corresponds to a parallel system with stochastic dependence.





A_1 : failure of the inner container

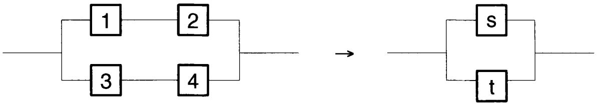
A_2 : failure of the outer container

Let the inner and the outer container be designed such that they have the same failure probability for the same gas pressure. Let the ratio of the failure probabilities be equal to the ratio of the gas pressures. Then the container has the following failure probability p :

$$p = P(A_1 \cap A_2) = P(A_1) \cdot P(A_2 \mid A_1) = p_1 \cdot p_1 \cdot V_2 / V_1 = p_1^2 V_2 / V_1$$

Example 3 : Mixed system with stochastically independent elements

A system of pumps consisting of two lines with two pumps each is installed to pump water out of an excavation. Each pump can handle the required amount of water. All the pumps have the same failure probability p_0, independent of the amount of water pumped. The failure probability of the system of pumps is calculated as follows :



line of pumps s : serial arrangement of pumps 1 and 2

$$\bar{A}_s = \bar{A}_1 \cap \bar{A}_2 \qquad q_s = (1-p_0)^2 \qquad p_s = 1-q_s = 2p_0 - p_0^2$$

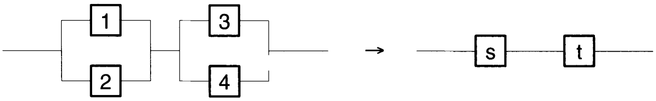
line of pumps t : serial arrangement of pumps 3 and 4

$$\bar{A}_t = \bar{A}_3 \cap \bar{A}_4 \qquad q_t = (1-p_0)^2 \qquad p_t = 1-q_t = 2p_0 - p_0^2$$

system of pumps : parallel arrangement of lines s and t

$$A = A_s \cap A_t \qquad p = p_s\, p_t = (2p_0 - p_0^2)^2 \approx 4p_0^2$$

As an alternative, consider a system of pumps consisting of two blocks with two pumps each. The failure probability of this system of pumps is calculated as follows :



block of pumps s : parallel arrangement of pumps 1 and 2

$$A_s = A_1 \cap A_2 \qquad p_s = p_0^2 \qquad q_s = 1 - p_0^2$$

block of pumps t : parallel arrangement of pumps 3 and 4

$$A_t = A_3 \cap A_4 \qquad p_t = p_0^2 \qquad q_t = 1 - p_0^2$$

system of pumps : serial arrangement of blocks s and t

$$\bar{A} = \bar{A}_s \cap \bar{A}_t \qquad q = q_s\, q_t \qquad p = 1-q = 1-(1-p_0^2)^2 \approx 2p_0^2$$

A comparison of the results shows that the failure probability of the first system of pumps is approximately twice as high as the failure probability of the second system of pumps.

10.3 RANDOM VARIABLES

10.3.1 INTRODUCTION

Random variable : In many applications the result of a stochastic experiment is determined by counting or measuring. The daily traffic census of vehicles at a certain location, annual measurements of the precipitation at a certain location and an experimental determination of the tensile strength of a certain grade of steel are typical examples. The result of these experiments is a real value. The random results of the experiment are described by a random variable which takes different numerical values.

In some applications, the result of a stochastic experiment is not represented by numerical values. For instance, the result of drawing a lot may be a "blank", a "prize" or a "first prize". This type of result often occurs in connection with a classification. The classification of wind velocities and the classification of damage states for buildings are typical examples. By assigning values to the possible results of an experiment, the random results may be described by a random variable which can take different numerical values.

Probability distribution and moments : The introduction of a random variable for the random result of an experiment allows the probability for the possible results to be represented as a function over the range of the random variable. This function is called the probability distribution of the random variable. The random character of an experiment is completely described by the probability distribution. Characteristic values for the random variable are calculated from the probability distribution. This calculation is based on the definition of the moments of a probability distribution. The fundamentals of probability distributions and their moments are treated in Sections 10.3.2 and 10.3.3.

Functional dependence : In applications, a random variable is often assumed to depend functionally on other random variables. For example, the force exerted by a spring is equal to the product of the spring constant and the displacement. If the spring constant and the displacement are independent random variables, then the resulting force is a dependent random variable. The probability distribution and the moments of a functionally dependent random variable may be determined from the probability distributions and moments of the independent random variables. The fundamentals for functional dependence on one or more independent random variables are treated in Sections 10.3.4 and 10.3.5.

Discrete and continuous distributions : Random variables may be discrete or continuous. Accordingly, discrete and continuous distributions are distinguished. Every distribution is based on a certain model and is specified by parameters. The discrete distributions mainly result from problems in game theory. Some of the continuous distributions result from limit considerations for the discrete distributions. The important discrete and continuous distributions as well as their technical applications are treated in Sections 10.3.6 and 10.3.7.

Tabulation : Various probability distributions for random variables lead to mathematical functions which can only be evaluated numerically. These distributions are therefore tabulated in mathematical handbooks. Such tabulations are used in the treatment of the examples.

10.3.2 PROBABILITY DISTRIBUTIONS

Introduction : A stochastic experiment is associated with a random variable which takes a real value for each of the random results. The axioms and rules of the calculus of probabilities for random events are applied to random variables. This leads to the definition of the distribution function for a random variable. In the case of a discrete random variable, a probability function is derived from the distribution function by the formation of differences. In the case of a continuous random variable, a density function is derived from the distribution function by differentiation. In the general case, the derivative of the distribution function may be represented as a combination of a probability function and a density function. The fundamentals of random variables and their distributions are compiled in the following.

Random variable : If S is the event space of an experiment, the corresponding random variable X is a mapping from the event set S to the set $\mathbb{R}$ of real numbers. Every elementary event $e \in S$ is assigned a real number $x \in \mathbb{R}$.

$$X : S \rightarrow \mathbb{R} \qquad X(e) = x$$

S	event set	X	random variable
e	elementary event	x	value of the random variable

Probability : The results of an experiment are described by associated ranges of the random variable X. A result $X < x$ means that the random variable X takes a real value which is less than a given real number x. The axioms and rules of the calculus of probabilities for events may therefore be transferred to random variables. In the limit $x \rightarrow -\infty$, the range $X < x$ is the impossible event; in the limit $x \rightarrow \infty$ it is the certain event :

$$\lim_{x \rightarrow -\infty} P(X < x) = 0 \qquad \lim_{x \rightarrow \infty} P(X < x) = 1$$

The ranges $X < x$ and $X \geq x$ describe complementary events, so that the following equation holds :

$$P(X < x) + P(X \geq x) = 1$$

The ranges $X < x_0$ and $x_0 \leq X < x_1$ describe incompatible events. Their union is $X \leq x_1$, and hence :

$$P(X < x_1) = P(X < x_0) + P(x_0 \leq X < x_1)$$

$$P(x_0 \leq X < x_1) = P(X < x_1) - P(X < x_0)$$

Distribution function : A distribution function $F_X(x)$ is introduced for the random variable X. The value of the distribution function is the probability $P(X < x)$; it lies in the interval [0,1].

$$F_X(x) := P(X < x) \qquad\qquad 0 \le F_X(x) \le 1$$

The properties of the distribution function $F_X(x)$ follow directly from the rules of the calculus of probabilities for a random variable. The distribution function $F_X(x)$ increases monotonically. It takes the value 0 for $x \to -\infty$ and the value 1 for $x \to \infty$.

$$\lim_{x \to -\infty} F_X(x) = 0 \qquad x_0 < x_1 \;\Rightarrow\; F_X(x_0) \le F_X(x_1) \qquad \lim_{x \to \infty} F_X(x) = 1$$

Classification of random variables : Random variables are classified according to the properties of their distribution functions :

(1) A random variable X is said to be discrete if the distribution function $F_X(x)$ is piecewise constant.

(2) A random variable X is said to be continuous if the distribution function $F_X(x)$ is continuous.

(3) If a distribution function $F_X(x)$ is neither continuous nor piecewise constant, then the random variable X is neither continuous nor discrete. In this case, the distribution function $F_X(x)$ may be represented as the sum of a continuous function and a piecewise constant function.

Since the derivatives of the distribution functions for different types of random variables differ significantly, they are referred to by different terms :

(1) The derivative of the distribution function of a discrete random variable is called a probability function.

(2) The derivative of the distribution function of a continuous random variable is called a density function.

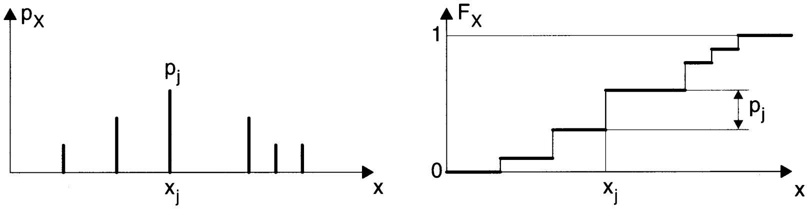
(3) The derivative of the distribution function of a general random variable, which may be neither discrete nor continuous, is called a generalized density function.

Discrete random variable : The distribution of a discrete random variable X is described by a probability function $p_X(x)$. The value of the probability function is the probability $P(X = x)$, which is non-zero only for discrete values x_j. The distribution function $F_X(x)$ is obtained by summing the probability function $p_X(x)$.

$$p_X(x) := P(X = x) \qquad\qquad 0 \le p_X(x) \le 1$$

$$F_X(x) := P(X < x) = \sum_{s<x} p_X(s) \qquad\qquad 0 \le F_X(x) \le 1$$

The probability function p_X is graphically represented by a bar diagram. The distribution function F_X is a step function. The relationship between the probability function and the distribution function is illustrated graphically.

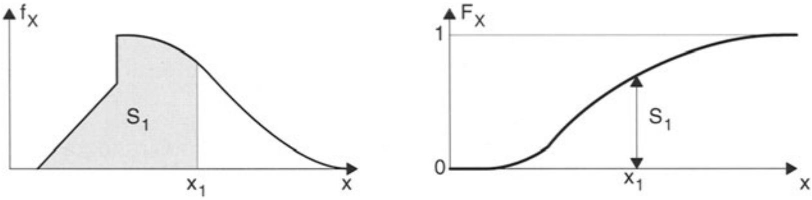


Continuous random variable : The distribution of a continuous random variable X is described by a density function, which is piecewise continuous. The limit of the probability density $P(x \le X < x + \Delta x) / \Delta x$ for $\Delta x \to 0$ is called the density function and is designated by $f_X(x)$. The distribution function $F_X(x)$ is calculated by integrating the density function f(x). Conversely, the density function $f_X(x)$ is obtained by differentiating the distribution function $F_X(x)$.

$$f_X(x) \quad := \lim_{\Delta x \to 0} \frac{P(x \le X < x + \Delta x)}{\Delta x} \qquad 0 \le f_X(x)$$

$$F_X(x) \ := \ P(X < x) \ = \int_{-\infty}^{x} f_X(s)\,ds \qquad 0 \le F_X(x) \le 1$$

The relationship between the density function and the distribution function is illustrated graphically.



General random variable : The distribution of a random variable X which is neither continuous nor discrete is described by a generalized density function $f_X(x)$ composed of the density function $f_0(x)$ for the continuous component and delta functions $\delta(x - x_j)$ with the probabilities $p_X(x_j)$ for the discrete values x_j. Using the rules for delta functions, the distribution function $F_X(x)$ is calculated by integrating the density function $f_X(x)$. Conversely, the density function $f_X(x)$ may be obtained by differentiating the distribution function $F_X(x)$.

$$f_X(x) \ := \ f_0(x) + \sum_j p_X(x_j)\,\delta(x - x_j)$$

$$F_X(x) \ := \ P(X < x) \ = \int_{-\infty}^{x} f_X(s)\,ds \qquad 0 \le F_X(x) \le 1$$

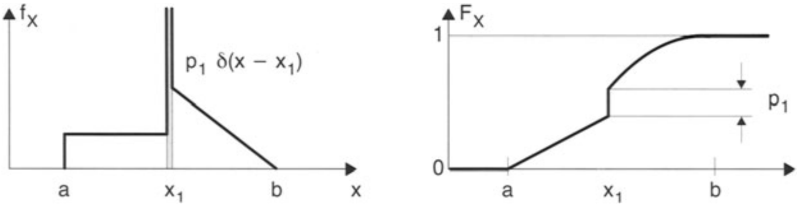
Integrating the delta function δ(x) yields the Heaviside function H(x). Conversely, differentiating the Heaviside function H(x) yields the delta function δ(x).

$$H(x) = \int_{-\infty}^{x} \delta(s)\, ds \qquad \delta(x) = \frac{dH(x)}{dx}$$

$$H(x) = 0 \quad \text{for} \quad x < 0 \qquad \delta(x) = 0 \quad \text{for} \quad x \neq 0$$

$$H(x) = 1 \quad \text{for} \quad x \geq 0$$

The relationship between a generalized density function and a general distribution function is illustrated graphically.



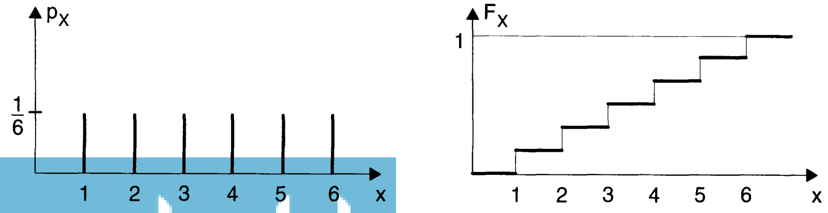
The distributions of discrete and continuous random variables may be considered as special cases of the generalized density function. This representation allows a uniform treatment of probability theory for random variables, without a distinction between discrete and continuous random variables. However, these differences must be taken into account in numerical calculations.

Example 1 : Discrete uniform distribution

Let a die with the numbers 1 to 6 be given. The discrete random variable X is the number thrown. Its range is the set of integers from 1 to 6. Each number occurs with the same probability 1 / 6 if the die is thrown. The random variable X is therefore uniformly distributed. The probability function $p_X(x)$ takes the value 1 / 6 for each of the integer values from 1 to 6.

$$p_X(x) = 1/6 \qquad x = 1, 2, 3, 4, 5, 6$$

The distribution function $F_X(x)$ is obtained by summing the probability function. Its value is 0 for $x < 1$ and 1 for $x > 6$. It is a step function.



Example 2 : Exponential distribution

Let a continuous random variable X be given whose density function is zero for $x < 0$ and decays exponentially with the rate $\lambda > 0$ for $x \geq 0$.

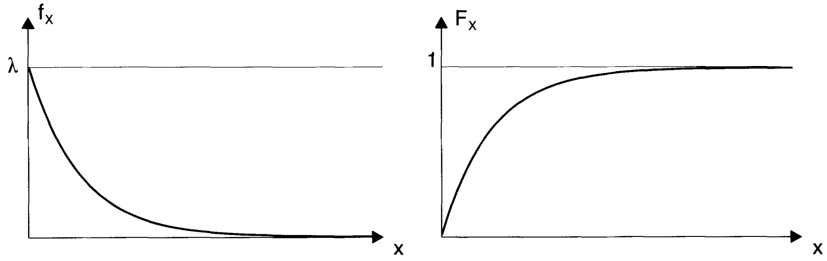
$$f_X(x) = C\, e^{-\lambda x} \qquad x \geq 0$$

The distribution function $F_X(x)$ is obtained by integrating the density function :

$$F_X(x) = \int_{-\infty}^{x} f_X(s)\, ds = C \int_{0}^{x} e^{-\lambda s}\, ds = -C\,(e^{-\lambda x} - 1)\,/\,\lambda$$

The constant of proportionality C is determined from the condition that the distribution function F_X tends to 1 for $x \to \infty$. The limit $F_X(x) = 1$ for $x \to \infty$ yields $C = \lambda$. This leads to the following density function and distribution function for the exponential distribution :

$$f_X(x) = \lambda\, e^{-\lambda x} \qquad F_X(x) = 1 - e^{-\lambda x} \qquad x \geq 0$$



10.3.3 MOMENTS

Introduction : Important properties of a random variable are described by characteristic values of the distribution of the random variable. The mean, the standard deviation and the skewness are typical characteristic values. The definition of the moments of a distribution forms the basis for the calculation of characteristic values. The moments are also called expectation values.

Moments : Let a random variable X with the density function $f_X(x)$ in general form be given. The k-th moment (moment of order k) is defined as follows :

$$E(X^k) := \int_{-\infty}^{\infty} x^k f_X(x)\, dx$$

The zeroth-order moment is given by the limit of the distribution function $F_X(x) = 1$ for $x \rightarrow \infty$. The first-order moment is the mean of X and is designated by μ_X.

$$E(X^0) = 1 \qquad E(X) = \mu_X$$

Central moments : The k-th central moment is the k-th moment with respect to the mean μ_X; it is designated by $D(X^k)$.

$$D(X^k) = \int_{-\infty}^{\infty} (x - \mu_X)^k f_X(x)\, dx$$

The central moments may be determined from the moments $E(X^k)$. The following formula is obtained by expanding the expression $(x - \mu_X)^k$ according to the binomial theorem :

$$D(X^k) = \sum_{j=0}^{k} \binom{k}{j} (-\mu_X)^j \; E(X^{k-j})$$

Evaluating this formula for the first, second and third central moment yields :

$$\begin{aligned} D(X\ ) &= \mu_X - \mu_X &&= 0 \\ D(X^2) &= E(X^2) - 2\mu_X^2 + \mu_X^2 &&= E(X^2) - \mu_X^2 \\ D(X^3) &= E(X^3) - 3\mu_X E(X^2) + 3\mu_X^3 - \mu_X^3 &&= E(X^3) - 3\mu_X E(X^2) + 2\mu_X^3 \end{aligned}$$

Characteristic values : The properties of a random variable X are described by the following characteristic values :

mean	:	μ_X	$=$	$E(X)$
variance	:	σ_X^2	$=$	$D(X^2)$
standard deviation	:	σ_X	$=$	$\sqrt{D(X^2)}$
variation coefficient	:	v_X	$=$	σ_X / μ_X
skewness	:	s_X	$=$	$D(X^3) / \sigma_X^3$

The mean μ_X is also called the expectation value of X. The variance σ_X^2 and the standard deviation σ_X characterize the extent to which X scatters around the mean μ_X. The variation coefficient v_X is the ratio of the standard deviation σ_X and the mean μ_X. The skewness s_X is a measure of the asymmetry of the generalized density function around the mean μ_X and is zero in the symmetric case.

Example 1 : Moments and characteristic values of the discrete uniform distribution

Consider the discrete random variable X with the discrete uniform distribution from Example 1 in Section 10.3.2.

$$p_X(x) = 1/6 \qquad\qquad x = 1,2,3,4,5,6$$

In the calculation of the moments for discrete random variables, the integral involving the generalized density function becomes a sum involving the probability function.

$$E(X) = \sum_{x=1}^{6} x \, p_X(x) = \frac{1}{6} \sum_{x=1}^{6} x = \frac{21}{6} = 3.5$$

$$E(X^2) = \sum_{x=1}^{6} x^2 p_X(x) = \frac{1}{6} \sum_{x=1}^{6} x^2 = \frac{91}{6} = 15.167$$

$$E(X^3) = \sum_{x=1}^{6} x^3 p_X(x) = \frac{1}{6} \sum_{x=1}^{6} x^3 = \frac{441}{6} = 73.5$$

The second and third central moments are calculated as follows using $\mu_X = E(X) = 3.5$:

$$D(X^2) = E(X^2) - \mu_X^2 = 2.917$$

$$D(X^3) = E(X^3) - 3\,\mu_X E(X^2) + 2\,\mu_X^3 = 0.000$$

The mean and the standard deviation are given by :

mean	:	μ_X	$=$	$E(X)$	$=$	3.500
standard deviation	:	σ_X	$=$	$\sqrt{D(X^2)}$	$=$	1.708

The skewness of the probability function is $s_X = 0$, since the distribution is symmetric with respect to its mean.

Example 2 : Moments and characteristic values of the exponential distribution

Consider the continuous random variable X with the exponential distribution from Example 2 in Section 10.3.2.

$$f_X(x) \quad = \quad \lambda\, e^{-\lambda x} \qquad\qquad \lambda > 0 \qquad\qquad x \geq 0$$

The k-th moments may be calculated recursively. The recursion formula is obtained through integration by parts :

$$E(X^k) \quad = \quad \int\limits_0^\infty x^k\, \lambda\, e^{-\lambda x}\, dx \quad = \quad \Big[-x^k\, e^{-\lambda x}\Big]_0^\infty \; + \; \frac{k}{\lambda} \int\limits_0^\infty x^{k-1}\, \lambda\, e^{-\lambda x}\, dx$$

$$E(X^k) \quad = \quad k\, E(X^{k-1}) \,/\, \lambda \qquad\qquad k \geq 1$$

With $E(X^0) = 1$, this yields the formula for the k-th moment :

$$E(X^k) \quad = \quad k\,!\,/\,\lambda^k \qquad\qquad k \geq 0$$

The second and third central moments are calculated as follows using $\mu_X = E(X) = 1\,/\,\lambda$:

$$D(X^2) \quad = \quad E(X^2) \; - \; \mu_X^2 \qquad\qquad = \quad (2-1)\,/\,\lambda^2 \qquad = \quad 1\,/\,\lambda^2$$

$$D(X^3) \quad = \quad E(X^3) \; - \; 3\,\mu_X\, E(X^2) \; + \; 2\,\mu_X^3 \quad = \quad (6-6+2)\,/\,\lambda^3 \quad = \quad 2\,/\,\lambda^3$$

The exponential distribution is thus characterized by the following values :

mean	:	μ_X	=	$E(X)$	=	$1\,/\,\lambda$
variance	:	σ_X^2	=	$D(X^2)$	=	$1\,/\,\lambda^2$
standard deviation	:	σ_X	=	$\sqrt{D(X^2)}$	=	$1\,/\,\lambda$
variation coefficient	:	v_X	=	$\sigma_X\,/\,\mu_X$	=	1
skewness	:	s_X	=	$D(X^3)\,/\,\sigma_X^3$	=	2

10.3.4 FUNCTIONS OF ONE RANDOM VARIABLE

Introduction : In technical applications, a random variable Y is often assumed to exhibit a deterministic dependence on a random variable X. The dependence is described by a function. The distribution and the moments for Y are to be determined from the distribution and the moments for X. The required fundamentals are treated in the following.

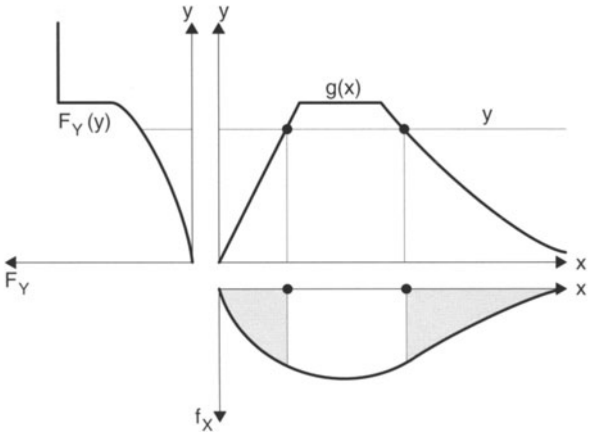
Function : Let every value x of the random variable X be assigned a unique value y of a random variable Y. This assignment is described by a function $y = g(x)$. The inverse function $x = g^{-1}(y)$ exists if every value y is also associated with a unique value x.

$$Y := g(X) \qquad\qquad y = g(x)$$

Distribution function : The probability that Y takes a value less than y is equal to the probability that the function g(X) takes a value less than y. The distribution function $F_Y(y)$ is therefore obtained by integrating the density function $f_X(x)$ over the range on which g(x) is less than y.

$$F_Y(y) = P(Y < y) = P(g(X) < y) = \int\limits_{g(x) < y} f_X(x)\,dx$$

The determination of the distribution function $F_Y(y)$ from the density function $f_X(x)$ using the function $y = g(x)$ is illustrated graphically.



If the inverse function $x = g^{-1}(y)$ exists, the relationship between the distribution functions F_X and F_Y may be determined explicitly. Two cases need to be considered, depending on whether the function $y = g(x)$ increases or decreases monotonically.

$$F_Y(y) = \int_{-\infty}^{g^{-1}(y)} f_X(x)\,dx = F_X(g^{-1}(y)) = F_X(x) \qquad g'(x) \ge 0$$

$$F_Y(y) = \int_{g^{-1}(y)}^{\infty} f_X(x)\,dx = 1 - F_X(g^{-1}(y)) = 1 - F_X(x) \qquad g'(x) \le 0$$

Probability function and density function : If the random variable X is discrete, the random variable Y is also discrete. If the random variable X is continuous, the random variable Y may be discrete or continuous, depending on the function $y = g(x)$. If $g(x)$ is a step function, the random variable Y is discrete. If $g(x)$ is piecewise invertible, the random variable Y is continuous.

The relationship between the distributions of X and Y can be determined explicitly if the inverse function $x = g^{-1}(y)$ exists. In this case, the probability function for a discrete random variable is given by :

$$p_Y(y) = p_X(g^{-1}(y)) = p_X(x)$$

The density function for a continuous random variable is obtained by differentiating the distribution function. A distinction between the cases of monotonically increasing and decreasing functions is avoided by using the absolute value.

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{dg^{-1}}{dy} \right| = f_X(x) \left| \frac{dx}{dy} \right|$$

Moments : The k-th moment of the random variable Y is calculated according to the following rule :

$$E(Y^k) = \int_{-\infty}^{\infty} y^k f_Y(y)\,dy = \int_{-\infty}^{\infty} g^k(x)\, f_X(x)\,dx$$

The rule holds for general functions $g(x)$ and for generalized density functions $f_X(x)$. The central moments $D(Y^k)$ are calculated from the moments $E(Y^k)$ according to the formulas in Section 10.3.3.

Transformation : If the function $y = g(x)$ has an inverse $x = g^{-1}(y)$, the random variable X may be mapped to Y using $g(X)$ and conversely the random variable Y may be mapped to X using $g^{-1}(Y)$. These mappings are called transformations of random variables.

Example 1 : Linear dependence

Let a random variable Y exhibit a linear dependence on a random variable X. Then the following rules of transformation hold :

$$\begin{array}{lll} Y = g(X) = aX + b & y = g(x) = ax + b & \\ X = g^{-1}(Y) = (Y - b) / a & x = g^{-1}(y) = (y - b) / a & a \neq 0 \end{array}$$

The distribution function $F_Y(y)$ is given by :

$$\begin{array}{ll} F_Y(y) = F_X(x) = F_X((y-b)/a) & a > 0 \\ F_Y(y) = 1 - F_X(x) = 1 - F_X((y-b)/a) & a < 0 \end{array}$$

If X is discretely distributed, then Y has the following probability function $p_Y(y)$:

$$p_Y(y) = p_X(x) = p_X((y-b)/a) \qquad a \neq 0$$

If X is continuously distributed, then Y has the following density function $f_Y(y)$:

$$f_Y(y) = \frac{1}{|a|} f_X(x) = \frac{1}{|a|} f_X((y-b)/a) \qquad a \neq 0$$

The first and second moments are determined as follows :

$$E(Y) = \int_{-\infty}^{\infty} (ax + b) \; f_X(x)\,dx = a \; E(X) + b$$

$$E(Y^2) = \int_{-\infty}^{\infty} (ax + b)^2 \, f_X(x)\,dx = a^2 E(X^2) + 2\,ab\,E(X) + b^2$$

This yields the following formulas for the mean and the variance :

$$\begin{array}{l} \mu_Y = E(Y) = a\mu_X + b \\ \sigma_Y^2 = E(Y^2) - \mu_Y^2 = a^2 E\,(X^2) + 2ab\,E(X) + b^2 - (a\mu_X + b)^2 \\ \sigma_Y^2 = a^2(E\,(X^2) - \mu_X^2) = a^2\,\sigma_X^2 \end{array}$$

Example 2 : Quadratic dependence

Let a random variable Y exhibit a purely quadratic dependence on a random variable X. Then the following rules hold :

$$Y = g(X) = X^2 \qquad\qquad y = g(x) = x^2$$

There is no inverse function $x = g^{-1}(y)$. The distribution function $F_Y(y)$ is therefore calculated as follows :

$$F_Y(y) = \int\limits_{x^2<y} f_X(x)\,dx = \int\limits_{-\sqrt{y}}^{\sqrt{y}} f_X(x)\,dx$$

$$F_Y(y) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$$

If the random variable X is continuous, the random variable Y is also continuous. The density function $f_Y(y)$ is obtained by differentiating the distribution function :

$$f_Y(y) = \frac{dF_Y}{dy} = \frac{1}{2}(f_X(\sqrt{y}) + f_X(-\sqrt{y})) \,/\, \sqrt{y}$$

The moments of the random variable Y are obtained from the moments of the random variable X :

$$E(Y^k) = \int\limits_{-\infty}^{\infty} y^k f_Y(y)\,dy = \int\limits_{-\infty}^{\infty} x^{2k} f_X(x)\,dx = E(X^{2k})$$

The mean is obtained as follows :

$$\mu_Y = E(Y) = E(X^2) = \mu_X^2 + \sigma_X^2$$

10.3.5 FUNCTIONS OF SEVERAL RANDOM VARIABLES

Introduction : In technical applications, a random variable is often assumed to depend deterministically on several independent random variables. This dependence is described by a multidimensional function. The distribution and the moments of the dependent random variable are to be determined from the distributions and moments of the independent random variables. The required fundamentals are treated mainly for functions of two independent random variables.

Independent random variables : Two random variables X and Y are said to be independent if the events $X < x$ and $Y < y$ are stochastically independent for all x, y.

Function : Every pair of values (x,y) of the independent random variables X,Y is assigned a value z of the random variable Z. This assignment is described by a two-dimensional function $z = g(x,y)$.

$$Z := g(X,Y) \qquad z = g(x,y)$$

Distribution function : Let the random variables X, Y be independent. Let their generalized density functions be $f_X(x)$, $f_Y(y)$. According to the product rule, the probability that X takes values between x and $x + dx$ and Y takes values between y and $y + dy$ is $f_X(x)\, f_Y(y)\, dx\, dy$. Integrating this probability over the range $g(x,y) < z$ yields the distribution function $F_Z(z)$:

$$F_Z(z) = P(Z < z) = P(g(X,Y) < z) = \iint\limits_{g(x,y) < z} f_X(x)\, f_Y(y)\, dx\, dy$$

Probability function and density function : The distribution of the random variable Z depends on the properties of the distributions for X, Y and the function g(x, y). If the random variable Z is continuous, the density function $f_Z(z)$ is obtained by differentiating the distribution function.

Moments : The k-th moment of the random variable Z is obtained according to the following rule :

$$E(Z^k) = \int_{-\infty}^{\infty} z^k\, f_Z(z)\, dz = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g^k(x,y)\, f_X(x)\, f_Y(y)\, dx\, dy$$

The rule is valid for general functions g(x,y) and for generalized density functions $f_X(x), f_Y(y)$. The central moments $D(Z^k)$ are obtained from the moments $E(Z^k)$ according to the formulas in Section 10.3.3.

Elementary dependences : The cases in which the random variable Z is the sum, the product, the minimum or the maximum of two independent random variables X and Y are of special importance in probability theory. These cases are treated in the following.

Sum Z := X + Y : The distribution function $F_Z(z)$ for the sum $Z = X + Y$ may be reduced from a double integral to a simple integral.

$$F_Z(z) = \iint\limits_{x+y<z} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) = \int\limits_{-\infty}^{\infty} \left\{ \int\limits_{-\infty}^{z-y} f_X(x)\, dx \right\} f_Y(y)\, dy$$

$$F_Z(z) = \int\limits_{-\infty}^{\infty} F_X(z-y)\, f_Y(y)\, dy$$

The simple integral is called a convolution integral. Interchanging X and Y as well as x and y yields the dual convolution integral. Differentiating the distribution function $F_Z(z)$ leads to the following convolution integral for the density function $f_Z(z)$:

$$f_Z(z) = \int\limits_{-\infty}^{\infty} f_X(z-y)\, f_Y(y)\, dy$$

The k-th moment is obtained as follows :

$$E(Z^k) = \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} (x+y)^k\, f_X(x)\, f_Y(y)\, dx\, dy$$

$$E(Z^k) = \sum_{j=0}^{k} \binom{k}{j} \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} x^{k-j}\, y^j\, f_X(x)\, f_Y(y)\, dx\, dy$$

$$E(Z^k) = \sum_{j=0}^{k} \binom{k}{j} E(X^{k-j})\, E(Y^j)$$

The mean and the variance are given by :

$$\begin{aligned} \mu_Z &= E(Z) &&= \mu_X + \mu_Y \\ \sigma_Z^2 &= E(Z^2) - \mu_Z^2 &&= \sigma_X^2 + \sigma_Y^2 \end{aligned}$$

Product Z := X • Y : As in the case of a sum, the double integral for the distribution function $F_Z(z)$ of the product $Z = X \cdot Y$ may be reduced to a simple integral.

$$F_Z(z) = \iint\limits_{x \cdot y<z} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) = \int\limits_{-\infty}^{0} \left\{ \int\limits_{z/y}^{\infty} f_X(x)\, dx \right\} f_Y(y)\, dy + \int\limits_{0}^{\infty} \left\{ \int\limits_{-\infty}^{z/y} f_X(x)\, dx \right\} f_Y(y)\, dy$$

$$F_Z(z) = \int\limits_{-\infty}^{0} (1 - F_X(z/y))\, f_Y(y)\, dy + \int\limits_{0}^{\infty} F_X(z/y)\, f_Y(y)\, dy$$

Differentiating the distribution function $F_Z(z)$ leads to the following convolution integral for the density function $f_Z(z)$:

$$f_Z(z) \;=\; \int_{-\infty}^{\infty} \frac{1}{|y|}\, f_X(z/y)\, f_Y(y)\, dy$$

The k-th moment of Z is equal to the product of the k-th moments of X and Y :

$$E(Z^k) \;=\; \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (xy)^k\, f_X(x)\, f_Y(y)\, dx\, dy \;=\; E(X^k)\, E(Y^k)$$

The mean and the variance are given by :

$$\begin{aligned} \mu_Z &= E(Z) &&= \mu_X\, \mu_Y \\ \sigma_Z^2 &= E(Z^2) - \mu_Z^2 &&= (\sigma_X^2 + \mu_X^2)\,(\sigma_Y^2 + \mu_Y^2) \;-\; \mu_Z^2 \end{aligned}$$

Maximum Z = max (X,Y) : The distribution function $F_Z(z)$ for the maximum Z = max(X,Y) may be reduced from the double integral to the product of the distribution functions F_X and F_Y :

$$F_Z(z) \;=\; \iint\limits_{\max(x,y) < z} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) \;=\; \int_{-\infty}^{z}\int_{-\infty}^{z} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) \;=\; F_X(z)\, F_Y(z)$$

Differentiating the distribution function $F_Z(z)$ leads to the following density function $f_Z(z)$:

$$f_Z(z) \;=\; f_X(z)\, F_Y(z) \;+\; F_X(z)\, f_Y(z)$$

The moments of Z can only be reduced to the moments of X,Y if the distributions of X, Y are known.

Minimum Z = min (X,Y) : As for the maximum, the double integral for the distribution function $F_Z(z)$ of the minimum Z = min (X,Y) may be reduced to an expression involving the distribution functions F_X and F_Y.

$$F_Z(z) \;=\; \iint\limits_{\min(x,y) < z} f_X(x)\, f_Y(y)\, dx\, dy \;=\; 1 - \iint\limits_{\min(x,y) \geq z} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) \;=\; 1 - \int_{z}^{\infty}\int_{z}^{\infty} f_X(x)\, f_Y(y)\, dx\, dy$$

$$F_Z(z) \;=\; 1 - (1 - F_X(z))\,(1 - F_Y(z))$$

Differentiating the distribution function $F_Z(z)$ leads to the following density function $f_Z(z)$:

$$f_Z(z) \quad = \quad f_X(z)\,(1 - F_Y(z)) \; + \; (1 - F_X(z))\, f_Y(z)$$

Generalized dependences : The elementary dependences of two independent random variables may be generalized to several random variables. The probability distribution and the moments of the dependent random variable are determined iteratively. In each step, a dependent random variable is considered which exhibits an elementary dependence on two independent random variables. This procedure leads to the following results for several random variables :

1. If the dependent random variable Z is a sum or a product of n independent random variables X_j, then the mean and the variance of Z are obtained as follows :

$$Z := \sum_{j=1}^{n} X_j \qquad \mu_Z = \sum_{j=1}^{n} \mu_{Xj} \qquad \sigma_Z^2 = \sum_{j=1}^{n} \sigma_{Xj}^2$$

$$Z := \prod_{j=1}^{n} X_j \qquad \mu_Z = \prod_{j=1}^{n} \mu_{Xj} \qquad \sigma_Z^2 = \prod_{j=1}^{n} (\sigma_{Xj}^2 + \mu_{Xj}^2) - \mu_Z^2$$

2. If the dependent random variable Z is a linear combination of n independent random variables X_j, then the mean and the variance of Z are calculated as follows, using the rules for linear transformations in Section 10.3.4 :

$$Z := \sum_{j=1}^{n} c_j\, X_j \qquad \mu_Z = \sum_{j=1}^{n} c_j\, \mu_{Xj} \qquad \sigma_Z^2 = \sum_{j=1}^{n} c_j^2\, \sigma_{Xj}^2$$

3. If the random variable Z is a maximum or a minimum of n independent random variables X_j, then the distribution function $F_Z(z)$ is obtained as follows :

$$Z := \max\,(X_1,\ldots,X_n) \qquad F_Z(z) = \prod_{j=1}^{n} F_{Xj}(z)$$

$$Z := \min\,(X_1,\ldots,X_n) \qquad F_Z(z) = 1 - \prod_{j=1}^{n} (1 - F_{Xj}(z))$$

Example 1 : Sum of two discrete random variables $Z = X + Y$

Two dice are thrown in a game. Each yields a number from 1 to 6. Let the number on the first die be the discrete random variable X, and let the number on the second die be the discrete random variable Y. Both random variables are uniformly distributed according to Example 1 in Section 10.3.2. The sum of the numbers on the two dice is the discrete random variable $Z = X + Y$. It takes integer values from 2 to 12. Its probability function is to be determined.

$$p_X(x) = 1/6 \qquad x = 1,2,3,4,5,6$$

$$p_Y(y) = 1/6 \qquad y = 1,2,3,4,5,6$$

The probability function $p_Z(z)$ may be determined by inspection. This is demonstrated for $z = 5$. The numbers on the dice add up to $z = 5$ for the pairs (1,4), (2,3), (3,2) and (4,1). Due to the stochastic independence of the two numbers, each pair of numbers occurs with the probability $\frac{1}{6} \cdot \frac{1}{6}$. There are 4 different pairs of numbers which yield $z = 5$, and hence the probability is $\frac{4}{36}$.

The probability function $p_Z(z)$ is formally obtained as a convolution of the probability functions $p_X(x)$ and $p_Y(y)$. For discrete random variables, the integral form of the convolution involving generalized density functions becomes a sum involving probability functions.

$$p_Z(z) = \sum_{y=-\infty}^{\infty} p_X(z-y)\, p_Y(y)$$

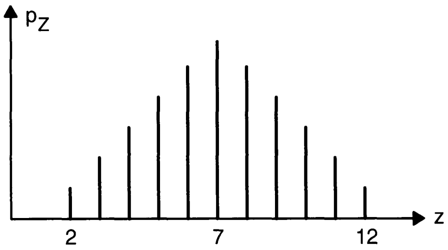
For $2 \le z \le 7$, the probability functions $p_X(z-y)$ and $p_Y(y)$ take non-zero values only for $1 \le y \le z-1$, and hence :

$$p_Z(z) = \sum_{y=1}^{z-1} p_X(z-y)\, p_Y(y) = \frac{1}{36} \sum_{y=1}^{z-1} 1 = \frac{1}{36}(z-1) \qquad 2 \le z \le 7$$

For $7 \le z \le 12$, the probability functions $p_X(z-y)$ and $p_Y(y)$ take non-zero values only for $z-6 \le y \le 6$, and hence :

$$p_Z(z) = \sum_{y=z-6}^{6} p_X(z-y)\, p_Y(y) = \frac{1}{36} \sum_{y=z-6}^{6} 1 = \frac{1}{36}(13-z) \qquad 7 \le z \le 12$$

The probability function $p_Z(z)$ of the discrete random variable Z has a triangular form over the range $2 \le z \le 12$.



Example 2 : Sum of two continuous random variables $Z = X + Y$

Let the random variable Z be the sum of random variables X and Y which are distributed exponentially according to Example 1 in Section 10.3.2 :

$$f_X(x) = \alpha\, e^{-\alpha x} \qquad \alpha > 0 \qquad x \geq 0$$

$$f_Y(y) = \beta\, e^{-\beta x} \qquad \beta > 0 \qquad y \geq 0$$

The density function is given by the following integral :

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(z-y)\, f_Y(y)\, dy$$

The function $f_X(z-y)$ takes non-zero values for $y \leq z$. The function $f_Y(y)$ takes non-zero values for $y \geq 0$. The integration range is therefore restricted to the interval from 0 to z.

$$f_Z(z) = \int_0^z \alpha\, e^{-\alpha(z-y)}\, \beta\, e^{-\beta y}\, dy$$

$$f_Z(z) = \alpha\beta\, e^{-\alpha z} \int_0^z e^{(\alpha-\beta)y}\, dy$$

$$f_Z(z) = \frac{\alpha\beta}{\alpha-\beta}\,(e^{-\beta z} - e^{-\alpha z}) \qquad \alpha \neq \beta$$

$$f_Z(z) = \alpha^2\, z\, e^{-\alpha z} \qquad \alpha = \beta$$

The mean μ_Z and the variance σ_Z^2 are given by :

$$\mu_X = 1/\alpha \qquad \mu_Y = 1/\beta \qquad \mu_Z = \mu_X + \mu_Y = 1/\alpha + 1/\beta$$

$$\sigma_X^2 = 1/\alpha^2 \qquad \sigma_Y^2 = 1/\beta^2 \qquad \sigma_Z^2 = \sigma_X^2 + \sigma_Y^2 = 1/\alpha^2 + 1/\beta^2$$

10.3.6 DISCRETE DISTRIBUTIONS

Introduction : Many problems in probability theory may be reduced to experiments in which a certain event either occurs or does not occur. In analogy with game theory, the occurrence of the event is called a success, and the non-occurrence of the event is called a failure. An experiment with the possible results success or failure is called an elementary experiment. In a series of elementary experiments, these experiments are assumed to be independent of each other. The different problems which arise if several elementary experiments are considered lead to different distributions for discrete random variables. The important discrete distributions and their parameters are treated in the following.

10.3.6.1 Bernoulli distribution

Model : An elementary experiment yields either a success or a failure. The discrete random variable X is assigned the value 0 for failure and the value 1 for success.

random variable $X = 0{,}1$

Success occurs with the probability p, and failure occurs with the probability $q = 1 - p$. The success probability p is the parameter of the distribution :

success probability $0 < p < 1$

failure probability $q = 1 - p$

Probability function : The Bernoulli distribution is described by the following probability function :

$$p_X(x) = p^x q^{1-x} \qquad x = 0{,}1$$

Moments : The k-th moment of the Bernoulli distribution for $k \geq 1$ is equal to the success probability.

$$E(X^k) = \sum_{x=0}^{1} x^k p^x q^{1-x} = p \qquad k \geq 1$$

The mean and the variance are given by :

$$\mu_X = E(X) = p$$
$$\sigma_X^2 = E(X^2) - \mu_X^2 = p - p^2 = pq$$

Remark : The Bernoulli distribution is the simplest distribution for a discrete random variable. It describes the elementary experiment and thus forms the theoretical foundation for other discrete distributions.

10.3.6.2 Binomial distribution

Model : An elementary experiment is repeated n times. Each experiment yields either a success or a failure. Let the experiments be independent of each other. The discrete random variable X is the number of successes in the n experiments. It can take the following values :

random variable	$X = 0,1,...,n$

The number n of elementary experiments and the success probability p are the parameters of the distribution :

number of experiments	$n > 0$
success probability	$0 < p < 1$
failure probability	$q = 1 - p$

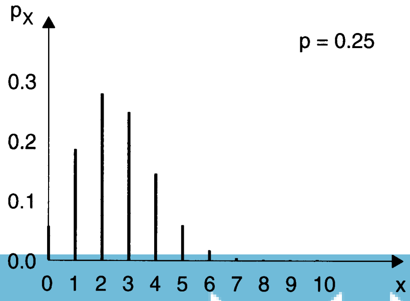
Probability function : For n experiments, there are $n!/(x!(n-x)!)$ different ordered n-tuples with exactly x successes and $(n-x)$ failures. According to the product rule, the probability for the occurrence of x successes and $(n-x)$ failures in an n-tuple is $p^x q^{n-x}$. The probability $p_X(x)$ for the occurrence of one of the possible n-tuples is determined using the sum rule :

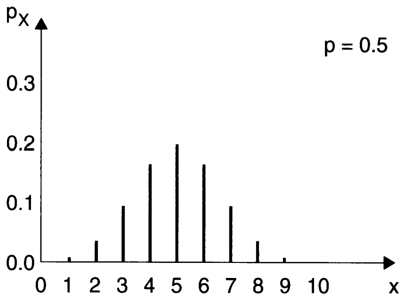
$$p_X(x) = \frac{n!}{x!(n-x)!} p^x q^{n-x} = \binom{n}{x} p^x q^{n-x}$$

The probability function $p_X(x)$ is calculated recursively according to the following rule :

$$p_X(x+1) = \frac{n-x}{x+1} \frac{p}{q} p_X(x) \qquad p_X(0) = q^n$$

The probability function has one or two maxima in the interval $[np - q, np + p]$. It is symmetric for $p = q = 0.5$. The probability function is illustrated graphically for $p = 0.25$ and $p = 0.50$ with $n = 10$.





Moments : The k-th moment of the binomial distribution is obtained as follows, using q = 1 − p :

$$E(X^k) = \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x}$$

Differentiating the k-th moment with respect to the parameter p yields a recursive equation for the moment of order (k + 1) :

$$\begin{aligned}
E'(X^k) &= \frac{d}{dp} E(X^k) = \frac{d}{dp} \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} x^k \binom{n}{x} \left[x p^{x-1} (1-p)^{n-x} - p^x (n-x) (1-p)^{n-x-1} \right] \\
&= \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{x}{p} - \frac{n-x}{1-p} \right] \\
&= \sum_{x=0}^{n} x^k \binom{n}{x} p^x (1-p)^{n-x} \left[\frac{x}{p(1-p)} - \frac{n}{1-p} \right] \\
E'(X^k) &= \frac{1}{p(1-p)} E(X^{k+1}) - \frac{n}{1-p} E(X^k) \\
E(X^{k+1}) &= np\, E(X^k) + p(1-p) E'(X^k)
\end{aligned}$$

The moments of higher order for k > 0 are calculated starting from the zeroth-order moment $E(X^0) = 1$ with $E'(X^0) = 0$. The first and second moments are :

$$\begin{aligned}
E(X) &= np\, E(X^0) + p(1-p)\, E'(X^0) = np \\
E(X^2) &= np\, E(X) + p(1-p)\, E'(X) = (np)^2 + np(1-p)
\end{aligned}$$

The mean and the variance are obtained from the first and second moments as follows :

$$\begin{aligned}
\mu_X &= E(X) &&= np \\
\sigma_X^2 &= E(X^2) - \mu_X^2 &&= np(1-p) = npq
\end{aligned}$$

Property : A discrete random variable X which has a binomial distribution with the parameters n, p is designated by X(n, p). The Bernoulli distribution is a special case of the binomial distribution with n = 1. According to the model of the binomial distribution, the random variable X(n, p) is the number of successes in n experiments. It is the sum of n random variables $X_j(1, p)$ which have a Bernoulli distribution. The random variable $X_j(1, p)$ is the result of the j-th experiment. Its value is 0 for a failure and 1 for a success.

$$X(n, p) = \sum_{j=1}^{n} X_j (1, p)$$

More generally, the sum of binomially distributed random variables with identical success probability p is also binomially distributed.

$$X(m, p) = \sum_{j=1}^{n} X_j\,(n_j, p) \qquad\qquad m = \sum_{j=1}^{n} n_j$$

Example : Consider the automated manufacture of a product. Let 1% of all products be faulty. The probabilities that exactly zero or one products in a sample of 10 products are faulty are calculated as follows :

$$p = 0.01 \qquad q = 0.99 \qquad n = 10$$

$$p_X(0) = \binom{10}{0}\ 0.01^0\ 0.99^{10} = 0.9044 = 90.44\%$$

$$p_X(1) = \binom{10}{1}\ 0.01^1\ 0.99^9 = 0.0914 = 9.14\%$$

The probability that several products in the sample are faulty is calculated as follows :

$$P(X > 1) = 1 - \sum_{s=0}^{1} p_X(s) = 0.0042 = 0.42\%$$

10.3.6.3 Pascal distribution

Model : An elementary experiment is repeated an arbitrary number of times. Each experiment yields either a success or a failure. Let the experiments be independent of each other. The discrete random variable X is the number of experiments up to and including the m-th success. It can take the following values :

random variable $\qquad X = m, m+1, ...$

The number m of successes and the success probability p are the parameters of the distribution :

number of successes	$m > 0$
success probability	$0 < p < 1$
failure probability	$q = 1 - p$

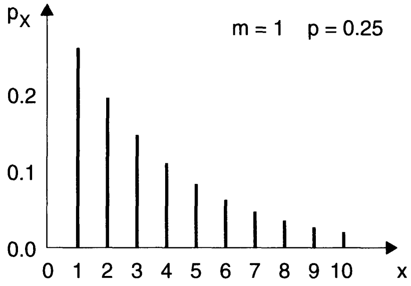
Probability function : The first $x-1$ experiments yield $m-1$ successes, and the x-th experiment yields the m-th success. The probability for $m-1$ successes in $x-1$ experiments is given by the binomial distribution. The probability for the occurrence of the m-th success in the x-th experiment is p. The product of these two probabilities is the probability $p_X(x)$:

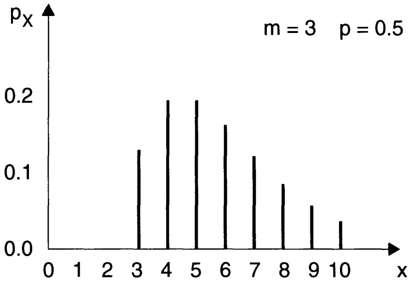
$$p_X(x) = \binom{x-1}{m-1}\ p^{m-1}\ q^{x-m}\ p = \binom{x-1}{m-1}\ p^m\ q^{x-m}$$

The probability function $p_X(x)$ is calculated recursively :

$$p_X(x+1) \;=\; \frac{x}{x-m+1}\, q\, p_X(x) \qquad\qquad p_X(m) \;=\; p^m$$

For $m = 1$, the probability function forms a geometric sequence which decays with the failure probability q. For $m > 1$, it possesses one or two maxima in the interval $[(m-1)/p,\ (m-1)/p+1]$. The probability function is illustrated graphically for $m = 1$ and $p = 0.25$ as well as for $m = 3$ and $p = 0.50$.





Moments : The k-th moment of the Pascal distribution is obtained as follows, using $q = 1 - p$:

$$E(X^k) \;=\; \sum_{x=m}^{\infty} x^k \binom{x-1}{m-1} p^m \,(1-p)^{x-m}$$

Differentiating the k-th moment with respect to the parameter p yields a recursive equation for the moment of order $(k+1)$:

$$\begin{aligned}
E'(X^k) \;&=\; \frac{d}{dp}\,E(X^k) \;=\; \frac{d}{dp}\sum_{x=m}^{\infty} x^k \binom{x-1}{m-1} p^m\,(1-p)^{x-m} \\
&=\; \sum_{x=m}^{\infty} x^k \binom{x-1}{m-1} \left[m\,p^{m-1}\,(1-p)^{x-m} - p^m\,(x-m)\,(1-p)^{x-m-1}\right] \\
&=\; \sum_{x=m}^{\infty} x^k \binom{x-1}{m-1} p^m\,(1-p)^{x-m} \left[\frac{m}{p} - \frac{x-m}{1-p}\right] \\
&=\; \sum_{x=m}^{\infty} x^k \binom{x-1}{m-1} p^m\,(1-p)^{x-m} \left[\frac{m}{p(1-p)} - \frac{x}{1-p}\right] \\
E'(X^k) \;&=\; \frac{m}{p(1-p)}\,E(X^k) \;-\; \frac{1}{1-p}\,E(X^{k+1}) \\
E(X^{k+1}) \;&=\; \frac{m}{p}\,E(X^k) \;-\; (1-p)\,E'(X^k)
\end{aligned}$$

The moments of higher order for $k > 0$ are calculated starting from the zeroth-order moment $E(X^0) = 1$ with $E'(X^0) = 0$. The first and second moments are :

$$E(X) = \frac{m}{p} E(X^0) - (1-p)\, E'(X^0) = \frac{m}{p}$$

$$E(X^2) = \frac{m}{p} E(X) - (1-p)\, E'(X) = \frac{m^2}{p^2} + (1-p)\frac{m}{p^2}$$

The mean and the variance are obtained from the first and second moments as follows :

$$\mu_X = E(X) = m/p$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = (1-p)\, m/p^2 = mq/p^2$$

Property : A discrete random variable X which has a Pascal distribution with the parameters m, p is designated by X(m, p). The geometric distribution is a special case of the Pascal distribution with $m = 1$. According to the model of the Pascal distribution, the random variable X(m, p) is the number of experiments up to and including the m-th success. It is the sum of m random variables $X_j(1, p)$ which are geometrically distributed. The random variable $X_j(1, p)$ is the number of experiments after the (j – 1)-th success up to and including the j-th success.

$$X(m,p) = \sum_{j=1}^{m} X_j(1,p)$$

More generally, the sum of random variables which have Pascal distributions with identical success probability p also has a Pascal distribution.

$$X(m,p) = \sum_{j=1}^{n} X_j(m_j, p) \qquad m = \sum_{j=1}^{n} m_j$$

Example : Construction work is to be carried out in a riverbed. It takes three years and needs to be protected against flooding. Let the protection be designed such that it fails in case of high water which on average occurs every 20 years. The probabilities that the protection first fails in the first, second or third year are calculated as follows :

$$p = 1/20 = 0.05 \qquad q = 0.95 \qquad m = 1$$

$$p_X(x) = p\, q^{x-1}$$

$$p_X(1) = 0.05 \cdot 0.95^0 = 0.0500 = 5.00\%$$

$$p_X(2) = 0.05 \cdot 0.95^1 = 0.0475 = 4.75\%$$

$$p_X(3) = 0.05 \cdot 0.95^2 = 0.0451 = 4.51\%$$

10.3.6.4 Poisson distribution

Model : Consider a period of time from 0 to t in which successes occur with an average success rate λ per unit of time. The discrete random variable X is the number of successes in the given period of time. It can take the following values :

random variable	$X = 0, 1, 2, \ldots$

The average success rate λ and the time t are the parameters of the distribution. They are combined into the average number ν of successes during the time t.

time	$t > 0$
average success rate per unit of time	$\lambda > 0$
average number of successes	$\nu = \lambda t$

Probability function : Let the period of time from 0 to t be divided into n intervals $\Delta t = t/n$ such that in each interval Δt either a success or a failure occurs. In each interval, the success probability is $p = \nu/n$ and the failure probability is $q = 1 - \nu/n$. The probability that x successes occur in the n intervals is given by the binomial distribution.

$$p_X(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!\,(n-x)!}\ \frac{\nu^x}{n^x}\ \frac{(1-\nu/n)^n}{(1-\nu/n)^x}$$

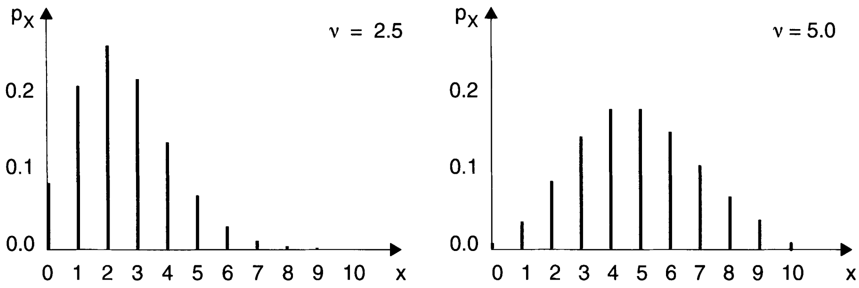
In the limit $n \to \infty$, the term $n! / (n^x (n-x)!)$ tends to 1, the term $(1-\nu/n)^x$ tends to 1 and the term $(1-\nu/n)^n$ tends to $e^{-\nu}$. This yields the probability function according to Poisson :

$$p_X(x) = \frac{\nu^x}{x!} e^{-\nu}$$

The probability function $p_X(x)$ is recursively calculated according to the following rule :

$$p_X(x+1) = \frac{\nu}{x+1} p_X(x) \qquad p_X(0) = e^{-\nu}$$

The probability function has one or two maxima in the interval $[\nu - 1, \nu]$. It is illustrated graphically for $\nu = 2.5$ and $\nu = 5.0$.



Moments : The k-th moment of the Poisson distribution is defined as follows :

$$E(X^k) = \sum_{x=0}^{\infty} x^k \frac{\nu^x}{x!} e^{-\nu}$$

Differentiating the k-th moment with respect to the parameter ν yields a recursive equation for the moment of order (k + 1) :

$$E'(X^k) = \frac{d}{d\nu} E(X^k) = \frac{d}{d\nu} \sum_{x=0}^{\infty} x^k \frac{\nu^x}{x!} e^{-\nu} = \sum_{x=0}^{\infty} x^k \frac{\nu^x}{x!} e^{-\nu} \left[\frac{x}{\nu} - 1\right]$$

$$E'(X^k) = \frac{1}{\nu} E(X^{k+1}) - E(X^k)$$

$$E(X^{k+1}) = \nu\,(E(X^k) + E'(X^k))$$

The moments of higher order for $k > 0$ are calculated starting from the zeroth-order moment $E(X^0) = 1$ with $E'(X^0) = 0$. The first and second moments are :

$$E(X) = \nu\,(E(X^0) + E'(X^0)) = \nu$$

$$E(X^2) = \nu\,(E(X\ ) + E'(X\ )) = \nu(\nu + 1)$$

The mean and the variance are calculated from the first and second moments as follows :

$$\mu_X = E(X) = \nu$$

$$\sigma_X^2 = E(X^2) - \mu_X^2 = \nu$$

Property : A discrete random variable X which has a Poisson distribution with the parameter ν is designated by X(ν). The sum of random variables which have Poisson distributions also has a Poisson distribution.

$$X(\nu) = \sum_{j=1}^{n} X_j(\nu_j) \qquad \nu = \sum_{j=1}^{n} \nu_j$$

This property is proved for n = 2 using the convolution of the two probability functions X_1 and X_2 according to Section 10.3.5. The probability that X takes the value x is equal to the probability that X_2 takes a value between 0 and x while X_1 takes the value $x - x_2$.

$$p_X(x) = \sum_{x_2=0}^{x} p_{X_1}(x - x_2)\, p_{X_2}(x_2)$$

$$p_X(x) = \sum_{x_2=0}^{x} \frac{\nu_1^{x-x_2} \quad \nu_2^{x_2}}{(x-x_2)!\; x_2!}\, e^{-(\nu_1+\nu_2)}$$

$$p_X(x) = \frac{e^{-(\nu_1+\nu_2)}}{x!} \sum_{x_2=0}^{x} \binom{x}{x_2} \nu_1^{x-x_2}\; \nu_2^{x_2}$$

The binomial theorem now yields :

$$p_X(x) = \frac{e^{-(\nu_1+\nu_2)}}{x!} (\nu_1 + \nu_2)^x$$

$$p_X(x) = \frac{\nu^x}{x!} e^{-\nu} \qquad \nu = \nu_1 + \nu_2$$

The property for n > 2 is proved by induction using the property for n = 2.

Example : On average 5 vehicles arrive on a given lane during the red phase of a traffic light. The probability that more than 8 vehicles arrive at the traffic light is calculated as follows :

$$\nu = 5.0$$

$$P(X>8) = 1 - e^{-5.0} \sum_{x=0}^{8} \frac{5.0^x}{x!}$$

$$P(X>8) = 1 - 0.932 = 0.068 = 6.8\%$$

The property that the sum of two random variables with Poisson distributions is again a random variable with a Poisson distribution is easily interpreted for the example of the traffic light. The number of vehicles which arrive at the traffic light during the red phase on two different lanes with ν_1 and ν_2 vehicles on average is equal to the number of vehicles which arrive at the traffic light during the red phase on one lane with $\nu_1 + \nu_2$ vehicles on average.

10.3.7 CONTINUOUS DISTRIBUTIONS

Introduction : Various problems in the probability theory of continuous random variables lead to different continuous distributions. Some of the continuous distributions may be derived from the discrete distributions by limit considerations. The important continuous distributions and their parameters are treated in the following.

10.3.7.1 Gamma distribution

Model : Consider a period of time from 0 to x in which successes occur with an average success rate λ per unit of time. The continuous random variable X is the time duration up to the m-th success. It can take the following values :

random variable	$X > 0.0$

The number m of successes and the average success rate λ are the parameters of the distribution :

number of successes	$m > 0$
average success rate per unit of time	$\lambda > 0$

Density function : Let the period of time from 0 to x be divided into n intervals $\Delta x = x/n$ such that in each interval either a success or a failure occurs. In each interval, the success probability is $p = \lambda x/n$, and the failure probability is therefore $q = 1 - \lambda x/n$. The probability that the m-th success occurs in the n-th interval is given by the Pascal distribution in Section 10.3.6.3.

$$p_X(x) = \binom{n-1}{m-1} p^m q^{n-m} = \frac{(n-1)!}{(m-1)!\,(n-m)!} \; \frac{\lambda \Delta x (\lambda x)^{m-1}}{n^{m-1}} \; \frac{(1-\lambda x/n)^n}{(1-\lambda x/n)^m}$$

The density function $f_X(x)$ is obtained from the quotient $p_X(x)/\Delta x$ in the limit $\Delta x \rightarrow 0$ and thus $n \rightarrow \infty$. In this limit the term $(n-1)!/(n^{m-1}(n-m)!)$ tends to 1, the term $(1-\lambda x/n)^m$ tends to 1 and the term $(1-\lambda x/n)^n$ tends to $e^{-\lambda x}$.

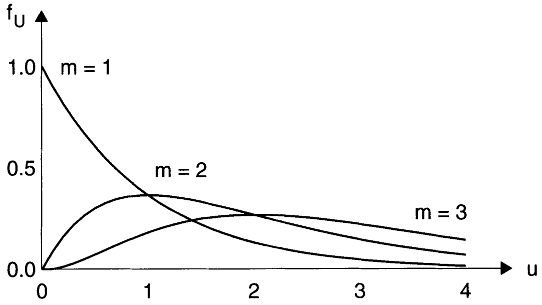
$$f_X(x) = \lim_{\Delta x \rightarrow 0} \frac{p_X(x)}{\Delta x} = \lambda \, \frac{(\lambda x)^{m-1}}{(m-1)!} \, e^{-\lambda x}$$

By a linear transformation from the random variable X to the standardized random variable U, the density function $f_X(x)$ is transformed to the standardized density function $f_U(u)$:

$$f_X(x) = \lambda \, f_U(u) \qquad u = \lambda x \qquad x = u/\lambda$$

$$f_U(u) = \frac{u^{m-1}}{(m-1)!} \, e^{-u}$$

For $m = 1$, the standardized density function $f_U(u)$ is a decaying exponential distribution. For $m > 1$, it has a maximum at $u = m - 1$. For $m = 2$, it has a point of inflection at $u = m$, and for $m > 2$ it has two points of inflection at $u = m - 1 \pm \sqrt{m-1}$. The standardized density function is illustrated graphically for $m = 1,2,3$.



Distribution function : The distribution functions $F_X(x)$ and $F_U(u)$ are obtained by integrating the corresponding density functions $f_X(x)$ and $f_U(u)$. The linear transformation yields :

$$F_X(x) = F_U(u) \qquad u = \lambda x \qquad x = u/\lambda$$

$$F_U(u) = \frac{1}{(m-1)!} \int_0^u s^{m-1} e^{-s} ds$$

Integration by parts yields the following expression for the standardized distribution function :

$$F_U(u) = \int_0^u e^{-s} ds = \left[-e^{-s}\right]_0^u = 1 - e^{-u} \qquad m = 1$$

$$F_U(u) = \int_0^u \frac{s^{m-1}}{(m-1)!} e^{-s} ds = -\left[\frac{s^{m-1}}{(m-1)!} e^{-s}\right]_0^u + \int_0^u \frac{s^{m-2}}{(m-1)!} (m-1)\, e^{-s} ds$$

$$= \int_0^u \frac{s^{m-2}}{(m-2)!} e^{-s} ds - \frac{u^{m-1}}{(m-1)!} e^{-u} \qquad m > 1$$

Applying the integration by parts recursively for $m > 1$ yields :

$$F_U(u) = 1 - \sum_{n=0}^{m-1} \frac{u^n}{n!} e^{-u}$$

Moments : The k-th moment of the standardized variable U is defined as follows :

$$E(U^k) = \int_0^\infty u^k \frac{u^{m-1}}{(m-1)!} e^{-u} \, du$$

Integrating the k-th moment by parts yields a recursive equation :

$$E(U^k) = \int_0^\infty \frac{u^{k+m-1}}{(m-1)!} e^{-u} \, du$$

$$= \left[-\frac{u^{k+m-1}}{(m-1)!} e^{-u} \right]_0^\infty + \int_0^\infty \frac{u^{k+m-2}}{(m-1)!} (k+m-1) \; e^{-u} \, du$$

$$E(U^k) = (k+m-1) \; E(U^{k-1})$$

The moments of higher order for $k > 0$ may be obtained starting from the moment $E(U^0) = 1$. The first and second moments are :

$$E(U) = m \; E(U^0) = m$$

$$E(U^2) = (m+1) \; E(U^1) = m(m+1)$$

The mean and the variance for the standardized random variable U are obtained from the first and second moments as follows :

$$\mu_U = E(U) = m$$

$$\sigma_U^2 = E(U^2) - \mu_U^2 = m$$

The mean and the variance for the random variable X are determined according to the rules for linear transformations :

$$x = u / \lambda$$

$$\mu_X = \mu_U / \lambda = m / \lambda$$

$$\sigma_X^2 = \sigma_U^2 / \lambda^2 = m / \lambda^2$$

Properties : The gamma distribution is derived from the Pascal distribution. The properties of the Pascal distribution may therefore be transferred to the gamma distribution. A continuous random variable X which has a gamma distribution with the parameters m and λ is designated by X(m, λ). The exponential distribution is a special case of the gamma distribution with $m = 1$. In analogy with the Pascal distribution, one obtains

$$X(m,\lambda) = \sum_{j=1}^{m} X_j \, (1, \lambda)$$

More generally, the sum of random variables which have gamma distributions with identical success rate λ also has a gamma distribution.

$$X(m,\lambda) = \sum_{j=1}^{n} X_j\,(m_j, \lambda) \qquad m = \sum_{j=1}^{n} m_j$$

Generalization : The gamma distribution is derived for integer values of m. It may, however, also be applied for real values of $m > 0$. The standardized density and distribution functions are calculated using the gamma functions, which are tabulated in mathematical handbooks.

$$f_U(u) = \frac{u^{m-1}}{\Gamma(m)}\, e^{-u} \qquad \Gamma(m) = \int_0^{\infty} u^{m-1}\, e^{-u}\, du$$

$$F_U(u) = \frac{\Gamma(m,u)}{\Gamma(m)} \qquad \Gamma(m,u) = \int_0^{u} s^{m-1}\, e^{-s}\, ds$$

Example : A device has an average lifetime of 5 years. The average failure rate of devices per year is $\lambda = 1/5$. The probabilities that the lifetime of a device is less than two years or greater than ten years, respectively, are calculated as follows using the exponential distribution :

$$\lambda = 0.2 \qquad m = 1 \qquad u = 0.2x \qquad F_U = 1 - e^{-u}$$

$$P(X \le 2) = F_X(2) = F_U(0.4) = 1 - e^{-0.4} = 0.330 = 33.0\%$$

$$P(X > 10) = 1 - F_X(10) = 1 - F_U(2.0) = 1 - 1 + e^{-2.0} = 0.135 = 13.5\%$$

10.3.7.2 Normal distribution

Model : Consider a continuous random variable X which is the sum of a large number of elementary random variables X_j. Let the elementary random variables be independent with identical distributions.

$$\text{random variable} \qquad X := \sum_{j=1}^{n} X_j \qquad n \to \infty$$

The mean μ_X and the standard deviation σ_X of the random variable X are the parameters of the distribution.

mean	μ_X
standard deviation	$\sigma_X > 0.0$

Density function : According to Section 10.3.6.2, a binomially distributed random variable is the sum of random variables with Bernoulli distributions. The summation constitutes a similarity between the model of the binomial distribution and the model of the normal distribution. The normal distribution may be derived from the binomial distribution. This derivation is shown in the following. The probability function $p_X(x)$ of the binomial distribution for integer variables $0 \leq x \leq n$ is recursively calculated as follows :

$$p_X(x+1) = \frac{n-x}{x+1}\frac{p}{q}\, p_X(x) \qquad\qquad p+q = 1$$

The difference of two consecutive function values is :

$$p_X(x+1) - p_X(x) = \frac{np-x-q}{(x+1)q}\, p_X(x)$$

The parameters n and p,q are replaced by the mean μ_X and the standard deviation σ_X of the binomial distribution :

$$\mu_X = np \qquad \sigma_X = \sqrt{npq}$$

$$p_X(x+1) - p_X(x) = \frac{\mu_X - x - \sigma_X^2/\mu_X}{(x+1)\sigma_X^2/\mu_X}\, p_X(x)$$

Using the following linear transformation, the integers $x \geq 0$ are mapped to real numbers u. The function $p_X(x)$ is transformed into $p_U(u)$:

$$u = (x-\mu_X)/\sigma_X \qquad \Delta u = 1/\sigma_X \qquad p_U(u) = p_X(x)$$

$$p_U(u+\Delta u) - p_U(u) + \Delta u\,\frac{u+\sigma_X/\mu_X}{1+1/\mu_X+u\sigma_X/\mu_X}\, p_U(u) = 0$$

Upon introduction of the density function $f_U(u) = p_U(u)/\Delta u$ this becomes :

$$\frac{f_U(u+\Delta u) - f_U(u)}{\Delta u} + \frac{u+\sigma_X/\mu_X}{1+1/\mu_X+u\sigma_X/\mu_X}\, f_U(u) = 0$$

In the limit $n \to \infty$, the terms Δu, $1/\mu_X$ and σ_X/μ_X tend to 0. This leads to a linear homogeneous differential equation for the density function with the following solution :

$$f'_U(u) + u f_U(u) = 0$$
$$f_U(u) = C\, e^{-u^2/2}$$

This limit consideration yields the standardized density function $f_U(u)$ of the normal distribution. The constant of proportionality C is $1/\sqrt{2\pi}$, so that the integral of $f_U(u)$ over the range $-\infty \leq u \leq \infty$ is equal to 1.

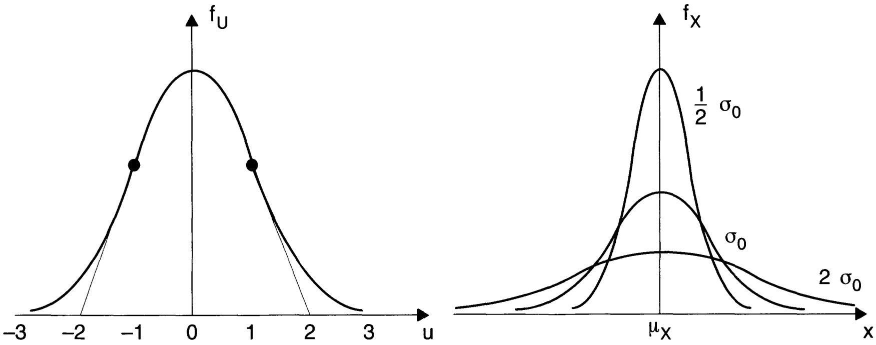
$$f_U(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

Using the following linear back transformation, the real numbers u are mapped to the real numbers x. The standardized density function $f_U(u)$ is transformed into the density function $f_X(x)$ of the normal distribution.

$$x = \mu_X + u\,\sigma_X \qquad\qquad f_X(x) = \frac{1}{\sigma_X} f_U(u)$$

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}$$

The graph of the standardized density function $f_U(u)$ is bell-shaped and symmetric. It has a maximum at $u = 0$ and two points of inflection at $u = \pm 1$. The graph of the density function $f_X(x)$ is a bell shape around the mean μ_X which becomes increasingly flat with increasing standard deviation σ_X. For $\sigma_X \to 0$, the density function $f_X(x)$ tends to the delta function $\delta(x - \mu_X)$. The density function $f_U(u)$ and the density function $f_X(x)$ with various standard deviations σ_X are shown.



Distribution function : The distribution functions $F_X(x)$ and $F_U(u)$ are obtained by integrating the corresponding density functions $f_X(x)$ and $f_U(u)$. The density functions cannot be integrated analytically. The integration must therefore be performed numerically. The results of the numerical integration of the standardized distribution functions $F_U(u)$ are tabulated in mathematical handbooks. The linear transformation between x and u yields :

$$F_X(x) = F_U(u) \qquad u = (x - \mu_X) / \sigma_X \qquad x = \mu_X + u\sigma_X$$

The distribution function $F_U(u)$ is usually tabulated only for positive values $u \geq 0$. For negative values $u < 0$, the distribution function $F_U(u)$ is calculated using the following formula, which follows from the symmetry of the density function $f_U(u)$.

$$F_U(-u) = 1 - F_U(u)$$

Moments : The k-th moment of the standardized normal distribution is defined as follows :

$$E(U^k) = C \int_{-\infty}^{\infty} u^k \, e^{-u^2/2} \, du \qquad C = \frac{1}{\sqrt{2\pi}}$$

Since the density function $f_U(u)$ is symmetric, all moments for odd k are zero. For even k, integrating the k-th moment by parts yields a recursive equation :

$$E(U^k) = C \int_{-\infty}^{\infty} u^k \, e^{-u^2/2} \, du$$

$$= C \left[\frac{u^{k+1}}{k+1} e^{-u^2/2} \right]_{-\infty}^{\infty} + C \int_{-\infty}^{\infty} \frac{u^{k+2}}{k+1} e^{-u^2/2} \, du$$

$$E(U^k) = \frac{1}{k+1} E(U^{k+2})$$

$$E(U^{k+2}) = (k+1) \, E(U^k)$$

The moments of higher order for even $k > 0$ are calculated starting from the moment $E(U^0) = 1$. The first and second moments are :

$$E(U) = 0$$

$$E(U^2) = E(U^0) = 1$$

The mean and the variance of the standardized random variable U are calculated from the first and second moments as follows :

$$\mu_U = E(U) = 0$$

$$\sigma_U^2 = E(U^2) - \mu_U^2 = 1$$

The mean and the variance of the random variable X are determined according to the rules for linear transformations; they correspond to the parameters of the normal distribution.

$$x = \mu_X + u \, \sigma_X$$

$$\mu_X = \mu_X + \mu_U \, \sigma_X = \mu_X$$

$$\sigma_X^2 = \sigma_U^2 \cdot \sigma_X^2 = \sigma_X^2$$

Properties : A continuous random variable X which has a normal distribution with the parameters μ_X and σ_X is designated by $X(\mu_X,\sigma_X)$. The sum of normally distributed random variables is also normally distributed.

$$X(\mu_X,\sigma_X) = \sum_{j=1}^{n} X_j \, (\mu_{Xj}, \sigma_{Xj})$$

$$\mu_X = \sum_{j=1}^{n} \mu_{Xj} \qquad \sigma_X^2 = \sum_{j=1}^{n} \sigma_{Xj}^2$$

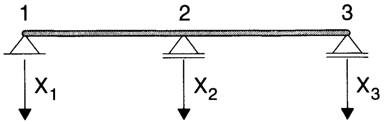
This property is proved for n = 2 using the convolution of the two density functions according to Section 10.3.5. The result for n > 2 then follows by induction. More generally, any linear combination of normally distributed random variables is also normally distributed.

$$X(\mu_X, \sigma_X) = \sum_{j=1}^{n} c_j\, X_j\, (\mu_{Xj}, \sigma_{Xj})$$

$$\mu_X = \sum_{j=1}^{n} c_j\, m_{Xj} \qquad\qquad \sigma_X^2 = \sum_{j=1}^{n} c_j^2\, \sigma_{Xj}^2$$

Central limit theorem : A binomially distributed random variable is the sum of n independent random variables with identical Bernoulli distributions. The normal distribution is derived from the binomial distribution for the limit $n \to \infty$. This relationship between the binomial distribution and the normal distribution is a special case of the central limit theorem. The central limit theorem states that the sum of independent random variables with different distributions tends to a normal distribution if the number of random variables tends to ∞. The contributions of the individual random variables to the sum are assumed to be uniformly small.

Example : Consider a bridge beam with two spans of equal length. Let the vertical displacements of the supports be normally distributed with the same mean $\mu = 0$ and the same standard deviation $\sigma = 0.5$ cm.



X_i displacement of support i

The probability that the displacement X_i at a support i is less than $x_i = 1.0$ cm is calculated as follows :

$$u = (x_i - \mu)/\sigma = (1.0 - 0.0) / 0.5 = 2.0$$

$$P(X_i < x_i) = F(x_i) = F_U(2.0) = 0.977$$

The bending moment B at the central support 2 depends linearly on the vertical displacements of the supports.

$$B = M_0(-X_1 + 2X_2 - X_3)$$

Since the vertical displacements X_i are normally distributed, the bending moment B is also normally distributed. The mean μ_B and the standard deviation σ_B are calculated as follows :

$$\mu_B = M_0(-\mu + 2\mu - \mu) = 0$$

$$\sigma_B = M_0(+\sigma^2 + 4\sigma^2 + \sigma^2)^{1/2} = M_0\, \sigma\, \sqrt{6} = 0.5\, M_0\, \sqrt{6}$$

The probability that the bending moment is less than $b = 2M_0$ is calculated as follows :

$$u = (b - \mu_B)/\sigma_B = (2M_0 - 0)/(0.5\,M_0\,\sqrt{6}) = 1.63$$

$$P(B < b) = F_B(b) = F_U(1.63) = 0.948 = 94.8\%$$

10.3.7.3 Logarithmic normal distribution

Model : Consider a continuous random variable Y which is the product of many elementary random variables $Y_j > 0$. Let the elementary random variables be independent with identical distributions.

random variable $\quad Y := \prod_{j=1}^{n} Y_j > 0 \quad n \rightarrow \infty$

The mean μ_Y and the standard deviation σ_Y are the parameters of the distribution.

mean $\quad \mu_Y > 0$

standard deviation $\quad \sigma_Y > 0$

Logarithmic transformation : The random variable X is introduced as the natural logarithm of the random variable Y. The product form of the elementary random variable Y_j is thereby transformed into a sum.

$$X = \ln Y = \ln \prod_{j=1}^{n} Y_j = \sum_{j=1}^{n} \ln Y_j$$

The sum form follows the model of the normal distribution. The random variable X is therefore normally distributed. The following transformation equations describe the logarithmic transformation :

$$x = \ln y \qquad \frac{dx}{dy} = \frac{1}{y}$$

$$y = e^x \qquad \frac{dy}{dx} = e^x$$

Distribution : The random variable X has a normal distribution with the mean μ_X and the standard deviation σ_X.

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2}$$

The logarithmic normal distribution of the random variable Y is determined from the normal distribution of the random variable X using the rules in Section 10.3.4.

$$f_Y(y) = f_X(x)\frac{dx}{dy} = \frac{1}{y} f_X(\ln y)$$

$$F_Y(y) = F_X(x) = F_X(\ln y)$$

Moments : The k-th moment of the random variable Y is obtained from the mean μ_X and the standard deviation σ_X of the random variable X according to the rules in Section 10.3.4.

$$E(Y^k) = \int_0^\infty y^k f_Y(y)\,dy = \int_{-\infty}^\infty e^{kx} f_X(x)\,dx$$

The integral expression yields the following formula for the k-th moment :

$$\begin{aligned} E(Y^k) &= \frac{1}{\sigma_X\sqrt{2\pi}} \int_{-\infty}^\infty e^{kx - \frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2} dx \\ &= \frac{1}{\sigma_X\sqrt{2\pi}} \int_{-\infty}^\infty e^{-\frac{1}{2}\left(\frac{x-\mu_X-k\sigma_X^2}{\sigma_X}\right)^2 + k\mu_X + k^2\sigma_X^2/2} dx \\ &= e^{k\mu_X + k^2\sigma_X^2/2} \end{aligned}$$

The mean and the variance are given by :

$$\mu_Y = E(Y) = e^{\mu_X + \sigma_X^2/2}$$

$$\sigma_Y^2 = E(Y^2) - \mu_Y^2 = \mu_Y^2\,(e^{\sigma_X^2} - 1)$$

The mean μ_Y and the standard deviation σ_Y are given as parameters of the logarithmic normal distribution. The parameters μ_X and σ_X required for the normal distribution of X are determined from μ_Y and σ_Y :

$$v_Y = \sigma_Y/\mu_Y$$

$$\mu_X = \ln\left(\mu_Y / \sqrt{1 + v_Y^2}\right)$$

$$\sigma_X^2 = \ln(1 + v_Y^2)$$

Properties : Since the logarithmic normal distribution is derived from the normal distribution, the properties of random variables with normal distribution may be transferred to random variables with logarithmic normal distribution. A random variable Y which has a logarithmic normal distribution with the parameters μ_Y and σ_Y is designated by $Y(\mu_Y,\sigma_Y)$. A product of random variables with logarithmic normal distributions also has a logarithmic normal distribution.

$$Y(\mu_Y, \sigma_Y) \quad = \quad \prod_{j=1}^{n} Y_j(\mu_{Yj}, \sigma_{Yj})$$

$$\mu_Y \quad = \quad \prod_{j=1}^{n} \mu_{Yj} \qquad\qquad \sigma_Y^2 \quad = \quad \prod_{j=1}^{n} (\sigma_{Yj}^2 + \mu_{Yj}^2) - \mu_Y^2$$

More generally, a product of powers of several random variables with logarithmic normal distributions also has a logarithmic normal distribution.

Example : A moment M is the product of the force F and the lever arm H. Let the force F and the lever arm H have logarithmic normal distributions with the means μ_F and μ_H. Let the variation coefficient for the force and the lever arm be v = 0.20. The probability that the moment M is less than 2.0 $\mu_F \mu_H$ is to be calculated.

Since the force F and the lever arm H have logarithmic normal distributions, the moment $M = P * H$ also has a logarithmic normal distribution. The mean μ_M, the variance σ_M^2 and the variation coefficient v_M are calculated as follows :

$$\mu_M \quad = \quad \mu_F \cdot \mu_H$$

$$\sigma_M^2 \quad = \quad (\sigma_F^2 + \mu_F^2)\,(\sigma_H^2 + \mu_H^2) - \mu_M^2$$

$$\sigma_M^2 \quad = \quad \mu_F^2\,(v^2+1)\,\mu_H^2\,(v^2+1) - \mu_M^2$$

$$\sigma_M^2 \quad = \quad \mu_M^2\,((v^2+1)^2 - 1) \quad = \quad 0.0816\,\mu_M^2$$

$$v_M \quad = \quad \sigma_M / \mu_M \quad = \quad 0.286$$

The mean μ and the standard deviation σ of the corresponding normal distribution are calculated as follows :

$$\mu \quad = \quad \ln\left(\mu_M / \sqrt{1+v_M^2}\right) \quad = \quad \ln \mu_M \; - \; \ln 1.04$$

$$\sigma^2 \quad = \quad \ln\,(1+v_M^2) \qquad = \quad \ln 1.0816 \quad = \quad 0.07844$$

$$\sigma \quad = \quad 0.280$$

The probability that the moment is less than 2.0 μ_M is determined using the standardized normal distribution.

$$P(M < 2.0\mu_M) \; = \; F_M(2.0\mu_M) \quad = \quad F_U(u)$$

$$u \; = \; (\ln\,(2.0\;\mu_M) \; - \; \mu)/\sigma \quad = \quad (\ln 2.0 + \ln \mu_M - \ln \mu_M + \ln 1.04)\;/\;\sigma$$

$$u \; = \; (\ln 2.0 \; + \; \ln 1.04)\,/\,0.280 \; = \; 2.62$$

$$F_U(u) \; = \; 0.9956 \; = \; 99.56\%$$

10.3.7.4 Maximum distributions

Model : Consider a continuous random variable X which is the maximum of a large number of elementary random variables X_j. Let the elementary random variables be independent with identical distributions.

random variable $X := \max(X_1, \dots, X_n) \qquad n \to \infty$

Distribution : The elementary random variables X_j have the same distribution function F(x). According to Section 10.3.4, the distribution function $F_X(x)$ of the maximum X is the product of all distribution functions of the elementary random variables.

$$F_X(x) = (F(x))^n$$

Since the maximum of all n random variables X_j is considered, their distribution function F(x) is approximated by the following function for large values of x :

$$F(x) = 1 - c \cdot g(x) \qquad 0 \le c \cdot g(x) \le 1$$

The function g(x) tends to 0 with increasing x. The constant c is eliminated by introducing as a parameter the value u which is exceeded with probability 1/n. This leads to :

$$F(x) = 1 - \frac{1}{n} \frac{g(x)}{g(u)}$$

By substituting F(x) into $F_X(x)$ and taking the limit $n \to \infty$, the following distribution function is obtained :

$$F_X(x) = \lim_{n \to \infty} (F(x))^n = \lim_{n \to \infty} \left(1 - \frac{1}{n} \frac{g(x)}{g(u)}\right)^n = e^{-g(x)/g(u)}$$

The density function $f_X(x)$ is calculated by differentiating the distribution function $F_X(x)$:

$$f_X(x) = -\frac{g'(x)}{g(u)} e^{-g(x)/g(u)} = -\frac{g'(x)}{g(u)} F_X(x)$$

Different forms of the function g(x) lead to different types of maximum distributions. The important types and their standardization are treated in the following.

Distribution type I : The function g(x) decays exponentially with the rate α for large values of x. This leads to the following distribution with the parameters u and α :

$$g(x) = e^{-\alpha x} \qquad \alpha > 0$$

$$F_X(x) = e^{-e^{-\alpha(x-u)}} \qquad f_X(x) = \alpha\, e^{-\alpha(x-u)}\, F_X(x)$$

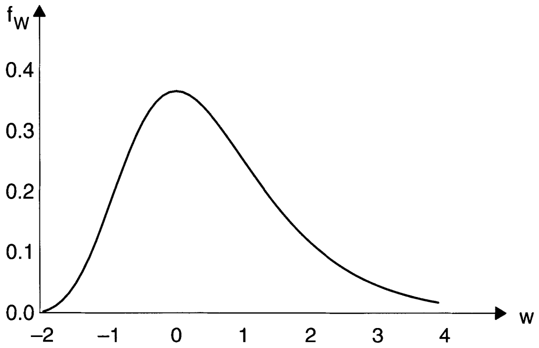
Using the following linear transformation, the distribution is standardized and reduced to the double exponential distribution :

$$w = \alpha(x-u) \qquad -\infty \le w \le \infty$$

$$F_W(w) = e^{-e^{-w}} \qquad f_W(w) = e^{-w}\, F_W(w)$$

$$F_X(x) = F_W(w) \qquad f_X(x) = \alpha\, f_W(w)$$

The standardized density function $f_W(w)$ has a maximum at $w = 0$ and two points of inflection at $w = \pm 0.9624$. Its graph is shown below.



Distribution type II : The function g(x) decays hyperbolically with the power α for large values $x > 0$. This leads to the following distribution with the parameters u and α :

$$g(x) = x^{-\alpha} \qquad x > 0,\ \alpha > 0$$

$$F_X(x) = e^{-(x/u)^{-\alpha}} \qquad f_X(x) = \frac{\alpha}{u}(x/u)^{-(\alpha+1)}\, F_X(x)$$

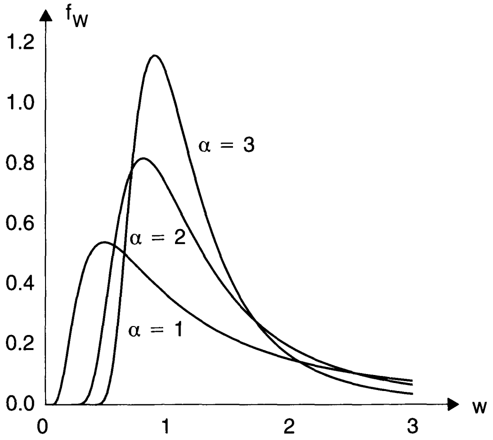
The distribution is standardized using the following linear transformation :

$$w = x/u \qquad w > 0$$

$$F_W(w) = e^{-w^{-\alpha}} \qquad f_W(w) = \alpha\, w^{-(\alpha+1)}\, F_W(w)$$

$$F_X(x) = F_W(w) \qquad f_X(x) = f_W(w)/u$$

The standardized density function $f_W(w)$ depends on the parameter α. It has a maximum at $w = (\alpha/(1+\alpha))^{1/\alpha}$. Its graph is shown for $\alpha = 1, 2, 3$.



Distribution type III : The function g(x) decays with the power α for large values $x \leq x_0$. The parameter x_0 is an upper bound for the values x. This leads to the following distribution with the parameters x_0, u and α :

$$g(x) = (x_0 - x)^{\alpha} \qquad x \leq x_0\,,\ \alpha > 0$$

$$F_X(x) = e^{-((x_0 - x)/(x_0 - u))^{\alpha}} \qquad f_X(x) = \frac{\alpha}{x_0 - u}\left(\frac{x_0 - x}{x_0 - u}\right)^{\alpha - 1} F_X(x)$$

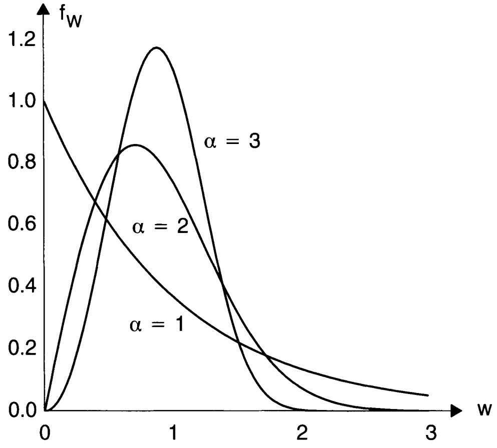
Using the following linear transformation, the distribution is standardized and reduced to the Weibull distribution :

$$w = (x_0 - x)/(x_0 - u) \qquad w \geq 0$$

$$F_W(w) = 1 - e^{-w^{\alpha}} \qquad f_W(w) = \alpha\, w^{\alpha - 1}\, e^{-w^{\alpha}}$$

$$F_X(x) = 1 - F_W(w) \qquad f_X(x) = \frac{1}{x_0 - u} f_W(w)$$

The Weibull distribution depends on the parameter α. For $\alpha = 1$, it coincides with the exponential distribution. For $\alpha > 1$, the standardized density function $f_W(w)$ has a maximum at $w = ((\alpha - 1)/\alpha)^{1/\alpha}$. Its graph is shown for $\alpha = 1, 2, 3$.



Moments : The k-th moment of a maximum distribution is calculated as follows, using the definition in Section 10.3.4 :

$$E(X^k) = \int_{-\infty}^{\infty} x^k \, f_X(x) \, dx = - \int_{-\infty}^{\infty} x^k \, \frac{g'(x)}{g\,(u)} \, e^{-g(x)\,/\,g(u)} \, dx$$

The integral expression is simplified by introducing the variable $t = g(x) / g(u)$ as the integration variable. Since $g(x)$ takes only positive values and tends to 0 with increasing x, this yields :

$$E(X^k) = \int_{0}^{\infty} x^k \, e^{-t} \, dt \qquad\qquad t = g(x) \,/\, g(u)$$

The moments of the various maximum distributions are first calculated for the standardized random variable W and then obtained for the random variable X according to the rules for linear transformations. The moments for W involve the gamma function, which is tabulated in mathematical handbooks and is defined as follows :

$$\Gamma(s) \;:=\; \int_{0}^{\infty} t^{s-1} \, e^{-t} \, dt$$

The k-th derivative of the gamma function with respect to s is obtained from this definition :

$$\Gamma^{(k)}(s) = \frac{d^k \, \Gamma(s)}{ds^k} = \int_{0}^{\infty} \frac{d^k}{ds^k} \, (t^{s-1}) \, e^{-t} \, dt = \int_{0}^{\infty} (\ln t)^k \, t^{s-1} \, e^{-t} \, dt$$

The first and second moments as well as the resulting means, variances and variation coefficients for the various maximum distributions are treated in the following.

Moments type I : The k-th moment for the standardized distribution of type I is given by the k-th derivative of the gamma function at s = 1 :

$$t = e^{-w} \qquad\qquad w = -\ln t$$

$$E(W^k) = \int_0^\infty w^k\, e^{-t}\, dt = \int_0^\infty (-\ln t)^k\, e^{-t}\, dt = (-1)^k\, \Gamma^{(k)}(1)$$

Using numerical values for the derivatives of the gamma function, the mean μ_W and the variance σ_W are obtained as :

$$\begin{aligned} \mu_W &= E(W) &&= -\Gamma'(1) = \gamma &&= 0.5772 \qquad \text{Euler's constant} \\ \sigma_W^2 &= E(W^2) - \mu_W^2 &&= \Gamma''(1) - \gamma^2 &&= \pi^2/6 = 1.645 \end{aligned}$$

The mean and the variance for the random variable X are determined using the rules for linear transformations :

$$\begin{aligned} x &= u + w/\alpha \\ \mu_X &= u + \mu_W/\alpha &&= u + 0.5772/\alpha \\ \sigma_X^2 &= \sigma_W^2 / \alpha^2 &&= 1.645/\alpha^2 \end{aligned}$$

Moments type II : The k-th moment for the standardized distribution of type II is given by the gamma function at a value depending on k and α. It exists only for k < α.

$$t = w^{-\alpha} \qquad\qquad w = t^{-1/\alpha}$$

$$E(W^k) = \int_0^\infty w^k\, e^{-t}\, dt = \int_0^\infty t^{-k/\alpha}\, e^{-t}\, dt = \Gamma(1-k/\alpha) \qquad\qquad k < \alpha$$

The mean μ_W and the variance σ_W^2 are given by :

$$\begin{aligned} \mu_W &= E(W) &&= \Gamma(1-1/\alpha) \\ \sigma_W^2 &= E(W^2) - \mu_W^2 &&= \Gamma(1-2/\alpha) - \Gamma^2(1-1/\alpha) \end{aligned}$$

The mean and the variance for the random variable X are determined according to the rules for linear transformations :

$$\begin{aligned} x &= u\, w \\ \mu_X &= u\, \mu_W &&= u\, \Gamma(1-1/\alpha) \\ \sigma_X^2 &= u^2 \sigma_W^2 &&= u^2\, (\Gamma(1-2/\alpha) - \Gamma^2(1-1/\alpha)) \end{aligned}$$

The variation coefficient v_X depends only on the parameter α :

$$v_X = \frac{\sigma_X}{\mu_X} = \sqrt{\frac{\Gamma\ (1-2/\alpha)}{\Gamma^2(1-1/\alpha)} - 1}$$

Moments type III : The k-th moment for the standardized distribution of type III is given by the gamma function at a value depending on k and α.

$$t = w^{\alpha} \qquad\qquad w = t^{1/\alpha}$$

$$E(W^k) = \int_0^{\infty} w^k \, e^{-t} \, dt = \int_0^{\infty} t^{k/\alpha} \, e^{-t} \, dt = \Gamma(1 + k/\alpha)$$

The mean μ_W and the variance σ_W^2 are :

$$\begin{aligned} \mu_W &= E(W) &&= \Gamma(1 + 1/\alpha) \\ \sigma_W^2 &= E(W^2) - \mu_W^2 &&= \Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha) \end{aligned}$$

The mean and the variance for the random variable X are determined according to the rules for linear transformations :

$$\begin{aligned} x &= x_0 - (x_0 - u) \, w \\ \mu_X &= x_0 - (x_0 - u) \, \mu_W &&= x_0 - (x_0 - u) \, \Gamma(1 + 1/\alpha) \\ \sigma_X^2 &= (x_0 - u)^2 \, \sigma_W^2 &&= (x_0 - u)^2 \, (\Gamma(1 + 2/\alpha) - \Gamma^2(1 + 1/\alpha)) \end{aligned}$$

The ratio $\sigma_X / (x_0 - \mu_X)$ depends only on the parameter α. It is called the variation coefficient and is designated by v_{X0}.

$$v_{X0} = \frac{\sigma_X}{x_0 - \mu_X} = \sqrt{\frac{\Gamma\,(1 + 2/\alpha)}{\Gamma^2(1 + 1/\alpha)} - 1}$$

Determination of the parameters : In practical calculations, the mean μ_X and the standard deviation σ_X are often given. The parameters α and u for the different maximum distributions are obtained by solving the equations which determine μ_X and σ_X for the parameters.

- For the maximum distribution of type I, α and u are determined as follows :

  $$\alpha = 1.283 \,/\, \sigma_X \qquad\qquad u = \mu_X - 0.5772 \,/\, \alpha$$

- For the maximum distribution of type II, the parameter α is determined as a function of the variation coefficient $v_X = \sigma_X / \mu_X$. The parameter α is specified for typical values of the variation coefficient. The parameter u is obtained from the formula for the mean :

v_X	0.40	0.30	0.20	0.10	0.05
α	4.18	5.18	7.30	13.6	26.4

$$u = \frac{\mu_X}{\Gamma(1 - 1/\alpha)}$$

- For the maximum distribution of type III, the parameter x_0 needs to be specified. The parameter α is determined as a function of the variation coefficient $v_{X0} = \sigma_X/(x_0 - \mu_X)$. The parameter α is specified for typical values of the variation coefficient. The parameter u is obtained from the formula for the mean.

v_{X0}	0.40	0.30	0.20	0.10	0.05
α	2.71	3.71	5.80	12.1	25.0

$$u = x_0 - \frac{x_0 - \mu_X}{\Gamma(1 + 1/\alpha)}$$

Example : In order to make realistic assumptions about the snow load on a building, the maximal snow levels at different locations are measured annually and evaluated statistically. While the mean μ_X is different at different locations, the variation coefficient may be assumed to be approximately constant. Let the variation coefficient v_X be 0.40. The probability for a value below the n-fold mean is to be calculated.

Under the assumption that the maximal snow levels have a maximum distribution of type I, the parameters u and α are determined as follows :

$$\alpha = 1.283/\sigma_X = 1.283/(v_X\,\mu_X) = 3.208/\mu_X$$

$$u = \mu_X - 0.5772/\alpha = 0.820\,\mu_X$$

The probability for a value below the n-fold mean μ_X is :

$$p(n) = P(X < n \cdot \mu_X) = F_X(n\,\mu_X)$$

$$p(n) = e^{-e^{-\alpha(n \cdot \mu_X - u)}} = e^{-e^{-3.208(n-0.82)}}$$

Under the assumption that the maximal snow levels have a maximum distribution of type II, the parameters u and α are determined as follows :

$$\alpha = 4.18 \quad \text{for} \quad v_X = 0.40$$

$$u = \mu_X/\Gamma(1-1/\alpha) = \mu_X/\Gamma(0.761) = \mu_X/1.211$$

The probability for a value below the n-fold mean μ_X is :

$$p(n) = P(X < n \cdot \mu_X) = F_X(n\,\mu_X)$$

$$p(n) = e^{-(n\mu_X/u)^{-\alpha}} = e^{-(1.211n)^{-4.18}}$$

The results for the maximum distributions of types I and II are compiled in the following table :

n	1.0	1.5	2.0	2.5
type I	57.0%	89.3%	97.8%	99.5%
type II	63.8%	92.1%	97.6%	99.0%

10.3.7.5 Minimum distributions

Model : Consider a continuous random variable Y which is the minimum of many elementary random variables Y_j . Let the elementary random variables be independent with identical distributions.

random variable $\quad Y := \min(Y_1, \ldots, Y_n) \quad n \to \infty$

If Y has a minimum, then X = –Y has a maximum. A minimum distribution for Y may therefore be reduced to a maximum distribution for X by linear transformation.

$$
\begin{aligned}
F_Y(y) &= 1 - F_X(x) = 1 - F_X(-y) \\
f_Y(y) &= f_X(x) \qquad = f_X(-y) \\
\mu_Y &= -\mu_X \\
\sigma_X^2 &= \sigma_Y^2 \\
v_Y &= v_X
\end{aligned}
$$

Example : The minimal water levels H in a harbor are measured annually and evaluated statistically. Let the average minimal water level be μ_H, and let the variation coefficient be $v_H = 0.20$. The probability that the minimal water level in a given year is less than 0.5 μ_H is to be calculated.

A minimum distribution of type III with the lower bound $h_0 = 0$ is assumed for the minimal water levels H. The probability is calculated using the maximum distribution of type III for X = –H.

$$x_0 = -h_0 = 0 \qquad \mu_X = -\mu_H \qquad v_X = v_H = 0.20 \qquad x = -0.5\ \mu_H$$

The parameters α and u of the maximum distribution of type III are calculated as follows using the formulas in the preceding section with $x_0 = 0$:

$$\alpha = 5.80 \qquad \text{for} \qquad v_X = 0.20$$

$$u = \mu_X / \Gamma(1 + 1/\alpha) = -\mu_H / \Gamma(1.172) = -1.080\ \mu_H$$

The probability that H is less than 0.5 μ_H is calculated as follows :

$$F_H(0.5\ \mu_H) = 1 - F_X(-0.5\ \mu_H) = 1 - e^{-((-0.5\ \mu_H)\ /\ (-1.080\ \mu_H))^{5.80}}$$

$$F_H(0.5\ \mu_H) = 1 - e^{-(0.5/1.08)^{5.80}} = 1 - 0.9886 = 1.14\%$$

10.4 RANDOM VECTORS

10.4.1 INTRODUCTION

Random vector : In many applications the result of an experiment is determined by counting or measuring several quantities. The hourly count of vehicles which turn left, drive straight on or turn right at a traffic light, the simultaneous measurement of annual precipitation at several locations or the simultaneous measurement of strain at different locations in a building are typical examples. The random results of such an experiment are described by a random vector with several random variables which can take different real values.

Probability distribution : The introduction of a random vector leads to a multidimensional probability distribution. Each random variable of the random vector has a one-dimensional probability distribution, which is a marginal distribution of the multidimensional probability distribution. If fixed values are given as conditions for some of the random variables in a random vector, a conditional probability distribution for the remaining random variables of the random vector is obtained. The fundamentals of probability distributions for random vectors are treated in Section 10.4.2.

Moments : The means of the individual random variables of the random vector are arranged in a vector of means. The variances of the individual random variables of the random vector are the diagonal elements of the covariance matrix. The non-diagonal elements of the covariance matrix are called covariances. If all covariances are zero, the random variables of the random vector are independent of each other. If some covariances are non-zero, there is a certain correlation between the random variables of the random vector. The fundamentals for the moments of multidimensional distributions are treated in Section 10.4.3.

Functional dependence : In many applications it is assumed that a random vector functionally depends on other random vectors. For example, the forces in the beams of a framework depend on the loads. If the vector of loads is a random vector, the vector of forces is also a random vector. Even for stochastically independent loads, the forces in the beams are correlated. The fundamentals for the functional dependence of random vectors are treated in Section 10.4.4.

Discrete and continuous distributions : As in the case of one-dimensional models, discrete and continuous multidimensional models are distinguished. In Sections 10.4.5 and 10.4.6, the multinomial distribution and the multinormal distribution are treated as generalizations of the binomial distribution and the normal distribution, respectively.

10.4.2 PROBABILITY DISTRIBUTIONS

Introduction : Each experiment is assigned a random vector whose values are real vectors. The random vector is represented by an uppercase boldface letter, a corresponding vector value is represented by a lowercase boldface letter. The rules for random variables may be transferred to random vectors. This leads to the definition of probability distributions for random vectors.

Random vector : An n-dimensional event space is associated with n random variables X_j, which are arranged in a random vector $\mathbf{X}$. The n real values x_j of the random variables are arranged in the vector $\mathbf{x}$. Every elementary event of the n-dimensional event space is assigned a unique vector value.

random vector $\mathbf{X}$
vector value $\mathbf{x} \in \mathbb{R}^n$

Probability : The events in an n-dimensional event space are described by corresponding ranges for the random vectors. The following definitions for complementary ranges are used :

$$\mathbf{X} < \mathbf{x} \quad :\Leftrightarrow \quad (X_1 < x_1) \cap (X_2 < x_2) \cap \ldots \cap (X_n < x_n)$$
$$\mathbf{X} \nless \mathbf{x} \quad :\Leftrightarrow \quad (X_1 \geq x_1) \cup (X_2 \geq x_2) \cup \ldots \cup (X_n \geq x_n)$$

Analogous definitions hold for other comparison operators. The axioms and rules of the calculus of probabilities may then be transferred to random vectors. In the limit $\mathbf{x} \to -\infty$, the range $\mathbf{X} < \mathbf{x}$ is the impossible event; in the limit $\mathbf{x} \to \infty$ it is the certain event :

$$\lim_{\mathbf{x} \to -\infty} P(\mathbf{X} < \mathbf{x}) = 0 \qquad \lim_{\mathbf{x} \to \infty} P(\mathbf{X} < \mathbf{x}) = 1$$

The ranges $\mathbf{X} < \mathbf{x}$ and $\mathbf{X} \nless \mathbf{x}$ describe complementary events, so that

$$P(\mathbf{X} < \mathbf{x}) + P(\mathbf{X} \nless \mathbf{x}) = 1$$

For $\mathbf{x}_0 < \mathbf{x}_1$, the ranges $\mathbf{X} < \mathbf{x}_0$ and $(\mathbf{X} < \mathbf{x}_1) \cap (\mathbf{X} \nless \mathbf{x}_0)$ describe incompatible events. Their union is $\mathbf{X} < \mathbf{x}_1$, and hence :

$$P(\mathbf{X} < \mathbf{x}_1) = P(\mathbf{X} < \mathbf{x}_0) + P((\mathbf{X} < \mathbf{x}_1) \cap (\mathbf{X} \nless \mathbf{x}_0))$$
$$P((\mathbf{X} < \mathbf{x}_1) \cap (\mathbf{X} \nless \mathbf{x}_0)) = P(\mathbf{X} < \mathbf{x}_1) - P(\mathbf{X} < \mathbf{x}_0)$$

Distribution function : A distribution function $F_{\mathbf{X}}(\mathbf{x})$ is introduced for the random vector. It is defined as the probability $P(\mathbf{X} < \mathbf{x})$ and takes values in the interval [0,1].

$$F_{\mathbf{X}}(\mathbf{x}) \; := \; P(\mathbf{X} < \mathbf{x}) \qquad\qquad 0 \le F_{\mathbf{X}}(\mathbf{x}) \le 1$$

The properties of the distribution functions $F_{\mathbf{X}}(\mathbf{x})$ follow directly from the rules of the calculus of probabilities for random vectors. The distribution function $F_{\mathbf{X}}(\mathbf{x})$ increases monotonically. It takes the value 0 for $\mathbf{x} \to -\infty$ and the value 1 for $\mathbf{x} \to \infty$.

$$\lim_{\mathbf{x} \to -\infty} F_{\mathbf{X}}(\mathbf{x}) = 0 \qquad \mathbf{x}_0 < \mathbf{x}_1 \;\Rightarrow\; F_{\mathbf{X}}(\mathbf{x}_0) \le F_{\mathbf{X}}(\mathbf{x}_1) \qquad \lim_{\mathbf{x} \to \infty} F_{\mathbf{X}}(\mathbf{x}) = 1$$

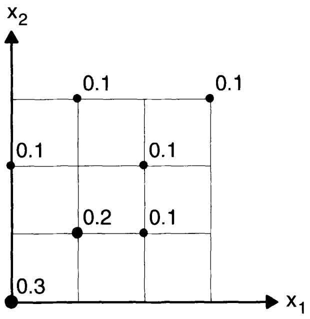
Like random variables, random vectors are classified according to the properties of their distribution function.

Discrete random vector : The distribution of a discrete random vector $\mathbf{X}$ is described by the probability function $p_{\mathbf{X}}(\mathbf{x})$. Its value is a probability $P(\mathbf{X} = \mathbf{x})$ which is non-zero only for discrete vectors $\mathbf{x}_m$. The distribution function $F_{\mathbf{X}}(\mathbf{x})$ is calculated by summing the probability function $p_{\mathbf{X}}(\mathbf{x})$.

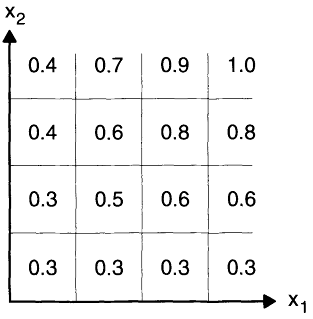
$$p_{\mathbf{X}}(\mathbf{x}) \; := \; P(\mathbf{X} = \mathbf{x}) \qquad\qquad 0 \le p_{\mathbf{X}}(\mathbf{x}) \le 1$$

$$F_{\mathbf{X}}(\mathbf{x}) \; := \; P(\mathbf{X} < \mathbf{x}) \; = \; \sum_{\mathbf{s} < \mathbf{x}} p_{\mathbf{X}}(\mathbf{s}) \qquad 0 \le F_{\mathbf{X}}(\mathbf{x}) \le 1$$

In two-dimensional space, the probability function is represented by a grid diagram with point values, and the distribution function is represented by a grid diagram with area values. Typical examples are shown below :



probability function



0.4	0.7	0.9	1.0
0.4	0.6	0.8	0.8
0.3	0.5	0.6	0.6
0.3	0.3	0.3	0.3

distribution function

Continuous random vector : The distribution of a continuous random vector **X** is described by the density function $f_{\mathbf{X}}(\mathbf{x})$, which is piecewise continuous. Its value is given by the probability $P(\mathbf{x} \le \mathbf{X} < \mathbf{x} + \Delta\mathbf{x})$ divided by the n-dimensional incremental volume element $\Delta V = \Delta x_1 \cdot \Delta x_2 \cdots \Delta x_n$, whose increments Δx_j all tend to 0. The distribution function $F_{\mathbf{X}}(\mathbf{x})$ is calculated by integrating the density function $f_{\mathbf{X}}(\mathbf{x})$. Conversely, the density function $f_{\mathbf{X}}(\mathbf{x})$ may be obtained by differentiating the distribution function $F_{\mathbf{X}}(\mathbf{x})$.

$$f_{\mathbf{X}}(\mathbf{x}) := \lim_{\substack{\Delta x_j \to 0 \\ j=1....n}} \frac{P(\mathbf{x} \le \mathbf{X} < \mathbf{x} + \Delta\mathbf{x})}{\Delta x_1 \cdot \Delta x_2 \cdots \Delta x_n} \qquad 0 \le f_{\mathbf{X}}(\mathbf{x})$$

$$F_{\mathbf{X}}(\mathbf{x}) := P(\mathbf{X} < \mathbf{x}) = \int\limits_{\mathbf{s}<\mathbf{x}} f_{\mathbf{X}}(\mathbf{s})\, dV \qquad 0 \le F_{\mathbf{X}}(\mathbf{x}) \le 1$$

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{\partial^n F_{\mathbf{X}}(\mathbf{x})}{\partial x_1 \partial x_2 \cdots \partial x_n}$$

General random vector : The distribution of a random vector **X** which is neither discrete nor continuous is described by a generalized density function $f_{\mathbf{X}}(\mathbf{x})$, which consists of the density function $f_0(\mathbf{x})$ for the continuous component and delta functions $\delta(\mathbf{x} - \mathbf{x}_j)$ with the probabilities $p_{\mathbf{X}}(\mathbf{x}_j)$ for the discrete vector values $\mathbf{x}_j$ of **X**. The distribution function $F_{\mathbf{X}}(\mathbf{x})$ is calculated by integrating the density function $f_{\mathbf{X}}(\mathbf{x})$ using the rules for delta functions. Conversely, the density function $f_{\mathbf{X}}(\mathbf{x})$ is obtained by differentiating the distribution function $F_{\mathbf{X}}(\mathbf{x})$.

$$f_{\mathbf{X}}(\mathbf{x}) := f_0(\mathbf{x}) + \sum_j p_{\mathbf{X}}(\mathbf{x}_j)\, \delta(\mathbf{x} - \mathbf{x}_j)$$

$$F_{\mathbf{X}}(\mathbf{x}) := P(\mathbf{X} < \mathbf{x}) = \int\limits_{\mathbf{s}<\mathbf{x}} f_{\mathbf{X}}(\mathbf{s})\, dV \qquad 0 \le F_{\mathbf{X}}(\mathbf{x}) \le 1$$

There are further types of density functions for random vectors. For example, a density function in two-dimensional space may contain a one-dimensional function along a given curve in addition to the two-dimensional density function for the continuous component. Such density functions must be formulated for the individual application. The distributions of discrete and continuous random vectors may be treated as special cases of the generalized density function.

Marginal distributions : The n-dimensional random vector **X** is decomposed into an m-dimensional subvector **Y** and an (n–m)-dimensional subvector **Z**. The density function $f_Y(\mathbf{y})$ is obtained from the density function $f_X(\mathbf{x}) = f_X(\mathbf{y},\mathbf{z})$ by integrating over the complete range (V_Z) of the random vector **Z**. The density function $f_Z(\mathbf{z})$ is obtained analogously.

$$f_Y(\mathbf{y}) \;=\; \int\limits_{(V_Z)} f_X(\mathbf{y},\mathbf{z})\, dV_Z$$

$$f_Z(\mathbf{z}) \;=\; \int\limits_{(V_Y)} f_X(\mathbf{y},\mathbf{z})\, dV_Y$$

The distributions of **Y** and **Z** are called m-dimensional and (n–m)-dimensional marginal distributions of **X**. Since $\binom{n}{m}$ different m-dimensional subvectors may be formed from an n-dimensional vector, an n-dimensional random vector has $\binom{n}{m}$ different m-dimensional marginal distributions.

Conditional distributions : Let the random vector **X** be decomposed into the subvectors **Y** and **Z**. The distribution for **Y** given that **Z** takes a given vector value **z** and the distribution for **Z** given that **Y** takes a given vector value **y** are called conditional distributions. Their density functions are designated by $f_{Y|Z}(\mathbf{y}\,|\,\mathbf{z})$ and $f_{Z|Y}(\mathbf{z}\,|\,\mathbf{y})$, respectively. They are calculated as follows :

$$f_{Y|Z}(\mathbf{y}\,|\,\mathbf{z}) \;=\; \frac{f_X(\mathbf{y},\mathbf{z})}{f_Z(\mathbf{z})}$$

$$f_{Z|Y}(\mathbf{z}\,|\,\mathbf{y}) \;=\; \frac{f_X(\mathbf{y},\mathbf{z})}{f_Y(\mathbf{y})}$$

The random vectors **Y** and **Z** are stochastically independent if the conditional distributions coincide with the marginal distributions. In this case, the density function $f_X(\mathbf{y},\mathbf{z})$ is equal to the product of the two marginal distributions $f_Y(\mathbf{y})$ and $f_Z(\mathbf{z})$.

$$\begin{aligned} f_{Y|Z}(\mathbf{y}\,|\,\mathbf{z}) &= f_Y(\mathbf{y}) \\ f_{Z|Y}(\mathbf{z}\,|\,\mathbf{y}) &= f_Z(\mathbf{z}) \\ f_X(\mathbf{y},\mathbf{z}) &= f_Y(\mathbf{y})\; f_Z(\mathbf{z}) \end{aligned}$$

Example 1 : Discrete two-dimensional distribution

A traffic census device counts the number of vehicles which pass in a certain period of time. The result is subject to errors. Let the real number of vehicles be X_1, and let the number registered by the device be X_2. The two-dimensional probability distribution $p_{\mathbf{X}}(x_1, x_2)$ is arranged in a matrix scheme. The one-dimensional marginal distribution for X_1 is the distribution for the real number of vehicles. The one-dimensional marginal distribution for X_2 is the distribution for the registered number of vehicles. The probability functions $p_{X_1}(x_1)$ and $p_{X_2}(x_2)$ of the two marginal distributions are calculated as row and column sums in the matrix scheme.

$p_{\mathbf{X}}$	$X_1 = 0$	$X_1 = 1$	$X_1 = 2$	$X_1 = 3$	$X_1 = 4$	p_{X_2}
$X_2 = 0$	0.25	0.03	0.01	0.00	0.00	0.29
$X_2 = 1$	0.00	0.30	0.02	0.01	0.00	0.33
$X_2 = 2$	0.00	0.00	0.20	0.02	0.00	0.22
$X_2 = 3$	0.00	0.00	0.00	0.10	0.01	0.11
$X_2 = 4$	0.00	0.00	0.00	0.00	0.05	0.05
p_{X_1}	0.25	0.33	0.23	0.13	0.06	1.00

The values of the conditional probability function $p_{X_1|X_2}(x_1|1)$ for X_1 given $X_2 = 1$ are proportional to the values in the row for $X_2 = 1$ in the matrix scheme. They are divided by $p_{X_2}(1)$ to make them add up to 1.

$$p_{X_1|X_2} \quad \begin{array}{|c|c|c|c|c|} \hline 0.00 & 0.30 & 0.02 & 0.01 & 0.00 \\ \hline \end{array} \cdot \frac{1}{0.33}$$

Example 2 : Stochastic independence

Let the two-dimensional exponential distribution for the continuous random vector **X** with the variables X_1 and X_2 and the following two-dimensional density function be given :

$$f_{\mathbf{X}}(x_1, x_2) = \lambda_1 \lambda_2 e^{-(\lambda_1 x_1 + \lambda_2 x_2)} \qquad x_1, x_2 \geq 0 \qquad \lambda_1, \lambda_2 > 0$$

The marginal distributions for X_1 and X_2 are calculated as follows :

$$f_{X_1}(x_1) = \int_0^\infty f_{\mathbf{X}}(x_1, x_2)\, dx_2 = \lambda_1 e^{-\lambda_1 x_1}$$

$$f_{X_2}(x_2) = \int_0^\infty f_{\mathbf{X}}(x_1, x_2)\, dx_1 = \lambda_2 e^{-\lambda_2 x_2}$$

Since the product of the marginal distributions yields the original two-dimensional exponential distribution, the random variables X_1 and X_2 are stochastically independent.

10.4.3 MOMENTS

Introduction : The definition of the moments of random variables may be generalized for random vectors. The moments of first and second order are especially important for random vectors; they are formulated using vector and matrix algebra. The moments of first order lead to the vector of means. The central moments of second order lead to the matrix of variances and covariances. This matrix allows statements about the linear correlation between the random variables of the random vector.

Mean : Let an n-dimensional random vector **X** with a generalized density function $f_{\mathbf{X}}(\mathbf{x})$ be given. The mean μ_j of a random variable X_j of the random vector **X** is the first-order moment $E(X_j)$, which is defined as follows :

$$\mu_j \;:=\; E(X_j) \;:=\; \int_{(V)} x_j \, f_{\mathbf{X}}(\mathbf{x}) \, dV \qquad\qquad j = 1,\ldots,n$$

The means of all random variables are arranged in the vector $\mathbf{m}_{\mathbf{X}}$:

$$\mathbf{m}_{\mathbf{X}} \;=\; \int_{(V)} \mathbf{x} \, f_{\mathbf{X}}(\mathbf{x}) \, dV$$

Variances and covariances : The variance σ_{jj} of the random variable X_j of the random vector **X** is the second-order central moment $D(X_j^2)$ with respect to the mean μ_j, which is defined as follows :

$$\sigma_{jj} \;:=\; D(X_j^2) \;:=\; \int_{(V)} (x_j - \mu_j)^2 \, f_{\mathbf{X}}(\mathbf{x}) \, dV$$

The covariance σ_{jk} of two random variables X_j and X_k is the second-order central moment $D(X_j \, X_k)$ with respect to the two means μ_j and μ_k, which is defined as follows :

$$\sigma_{jk} \;:=\; D(X_j \, X_k) \;:=\; \int_{(V)} (x_j - \mu_j) \, (x_k - \mu_k) \, f_{\mathbf{X}}(\mathbf{x}) \, dV \qquad\qquad j \neq k$$

The covariances σ_{jk} and σ_{kj} are identical. The variances and covariances for all pairs of random variables are arranged in the symmetric covariance matrix $\mathbf{V}_{\mathbf{X}}$.

$$\mathbf{V}_{\mathbf{X}} \;=\; \int_{(V)} (\mathbf{x} - \mathbf{m}_{\mathbf{X}}) \, (\mathbf{x} - \mathbf{m}_{\mathbf{X}})^T \, f_{\mathbf{X}}(\mathbf{x}) \, dV$$

The covariance matrix is positive semidefinite. Its quadratic form $Q = \mathbf{z}^T \mathbf{V}_{\mathbf{X}} \, \mathbf{z}$ is non-negative for an arbitrary vector $\mathbf{z} \neq \mathbf{0}$. This is proved as follows :

$$Q \;=\; \mathbf{z}^T \mathbf{V_X} \mathbf{z} \;=\; \int\limits_{(V)} \mathbf{z}^T (\mathbf{x}-\mathbf{m_X})\; (\mathbf{x}-\mathbf{m_X})^T \mathbf{z}\; f_{\mathbf{X}}(\mathbf{x})\; dV$$

$$Q \;=\; \int\limits_{(V)} (\mathbf{z}^T(\mathbf{x}-\mathbf{m_X}))^2\; f_{\mathbf{X}}(\mathbf{x})\; dV \;\geq\; 0$$

The quadratic form Q is the integral of the product of two non-negative functions. Hence it cannot take negative values. If a random vector **Y** is an m-dimensional subvector of the n-dimensional random vector **X**, then $\mathbf{V_Y}$ is a submatrix of $\mathbf{V_X}$ consisting of the variances and covariances for the random variables of **X** contained in **Y**. Each covariance submatrix in an m-dimensional subspace is symmetric and positive semidefinite.

According to the rules of linear algebra, the determinant of a positive semidefinite matrix is non-negative. Applying this rule to the covariance matrix with its covariance submatrices in the various subspaces leads to restrictions on the variances and covariances. The determinants for the covariance submatrices in the subspaces of dimension m = 1, 2, 3 are readily calculated analytically. They yield the following restrictions :

$$\det \begin{array}{|c|}\hline \sigma_{jj} \\ \hline\end{array} \;\geq\; 0 \qquad\qquad \sigma_{jj} \geq 0$$

$$\det \begin{array}{|c|c|}\hline \sigma_{jj} & \sigma_{jk} \\ \hline \sigma_{jk} & \sigma_{kk} \\ \hline\end{array} \;\geq\; 0 \qquad\qquad \sigma_{jj}\,\sigma_{kk} - \sigma_{jk}^2 \;\geq\; 0$$

$$\det \begin{array}{|c|c|c|}\hline \sigma_{jj} & \sigma_{jk} & \sigma_{jq} \\ \hline \sigma_{jk} & \sigma_{kk} & \sigma_{kq} \\ \hline \sigma_{jq} & \sigma_{kq} & \sigma_{qq} \\ \hline\end{array} \;\geq\; 0$$

$$\sigma_{jj}\;\sigma_{kk}\;\sigma_{qq} + 2\,\sigma_{jk}\;\sigma_{jq}\;\sigma_{kq} - \sigma_{jj}\;\sigma_{kq}^2 - \sigma_{kk}\,\sigma_{jq}^2 - \sigma_{qq}\;\sigma_{jk}^2 \geq 0$$

According to the rules of linear algebra, the non-negative determinant of a positive semidefinite matrix is bounded from above by the product of all diagonal elements. The covariance matrix therefore satisfies :

$$0 \;\leq\; \det \mathbf{V_X} \;\leq\; \prod_{j=1}^{n} \sigma_{jj}$$

Correlation : The correlation factor ϱ_{jk} of two random variables X_j and X_k is calculated from the covariance σ_{jk} and the standard deviations σ_j and σ_k :

$$\varrho_{jk} := \frac{\sigma_{jk}}{\sigma_j \, \sigma_k} \qquad \sigma_j = \sqrt{\sigma_{jj}} > 0 \qquad \sigma_k = \sqrt{\sigma_{kk}} > 0$$

The correlation factors ϱ_{jj} are 1. The correlation factors ϱ_{jk} and ϱ_{kj} are equal. The correlation factors are arranged in the symmetric correlation matrix $\mathbf{R_X}$. It is obtained from the covariance matrix $\mathbf{V_X}$ by dividing each row j and each column j by the positive standard deviation σ_j. These operations are formulated in matrix notation as follows, using the diagonal matrix $\mathbf{S_X}$ of the standard deviations :

$$\mathbf{R_X} = \mathbf{S_X^{-1}} \, \mathbf{V_X} \, \mathbf{S_X^{-1}} \qquad\qquad \mathbf{V_X} = \mathbf{S_X} \, \mathbf{R_X} \, \mathbf{S_X}$$

$\mathbf{S_X}$ diagonal matrix of the standard deviations σ_j

The correlation matrix with diagonal elements 1 is a normalized covariance matrix. It is positive semidefinite. The non-negative determinants of its submatrices in the subspaces of dimension m = 2, 3 yield the following restrictions on the correlation factors :

$$\det \begin{vmatrix} 1 & \varrho_{jk} \\ \varrho_{jk} & 1 \end{vmatrix} \geq 0 \qquad\qquad \varrho_{jk}^2 \leq 1$$

$$\det \begin{vmatrix} 1 & \varrho_{jk} & \varrho_{jq} \\ \varrho_{jk} & 1 & \varrho_{kq} \\ \varrho_{jq} & \varrho_{kq} & 1 \end{vmatrix} \geq 0 \qquad\qquad \varrho_{jk}^2 + \varrho_{jq}^2 + \varrho_{kq}^2 \leq 1 + 2 \, \varrho_{jk} \, \varrho_{jq} \, \varrho_{kq}$$

The non-negative determinant of the correlation matrix is bounded from above by the product of all diagonal elements. Since all diagonal elements are 1, this implies :

$$0 \leq \det \mathbf{R_X} \leq 1$$

Linear dependence : The determinant of the correlation matrix is 0 if the correlation matrix $\mathbf{R_X}$ is singular and does not have an inverse. In this case the random variables of the random vector **X** are said to be linearly dependent. The determinant of the correlation matrix is 1 if the correlation matrix $\mathbf{R_X}$ is equal to the identity matrix **I**. In this case the random variables of the random vector **X** are said to be linearly independent.

$$\text{linear dependence} \quad :\Leftrightarrow \quad \det \mathbf{R_X} = 0$$
$$\text{linear independence} \quad :\Leftrightarrow \quad \det \mathbf{R_X} = 1$$

If the random variables of the random vector **X** are stochastically independent, then they are also linearly independent. If the random variables are linearly independent, they may nevertheless exhibit a non-linear dependence. In this case, moments of higher order need to be considered.

Degree of linear dependence : The eigenvalues λ of the positive semidefinite correlation matrix $\mathbf{R}_\mathbf{X}$ are positive or zero. The product of all eigenvalues is equal to the determinant of the correlation matrix. Since the determinant only takes values in the interval [0,1], the least eigenvalue λ_{min} must also lie in the interval [0,1].

$$0 \le \lambda_{\text{min}} \le 1$$

If the random variables are linearly dependent, the correlation matrix is singular and the least eigenvalue is zero. If the random variables are almost linearly dependent, the correlation matrix is almost singular and the least eigenvalue is almost zero. If the random variables are linearly independent, the correlation matrix is equal to the identity matrix and the least eigenvalue is one. Like the determinant of $\mathbf{R}_\mathbf{X}$, the least eigenvalue of $\mathbf{R}_\mathbf{X}$ is therefore a suitable measure of the degree of linear dependence of the random variables of a random vector.

Example : Moments of a discrete two-dimensional distribution

Example 1 of Section 10.4.2 shows a discrete two-dimensional distribution obtained by checking a traffic census device. The random variable X_1 is the real number of vehicles. The random variable X_2 is the number of vehicles registered by the device. The means and the variances and covariances for the discrete distribution are obtained by summation.

$$\mu_j = \sum_{x_1=0}^{4} \sum_{x_2=0}^{4} x_j \, p_\mathbf{X}(x_1,x_2) \qquad j = 1,2$$

$$\sigma_{jk} = \sum_{x_1=0}^{4} \sum_{x_2=0}^{4} (x_j - \mu_j)(x_k - \mu_k) \, p_\mathbf{X}(x_1,x_2) \qquad j,k = 1,2$$

The calculated numerical values for the means, variances and covariances are arranged in the vector $\mathbf{m}_\mathbf{X}$ and the matrix $\mathbf{V}_\mathbf{X}$.

$$\mathbf{m}_\mathbf{X} = \begin{array}{|c|}\hline 1.420 \\ \hline 1.300 \\ \hline \end{array} \qquad \mathbf{V}_\mathbf{X} = \begin{array}{|c|c|}\hline 1.364 & 1.264 \\ \hline 1.264 & 1.310 \\ \hline \end{array}$$

The standard deviations $\sigma_j = \sqrt{\sigma_{jj}}$ are arranged in the diagonal matrix $\mathbf{S}_\mathbf{X}$. The correlation matrix $\mathbf{R}_\mathbf{X}$ is obtained from the covariance matrix $\mathbf{V}_\mathbf{X}$ by dividing each element σ_{ij} by $\sigma_i \, \sigma_j$.

$$\mathbf{S}_\mathbf{X} = \begin{array}{|c|c|}\hline 1.168 & 0.000 \\ \hline 0.000 & 1.145 \\ \hline \end{array} \qquad \mathbf{R}_\mathbf{X} = \begin{array}{|c|c|}\hline 1.000 & 0.946 \\ \hline 0.946 & 1.000 \\ \hline \end{array}$$

The determinant and the two eigenvalues λ_1 and λ_2 of the correlation matrix are :

$$\det \mathbf{R}_\mathbf{X} = 0.105 \qquad \lambda_1 = 0.054 \qquad \lambda_2 = 1.946$$

The least eigenvalue of the correlation matrix is nearly zero. The random variables X_1 and X_2 are therefore nearly linearly dependent.

10.4.4 FUNCTIONS OF A RANDOM VECTOR

Introduction : The deterministic dependence of random variables may be generalized for random vectors. The dependence is described by a set of multidimensional functions. The relationships between the probability distributions and moments of deterministically dependent random vectors are treated in the following.

Functions : Let a unique vector value **y** of the random vector **Y** be assigned to every vector value **x** of the random vector **X**. This assignment is described by a set of functions **y** = **g**(**x**). If conversely every vector value **y** is associated with a unique vector value **x**, then there is also a set of inverse functions **x** = **h**(**y**).

$$\mathbf{Y} := \mathbf{g}(\mathbf{X}) \qquad\qquad \mathbf{y} := \mathbf{g}(\mathbf{x})$$

Distribution function : The probability that **Y** takes a vector value less than **y** is equal to the probability that **g**(**X**) takes a vector value less than **y**. The distribution function $F_Y(\mathbf{y})$ is therefore calculated by integrating the density function $f_X(\mathbf{x})$ over the range for which **g**(**x**) is less than **y**.

$$F_Y(\mathbf{y}) = P(\mathbf{Y} < \mathbf{y}) = P(\mathbf{g}(\mathbf{X}) < \mathbf{y}) = \int\limits_{\mathbf{g}(\mathbf{x}) < \mathbf{y}} f_X(\mathbf{x})\, dV_X$$

Density function : If the random vector **Y** is continuous, the density function $f_Y(\mathbf{y})$ is obtained from the distribution function $F_Y(\mathbf{y})$ by partial differentiation. If the continuous random vectors **X**, **Y** have the same dimension n and the functions **y** = **g**(**x**) possess inverse functions **x** = **h**(**y**), then the density function $f_Y(\mathbf{y})$ may be obtained directly from the density function $f_X(\mathbf{x})$ using the determinant of the Jacobian matrix **J**.

$$f_Y(\mathbf{y}) = |\det \mathbf{J}|\, f_X(\mathbf{h}(\mathbf{y}))$$

$$\mathbf{J} = \begin{array}{|c|c|c|} \hline \dfrac{\partial h_1}{\partial y_1} & \cdots & \dfrac{\partial h_n}{\partial y_1} \\ \hline \vdots & & \vdots \\ \hline \dfrac{\partial h_1}{\partial y_n} & \cdots & \dfrac{\partial h_n}{\partial y_n} \\ \hline \end{array}$$

Moments : The vector of means and the covariance matrix for the random vector **Y** are determined as follows :

$$\mathbf{m}_Y = \int_{(V_Y)} \mathbf{y}\, f_Y(\mathbf{y})\, dV_Y = \int_{(V_X)} \mathbf{g}(\mathbf{x})\, f_X(\mathbf{x})\, dV_X$$

$$\mathbf{V}_Y = \int_{(V_Y)} (\mathbf{y} - \mathbf{m}_Y)\, (\mathbf{y} - \mathbf{m}_Y)^T\, f_Y(\mathbf{y})\, dV_Y$$

$$\mathbf{V}_Y = \int_{(V_X)} (\mathbf{g}(\mathbf{x}) - \mathbf{m}_Y)\, (\mathbf{g}(\mathbf{x}) - \mathbf{m}_Y)^T\, f_X(\mathbf{x})\, dV_X$$

Example : Linear dependence

Let a random vector **Y** depend linearly on a random vector **X**, that is

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \mathbf{A}\,\mathbf{X} + \mathbf{b} \qquad \mathbf{y} = \mathbf{g}(\mathbf{x}) = \mathbf{A}\,\mathbf{x} + \mathbf{b}$$

If the matrix **A** is regular, it may be inverted to obtain

$$\mathbf{X} = \mathbf{h}(\mathbf{Y}) = \mathbf{A}^{-1}(\mathbf{Y} - \mathbf{b}) \qquad \mathbf{x} = \mathbf{h}(\mathbf{y}) = \mathbf{A}^{-1}(\mathbf{y} - \mathbf{b})$$

If the random vector **X** is continuous, the random vector **Y** is also continuous. The density function $f_Y(\mathbf{y})$ may be obtained directly from the density function $f_X(\mathbf{x})$. The Jacobian matrix is $\mathbf{A}^{-1}$.

$$f_Y(\mathbf{y}) = |\det \mathbf{A}^{-1}|\, f_X(\mathbf{x}) = f_X\,(\mathbf{A}^{-1}(\mathbf{y} - \mathbf{b}))\, /\, |\det \mathbf{A}|$$

The moments of the random vector **Y** are obtained directly from the moments of the random vector **X**. The vector of means and the covariance matrix are given by :

$$\mathbf{m}_Y = \int_{(V)} (\mathbf{A}\,\mathbf{x} + \mathbf{b})\, f_X(\mathbf{x})\, dV = \mathbf{A}\mathbf{m}_X + \mathbf{b}$$

$$\mathbf{V}_Y = \int_{(V)} (\mathbf{A}\,\mathbf{x} + \mathbf{b} - \mathbf{m}_Y)\, (\mathbf{A}\,\mathbf{x} + \mathbf{b} - \mathbf{m}_Y)^T\, f_X(\mathbf{x})\, dV$$

$$\mathbf{V}_Y = \int_{(V)} \mathbf{A}\,(\mathbf{x} - \mathbf{m}_X)\,(\mathbf{x} - \mathbf{m}_X)^T\, \mathbf{A}^T\, f_X(\mathbf{x})\, dV$$

$$\mathbf{V}_Y = \mathbf{A}\,\mathbf{V}_X\,\mathbf{A}^T$$

10.4.5 MULTINOMIAL DISTRIBUTION

The basic multidimensional distribution for discrete random vectors is the multinomial distribution. It is a generalization of the binomial distribution treated in Section 10.3.6.2.

Model : An experiment is repeated n times. Exactly one of m possible events A_j occurs in each experiment. The discrete random variable X_k is the number of events in n experiments with the same value A_k. The discrete random variables X_j are arranged in a random vector. They are linearly dependent, since the sum of all random variables X_j must be n.

$$\text{random vector} \qquad \mathbf{X} \geq \mathbf{0}$$

$$\text{condition} \qquad n = \sum_{j=1}^{m} X_j$$

Let the possible events be incompatible and stochastically independent. Each event A_j is assigned a probability of occurrence $p_j = P(A_j)$. The union of all possible events is the certain event, so that the sum of the probabilities of occurrence for all events is one. The probabilities of occurrence for the m events are arranged in a vector **p**.

$$\text{probabilities} \qquad \mathbf{p} \geq \mathbf{0}$$

$$\text{condition} \qquad 1 = \sum_{j=1}^{m} p_j$$

Probability function : For n experiments there are $n! \,/\, (x_1!\, x_2! \ldots x_m!)$ different ordered n-tuples with x_j events A_j. The probability for the occurrence of the events A_j within an n-tuple is determined using the product rule. The probability $p_{\mathbf{X}}(\mathbf{x})$ for the occurrence of one of the possible n-tuples is determined using the sum rule :

$$p_{\mathbf{X}}(\mathbf{x}) = \frac{n!}{x_1!\, x_2! \ldots x_m!}\, p_1^{x_1}\, p_2^{x_2} \ldots p_m^{x_m}$$

$$p_{\mathbf{X}}(\mathbf{x}) = n! \prod_{j=1}^{m} p_j^{x_j} / x_j! \qquad\qquad n = \sum_{j=1}^{m} x_j$$

For the special case $m = 2$ with $x_1 = x$ and $x_2 = n - x$ and with $p_1 = p$ and $p_2 = q = 1 - p$, this is the binomial distribution from Section 10.3.6.2 :

$$p_X(x) = \frac{n!}{x!\,(n-x)!}\, p^x q^{n-x} = \binom{n}{x} p^x q^{n-x}$$

Every marginal distribution of a multinomial distribution is also a multinomial distribution. The marginal distribution for a random variable X_j of the random vector is a binomial distribution.

Moments : The means and the variances and covariances are :

$$\mu_j = n\,p_j$$

$$\sigma_{jj} = n\,p_j\,(1 - p_j)$$

$$\sigma_{jk} = -\,n\,p_j\,p_k \qquad\qquad j \neq k$$

Thus the vector $\mathbf{m}_X$ of means and the covariance matrix $\mathbf{V}_X$ are given by :

$$\mathbf{m}_X = n\,\mathbf{p}$$

$$\mathbf{V}_X = n\,(\mathbf{D} - \mathbf{p}\mathbf{p}^T)$$

$\mathbf{D}$ diagonal matrix with the probabilities p_j

Example : The vehicles on a street reach a junction. On average, 20% turn left and 30% turn right, while the remaining 50% drive straight on. If 8 vehicles are observed, the probability that 4 turn left, 3 drive straight on and 1 turns right is calculated as follows :

$$p_{\mathbf{X}}(4,3,1) = \frac{8!}{4!\ 3!\ 1!}\,0.2^4\,0.5^3\,0.3^1 = 0.0168 = 1.68\%$$

10.4.6 MULTINORMAL DISTRIBUTION

The basic multidimensional distribution for continuous random vectors is the multinormal distribution. It is a generalization of the normal distribution treated in Section 10.3.7.2.

Model : For the multinormal distribution, an n-dimensional continuous random vector **X** with the vector $\mathbf{m}_X$ of means and the positive definite covariance matrix $\mathbf{V}_X$ is considered.

Density function : The random vector **X** has a multinormal distribution if its density function has the following form :

$$f_X(\mathbf{x}) = C\, e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m}_X)^T \mathbf{V}_X^{-1}(\mathbf{x}-\mathbf{m}_X)}$$

$$C = 1 / \sqrt{(2\pi)^n \det \mathbf{V}_X}$$

The general normal distribution may be standardized by a linear transformation of the random vector **X**. The transformation rule contains the vector $\mathbf{m}_X$ of the means and the diagonal matrix $\mathbf{S}_X$ of the standard deviations. The standardized multinormal distribution depends only on the correlation matrix $\mathbf{R}_X$. The rules for linear transformations in Section 10.4.4 yield :

$$\mathbf{U} = \mathbf{S}_X^{-1}(\mathbf{X} - \mathbf{m}_X) \qquad \mathbf{X} = \mathbf{S}_X \mathbf{U} + \mathbf{m}_X$$

$$f_U(\mathbf{u}) = C' e^{-\frac{1}{2}\mathbf{u}^T \mathbf{R}_X^{-1}\mathbf{u}} \qquad f_X(\mathbf{x}) = f_U(\mathbf{u}) / \det \mathbf{S}_X$$

$$C' = 1 / \sqrt{(2\pi)^n \det \mathbf{R}_X}$$

Every marginal distribution and every conditional distribution of a multinormal distribution is also a multinormal distribution. The marginal distribution for a random variable X_j of the random vector is a normal distribution.

Properties : Let a random vector **Y** depend linearly on a random vector **X**. If the random vector **X** has a multinormal distribution, then the random vector **Y** also has a multinormal distribution. The vector $\mathbf{m}_Y$ of means and the covariance matrix $\mathbf{V}_Y$ are obtained from the vector $\mathbf{m}_X$ of means and the covariance matrix $\mathbf{V}_X$ according to the rules in Section 10.4.4.

Example : The vertical displacements of the supports of a bridge beam are considered in the example in Section 10.3.7.2. The bending moment B at the central support depends linearly on the vertical displacements X_1, X_2, X_3 at the left, central and right support.

$$B = M_0(-X_1 + 2X_2 - X_3)$$

Let the vertical displacements X_1, X_2, X_3 possess a multinormal distribution, and let them have the same mean $\mu = 0$ and the same standard deviation $\sigma = 0.5$. In contrast to the example in Section 10.3.7.2, the vertical displacements are not assumed to be stochastically independent. Rather, a correlation with the factor ϱ between the vertical displacements of any two neighboring supports is assumed. The means, correlations and variances of the vertical displacements are given by :

$$\begin{aligned} \mathbf{m}_X &= \mathbf{0} \\ \mathbf{S}_X &= \sigma \mathrm{I} \\ \mathbf{V}_X &= \sigma^2 \mathbf{R}_X \end{aligned} \qquad \mathbf{R}_X = \begin{array}{|c|c|c|} \hline 1 & \varrho & 0 \\ \hline \varrho & 1 & \varrho \\ \hline 0 & \varrho & 1 \\ \hline \end{array}$$

The multinormal distribution requires that the correlation matrix $\mathbf{R}_X$ be positive definite. The admissible range for the correlation factor ϱ is determined from the condition that the determinant of $\mathbf{R}_X$ takes only positive values :

$$\det \mathbf{R}_X = 1 - 2\varrho^2 > 0 \qquad \varrho^2 < \frac{1}{2} \qquad -\sqrt{2}/2 < \varrho < \sqrt{2}/2$$

Since the bending moment B is a linear combination of the vertical displacements, whose distribution is multinormal, it possesses a normal distribution. The mean and the standard deviation of B are calculated as follows :

$$B = \mathbf{a}^T \mathbf{X} \qquad \mathbf{a}^T = \begin{array}{|c|c|c|} \hline -1 & 2 & -1 \\ \hline \end{array} \; M_0$$

$$\mu_B = \mathbf{a}^T \mathbf{m}_X = 0$$

$$\sigma_B^2 = \mathbf{a}^T \mathbf{V}_X \mathbf{a} = \sigma^2 \mathbf{a}^T \mathbf{R}_X \mathbf{a} = \sigma^2 M_0^2 (6 - 8\varrho)$$

$$\sigma_B = \sigma M_0 \sqrt{6 - 8\varrho} = 0.5 M_0 \sqrt{6 - 8\varrho}$$

The probability that the bending moment is less than 2 M_0 is calculated using the standardized normal distribution. For the case $\varrho = 0$ in which the vertical displacements are stochastically independent, the result in Section 10.3.7.2 is obtained. For the two limiting cases in which the vertical displacements are linearly dependent, one obtains the following results :

$$\begin{aligned} \varrho &= -\sqrt{2}/2 &: \quad P(B < 2M_0) &= 87.9\% \\ \varrho &= 0 &: \quad P(B < 2M_0) &= 94.8\% \\ \varrho &= +\sqrt{2}/2 &: \quad P(B < 2M_0) &\approx 100.0\% \end{aligned}$$

10.5 RANDOM PROCESSES

10.5.1 INTRODUCTION

A random process describes random time-dependent states. A random process is also called a stochastic process. Every stochastic process has a certain time domain and a certain space of states. Each of these sets may be discrete or continuous, and the stochastic processes are classified accordingly. The basic definitions, the classification and simple exemplary applications of stochastic processes are treated in the following.

Time domain : Let a process depend on a parameter t which takes real values. In many applications t represents time. A totally ordered set T which contains each of the possible time points t_k is called a time domain. If the time points are countable, the time domain is said to be discrete. If they are not countable, the time domain is said to be continuous.

$$T := \{t_1, t_2, \dots\} \subseteq \mathbb{R} \qquad\qquad t_1 < t_2 < \dots$$

t_k time point

Space of states : A process may be in various elementary states. A set S which contains each of the possible elementary states e_j is called a space of states. If the elementary states are countable, the space of states is said to be discrete. If they are not countable, the space of states is said to be continuous.

$$S := \{e_1, e_2, \dots, e_m\}$$

e_j elementary state

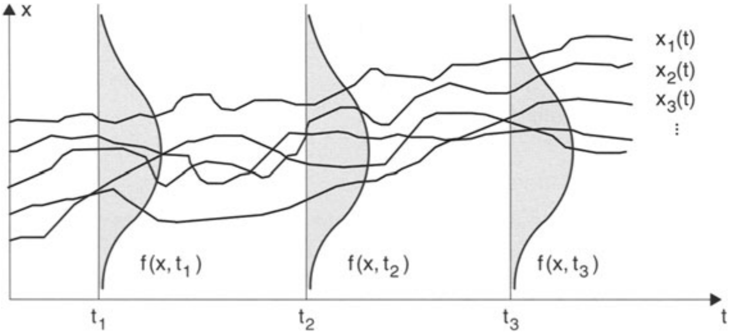
Random function : A function which for every time point $t \in T$ maps the space of states S to the set $\mathbb{R}$ of real numbers is called a random function and is designated by X(t). The random function describes the random process.

$$X(t) : S \rightarrow \mathbb{R} \qquad\qquad t \in T$$

If the time-dependent course of states of a random process is recorded in an experiment, the recorded function x(t) is called a realization (trajectory) of the random function X(t). If the experiment is repeated several times, different realizations $x_1(t)$, $x_2(t), \dots$ are obtained. The set of all possible realizations forms a functional space.

If a time point $t = t_1$ is considered, $X(t_1)$ is a random variable which takes values $x_1(t_1), x_2(t_1), \ldots$ in the experiments. The random variable $X(t_1)$ is described by a probability distribution. If several different time points $t = t_1, t_2, \ldots$ are considered, $X(t_1), X(t_2), \ldots$ are different random variables which are generally stochastically dependent. The random variables $X(t_1), X(t_2), \ldots$ may be arranged in a random vector and described by a multidimensional probability distribution.

The following diagram shows realizations and density functions for fixed points in time for a random process with a continuous space of states and a continuous time domain.



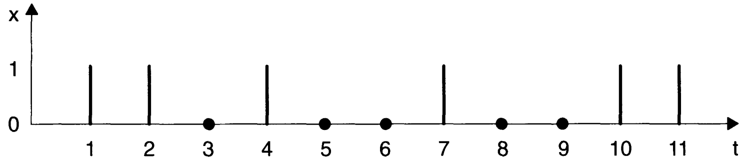
Classification : Random processes are classified with respect to the discreteness or continuity of their time domain and space of states :

- random processes in discrete time with a discrete space of states
- random processes in continuous time with a discrete space of states
- random processes in discrete time with a continuous space of states
- random processes in continuous time with a continuous space of states

Different classes of random processes are described in the following examples.

Example 1 : Bernoulli process

A coin is tossed several times. For each toss of the coin, the result is either "heads" or "tails". The result "heads" is assigned the value 0, and the result "tails" is assigned the value 1. The random variable $X(t)$ is the result of the t-th toss. This random process is called a Bernoulli process. Its time domain is discrete and contains the integer values $t > 0$. Its space of states is discrete and consists of two states. Since the result of a toss of the coin is independent of the results of the preceding tosses, the states of the Bernoulli process for different instants are independent. A possible realization of this process is illustrated graphically below.

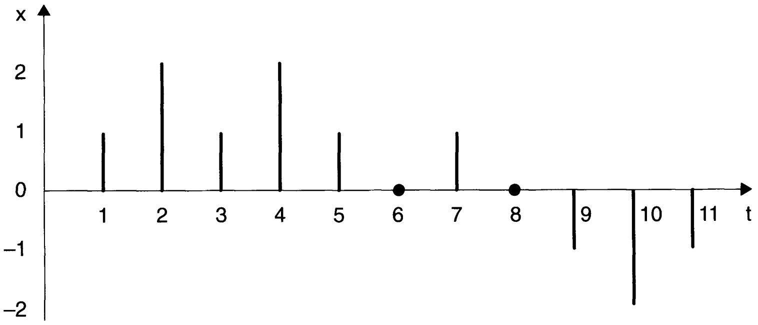


Example 2 : Bernoulli process

The precipitation at a certain location is observed on consecutive days. On a given day, the weather is either ”dry” or ”wet”. ”Dry” weather is assigned the value 0, and ”wet” weather is assigned the value 1. The random variable X(t) is the weather on the t-th day. This random process is a Bernoulli process, as in Example 1. However, in contrast to Example 1, the states of this Bernoulli process at different times are not independent, since the weather on one day depends on the weather on the preceding days to a certain degree.

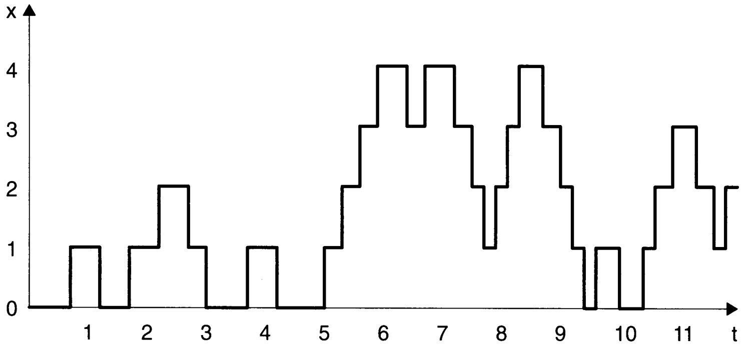
Example 3 : Simple random walk

In a one-dimensional space with the coordinate x, a particle starts out at the position $x = 0$. This particle moves in steps of length 1 in the positive or negative x-direction. The direction of motion is random. The random variable X(t) is the position of the particle after t steps. This random process is called a simple random walk. It has a discrete time domain and a discrete space of states. A possible realization of a random walk is illustrated below.



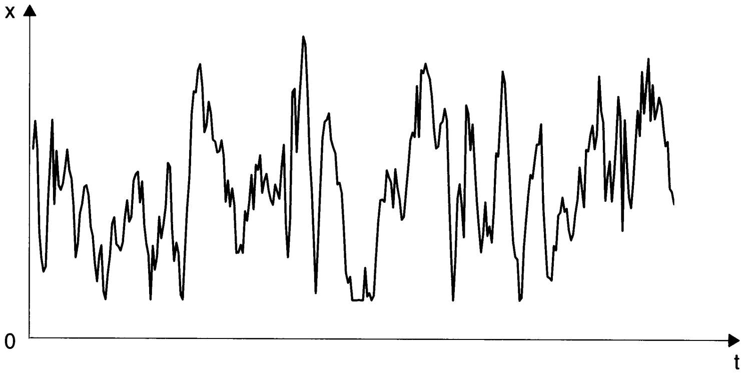
Example 4 : Queue

Customers arrive at a counter in randomly fluctuating time intervals. Let the time required to serve a customer also be random. The random variable X(t) is the number of waiting customers. This random process is called a queueing process. It has a continuous time domain and a discrete space of states. A possible realization of a queueing process is illustrated below.



Example 5 : Wind measurements

Wind velocities are measured at a certain location and recorded as a function of time. The wind velocities vary randomly. The random variable X(t) is the wind velocity at time t. This random process has a continuous time domain and a continuous space of states. A possible realization of this random process is illustrated below.



10.5.2 FINITE MARKOV PROCESSES IN DISCRETE TIME

10.5.2.1 Introduction

A random process with a finite discrete space of states and a discrete time domain is a finite Markov process. At each time point, the Markov process is in one of the possible states with a certain probability. Every state changes to a new state at the next time point with a certain transition probability. The state probabilities at this time point are calculated from the state probabilities at the preceding time point and the transition probabilities according to the rules of the calculus of probabilities. The basic definitions and rules for Markov processes of this kind are treated in Section 10.5.2.2.

If the transition probabilities are the same for each time point, the Markov process is said to be homogeneous. The long-term behavior of homogeneous Markov processes is of central importance in practical applications. Structural and spectral analysis lead to qualitative and quantitative statements, respectively, about the properties of the long-term behavior. They are treated in Sections 10.5.2.3 and 10.5.2.4. In practical applications, it is often important to know when a state can be reached for the first time. The methods for solving this problem are treated in Section 10.5.2.5.

10.5.2.2 States and transitions

Introduction : A Markov process is described by a finite set of states and a finite set of possible state transitions. It may be represented by a graph. At every time point, the states and transitions are associated with probabilities. The state probabilities are arranged in a stochastic vector, the transition probabilities are arranged in a stochastic matrix. The rules of calculation for Markov processes may then be formulated on the basis of vector and matrix algebra.

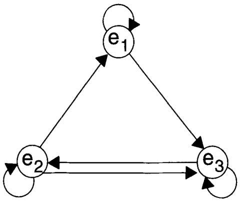
State and weight : A finite Markov process can be in different elementary states e_j. The set of all possible elementary states is the space of states S. Let the number of possible elementary states be finite. Each elementary state $e_j \in S$ is assigned a real number $x_j \in \mathbb{R}$ as a weight.

$$S = \{e_1, e_2, ..., e_m\}$$

Transition : The state of a finite Markov process changes in steps. In each step, the process makes a transition from a state e_j to a state e_k. The transition is described by the ordered pair (e_j, e_k) of states. The set of all possible transitions is a binary relation R in the space of states S, that is a subset of the cartesian product $S \times S$.

$$R \subseteq S \times S$$

Transition graph : The possible transitions are often represented by a directed graph (S ; R) with the set of states S and the transition relation R. Every state is a vertex, and every transition is a directed edge of the transition graph. A transition graph for the space of states $S = \{e_1, e_2, e_3\}$ is shown as an example.



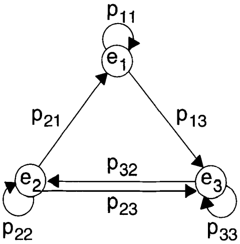
State probabilities : At time t_n, a finite Markov process has certain probabilities $p_{j,n}$ of being in the different states e_j. The state probability $p_{j,n}$ is the probability that the random variable $X(t_n)$ takes the value x_j. The state probabilities for all possible states form the probability function for the random variable $X(t_n)$. Their sum is 1.

$$p_{j,n} := P(X(t_n) = x_j) \geq 0 \qquad \sum_{j=1}^{m} p_{j,n} = 1$$

Transition probabilities : A state e_j at time t_n makes a transition to a state e_k at time t_{n+1} with a certain probability p_{jk}. The transition probability p_{jk} is the probability that the random variable $X(t_{n+1})$ takes the value x_k given that the random variable $X(t_n)$ takes the value x_j. The transition probabilities from a state e_j to all possible states e_k form a probability function. Their sum is 1.

$$p_{jk} := P(X(t_{n+1}) = x_k \mid X(t_n) = x_j) \geq 0 \qquad \sum_{k=1}^{m} p_{jk} = 1$$

The transition probabilities may be represented as weights in the transition graph. Every edge for a transition from the state e_j to the state e_k is weighted with a transition probability $0 < p_{jk} \leq 1$. The sum of the transition probabilities for all edges emanating from a vertex is 1. A weighted transition graph is shown as an example.



$$\begin{aligned} p_{11} \qquad\quad &+ p_{13} = 1 \\ p_{21} + p_{22} &+ p_{23} = 1 \\ p_{32} &+ p_{33} = 1 \end{aligned}$$

Rule of calculation : If the state probabilities at time t_n and the transition probabilities from t_n to t_{n+1} are known, the state probabilities at time t_{n+1} are calculated according to the total probability theorem in Section 10.2.4.

$$P(X(t_{n+1}) = x_k) = \sum_{j=1}^{m} P(X(t_{n+1}) = x_k \mid X(t_n) = x_j) * P(X(t_n) = x_j)$$

$$p_{k,n+1} = \sum_{j=1}^{m} p_{jk} \; p_{j,n}$$

Stochastic vectors and matrices : The rules of calculation for finite Markov processes are represent concisely using stochastic vectors and matrices. A stochastic vector **p** contains probabilities whose sum is 1. A stochastic matrix **P** is quadratic and contains probabilities whose sum for each row is 1. The sum conditions for stochastic vectors and matrices are conveniently formulated using the one vector **e**.

stochastic vector	:	$\mathbf{p} \geq \mathbf{0}$	with	$\mathbf{e}^T\mathbf{p} = 1$
stochastic matrix	:	$\mathbf{P} \geq \mathbf{0}$	with	$\mathbf{P}\,\mathbf{e} = \mathbf{e}$

$$\mathbf{p} = \begin{array}{|c|}\hline p_1 \\ \hline p_2 \\ \hline \vdots \\ \hline p_m \\ \hline\end{array} \qquad \mathbf{P} = \begin{array}{|c|c|c|c|}\hline p_{11} & p_{12} & \cdots & p_{1m} \\ \hline p_{21} & p_{22} & \cdots & p_{2m} \\ \hline \vdots & \vdots & & \vdots \\ \hline p_{m1} & p_{m2} & \cdots & p_{mm} \\ \hline\end{array} \qquad \mathbf{e} = \begin{array}{|c|}\hline 1 \\ \hline 1 \\ \hline \vdots \\ \hline 1 \\ \hline\end{array}$$

The state probabilities $p_{j,n}$ of a finite Markov process at time t_n are arranged in a stochastic vector $\mathbf{p}_n$, which corresponds to a probability function. The transition probabilities are arranged in a stochastic matrix **P**, called the transition matrix. This leads to the following rule of calculation :

$$\mathbf{p}_{n+1} = \mathbf{P}^T \mathbf{p}_n$$

Homogeneous Markov process : A finite Markov process is said to be inhomogeneous if the transition probabilities are time-dependent. It is said to be homogeneous if the transition probabilities are time-independent. A homogeneous Markov process is completely described by the initial distribution $\mathbf{p}_0$ at the initial time t_0 and the constant transition matrix **P** for the time domain under consideration. The distributions $\mathbf{p}_1, \mathbf{p}_2, \ldots$ for the later time points $t_1, t_2, \ldots$ are obtained recursively according to the above rule of calculation.

Limit distribution : Starting from an initial distribution $\mathbf{p}_0$, the distributions $\mathbf{p}_n$ of a homogeneous Markov process may tend to a limit distribution $\mathbf{p}_\infty$ in the limit $n \to \infty$.

$$\lim_{n \to \infty} \mathbf{p}_n = \mathbf{p}_\infty$$

A limit distribution need not exist. If a limit distribution does exist, it may or may not depend on the initial distribution $\mathbf{p}_0$. The conditions for the existence of a limit distribution and for its dependence on the initial distribution are treated in the following sections.

Equilibrium distribution : A homogeneous Markov process is in equilibrium at a given time point if the distribution does not change at the next time point. Such a distribution is called an equilibrium distribution (stationary distribution) and is designated by $\overset{*}{\mathbf{p}}$. The rule of calculation yields the following equilibrium condition :

$$\overset{*}{\mathbf{p}} = \mathbf{P}^{\mathrm{T}}\overset{*}{\mathbf{p}}$$

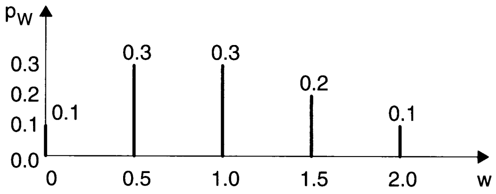
A distribution $\overset{*}{\mathbf{p}}$ is determined from the equilibrium condition and the distribution conditions. This leads to a homogeneous system of linear equations with constraints :

$$(\mathbf{P}^{\mathrm{T}} - \mathbf{I})\overset{*}{\mathbf{p}} = \mathbf{0} \qquad \mathbf{e}^{\mathrm{T}}\overset{*}{\mathbf{p}} = 1 \qquad \overset{*}{\mathbf{p}} \geq \mathbf{0}$$

If the distributions for a homogeneous Markov process tend to a limit distribution $\mathbf{p}_{\infty}$ with increasing time, then this limit distribution is an equilibrium distribution. If the initial distribution $\mathbf{p}_0$ is an equilibrium distribution, the homogeneous Markov process is a stationary process, that is its distribution is time-independent.

Example : Water management for a reservoir

The water management for a reservoir used for agricultural irrigation is treated in a simplified form as a homogeneous Markov process. The irrigation annually requires an amount V of water. Let the reservoir be designed for a capacity V. The amount W of water which the reservoir collects from affluents and precipitation is described as a multiple of the required amount V. The annual input W is a random variable, for which the following simplified probability function is assumed :



An amount V of water is annually extracted from the reservoir if this is possible. If the capacity V of the reservoir is exceeded, the excess amount of water is released. Due to this procedure and the simplifying assumptions about the input, the reservoir is either empty, half-full or full at the end of each year. The amount Q of water in the reservoir is referred to the capacity V.

empty reservoir	Q = 0.0
half-full reservoir	Q = 0.5
full reservoir	Q = 1.0

Let the annual inputs in consecutive years be stochastically independent. Under this assumption, the probabilities for the transitions of the states of the reservoir in two consecutive years n and n + 1 are determined from the probability function for the input W.

- Let the reservoir be empty. It remains empty if it receives an input $W \le 1.0$. It becomes half-full if it receives the input $W = 1.5$. It becomes full if it receives an input $W \ge 2.0$.

$$P(Q(n+1) = 0.0 \mid Q(n) = 0.0) = P(W \le 1.0) = 0.70$$

$$P(Q(n+1) = 0.5 \mid Q(n) = 0.0) = P(W = 1.5) = 0.20$$

$$P(Q(n+1) = 1.0 \mid Q(n) = 0.0) = P(W \ge 2.0) = 0.10$$

- Let the reservoir be half-full. It becomes empty if it receives an input $W \le 0.5$. It remains half-full if it receives the input $W = 1.0$. It becomes full if it receives an input $W \ge 1.5$.

$$P(Q(n+1) = 0.0 \mid Q(n) = 0.5) = P(W \le 0.5) = 0.40$$

$$P(Q(n+1) = 0.5 \mid Q(n) = 0.5) = P(W = 1.0) = 0.30$$

$$P(Q(n+1) = 1.0 \mid Q(n) = 0.5) = P(W \ge 1.5) = 0.30$$

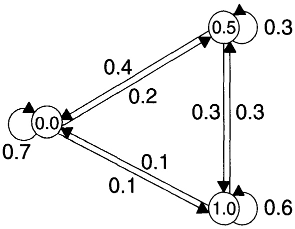
- Let the reservoir be full. It becomes empty if it receives the input $W = 0.0$. It becomes half-full if it receives the input $W = 0.5$. It remains full if it receives an input $W \ge 1.0$.

$$P(Q(n+1) = 0.0 \mid Q(n) = 1.0) = P(W \le 0.0) = 0.10$$

$$P(Q(n+1) = 0.5 \mid Q(n) = 1.0) = P(W = 0.5) = 0.30$$

$$P(Q(n+1) = 1.0 \mid Q(n) = 1.0) = P(W \ge 1.0) = 0.60$$

The possible state transitions and the transition probabilities between the states $Q = 0.0, 0.5, 1.0$ are shown in the transition graph and in the transition matrix **P**.



$$\mathbf{P} = \begin{array}{|c|c|c|c}
\multicolumn{1}{c}{0.0} & \multicolumn{1}{c}{0.5} & \multicolumn{1}{c}{1.0} & \\
\hline
0.7 & 0.2 & 0.1 & 0.0 \\
\hline
0.4 & 0.3 & 0.3 & 0.5 \\
\hline
0.1 & 0.3 & 0.6 & 1.0 \\
\hline
\end{array}$$

In the first year n = 0 after the reservoir is built, it is in the empty state Q = 0.0. The initial distribution $\mathbf{p}_0$ is thus given by the unit vector for Q = 0.0. The distribution $\mathbf{p}_n$ for the different states in the subsequent years is calculated recursively according to the rule of calculation for homogeneous Markov processes.

$$\mathbf{p}_{n+1} = \mathbf{P}^T \mathbf{p}_n \qquad \begin{array}{cccccc} \mathbf{p}_0 & \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & & \mathbf{p}_\infty \\ \begin{bmatrix} 1.000 \\ 0.000 \\ 0.000 \end{bmatrix} & \begin{bmatrix} 0.700 \\ 0.200 \\ 0.100 \end{bmatrix} & \begin{bmatrix} 0.580 \\ 0.230 \\ 0.190 \end{bmatrix} & \begin{bmatrix} 0.517 \\ 0.242 \\ 0.241 \end{bmatrix} & \cdots & \begin{bmatrix} 0.442 \\ 0.256 \\ 0.302 \end{bmatrix} \end{array}$$

$$\begin{array}{cccccccc} \begin{bmatrix} 0.7 & 0.4 & 0.1 \\ 0.2 & 0.3 & 0.3 \\ 0.1 & 0.3 & 0.6 \end{bmatrix} & \begin{bmatrix} 0.700 \\ 0.200 \\ 0.100 \end{bmatrix} & \begin{bmatrix} 0.580 \\ 0.230 \\ 0.190 \end{bmatrix} & \begin{bmatrix} 0.517 \\ 0.242 \\ 0.241 \end{bmatrix} & \begin{bmatrix} 0.483 \\ 0.248 \\ 0.269 \end{bmatrix} & \cdots & \begin{bmatrix} 0.442 \\ 0.256 \\ 0.302 \end{bmatrix} \\ \mathbf{P}^T & \mathbf{p}_1 & \mathbf{p}_2 & \mathbf{p}_3 & \mathbf{p}_4 & & \mathbf{p}_\infty \end{array}$$

In the limit n → ∞, the distribution tends to a fixed probability distribution for the different states of the reservoir. In this limit, the reservoir is empty with probability 44.2%, half-full with probability 25.6% and full with probability 30.2%.

The distribution in the limit n → ∞ is an equilibrium distribution. It may be determined directly as the solution of a system of linear equations :

$$(\mathbf{P}^T - \mathrm{I})\overset{*}{\mathbf{p}} = \mathbf{0}$$

$$\begin{bmatrix} -0.3 & 0.4 & 0.1 \\ 0.2 & -0.7 & 0.3 \\ 0.1 & 0.3 & -0.4 \end{bmatrix} * \begin{bmatrix} \overset{*}{p}_1 \\ \overset{*}{p}_2 \\ \overset{*}{p}_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \qquad \overset{*}{\mathbf{p}} = \frac{c}{13} \begin{bmatrix} 19 \\ 11 \\ 13 \end{bmatrix} \qquad \overset{*}{\mathbf{p}} = \frac{1}{43} \begin{bmatrix} 19 \\ 11 \\ 13 \end{bmatrix}$$

The system of linear equations has non-negative solutions proportional to a constant c. The constant c is determined such that the sum of all solutions is 1. This leads to the state probabilities $\overset{*}{\mathbf{p}}$ for the equilibrium distribution. The homogeneous Markov process has exactly one equilibrium distribution.

10.5.2.3 Structural analysis

Introduction : A homogeneous Markov process is represented by the transition graph for the possible changes of state. The analysis of the structure of a Markov process is based on graph theory and does not depend on the transition probabilities. The essential structural properties of homogeneous Markov processes are treated in the following.

Reachability : A state y is said to be reachable from a state x if there is an edge sequence from x to y in the transition graph. The length of the edge sequence is the number of its edges. If x and y are identical, then there exists an edge sequence of length 0.

Strong connectedness : Two states x and y are said to be strongly connected if x is reachable from y and y is reachable from x in the transition graph. The strong connectedness relation is an equivalence relation. It partitions the space of states into equivalence classes. For Markov processes, such an equivalence class is called a class of states.

Class of states : Every class of states is represented by a state x. A state y belongs to the class with the representative x if x and y are strongly connected. Every state in a class is strongly connected with every state in the same class. None of the states in a class is strongly connected with a state in another class.

Reduction : The space of states S is partitioned into classes of states. These classes are disjoint subsets of S. If there is at least one transition from a state in class A to a state in class B, then there is also a class transition from A to B. In the reduced transition graph, the classes of states are represented as vertices and the class transitions are represented as directed edges. The definition of reachability is transferred to classes of states in the reduced transition graph. Two different classes of states are not mutually reachable and are therefore not strongly connected. The reduced transition graph does not contain cycles.

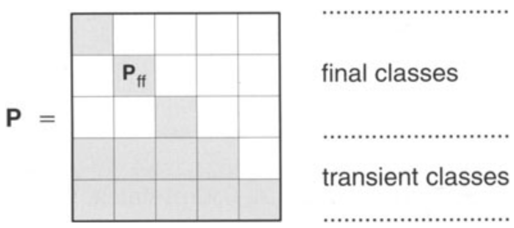
Transient class of states : A class of states is said to be transient if at least one other class is reachable from this class in the reduced transition graph. If a process ever leaves a transient class, it can never return to that class.

Final class of states : A class of states is said to be final if no other class is reachable from this class in the reduced transition graph. If a process ever reaches a final class, it can never leave that class.

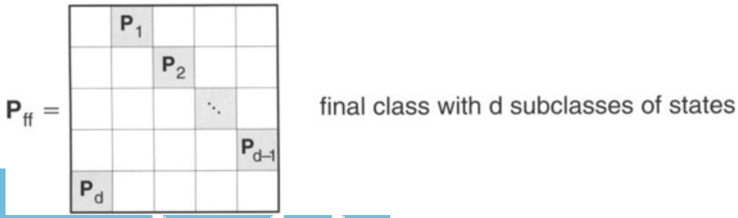
Periodicity of final classes of states : Within a final class, a process may return to a state along various edge sequences. Each of these edge sequences is a cycle. The greatest common divisor of the lengths of all possible cycles is called the period and is designated by d. The period is the same for all states in a final class, and is therefore a property of the final class. A final class is said to be aperiodic if $d = 1$ and periodic if $d > 1$.

A final class of states with a period $d > 1$ may be partitioned into d disjoint subclasses of states. Every subclass of states is represented by a state x. A state y belongs to the subclass of states with the representative x if and only if y can be reached from x along an edge sequence whose length is a positive integer multiple of the period d. There are no transitions between two states in the same subclass. The subclasses of states and their class transitions form an elementary cycle. If a process leaves a state in a subclass, a state in the same subclass is reached again after exactly d steps.

Matrix structure : A submatrix scheme for the transition matrix **P** is obtained by arranging the states of a Markov process in subvectors according to their class. Since no other class $k \neq f$ is reachable from a final class f and two different transient classes t_1, t_2 are not mutually reachable, the transition matrix has the following standard submatrix structure for a suitable arrangement of the classes :

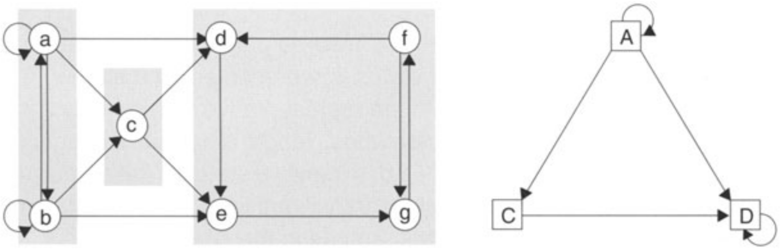


A final class f with the period $d > 1$ is partitioned into d disjoint subclasses. A submatrix scheme for the transition matrix $\mathbf{P}_{ff}$ is obtained by arranging the states in subvectors according to their subclass. Since exactly one other subclass $s_2 \neq s_1$ is reachable from a subclass s_1, the transition matrix of a final class has the following standard submatrix structure for a suitable arrangement of the subclasses :



Example : Structural analysis of a transition graph

Let the following transition graph for a homogeneous Markov process be given. The vertices of the transition graph are the possible states. The edges of the transition graph are the possible transitions between two states.

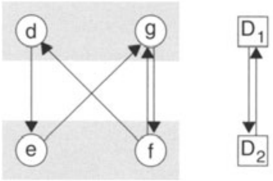


space of states	$S = \{a, b, c, d, e, f, g\}$		
classes of states	$A = \{a, b\}$	$C = \{c\}$	$D = \{d, e, f, g\}$

The vertex set of the transition graph is partitioned into classes of states. The state a is chosen as a representative of the class A. All states from a to g are reachable from the state a. The state a is reachable from the states a, b. Since the states a, b are mutually reachable, they form the class A. The state c is chosen as a representative of the class C. The states c to g are reachable from the state c. The state c is reachable only from itself. Thus it forms the class C by itself. The state d is chosen as a representative of the class D. The states d to g are reachable from the state d. The state d is reachable from all states. Since the states d, e, f, g are mutually reachable, they form the class D.

The space of states S is partitioned into the classes A, C, D of states. The reduced transition graph with the classes A, C, D is shown above. The classes A and C are transient, since the class D can be reached from them. The class D is final, since no other class can be reached from it.

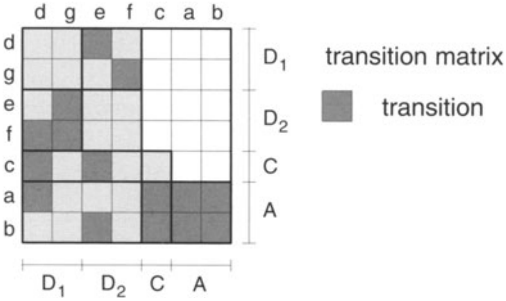
The period of the final class D is to be determined. For this purpose, the vertex d is chosen as a representative of D, and the possible ways of returning to this vertex are determined. A return to the vertex d may be accomplished along the cycles < d, e, g, f, d >, < d, e, g, f, g, f, d >, < d, e, g, f, d, e, g, f, d >,... of length 4, 6, 8, ... The greatest common divisor of these lengths is the period 2. The final class D may therefore be partitioned into two subclasses D_1 and D_2. The state d is chosen as a representative of the subclass D_1. From the state d, the states g and d can be reached via edge sequences whose lengths are a positive integer multiple of $d = 2$. Thus the states d,g form the subclass D_1. The two remaining states e,f form the subclass D_2. In a reduced transition graph for the final class D, the two subclasses form an elementary cycle.



final class of states $D = \{d, e, f, g\}$

subclasses of states $D_1 = \{d, g\}$ $D_2 = \{e, f\}$

The transition matrix associated with the transition graph assumes a standardized form if the states are arranged according to their classes and subclasses :



Structural analysis : A finite homogeneous Markov process is said to be irreducible if its space of states cannot be partitioned into several classes of states. It is said to be reducible if its space of states can be partitioned into several classes of states. An irreducible finite Markov process possesses exactly one final class of states. A reducible finite Markov process possesses at least one final class of states and zero, one or more transient classes of states.

The structural analysis leads to the following important consequences for the long-term behavior of reducible finite Markov processes :

- The sum of the probabilities for all states in a transient class tends to 0 with increasing time, since a transient class cannot be reached once the process has left it.
- The sum of the probabilities for all states in a final class tends to a fixed value with increasing time, since a final class is never left once the process has reached it.

10.5.2.4 Spectral analysis

Introduction : The rule of calculation for homogeneous Markov processes suggests an iterative procedure for determining the state probabilities in consecutive time steps. The long-term behavior of the process depends on the convergence behavior of the iterative procedure. This convergence behavior is determined by the eigenvalues and eigenvectors of the transition matrix. The determination of the eigenvalues and eigenvectors leads to a spectral analysis. The spectral analysis is treated in the following for the special case of a transition matrix without multiple eigenvalues. The results of the spectral analysis for the general case of the long-term behavior of the process are compiled.

Eigenvalue problem : The eigenvalue problems for the quadratic stochastic matrices $\mathbf{P}$ and $\mathbf{P}^T$ are :

$$\mathbf{P}\mathbf{y} = \lambda\mathbf{y} \qquad\qquad \mathbf{P}^T\mathbf{z} = \lambda\mathbf{z}$$

The matrices $\mathbf{P}$ and $\mathbf{P}^T$ have the same eigenvalues. However, they generally have different eigenvectors $\mathbf{y}$ and $\mathbf{z}$ for the eigenvalues λ. The eigenvectors $\mathbf{y}$ are called right eigenvectors of $\mathbf{P}$. The eigenvectors $\mathbf{z}$ are called left eigenvectors of $\mathbf{P}$, since the eigenvalue problem for $\mathbf{P}^T$ is transformed into $\mathbf{z}^T\mathbf{P} = \lambda\mathbf{z}^T$ by transposition and in this form the eigenvector $\mathbf{z}^T$ appears to the left of $\mathbf{P}$.

Eigenvalues : According to the rules of algebra, the absolute value $|\lambda|$ of each eigenvalue is less than or equal to a matrix norm $\|\mathbf{P}\|$. Since the sum of the absolute values of all elements of a stochastic matrix $\mathbf{P}$ for every row is 1, the row norm $\|\mathbf{P}\|$ is 1. Hence all eigenvalues of $\mathbf{P}$ are less than or equal to 1. The eigenvalue λ_1 with the largest absolute value is exactly 1, since by definition the equation $\mathbf{P}\,\mathbf{e} = 1\,\mathbf{e}$ holds for a stochastic matrix.

$$|\lambda| \le 1 \qquad\qquad \lambda_1 = 1$$

The eigenvalues λ_j of the matrix $\mathbf{P}$ are determined as the zeros of the characteristic polynomial $\det(\mathbf{P} - \lambda\mathbf{I}) = 0$ with the identity matrix $\mathbf{I}$. The eigenvalues may be simple or multiple, and they may be real or occur in conjugate complex pairs.

Eigenvectors : If all eigenvalues are simple, then for every eigenvalue λ_j there is a right eigenvector $\mathbf{y}_j$ and a left eigenvector $\mathbf{z}_j$. They are solutions of the following homogeneous systems of linear equations :

$$(\mathbf{P} - \lambda_j\mathbf{I})\mathbf{y}_j = \mathbf{0} \qquad\qquad (\mathbf{P}^T - \lambda_j\mathbf{I})\mathbf{z}_j = \mathbf{0}$$

The right eigenvector $\mathbf{y}_1$ of the stochastic matrix $\mathbf{P}$ belonging to the maximal eigenvalue $\lambda_1 = 1$ is the one vector $\mathbf{e}$, since $\mathbf{P}\mathbf{e} = 1\mathbf{e}$.

$$\mathbf{y}_1 = \mathbf{e} \qquad\qquad \lambda_1 = 1$$

The eigenvectors $\mathbf{y}_j$ and $\mathbf{z}_j$ for each eigenvalue λ_j are normalized such that $\mathbf{y}_j^T \mathbf{z}_j = 1$. The eigenvectors $\mathbf{y}_j$ and $\mathbf{z}_k$ belonging to different eigenvalues λ_j and λ_k are orthogonal, so that

$$\mathbf{y}_j^T \mathbf{z}_k = \delta_{jk} \qquad\qquad \delta_{jk} = \begin{cases} 0 \text{ for } j \neq k \\ 1 \text{ for } j = k \end{cases}$$

Spectral decomposition : If all eigenvalues are simple, the matrices **P** and $\mathbf{P}^T$ have the following spectral decompositions :

$$\mathbf{P} = \sum_{j=1}^{m} \lambda_j \mathbf{y}_j \mathbf{z}_j^T \qquad\qquad \mathbf{P}^T = \sum_{j=1}^{m} \lambda_j \mathbf{z}_j \mathbf{y}_j^T$$

If the spectral decompositions of **P** and $\mathbf{P}^T$ are substituted into the equations for the eigenvectors, these equations are satisfied due to the orthonormality properties of the eigenvectors. The following spectral decompositions for the n-th powers of the matrices **P** and $\mathbf{P}^T$ are obtained using these orthonormality properties :

$$\mathbf{P}^n = \sum_{j=1}^{m} \lambda_j^n \mathbf{y}_j \mathbf{z}_j^T \qquad\qquad (\mathbf{P}^n)^T = \sum_{j=1}^{m} \lambda_j^n \mathbf{z}_j \mathbf{y}_j^T$$

If apart from the maximal eigenvalue $\lambda_1 = 1$ the absolute values of all remaining eigenvalues λ_j for $j > 1$ are less than 1, the powers λ_j^n tend to 0 for $n \to \infty$. Hence the matrices $\mathbf{P}^n$ and $(\mathbf{P}^n)^T$ tend to fixed limits :

$$\mathbf{P}^\infty = \mathbf{e}\,\mathbf{z}_1^T \qquad\qquad (\mathbf{P}^\infty)^T = \mathbf{z}_1\,\mathbf{e}^T$$

Spectral analysis : The stochastic vector $\mathbf{p}_n$ at time t_n for a homogeneous Markov process with the transition matrix **P** is calculated iteratively starting with the stochastic vector $\mathbf{p}_0$ at time t_0 :

$$\begin{aligned} \mathbf{p}_1 &= \mathbf{P}^T \mathbf{p}_0 & & \\ \mathbf{p}_2 &= \mathbf{P}^T \mathbf{p}_1 &=& \ (\mathbf{P}^2)^T \mathbf{p}_0 \\ &\vdots & & \\ \mathbf{p}_n &= \mathbf{P}^T \mathbf{p}_{n-1} &=& \ (\mathbf{P}^n)^T \mathbf{p}_0 \end{aligned}$$

If the eigenvalues of the transition matrix **P** are simple, then substituting the spectral decomposition of $(\mathbf{P}^n)^T$ into this rule yields the following sum for the stochastic vector $\mathbf{p}_n$:

$$\mathbf{p}_n = (\mathbf{P}^n)^T \mathbf{p}_0 = \sum_{j=1}^{m} \lambda_j^n \mathbf{z}_j \mathbf{y}_j^T \mathbf{p}_0$$

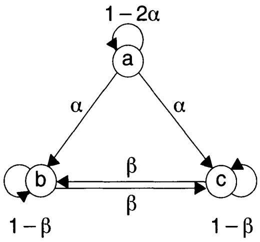
If the absolute values of all eigenvalues λ_j for $j > 1$ are less than 1, the stochastic vector $\mathbf{p}_n$ tends to the left eigenvector $\mathbf{z}_1$ of **P** for $n \to \infty$ independent of $\mathbf{p}_0$.

$$\mathbf{p}_\infty = (\mathbf{P}^\infty)^T \mathbf{p}_0 = \mathbf{z}_1 \mathbf{e}^T \mathbf{p}_0 = \mathbf{z}_1$$

The left eigenvector $\mathbf{z}_1$ is a stochastic vector, since $\mathbf{y}_1^T \mathbf{z}_1 = \mathbf{e}^T \mathbf{z}_1 = 1$ due to normalization.

Example : Spectral analysis

Let the homogeneous Markov process with the illustrated weighted transition graph and the corresponding transition matrix **P** be given.



$$\mathbf{P} \;=\; \begin{array}{|c|c|c|c} a & b & c & \\ \hline 1-2\alpha & \alpha & \alpha & a \\ \hline 0 & 1-\beta & \beta & b \\ \hline 0 & \beta & 1-\beta & c \\ \hline \end{array}$$

$$0 < \alpha < \frac{1}{2} \qquad 0 < \beta \le 1$$

The space of states of the homogeneous Markov process consists of a transient class of states {a} and a final class of states {b, c}. The zeros of the characteristic polynomial are the eigenvalues λ_j of the transition matrix **P**.

$$\det(\mathbf{P} - \lambda\mathbf{I}) \;=\; (1 - 2\alpha - \lambda)\,((1 - \beta - \lambda)^2 - \beta^2) \;=\; 0$$

$$\text{eigenvalues} \;:\; \lambda_1 \;=\; 1 \qquad \lambda_2 \;=\; 1 - 2\beta \qquad \lambda_3 \;=\; 1 - 2\alpha$$

The right and left eigenvectors $\mathbf{y}_j$ and $\mathbf{z}_j$ for the eigenvalues λ_j are determined by solving the homogeneous systems of linear equations $(\mathbf{P} - \lambda_j\,\mathbf{I})\,\mathbf{y}_j = \mathbf{0}$ and $(\mathbf{P}^T - \lambda_j\mathbf{I})\,\mathbf{z}_j = \mathbf{0}$.

$$\lambda_1 = 1 \qquad \mathbf{y}_1^T = \begin{array}{|c|c|c|}\hline 1.0 & 1.0 & 1.0 \\ \hline\end{array} \qquad \mathbf{z}_1^T = \begin{array}{|c|c|c|}\hline 0.0 & 0.5 & 0.5 \\ \hline\end{array}$$

$$\lambda_2 = 1 - 2\beta \qquad \mathbf{y}_2^T = \begin{array}{|c|c|c|}\hline 0.0 & 1.0 & -1.0 \\ \hline\end{array} \qquad \mathbf{z}_2^T = \begin{array}{|c|c|c|}\hline 0.0 & 0.5 & -0.5 \\ \hline\end{array}$$

$$\lambda_3 = 1 - 2\alpha \qquad \mathbf{y}_3^T = \begin{array}{|c|c|c|}\hline 1.0 & 0.0 & 0.0 \\ \hline\end{array} \qquad \mathbf{z}_3^T = \begin{array}{|c|c|c|}\hline 1.0 & -0.5 & -0.5 \\ \hline\end{array}$$

The right and left eigenvectors are orthonormal. The scalar products $\mathbf{y}_j^T\mathbf{z}_j$ are 1. The scalar products $\mathbf{y}_j^T\mathbf{z}_k$ are 0 for $j \ne k$. The spectral decomposition for the transition matrix **P** is given by :

$$\mathbf{P} \;=\; \sum_{j=1}^{3} \lambda_j\,\mathbf{y}_j\,\mathbf{z}_j^T \;=\; \mathbf{y}_1\,\mathbf{z}_1^T + (1 - 2\beta)\,\mathbf{y}_2\,\mathbf{z}_2^T + (1 - 2\alpha)\,\mathbf{y}_3\,\mathbf{z}_3^T$$

$$\mathbf{P} \;=\; \begin{array}{|c|c|c|}\hline 0.0 & 0.5 & 0.5 \\ \hline 0.0 & 0.5 & 0.5 \\ \hline 0.0 & 0.5 & 0.5 \\ \hline\end{array} \;+\; (1{-}2\beta)\,\begin{array}{|c|c|c|}\hline 0.0 & 0.0 & 0.0 \\ \hline 0.0 & 0.5 & -0.5 \\ \hline 0.0 & -0.5 & 0.5 \\ \hline\end{array} \;+\; (1{-}2\alpha)\,\begin{array}{|c|c|c|}\hline 1.0 & -0.5 & -0.5 \\ \hline 0.0 & 0.0 & 0.0 \\ \hline 0.0 & 0.0 & 0.0 \\ \hline\end{array}$$

The spectral decomposition for the power $\mathbf{P}^n$ of the transition matrix is given by :

$$\mathbf{P}^n \;=\; \sum_{j=1}^{3} \lambda_j^n\,\mathbf{y}_j\,\mathbf{z}_j \;=\; \mathbf{y}_1\,\mathbf{z}_1^T + (1 - 2\beta)^n\,\mathbf{y}_2\,\mathbf{z}_2^T + (1 - 2\alpha)^n\,\mathbf{y}_3\,\mathbf{z}_3^T$$

For $0 < \beta < 1$, the final class {b, c} is aperiodic. The absolute values of the eigenvalues $\lambda_2 = 1 - 2\beta$ and $\lambda_3 = 1 - 2\alpha$ are less than 1. The powers λ_2^n, λ_3^n tend to 0 for $n \rightarrow \infty$, so that the matrix $\mathbf{P}^n$ tends to the matrix $\mathbf{P}^{\infty}$:

$$\mathbf{P}^{\infty} = \mathbf{y}_1 \mathbf{z}_1^T$$

For $\beta = 1$, the final class {b, c} is periodic with the period $d = 2$. The absolute values of the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = -1$ are equal to 1. The matrix $\mathbf{P}^n$ does not tend to a fixed limit for $n \rightarrow \infty$. For sufficiently large n, it alternates in consecutive steps :

$$\mathbf{P}^n = \mathbf{y}_1 \mathbf{z}_1^T + (-1)^n \mathbf{y}_2 \mathbf{z}_2^T$$

General spectral analysis : In the general case, the transition matrix **P** of a homogeneous Markov process has simple and multiple eigenvalues. The theoretical foundations for eigenvalue problems with multiple eigenvalues are a part of linear algebra. The results required for the structural analysis are compiled in the following. If the space of states of the homogeneous Markov process contains several classes of states, then by virtue of the structure of the transition matrix **P** the eigenvalues may be determined classwise. The essential results are :

- The absolute values of all eigenvalues of a transition matrix $\mathbf{P}_T$ for the states within a transient class of states are less than 1.
- A transition matrix $\mathbf{P}_F$ for the states within a final class of states has an eigenvalue 1.
- If a final class of states is aperiodic, then apart from the eigenvalue 1 the transition matrix $\mathbf{P}_F$ has only eigenvalues with absolute values less than 1.
- If a final class of states is periodic with a period $d > 1$, then the transition matrix $\mathbf{P}_F$ has exactly d eigenvalues with the absolute value 1. The characteristic equation $\det(\mathbf{P}_F - \lambda\mathbf{I})$ contains a factor $(\lambda^d - 1)$.

In the general spectral decomposition, the n-th power of the transition matrix **P** also depends on the n-th powers of its eigenvalues λ. In the limit $n \rightarrow \infty$, the components with the eigenvalues $|\lambda| < 1$ tend to zero. The general spectral analysis of a homogeneous Markov process leads to the following essential results :

- If a Markov process contains exactly one final class of states and this class is aperiodic, then there is a unique limit distribution $\mathbf{p}_\infty$ which is independent of the initial distribution $\mathbf{p}_0$.
- If a Markov process contains several final classes of states and these classes are aperiodic, then there are different limit distributions $\mathbf{p}_\infty$ depending on the initial distribution $\mathbf{p}_0$.
- If a Markov process contains a final class of states which is periodic, then there is no limit distribution $\mathbf{p}_\infty$. The long-term behavior of the process is periodic.

10.5.2.5 First passage

Introduction : In practical applications, one is often interested in the probability that a final state z is reached for the first time from an initial state a at time t. This is called the first passage problem. The solution of this problem leads to a distribution as a function of time. Since the time domain is discrete, the distribution is also discrete. The basic definitions for the first passage problem for finite homogeneous Markov processes as well as a suitable method for calculating the time-dependent distribution and its moments are treated in the following.

Passage time : The passage time T_{ij} is the number of time steps which a homogeneous Markov process takes to reach a state j from a state $i \neq j$ for the first time. The possible values of the passage time T_{ij} may be determined directly from the transition graph. If the transition graph contains a path of length n from the state i via $n-1$ intermediate states $k \neq j$ to the state j, then n is a possible passage time.

Probability function of the passage time : If a state j can be reached from another state i, the passage time T_{ij} takes natural numbers $n \in \mathbb{N}'$ as values. The probability $P(T_{ij} = n)$ is designated by $f_{ij}(n)$. For $n = 1$, the probability $f_{ij}(1)$ coincides with the transition probability p_{ij}. For $n > 1$, the probability $f_{ij}(n)$ is calculated recursively according to the total probability theorem. In a single step, the process makes a transition from the state i to the states $k \neq j$ with the transition probabilities p_{ik} and reaches the state j for the first time in $n-1$ steps with the probabilities $f_{kj}(n-1)$.

$$\begin{array}{llll} f_{ij}(1) & = p_{ij} & n = 1 & i \neq j \\ f_{ij}(n) & = \sum\limits_{k \neq j} p_{ik}\, f_{kj}(n-1) & n > 1 & i \neq j \end{array}$$

This recursive rule is conveniently formulated using vector and matrix notation. For a given state j and all states $i \neq j$, the probabilities $f_{ij}(n)$ and p_{ij} are arranged in the vectors $\mathbf{f}_j(n)$ and $\mathbf{p}_j$. The matrix $\mathbf{P}_j$ is obtained from the matrix $\mathbf{P}$ by deleting the row j and the column j of $\mathbf{P}$.

$$\mathbf{f}_j(n) = \mathbf{P}_j\, \mathbf{f}_j(n-1) \qquad \mathbf{f}_j(1) = \mathbf{p}_j$$

For $n = 1, 2, \ldots$, the discrete function $f_{ij}(n)$ corresponds to a probability function for the passage time T_{ij}. However, in the general case it does not satisfy the condition that the sum of all probabilities $f_{ij}(n)$ for $n > 0$ are one. If the state j is not reachable from the state i, then $f_{ij}(n) = 0$ for all $n > 0$.

Moments of the passage time : The moments of the passage time T_{ij} are determined from the probabilities $f_{ij}(n)$ with $n > 0$. The k-th moment is defined as follows :

$$m_{ij}(k) := \sum_{n=1}^{\infty} n^k f_{ij}(n) \qquad\qquad k \geq 0$$

The moments may be calculated recursively without explicitly determining the probabilities $f_{ij}(n)$. For a given state j and all states $i \neq j$, the moments $m_{ij}(k)$ are arranged in a vector $\mathbf{m}_j(k)$. The moments are defined as follows :

$$\mathbf{m}_j(k) = \sum_{n=1}^{\infty} n^k \mathbf{f}_j(n) = \mathbf{p}_j + \sum_{n=2}^{\infty} n^k \mathbf{f}_j(n)$$

The recursive rule $\mathbf{f}_j(n) = \mathbf{P}_j \mathbf{f}_j(n-1)$ for $n > 1$ yields :

$$\mathbf{m}_j(k) = \mathbf{p}_j + \sum_{n=2}^{\infty} n^k \mathbf{P}_j \mathbf{f}_j(n-1) = \mathbf{p}_j + \mathbf{P}_j \sum_{n=1}^{\infty} (n+1)^k \mathbf{f}_j(n)$$

Using the binomial theorem for $(n+1)^k$ and the definitions of the moments, one obtains :

$$\mathbf{m}_j(k) = \mathbf{p}_j + \mathbf{P}_j \sum_{n=1}^{\infty} \sum_{q=0}^{k} \binom{k}{q} n^q \mathbf{f}_j(n) = \mathbf{p}_j + \mathbf{P}_j \sum_{q=0}^{k} \binom{k}{q} \mathbf{m}_j(q)$$

Thus the following system of linear equations is obtained, which determines the k-th moments of the passage time in terms of the q-th moments with $0 \leq q < k$:

$$(\mathbf{I} - \mathbf{P}_j)\, \mathbf{m}_j(k) = \mathbf{p}_j + \mathbf{P}_j \sum_{q=0}^{k-1} \binom{k}{q} \mathbf{m}_j(q)$$

If the state j cannot be reached from the state i, then since $f_{ij}(n) = 0$ for $n > 0$ all moments $m_{ij}(k)$ for $k \geq 0$ are zero.

Passage probability : The probability $P(T_{ij} > 0)$ is called a passage probability and is designated by $\overset{*}{f}_{ij}$. It is the sum of all probabilities $P(T_{ij} = n) = f_{ij}(n)$ for $n > 0$, and hence the zeroth-order moment $m_{ij}(0)$.

$$\overset{*}{f}_{ij} := P(T_{ij} > 0) = m_{ij}(0)$$

Since $\overset{*}{\mathbf{f}}_j = \mathbf{m}_j(0)$, the passage probabilities are determined in vector form according to the rule for the zeroth-order moments :

$$(\mathbf{I} - \mathbf{P}_j)\, \overset{*}{\mathbf{f}}_j = \mathbf{p}_j$$

If the transition graph contains no path from the state i to the state j, then j cannot be reached from i and $\overset{*}{f}_{ij} = 0$. If there is at least one path from the state i to the state j, then j can be reached from i and $\overset{*}{f}_{ij} > 0$. If every sufficiently long path from the state i leads to the state j, then j is always reached from i and $\overset{*}{f}_{ij} = 1$. The probability that the state j is not reached from the state i is $1 - \overset{*}{f}_{ij}$.

Average passage time : If a state j is always reached from a state i, then $\overset{*}{f}_{ij} = 1$, and hence $f_{ij}(n)$ for $n > 0$ is a probability function. In this case, the first moment $m_{ij}(1)$ is the mean μ_{ij} of the passage time and is called the average passage time.

$$\mu_{ij} := m_{ij}(1) \quad \text{for} \quad \overset{*}{f}_{ij} = 1$$

Since $\boldsymbol{\mu}_j = \mathbf{m}_j(1)$, the means are determined in vector form according to the rule for the first moments :

$$(\mathbf{I} - \mathbf{P}_j)\,\boldsymbol{\mu}_j = \mathbf{p}_j + \mathbf{P}_j\,\overset{*}{\mathbf{f}}_j \qquad \Leftrightarrow \qquad (\mathbf{I} - \mathbf{P}_j)\boldsymbol{\mu}_j = \overset{*}{\mathbf{f}}_j$$

Recurrence time : The recurrence time T_{jj} is the number of time steps which a homogeneous Markov process takes to return to a state j for the first time after leaving it. The possible values of the recurrence times are determined directly in the transition graph. If the transition graph contains a cycle of length n from state j via $n-1$ intermediate states $i \neq j$ to state j, then n is a possible recurrence time.

Probability function of the recurrence time : If a state j can be reached after the process has left it, the recurrence time is a natural number $n \in \mathbb{N}'$. The probability $P(T_{jj} = n)$ is designated by $f_{jj}(n)$. For $n = 1$, the probability $f_{jj}(1)$ coincides with the transition probability p_{jj}. For $n > 1$, the probability $f_{jj}(n)$ is calculated from the probabilities for the passage times according to the total probability theorem. In a single step, the process makes a transition from the state j to the states $i \neq j$ with the transition probabilities p_{ji} and returns to the state j for the first time in $n-1$ steps with the probabilities $f_{ij}(n-1)$.

$$\begin{aligned} f_{jj}(1) &= p_{jj} && n = 1 \\ f_{jj}(n) &= \sum_{i \neq j} p_{ji}\, f_{ij}(n-1) && n > 0 \end{aligned}$$

This rule is conveniently formulated in vector notation. For a given state j and all states $i \neq j$, the probabilities p_{ji} and $f_{ij}(n-1)$ are arranged in the vectors $\mathbf{q}_j^T$ and $\mathbf{f}_j(n-1)$.

$$f_{jj}(1) = p_{jj} \qquad f_{jj}(n) = \mathbf{q}_j^T \mathbf{f}_j(n-1) \qquad n > 1$$

The discrete function $f_{jj}(n)$ for $n = 1,2,...$ corresponds to a probability function for the recurrence time T_{jj}. However, in the general case it does not satisfy the condition that the sum of all probabilities $f_{jj}(n)$ for $n > 0$ is one. If the state j cannot be reached again after the process has left it, then $f_{jj}(n) = 0$ for every $n > 0$.

Moments of the recurrence time : The moments of the recurrence time T_{jj} are determined from the probabilities $f_{jj}(n)$ with $n > 0$. The k-th moment is defined as follows :

$$m_{jj}(k) := \sum_{n=1}^{\infty} n^k f_{jj}(n)$$

The moments of the recurrence time are calculated directly from the moments of the passage times. The rule of calculation is derived from the definition of the moments :

$$m_{jj}(k) = \sum_{n=1}^{\infty} n^k f_{jj}(n) = p_{jj} + \sum_{n=2}^{\infty} n^k f_{jj}(n)$$

Applying the rule $f_{jj}(n) = \mathbf{q}_j^T \mathbf{f}_j(n-1)$ for $n > 0$ yields :

$$m_{jj}(k) = p_{jj} + \sum_{n=2}^{\infty} n^k \mathbf{q}_j^T \mathbf{f}_j(n-1) = p_{jj} + \mathbf{q}_j^T \sum_{n=1}^{\infty} (n+1)^k \mathbf{f}_j(n)$$

Using the binomial theorem for $(n+1)^k$ and the vector $\mathbf{m}_j(k)$ of the moments of the passage time, one obtains :

$$m_{jj}(k) = p_{jj} + \mathbf{q}_j^T \sum_{n=1}^{\infty} \sum_{q=0}^{k} \binom{k}{q} n^q \mathbf{f}_j(n)$$

$$m_{jj}(k) = p_{jj} + \mathbf{q}_j^T \sum_{q=0}^{k} \binom{k}{q} \mathbf{m}_j(k)$$

Recurrence probability : The probability $P(T_{jj} > 0)$ is called a recurrence probability and is designated by $\overset{*}{f}_{jj}$. It is the sum of all probabilities $P(T_{jj} = n) = f_{jj}(n)$ for $n > 0$, and hence the zeroth-order moment $m_{jj}(0)$.

$$\overset{*}{f}_{jj} := P(T_{jj} > 0) = m_{jj}(0)$$

The recurrence probability is determined from the passage probabilities $\overset{*}{\mathbf{f}}_j = \mathbf{m}_j(0)$ in vector form as follows :

$$\overset{*}{f}_{jj} = p_{jj} + \mathbf{q}_j^T \overset{*}{\mathbf{f}}_j$$

If the transition graph does not contain a cycle through the state j, then the process cannot return to j and $\overset{*}{f}_{jj} = 0$. If the transition graph contains at least one cycle through the state j, then the process can return to j and $\overset{*}{f}_{jj} > 0$. If the transition graph contains only cycles through the state j, then the process always returns to j and $\overset{*}{f}_{jj} = 1$. The probability that the state j is not reached again is $1 - \overset{*}{f}_{jj}$.

Average recurrence time : If the process always returns to a state j, then $\overset{*}{f}_{jj} = 1$, and hence $f_{jj}(n)$ for $n > 0$ is a probability function. In this case, the first moment $m_{jj}(1)$ is the mean μ_{jj} of the recurrence time and is called the average recurrence time.

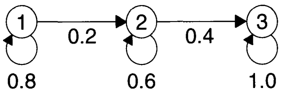
$$\mu_{jj} := m_{jj}(1) \quad \text{for} \quad \overset{*}{f}_{jj} = 1$$

The average recurrence time is obtained from the passage probabilities $\overset{*}{\mathbf{f}}_j = \mathbf{m}_j(0)$ and the average passage times $\boldsymbol{\mu}_j = \mathbf{m}_j(1)$ in vector form :

$$\mu_{jj} = p_{jj} + \mathbf{q}_j^T(\overset{*}{\mathbf{f}}_j + \boldsymbol{\mu}_j) = \overset{*}{f}_{jj} + \mathbf{q}_j^T \boldsymbol{\mu}_j$$

Example : Deterioration process

The surface of a road is damaged over the years by various random influences. The damage is described in a simplified manner by the deterioration states 1,2,3. The deterioration process is treated as a homogeneous Markov process with the illustrated weighted transition graph and the corresponding transition matrix.



$$\mathbf{P} = \begin{array}{|c|c|c|}\hline 0.8 & 0.2 & 0 \\ \hline 0 & 0.6 & 0.4 \\ \hline 0 & 0 & 1 \\ \hline \end{array} = \begin{array}{|c|c|}\hline \mathbf{P}_3 & \mathbf{p}_3 \\ \hline \mathbf{q}_3^T & 1 \\ \hline \end{array}$$

State 3 can first be reached from state 1 in the years $T_{13} = 2, 3, 4, \ldots$; it can first be reached from state 2 in the years $T_{23} = 1, 2, 3, \ldots$. The probabilities $f_{13}(n)$, $f_{23}(n)$ are arranged in the vector $\mathbf{f}_3(n)$ and calculated recursively using the transition matrix $\mathbf{P}_3$ and the transition vector $\mathbf{p}_3$.

$$\mathbf{f}_3(n) = \mathbf{P}_3\, \mathbf{f}_3(n-1) \qquad \qquad \mathbf{f}_3(1) = \mathbf{p}_3$$

The calculation of the vectors $\mathbf{f}_3(n)$ is shown :

		$\mathbf{p}_3 = \mathbf{f}_3(1)$	$\mathbf{f}_3(2)$	$\mathbf{f}_3(3)$	$\mathbf{f}_3(4)$	$\mathbf{f}_3(5)$	$\mathbf{f}_3(6)$
		0.000	0.080	0.112	0.118	0.112	0.100
		0.400	0.240	0.144	0.086	0.052	0.031
0.8	0.2	0.080	0.112	0.118	0.112	0.100	0.086
0	0.6	0.240	0.144	0.086	0.052	0.031	0.019
$\mathbf{P}_3$		$\mathbf{f}_3(2)$	$\mathbf{f}_3(3)$	$\mathbf{f}_3(4)$	$\mathbf{f}_3(5)$	$\mathbf{f}_3(6)$	$\mathbf{f}_3(7)$

The zeroth-order moments $\overset{*}{f}_{i3}$ and the first-order moments μ_{i3} for $i = 1, 2$ are arranged in the vectors $\overset{*}{\mathbf{f}}_3$ and $\boldsymbol{\mu}_3$. They are calculated by solving systems of linear equations :

$$(\mathbf{I} - \mathbf{P}_3) * \overset{*}{\mathbf{f}}_3 = \mathbf{p}_3 \qquad\qquad (\mathbf{I} - \mathbf{P}_3) * \boldsymbol{\mu}_3 = \overset{*}{\mathbf{f}}_3$$

$$\begin{bmatrix} 0.2 & -0.2 \\ 0 & 0.4 \end{bmatrix} * \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 0.0 \\ 0.4 \end{bmatrix} \qquad\qquad \begin{bmatrix} 0.2 & -0.2 \\ 0 & 0.4 \end{bmatrix} * \begin{bmatrix} 7.5 \\ 2.5 \end{bmatrix} = \begin{bmatrix} 1.0 \\ 1.0 \end{bmatrix}$$

The passage probabilities from state 1 to state 3 and from state 2 to state 3 are $\overset{*}{f}_{13} = \overset{*}{f}_{23} = 1.0$. The average passage time is $\mu_{13} = 7.5$ years from state 1 to state 3 and $\mu_{23} = 2.5$ years from state 2 to state 3.

Process behavior : The behavior of a homogeneous Markov process with respect to first passage from a state j to a state i may be analyzed on the basis of the structural properties and the probabilities $\overset{*}{f}_{ij}$. The recurrence behavior, the passage behavior and the absorption behavior are considered.

Recurrence behavior : The recurrence probability allows the states of the homogeneous Markov process to be classified. A state j with recurrence probability $\overset{*}{f}_{jj} < 1$ is said to be transient and belongs to a transient class of states. A state j with recurrence probability $\overset{*}{f}_{jj} = 1$ is said to be recurrent and belongs to a final class of states.

$$\text{state j is transient} \quad :\Leftrightarrow \quad \overset{*}{f}_{jj} < 1$$

$$\text{state j is recurrent} \quad :\Leftrightarrow \quad \overset{*}{f}_{jj} = 1$$

Passage behavior : The passage behavior is determined by the passage probabilities $\overset{*}{f}_{ij}$ for two different states $i \neq j$. The following cases are distinguished :

- The states i and j are both recurrent. They may belong either to the same final class of states or to two different final classes of states. In the first case $\overset{*}{f}_{ij} = \overset{*}{f}_{ji} = 1$, since the process cannot leave the final class and the states are always mutually reachable within their class. In the second case $\overset{*}{f}_{ij} = \overset{*}{f}_{ji} = 0$, since the states belong to two different final classes of states and are therefore not reachable from each other.
- The states i and j are both transient. They may belong either to the same transient class of states or to two different transient classes of states. In the first case $\overset{*}{f}_{ij} > 0$ and $\overset{*}{f}_{ji} > 0$, since states within the same class are mutually reachable. In the second case $\overset{*}{f}_{ij} = 0$ or $\overset{*}{f}_{ji} = 0$, since i cannot be reached from j or j cannot be reached from i.
- The state i is transient and the state j is recurrent. Then $\overset{*}{f}_{ji} = 0$, since i cannot be reached from j.

Absorption behavior : After a final class of states J has first been reached from a transient state i, the process cannot leave the class. The probability that this happens is called an absorption probability and is designated by $\overset{*}{f}_{iJ}$. It is a property of the class J, that is $\overset{*}{f}_{ij} = \overset{*}{f}_{iJ}$ for every recurrent state $j \in J$. The absorption probabilities $\overset{*}{f}_{iJ}$ are calculated for a reduced homogeneous Markov process.

The homogeneous Markov process is reduced by mapping each recurrent state $j \in J$ to its final class J. The transition probabilities for this reduction are determined as follows :

- The probability p_{iJ} for the transition from a transient state i to a final class J is the sum of the probabilities p_{ij} for all states $j \in J$. Since there are no transitions from a recurrent state $j \in J$ to a transient state i, the transition probability p_{Ji} is zero.

$$p_{iJ} = \sum_{j \in J} p_{ij} \qquad\qquad p_{Ji} = 0$$

- Since a transition from a recurrent state $j \in J$ in a final class J always leads to a state in the same class, the transition probability p_{JJ} is one and the transition probability p_{JK} for $K \neq J$ is zero.

$$p_{JK} = \delta_{JK} = \begin{cases} 0 \text{ for } K \neq J \\ 1 \text{ for } K = J \end{cases}$$

Thus the transition matrix $\mathbf{P}_R$ of a reduced homogeneous Markov process has the following special structure :

$$\mathbf{P}_R = \begin{array}{|c|c|}\hline \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \\ \hline \end{array} \begin{array}{ll} \leftarrow i & \text{transient states} \\ \leftarrow J & \text{final classes of states} \end{array}$$

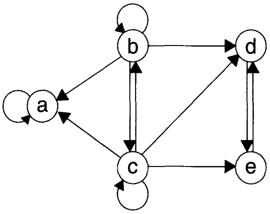
If the absorption probabilities $\overset{*}{f}_{iJ}$ for all transient states i are arranged in a vector $\overset{*}{\mathbf{f}}_J$, the special structure of the transition matrix $\mathbf{P}_R$ leads to the following system of linear equations for $\overset{*}{\mathbf{f}}_J$:

$$(\mathbf{I} - \mathbf{Q})\,\overset{*}{\mathbf{f}}_J = \mathbf{r}_J \qquad\qquad \mathbf{r}_J \;\; \text{column of } \mathbf{R} \text{ for class of states J}$$

The system of linear equations for $\overset{*}{\mathbf{f}}_J$ has a unique solution. The matrix $\mathbf{I} - \mathbf{Q}$ is regular, since the matrix $\mathbf{Q}$ for the transitions between the transient states has only eigenvalues with $|\lambda| < 1$ and the determinant $\det(\mathbf{I} - \mathbf{Q})$ is therefore non-zero.

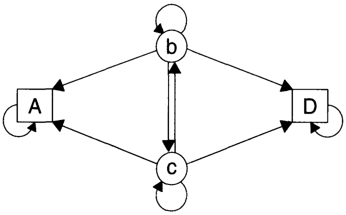
Example : Absorption behavior

Let the homogeneous Markov process with the illustrated transition graph and the corresponding transition matrix **P** be given.



$$\mathbf{P} = \frac{1}{5}\begin{array}{|c|c|c|c|c|l}
\multicolumn{1}{c}{a} & \multicolumn{1}{c}{b} & \multicolumn{1}{c}{c} & \multicolumn{1}{c}{d} & \multicolumn{1}{c}{e} & \\ \hline
5 & 0 & 0 & 0 & 0 & a \\ \hline
1 & 2 & 1 & 1 & 0 & b \\ \hline
1 & 1 & 1 & 1 & 1 & c \\ \hline
0 & 0 & 0 & 0 & 5 & d \\ \hline
0 & 0 & 0 & 5 & 0 & e \\ \hline
\end{array}$$

The space of states of the Markov process consists of the final class of states A = {a}, the final class of states D = {d, e} and the transient class of states T = {b, c}. The states a, d, e in the two final classes are recurrent. The states b, c in the transient class are transient. Reducing the process to its transient states and its final classes of states leads to the following transition graph with the reduced transition matrix $\mathbf{P}_R$.



$$\mathbf{P}_R = \frac{1}{5}\begin{array}{|c|c|c|c|l}
\multicolumn{1}{c}{b} & \multicolumn{1}{c}{c} & \multicolumn{1}{c}{A} & \multicolumn{1}{c}{D} & \\ \hline
2 & 1 & 1 & 1 & b \\ \hline
1 & 1 & 1 & 2 & c \\ \hline
0 & 0 & 5 & 0 & A \\ \hline
0 & 0 & 0 & 5 & D \\ \hline
\end{array} = \begin{array}{|c|c|} \hline \mathbf{Q} & \mathbf{R} \\ \hline \mathbf{0} & \mathbf{I} \\ \hline \end{array}$$

The absorption probabilities $\overset{*}{f}_{bD}$, $\overset{*}{f}_{cD}$ are arranged in a vector $\overset{*}{\mathbf{f}}_D$ and determined by solving a system of linear equations :

$$(\mathbf{I} - \mathbf{Q}) \quad * \quad \overset{*}{\mathbf{f}}_D \quad = \quad \mathbf{r}_D$$

$$\begin{array}{|c|c|} \hline 3/5 & -1/5 \\ \hline -1/5 & 4/5 \\ \hline \end{array} * \begin{array}{|c|} \hline 6/11 \\ \hline 7/11 \\ \hline \end{array} = \begin{array}{|c|} \hline 1/5 \\ \hline 2/5 \\ \hline \end{array}$$

The final class D is reached with the absorption probability 6/11 from the state b and with the absorption probability 7/11 from the state c.

10.5.2.6 Processes of higher order

Introduction : The preceding sections deal with finite Markov processes of first order. Markov processes of higher order may be reduced to Markov processes of first order. Thus the fundamentals treated here may also be applied to Markov processes of higher order.

Process of first order : A stochastic process is called a Markov process of first order if the state of the process at the future time t_{n+1} depends only on the state at the present time t_n. Such a process is also called a process without memory.

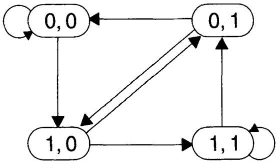
Process of r-th order : A stochastic process is called a Markov process of r-th order if the state of the process at the future time t_{n+1} depends on the state at the present time t_n and on the states at the past $r-1$ times $t_{n-1},\dots,t_{n-r+1}$. Such a process with $r>1$ is also called a process with memory.

State : A Markov process of r-th order with the space of states S is reduced to a Markov process of first order with the space of states S^r. The space of states S^r contains all ordered r-tuples of the space of states S. An ordered r-tuple contains the states which the process can be in at the times t_n, $t_{n-1},\dots,$ t_{n-r+1}.

Transition : The possible transitions in the space of states S^r are restricted. A transition from the tuple $a \in S^r$ at time t_n to the tuple $b \in S^r$ at time t_{n+1} is only possible if the first $r-1$ states of a coincide with the last $r-1$ states of b. The transition probabilities for the impossible transitions are zero.

Example : Markov process of second order

Let a Markov process of second order with the space of states S containing the states 0 and 1 be given. It is reduced to a Markov process of first order with the space of states S^2. The possible transitions in the space of states S^2 are shown in the transition graph.



$$S = \{0,1\}$$

$$S^2 = S \times S$$

$$S^2 = \{(0,0),(0,1),(1,0),(1,1)\}$$

10.5.3 FINITE MARKOV PROCESSES IN CONTINUOUS TIME

10.5.3.1 Introduction

Finite Markov processes with a finite discrete space of states and a discrete time domain are treated in the preceding section. Based on that material, this section considers finite Markov processes in continuous time. At a given time t, the Markov process has certain probabilities for being in the various states. At time t + dt, every state continuously changes into a new state with a certain transition rate. The probabilities for the different possible states of the process are therefore continuous time-dependent functions.

If the transition rates of a Markov process are constant, the Markov process is said to be homogeneous. The basic definitions and methods for homogeneous Markov processes in continuous time are analogous to the definitions and methods for homogeneous Markov processes in discrete time. In Sections 10.5.3.2 and 10.5.3.3 they are derived and applied to typical examples.

The theory of queues is an important area of application for homogeneous Markov processes in continuous time. Queue models consist of an arrival process and a service process. The fundamentals of the theory of queues including the definition of models are treated in Sections 10.5.3.4 and 10.5.3.5.

10.5.3.2 States and transition rates

Introduction : Finite Markov processes in continuous time may be derived directly from finite Markov processes in discrete time. For this purpose, the change of state of the process within an infinitesimal time increment is considered. This approach leads to an initial value problem with a linear system of first order differential equations for the time-dependent state probabilities of the process.

State probability : The finite Markov process in continuous time is described by a random variable X(t) which can take different real values x_j for the possible states e_j of the process at time t. The state probability $p_j(t)$ is the probability that the process is in the state e_j at time t, and hence the probability that the random variable X(t) takes the value x_j. The state probabilities for all possible states form the probability function for the random variable X(t). They are conveniently arranged in a stochastic vector **p**(t). Their sum is 1.

$$p_j(t) \;=\; P(X(t) = x_j)$$

$$\mathbf{p}(t) \;\geq\; \mathbf{0} \qquad\qquad \mathbf{e}^T\mathbf{p}(t) \;=\; 1$$

Incremental transition : Let a homogeneous Markov process with the state probabilities $\mathbf{p}(t)$ at time t be given. The state probabilities $\mathbf{p}(t+\Delta t)$ at time $t+\Delta t$ are calculated using the transition probabilities $\mathbf{P}(\Delta t)$, which for a homogeneous Markov process depend only on the time increment Δt and not on the time t.

$$\mathbf{p}(t+\Delta t) \;=\; \mathbf{P}^{T}(\Delta t)\,\mathbf{p}(t) \tag{1}$$

The changes in the state probabilities are referred to the time increment Δt :

$$\frac{1}{\Delta t}(\mathbf{p}(t+\Delta t)-\mathbf{p}(t)) \;=\; \frac{1}{\Delta t}(\mathbf{P}(\Delta t)-\mathbf{I})^{T}\,\mathbf{p}(t) \tag{2}$$

Infinitesimal transition : In the limit $\Delta t \to 0$, the incremental transition leads to an infinitesimal transition. Equation (1) implies the following continuity condition for the state probabilities $\mathbf{p}(t)$:

$$\lim_{\Delta t\to 0}\mathbf{p}(t+\Delta t) \;=\; \lim_{\Delta t\to 0}\mathbf{P}^{T}(\Delta t)\,\mathbf{p}(t)$$

$$\lim_{\Delta t\to 0}\mathbf{P}(\Delta t) \;=\; \mathbf{I} \qquad\qquad \text{continuity condition}$$

$$\lim_{\Delta t\to 0}\mathbf{p}(t+\Delta t) \;=\; \mathbf{p}(t) \qquad\qquad \text{continuity}$$

Equation (2) implies the transition condition for the state probabilities $\mathbf{p}(t)$ in form of a system of first order differential equations with a constant transition matrix $\mathbf{G}$.

$$\mathbf{p}'(t) \;=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{p}(t+\Delta t)-\mathbf{p}(t)) \;=\; \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{P}(\Delta t)-\mathbf{I})^{T}\,\mathbf{p}(t)$$

$$\mathbf{G} := \lim_{\Delta t\to 0}\frac{1}{\Delta t}(\mathbf{P}(\Delta t)-\mathbf{I}) \qquad\qquad \text{transition matrix}$$

$$\mathbf{p}'(t) \;=\; \mathbf{G}^{T}\,\mathbf{p}(t) \qquad\qquad \text{transition condition}$$

The transition matrix $\mathbf{G}$ is called the generator of the homogeneous Markov process. Its elements may be interpreted as transition rates.

Transition rates : A homogeneous Markov process in continuous time is described by transition rates. The probability $p_{ij}(\Delta t)$ for a transition from the state i to the state $j \neq i$ is proportional to the time increment Δt, with the non-negative transition rate g_{ij} acting as the constant of proportionality :

$$p_{ij}(\Delta t) \;=\; g_{ij}\,\Delta t + o(\Delta t) \;\le\; 1 \qquad\qquad i \neq j$$

$o(\Delta t)$ terms of higher order in Δt

If the time increment Δt is sufficiently small, the higher-order terms $o(\Delta t)$ are negligible compared to $g_{ij}\Delta t$. The probabilities for transitions from state i to all states j in the time increment Δt must add up to 1. This condition yields the probability $p_{ii}(\Delta t)$ with the non-positive transition rate g_{ii}.

$$\begin{aligned} p_{ii}(\Delta t) &= 1 - \sum_{j \neq i} p_{ij}(\Delta t) = 1 - \sum_{j \neq i} g_{ij} \Delta t + o(\Delta t) \\ p_{ii}(\Delta t) &= 1 + g_{ii} \Delta t + o(\Delta t) \\ g_{ii} &= - \sum_{j \neq i} g_{ij} \end{aligned}$$

In the limit $\Delta t \rightarrow 0$, the transition probabilities $p_{ij}(\Delta t)$ tend to 0 for $i \neq j$ and to 1 for $i = j$. Hence the continuity condition is satisfied. The transition rates g_{ij} are non-positive for $i = j$ and non-negative for $i \neq j$. The sum of the rates for the transitions from a state i to all states j is 0.

$$\begin{aligned} & -\infty < g_{ii} \leq 0 \\ & 0 \leq g_{ij} < \infty \qquad \text{for} \quad i \neq j \\ & \sum_j g_{ij} = 0 \end{aligned}$$

Transition graph : A finite homogeneous Markov process in continuous time may be represented by a transition graph with the transition rates as weights. Its structure is analyzed using the graph-theoretical methods in Section 10.5.2.3. The space of states of the process may be partitioned into classes of states, which are either transient or final. It contains at least one final class of states. By virtue of the continuity condition for the states, every final class of states is aperiodic. The long-term behavior of the process is therefore not periodic.

Homogeneous Markov process : A homogeneous Markov process in continuous time is completely described by the initial distribution $\mathbf{p}_0$ at time $t = 0$ and the generator **G** with the transition rates. This description leads to the following initial value problem :

$$\mathbf{p}'(t) = \mathbf{G}^T \mathbf{p}(t) \qquad \mathbf{p}(t = 0) = \mathbf{p}_0 \qquad t \geq 0$$

The solution of this initial value problem is given by

$$\mathbf{p}(t) = \mathbf{P}^T(t)\, \mathbf{p}_0 \qquad \mathbf{P}(t) = e^{t\mathbf{G}} := \sum_{k=0}^{\infty} \frac{1}{k!} (t\mathbf{G})^k$$

The matrix **P**(t) is the transition matrix for the state probabilities of the process in the time interval from 0 to t. It may be represented as an exponential function with the generator **G**. Its derivative with respect to time is given by

$$\mathbf{P}'(t) = \mathbf{G}\mathbf{P}(t) = \mathbf{P}(t)\mathbf{G}$$

Limit distribution : Starting from an initial distribution $\mathbf{p}_0$, the calculated distributions $\mathbf{p}(t)$ of a homogeneous Markov process tend to a limit distribution $\mathbf{p}_\infty$ in the limit $t \to \infty$. The limit distribution may or may not depend on the initial distribution $\mathbf{p}_0$.

$$\mathbf{p}_\infty = \lim_{t \to \infty} \mathbf{p}(t)$$

If a homogeneous Markov process possesses exactly one final class of states, the limit distribution is independent of the initial distribution $\mathbf{p}_0$. If it possesses several final classes of states, the limit distribution depends on the initial distribution $\mathbf{p}_0$.

Equilibrium distribution : A homogeneous Markov process is in equilibrium if the distribution $\mathbf{p}(t)$ is independent of time. This is only possible if the derivative $\mathbf{p}'(t)$ is zero. Such a distribution $\overset{*}{\mathbf{p}}$ is called an equilibrium distribution (stationary distribution). The rules of calculation yield the following equilibrium condition :

$$\mathbf{G}^T \overset{*}{\mathbf{p}} = \mathbf{0}$$

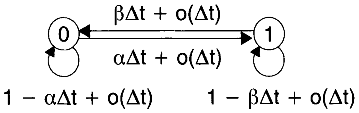
A distribution $\overset{*}{\mathbf{p}}$ may be determined directly from the equilibrium condition and the distribution conditions. This leads to a homogeneous system of linear equations with constraints.

$$\mathbf{G}^T \overset{*}{\mathbf{p}} = \mathbf{0} \qquad \mathbf{e}^T \overset{*}{\mathbf{p}} = 1 \qquad \overset{*}{\mathbf{p}} \geq \mathbf{0}$$

Every limit distribution $\mathbf{p}_\infty$ is an equilibrium distribution. If the initial distribution $\mathbf{p}_0$ is an equilibrium distribution, the homogeneous Markov process is a stationary process, that is its distribution is independent of time.

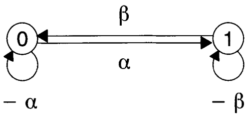
Example : Device states

Assume that a device is either in a broken state 0 or in an intact state 1 at time t. The transitions between these states are treated as a homogeneous Markov process. The average repair time for a broken device is T_0. The rate for the transition from the broken state to the intact state is therefore $\alpha = 1/T_0$. The average lifetime for an intact device is T_1. The rate for the transition from the intact state to the broken state is therefore $\beta = 1/T_1$. The homogeneous Markov process with the transition probabilities for a time increment Δt may be represented as follows :



$$\mathbf{P}(\Delta t) = \begin{array}{|c|c|} \hline 1 - \alpha\Delta t & \alpha\Delta t \\ \hline \beta\Delta t & 1 - \beta\Delta t \\ \hline \end{array} + \mathbf{o}(\Delta t)$$

The limit Δt → 0 leads to a homogeneous Markov process in continuous time $t \geq 0$ with the generator **G**, which is graphically represented as follows :



$$\mathbf{G} = \begin{array}{|c|c|}\hline -\alpha & \alpha \\ \hline \beta & -\beta \\ \hline \end{array}$$

This leads to the following initial value problem for the state probabilities $p_0(t)$ and $p_1(t)$:

$$\begin{array}{ccccccc} \mathbf{p}'(t) & = & \mathbf{G}^T & * & \mathbf{p}(t) & \qquad & \mathbf{p}(t=0) = \mathbf{p}(0) \\ \begin{array}{|c|}\hline p_0'(t) \\ \hline p_1'(t) \\ \hline \end{array} & = & \begin{array}{|c|c|}\hline -\alpha & \beta \\ \hline \alpha & -\beta \\ \hline \end{array} & * & \begin{array}{|c|}\hline p_0(t) \\ \hline p_1(t) \\ \hline \end{array} & & \end{array}$$

The solution of this initial value problem under the constraint $p_0(t) + p_1(t) = 1$ is given by

$$p_0(t) = p_0(0)\, e^{-(\alpha+\beta)t} + \frac{\beta}{\alpha+\beta}\left(1 - e^{-(\alpha+\beta)t}\right)$$

$$p_1(t) = p_1(0)\, e^{-(\alpha+\beta)t} + \frac{\alpha}{\alpha+\beta}\left(1 - e^{-(\alpha+\beta)t}\right)$$

In the limit $t \to \infty$, the state probabilities tend to the following limits, which are independent of the state probabilities at time $t = 0$:

$$p_0(\infty) = \frac{\beta}{\alpha+\beta}$$

$$p_1(\infty) = \frac{\alpha}{\alpha+\beta}$$

The limit distribution for $t \to \infty$ coincides with the equilibrium distribution, which is calculated by solving a constrained homogeneous system of linear equations :

$$\begin{array}{ccccc} \mathbf{G}^T & * & \overset{*}{\mathbf{p}} & = & \mathbf{0} \\ \begin{array}{|c|c|}\hline -\alpha & \beta \\ \hline \alpha & -\beta \\ \hline \end{array} & * & \begin{array}{|c|}\hline \overset{*}{p}_0 \\ \hline \overset{*}{p}_1 \\ \hline \end{array} & = & \begin{array}{|c|}\hline 0 \\ \hline 0 \\ \hline \end{array} \end{array}$$

$$\overset{*}{p}_0 + \overset{*}{p}_1 = 1$$

$$\overset{*}{p}_0 = \beta/(\alpha+\beta)$$

$$\overset{*}{p}_1 = \alpha/(\alpha+\beta)$$

10.5.3.3 First passage

Introduction : The first passage problem for finite homogeneous Markov processes in discrete time is treated in Section 10.5.2.5. The principles may be transferred to finite homogeneous Markov processes in continuous time. In the case of discrete time, the time-dependent distribution for the first passage from an initial state a to a final state z is discrete. In the case of continuous time it is continuous. A suitable method for determining the continuous distribution and its moments is described in the following.

Passage time : The passage time T_{ij} is the time which a homogeneous Markov process in continuous time takes to reach a state j from another state i for the first time. The passage time is a continuous random variable which takes values $t \geq 0$.

Density function of the passage time : If a state j can be reached from a state i, the passage time T_{ij} takes real values $t \geq 0$. It is described by the probability density $f_{ij}(t)$ for $t \geq 0$. If the time axis $t \geq 0$ is divided into sufficiently small time increments Δt, then $\Delta t \, f_{ij}(t)$ is the probability that the state j is reached for the first time from the state i in the time interval from $t - \Delta t$ to t. The recursive rule of calculation for homogeneous Markov processes in discrete time from Section 10.5.2.5 may thus be applied. With the incremental transition probabilities $p_{ij}(\Delta t)$, this yields :

$$\Delta t \; f_{ij}(\Delta t) \;=\; p_{ij}(\Delta t) \qquad\qquad i \neq j$$

$$\Delta t \; f_{ij}(t) \;=\; \sum_{k \neq j} p_{ik}(\Delta t)\; \Delta t \; f_{kj}(t - \Delta t) \qquad\qquad i \neq j$$

Transforming the equations yields :

$$f_{ij}(\Delta t) \;=\; \frac{1}{\Delta t}\, p_{ij}\,(\Delta t)$$

$$\frac{1}{\Delta t}\,(f_{ij}(t) - f_{ij}(t - \Delta t)) \;=\; \frac{1}{\Delta t}\,(\sum_{k \neq j} p_{ik}(\Delta t)\; f_{kj}(t - \Delta t) \;-\; f_{ij}(t - \Delta t))$$

Taking the limit $\Delta t \to 0$ for $i \neq j$ and using the definitions of the transition rates g_{ij} leads to the following initial value problem :

$$f'_{ij}(t) \;=\; \sum_{k \neq j} g_{ik}\; f_{kj}(t) \qquad\qquad f_{ij}\,(0) \;=\; g_{ij} \qquad\qquad i \neq j$$

The rules of calculation may be formulated concisely in vector and matrix notation. For a given state j and all states $i \neq j$, the probability densities $f_{ij}(t)$ and transition rates g_{ij} are arranged in the vectors $\mathbf{f}_j(t)$ and $\mathbf{g}_j$. The matrix $\mathbf{G}_j$ is obtained from the generator $\mathbf{G}$ by deleting row j and column j of $\mathbf{G}$.

$$\mathbf{f}'_j\,(t) \;=\; \mathbf{G}_j\; \mathbf{f}_j(t) \qquad\qquad \mathbf{f}_j\,(0) \;=\; \mathbf{g}_j$$

For $t \geq 0$, the solutions $f_{ij}(t)$ of the initial value problem correspond to a density function for the passage time T_{ij}. However, in the general case this function does not satisfy the condition that the integral of $f_{ij}(t)$ over $t \geq 0$ is one. If the state j cannot be reached from the state i, then $f_{ij}(t) = 0$ for $t \geq 0$.

Moments of the passage time : The moments of the passage time T_{ij} are determined from the probability density $f_{ij}(t)$ with $t \geq 0$. The k-th moment is defined as follows :

$$m_{ij}(k) := \int_0^\infty t^k f_{ij}(t)\, dt \qquad k \geq 0$$

The moments may be calculated directly without explicitly determining the probability densities $f_{ij}(t)$. The rule of calculation is concisely formulated in vector and matrix notation. For a given state j and every state $i \neq j$, the moments $m_{ij}(k)$ are arranged in the vector $\mathbf{m}_j(k)$. The definition of the moments then takes the form :

$$\mathbf{m}_j(k) = \int_0^\infty t^k \mathbf{f}_j(t)\, dt$$

Multiplying the equation with the matrix $\mathbf{G}_j$ from the left and substituting the rule of calculation $\mathbf{f}_j'(t) = \mathbf{G}_j \mathbf{f}_j(t)$ yields :

$$\mathbf{G}_j \mathbf{m}_j(k) = \mathbf{G}_j \int_0^\infty t^k \mathbf{f}_j(t)\, dt = \int_0^\infty t^k \mathbf{G}_j \mathbf{f}_j(t)\, dt = \int_0^\infty t^k \mathbf{f}_j'(t)\, dt$$

Direct integration for $k = 0$ and integration by parts for $k > 0$ now yields :

$$\mathbf{G}_j \mathbf{m}_j(0) = \int_0^\infty \mathbf{f}_j'(t)\, dt = \mathbf{f}_j(\infty) - \mathbf{f}_j(0) = \mathbf{f}_j(\infty) - \mathbf{g}_j$$

$$\mathbf{G}_j \mathbf{m}_j(k) = \int_0^\infty t^k \mathbf{f}_j'(t)\, dt = [t^k \mathbf{f}_j(t)]_0^\infty - k \int_0^\infty t^{k-1} \mathbf{f}_j(t)\, dt \qquad k > 0$$

Since the probability densities for $t \rightarrow \infty$ tend to zero exponentially, the following equations for the moments of the passage times are obtained :

$$\mathbf{G}_j \mathbf{m}_j(0) = -\mathbf{g}_j$$

$$\mathbf{G}_j \mathbf{m}_j(k) = -k\, \mathbf{m}_j(k-1) \qquad k > 0$$

If the state j cannot be reached from the state i, then since $f_{ij}(t) = 0$ for $t \geq 0$ all moments $m_{ij}(k)$ for $k \geq 0$ are also zero.

Passage probability : The probability $P(T_{ij} \geq 0)$ is called a passage probability and is designated by $\overset{*}{f}_{ij}$. It is the integral of the probability density $f_{ij}(t)$ over $t \geq 0$, and hence the zeroth-order moment.

$$\overset{*}{f}_{ij} \ := \ P(T_{ij} \geq 0) \ = \ m_{ij}(0)$$

The passage probabilities are determined as follows according to the rule of calculation for the zeroth-order moments $\overset{*}{\mathbf{f}}_j = \mathbf{m}_j(0)$:

$$\mathbf{G}_j \, \overset{*}{\mathbf{f}}_j \ = \ -\mathbf{g}_j$$

If the transition graph contains no path from the state i to the state j, then j cannot be reached from i and $\overset{*}{f}_{ij} = 0$. If there is at least one path from state i to state j, then j can be reached from i and $\overset{*}{f}_{ij} > 0$. If every sufficiently long path from state i leads to state j, then j is always reached from i and $\overset{*}{f}_{ij} = 1$. The probability that state j is not reached from state i is $1 - \overset{*}{f}_{ij}$.

Average passage time : If state j is always reached from state i, then $\overset{*}{f}_{ij} = 1$, and hence $f_{ij}(t)$ for $t > 0$ is a density function. In this case the first moment $m_{ij}(1)$ is the mean μ_{ij} of the passage time and is called the average passage time.

$$\mu_{ij} \ := \ m_{ij}(1) \quad \text{for} \quad \overset{*}{f}_{ij} \ = \ 1$$

The means are determined in the vector form $\boldsymbol{\mu}_j = \mathbf{m}_j(1)$ according to the rule of calculation for the first moments :

$$\mathbf{G}_j \, \boldsymbol{\mu}_j \ = \ -\overset{*}{\mathbf{f}}_j$$

Recurrence time : The recurrence time T_{jj} is the time which a homogeneous Markov process in continuous time takes to return to a state j for the first time after leaving it. The homogeneous Markov process leaves the state j for the states $i \neq j$ with the probabilities p_{ji}, which are calculated from the transition rates as follows :

$$p_{ji} \ = \ -\frac{g_{ji}}{g_{jj}} \qquad\qquad g_{jj} \ = \ -\sum_{i \neq j} g_{ij} \qquad\qquad i \neq j$$

Density function of the recurrence time : If a state j can be reached again after the process has left it, the recurrence time T_{jj} takes real values $t \geq 0$. It is described by the probability density $f_{jj}(t)$ and calculated from the probability densities $f_{ij}(t)$ for the passage times T_{ij} with $i \neq j$. The homogeneous Markov process leaves the state j for the states $i \neq j$ with the probabilities p_{ji} and then returns to the original state for the first time. Hence the probability density $f_{jj}(t)$ is the weighted sum of the probability densities $f_{ij}(t)$ with the probabilities p_{ji} as weight factors.

$$f_{jj}(t) \ = \ \sum_{i \neq j} p_{ji} \ f_{ij}(t)$$

This rule of calculation is concisely formulated in vector notation. For a given state j and all states $i \neq j$, the probabilities p_{ji} and $f_{ij}(t)$ are arranged in the vectors $\mathbf{q}_j^T$ and $\mathbf{f}_j(t)$.

$$f_{jj}(t) \;=\; \mathbf{q}_j^T \, \mathbf{f}_j(t)$$

The probability density $f_{jj}(t)$ corresponds to a density function for the recurrence time. However, in the general case it does not satisfy the condition that the integral of $f_{jj}(t)$ over $t \geq 0$ is one. If the process cannot leave the state j or the state j cannot be reached after the process has left it, then $f_{jj}(t) = 0$ for $t > 0$.

Moments of the recurrence time : The moments of the recurrence time T_{jj} are obtained from the probability density $f_{jj}(t)$ with $t \geq 0$. The k-th moment is defined as follows :

$$m_{jj}(k) \;:=\; \int_0^\infty t^k \, f_{jj}(t) \, dt$$

The moments of the recurrence time may be calculated directly from the moments of the passage times. Using the rule $f_{jj}(t) = \mathbf{q}_j^T \, \mathbf{f}_j(t)$, one obtains :

$$m_{jj}(k) \;=\; \mathbf{q}_j^T \int_0^\infty t^k \, \mathbf{f}_j(t) \, dt \;=\; \mathbf{q}_j^T \, \mathbf{m}_j(k)$$

Recurrence probability : The probability $P(T_{jj} \geq 0)$ is called a recurrence probability and is designated by $\overset{*}{f}_{jj}$. It is the integral of the probability density $f_{jj}(t)$ over $t \geq 0$, and hence the zeroth-order moment.

$$\overset{*}{f}_{jj} \;:=\; P(T_{jj} \geq 0) \;=\; \int_0^t f_{jj}(t) \, dt \;=\; m_{jj}(0)$$

The recurrence probability is determined as follows from the passage probabilities $\overset{*}{\mathbf{f}}_j = \mathbf{m}_j(0)$ in vector form :

$$\overset{*}{f}_{jj} \;=\; \mathbf{q}_j^T \, \overset{*}{\mathbf{f}}_j$$

If the transition graph does not contain a cycle through the state j, then the process cannot return to j and $\overset{*}{f}_{jj} = 0$. If the transition graph contains at least one cycle through the state j, then the process can return to j and $\overset{*}{f}_{jj} > 0$. If the transition graph contains only cycles through the state j, then the process always returns to j and $\overset{*}{f}_{jj} = 1$. The probability that the state j is not reached again is $1 - \overset{*}{f}_{jj}$.

Average recurrence time : If the process always returns to the state j, then $\overset{*}{f}_{jj} = 1$, and hence $f_{jj}(t)$ for $t \geq 0$ is a density function. In this case the first moment $m_{jj}(1)$ is the mean μ_{jj} of the recurrence time and is called the average recurrence time.

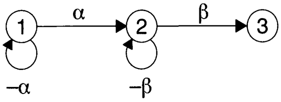
$$\mu_{jj} := m_{jj}(1) \quad \text{for} \quad \overset{*}{f}_{jj} = 1$$

The average recurrence time is determined as follows from the average passage times $\boldsymbol{\mu}_j = \mathbf{m}_j(1)$ in vector form :

$$\mu_{jj} = \mathbf{q}_j^T \boldsymbol{\mu}_j$$

Example : Deterioration process

The example of a simple deterioration process with the deterioration states 1,2,3 treated in Section 10.5.2.5 may also be considered as a homogeneous Markov process in continuous time. The transition graph with the transition rates and the corresponding generator **G** for this process are given.



$$\mathbf{G} = \begin{array}{ccc} 1 & 2 & 3 \\ \hline -\alpha & \alpha & 0 \\ 0 & -\beta & \beta \\ 0 & 0 & 0 \end{array} = \begin{array}{|c|c|} \hline \mathbf{G}_3 & \mathbf{g}_3 \\ \hline \mathbf{0}^T & 0 \\ \hline \end{array}$$

The deterioration state 3 is first reached from the deterioration state 1 after the time T_{13}; it is first reached from the deterioration state 2 after the time T_{23}. The probability densities $f_{13}(t)$ and $f_{23}(t)$ are arranged in a vector $\mathbf{f}_3(t)$. The following initial value problem with the matrix $\mathbf{G}_3$ and the vector $\mathbf{g}_3$ is obtained :

$$\mathbf{f}_3^{\prime}(t) = \mathbf{G}_3 * \mathbf{f}_3(t) \qquad \mathbf{f}_3(0) = \mathbf{g}_3$$

$$\begin{bmatrix} f_{13}^{\prime}(t) \\ f_{23}^{\prime}(t) \end{bmatrix} = \begin{bmatrix} -\alpha & \alpha \\ 0 & -\beta \end{bmatrix} * \begin{bmatrix} f_{13}(t) \\ f_{23}(t) \end{bmatrix} \qquad \begin{bmatrix} f_{13}(0) \\ f_{23}(0) \end{bmatrix} = \begin{bmatrix} 0 \\ \beta \end{bmatrix}$$

The equations for the probability densities are considered in component representation :

$$\begin{aligned} f_{13}^{\prime}(t) &= -\alpha\, f_{13}(t) + \alpha\, f_{23}(t) & \qquad f_{13}(0) &= 0 \\ f_{23}^{\prime}(t) &= -\beta\, f_{23}(t) & \qquad f_{23}(0) &= \beta \end{aligned}$$

The solutions for $\alpha \neq \beta$ are :

$$\begin{aligned} f_{13}(t) &= (e^{-\alpha t} - e^{-\beta t}) \,/\, (1/\alpha - 1/\beta) \\ f_{23}(t) &= \beta e^{-\beta t} \end{aligned}$$

The zeroth-order moments $\overset{*}{f}_{i3}$ and the first-order moments μ_{i3} for $i = 1, 2$ are arranged in the vectors $\overset{*}{\mathbf{f}}_3$ and $\boldsymbol{\mu}_3$. They are calculated by solving systems of linear equations :

$$\begin{array}{ccccc} \mathbf{G}_3 & * & \overset{*}{\mathbf{f}}_3 & = & -\mathbf{g}_3 \\ \begin{bmatrix}-\alpha & \alpha \\ 0 & -\beta\end{bmatrix} & * & \begin{bmatrix}1 \\ 1\end{bmatrix} & = & \begin{bmatrix}0 \\ -\beta\end{bmatrix} \end{array} \qquad \begin{array}{ccccc} \mathbf{G}_3 & * & \boldsymbol{\mu}_3 & = & -\overset{*}{\mathbf{f}}_3 \\ \begin{bmatrix}-\alpha & \alpha \\ 0 & -\beta\end{bmatrix} & * & \begin{bmatrix}1/\alpha + 1/\beta \\ 1/\beta\end{bmatrix} & = & \begin{bmatrix}-1 \\ -1\end{bmatrix} \end{array}$$

The average passage time from state 1 to state 3 is $\mu_{13} = 1/\alpha + 1/\beta$; the average passage time from state 2 to state 3 is $\mu_{23} = 1/\beta$. The probability density for the return of state 3 to itself is zero, since the process cannot leave state 3.

10.5.3.4 Queues

Introduction : Queue problems arise in various forms in different areas of application. A simple example is furnished by customers being served at a counter. Customers arrive at the counter in random time intervals and wait to be served. Each customer is served in a random time. After being served, the customer leaves the counter. The steady arrival of customers at the counter and the steady departure of customers lead to a queue whose length changes over time. Queue problems of this type may be treated as Markov processes.

Queue models : Mathematical models for treating queue problems are called queue models. They consist of an arrival model and a service model. A queue model is described by the type of the arrival and service processes, the number of service stations, the capacity of the waiting room and the service rule. Abbreviations of the following form are used in the literature for the various queue models :

queue model	A / B / s / m / R
type of arrival process	A
type of service process	B
number of service stations	s
capacity of the waiting room	m
service rule	R

The arrival process and the service process may both be a Markov process, a generalized stochastic process or a deterministic process. Accordingly, the type of arrival or service process is designated by the letters M, G or D. The service model possesses one or more service stations. The waiting capacity is bounded or unbounded. The following rules are used for the dependence of the order of service on the order of arrival :

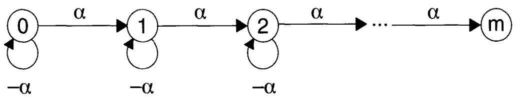
- Customers are served in order of arrival (FIFO : first in first out).
- Customers are served in reverse order of arrival (LIFO : last in first out).
- Customers are served in random order independent of the order of arrival (SIRO : service in random order).

The following treatment of queue models is confined to Markov processes for arrival and service, and to service in the order of arrival. The service rule is not specified in the abbreviations in the following.

Queue : A queue is a sequence of customers who are being served or are waiting to be served. The number of customers in a queue is called the length of the queue. The length is a natural number and may be bounded by a given waiting capacity. The length of a queue at time t is a discrete random variable and is designated by X(t).

Waiting time : The time for which a customer who is currently in position n in the queue must wait to leave the queue after having been served is a continuous random variable T_n; it is called the waiting time of the n-th customer. The waiting time of the first customer in the queue who is being served is also called the service time.

Arrival model : Assume that no customers are waiting at a given service station at time $t = 0$. Over time, customers arrive at the service station with the rate of arrival α. Assume that the service station is closed, so that no customers are served and a queue forms. Let the waiting capacity be restricted to m customers. The length of the queue at time t is a random variable X(t) which takes natural values in the interval [0,m]. This arrival process is a homogeneous Markov process in continuous time $t \geq 0$ and is represented by the following transition graph.



α rate of arrival

Queue : The probability that the queue in the arrival process has length n at time t is $p_n(t)$. The probabilities $p_n(t)$ for $0 \leq n \leq m$ are determined according to Section 10.5.3.2 by solving the following initial value problem :

$$\begin{array}{lll} p'_0(t) = -\alpha\, p_0(t) & p_0(0) = 1 & \\ p'_n(t) = -\alpha\, p_n(t) + \alpha\, p_{n-1}(t) & p_n(0) = 0 & 0 < n < m \\ p'_m(t) = \qquad\qquad \alpha\, p_{m-1}(t) & p_m(0) = 0 & \end{array}$$

The solutions of this initial value problem are given by :

$$p_n(t) = \frac{(\alpha t)^n}{n!} e^{-\alpha t} \qquad 0 \le n < m$$

$$p_m(t) = 1 - e^{-\alpha t} \sum_{n=0}^{m-1} \frac{(\alpha t)^n}{n!}$$

The number of waiting customers is equal to the waiting capacity with probability $p_m(t)$. The probability that an arriving customer is turned away is also given by $p_m(t)$. This probability tends to zero in the limit $m \to \infty$.

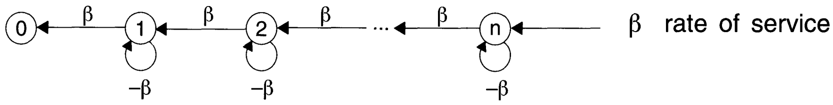
$$p_\infty(t) = \lim_{m \to \infty} p_m(t) = 1 - e^{-\alpha t} \sum_{n=0}^{\infty} \frac{(\alpha t)^n}{n!} = 1 - e^{-\alpha t} e^{\alpha t} = 0$$

For an unrestricted waiting capacity $m = \infty$, the probabilities $p_n(t)$ with $n \ge 0$ have a Poisson distribution as described in Section 10.3.6.4. The mean of the Poisson distribution is αt; it is called the average length $\mu(t)$ of the queue at time t.

$$\mu(t) = \alpha t \qquad m = \infty$$

A Markov process which leads to a Poisson distribution is also called a Poisson process. An arrival process with constant rate of arrival and unrestricted waiting capacity is a Poisson process.

Service model : Assume that several customers are waiting at a given service station at time $t = 0$. Over time, the customers are served in the order of their arrival with the rate of service β. Like the arrival process, the service process is a homogeneous Markov process in continuous time $t \ge 0$. It is represented by the following transition graph.



Waiting time : The waiting time T_n of the n-th customer in a queue is the time which the service process takes to reach the state 0 from the state n for the first time. The density functions $f_n(t)$ for the waiting times T_n are determined according to Section 10.5.3.3 by solving the following initial value problem :

$$f_1'(t) = -\beta\, f_1(t) \qquad f_1(0) = \beta$$

$$f_n'(t) = -\beta\, f_n(t) + \beta\, f_{n-1}(t) \qquad f_n(0) = 0 \qquad n > 1$$

The solutions of this initial value problem are given by :

$$f_n(t) = \beta \frac{(\beta t)^{n-1}}{(n-1)!} e^{-\beta t} \qquad n > 0$$

According to Section 10.3.7.1, the density function $f_n(t)$ for the waiting time T_n is a gamma distribution. The mean n/β of the gamma distribution is called the average waiting time μ_n of the n-th customer.

$$\mu_n = n/\beta$$

Service time : The waiting time T_1 of the first customer in a queue is called the service time. It is exponentially distributed. The mean $1/\beta$ of the exponential distribution is called the average service time.

$$f_1(t) = \beta e^{-\beta t} \qquad t \geq 0$$

$$\mu_1 = 1/\beta$$

The probability that the service time is $T_1 \geq t$ is calculated as follows :

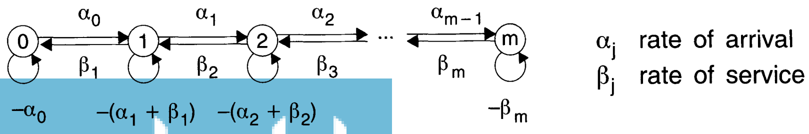
$$P(T_1 \geq t) = \int_t^{\infty} f_1(\tau) d\tau = e^{-\beta t}$$

If the customer is still being served at time $t \geq 0$, then $\overline{T}_1 = T_1 - t$ is the remaining service time. The probability that the remaining service time is $\overline{T}_1 \geq s$ is determined as follows :

$$P(\overline{T}_1 \geq s) = P(T_1 \geq t + s \mid T_1 \geq t) = \frac{e^{-\beta(t+s)}}{e^{-\beta t}} = e^{-\beta s} = P(T_1 \geq s)$$

Since $P(\overline{T}_1 \geq s) = P(T_1 \geq s)$, the service time T_1 and the remaining service time $\overline{T}_1$ have the same distribution. Thus the remaining service time $\overline{T}_1$ of the customer does not depend on how long the customer has already been served. This property is a consequence of the "lack of memory" of the process.

Queue model : A queueing process consists of an arrival process and a service process. Let the waiting capacity be restricted to m customers. The length of the queue at time t is a random variable X(t) which takes natural values in the interval [0,m]. The rates of arrival α_n and the rates of service β_n generally depend on the length n of the queue. The queueing process is a homogeneous Markov process in continuous time $t \geq 0$. It is represented by the following transition graph :



Stationary queue : The queueing process has a unique equilibrium distribution. It coincides with the limit distribution to which the process tends for $t \to \infty$ independent of the initial distribution for $t = 0$. If the initial distribution for $t = 0$ coincides with the equilibrium distribution, the process is stationary. The equilibrium distribution is therefore also called the stationary distribution of the queue. The probabilities $\overset{*}{p}_n$ of the equilibrium distribution are the probabilities that the queue has length n in the stationary case. According to Section 10.5.3.2, they satisfy the following system of linear equations :

$$\begin{array}{lll} -\alpha_0 \overset{*}{p}_0 + \beta_1 \overset{*}{p}_1 & = 0 & \\ \alpha_{n-1} \overset{*}{p}_{n-1} - (\alpha_n + \beta_n)\overset{*}{p}_n + \beta_{n+1} \overset{*}{p}_{n+1} & = 0 & 0 < n < m \\ \alpha_{m-1} \overset{*}{p}_{m-1} - \beta_m \overset{*}{p}_m & = 0 & \end{array}$$

The solutions of this system of equations are given by :

$$\overset{*}{p}_n = \overset{*}{p}_{n-1} \frac{\alpha_{n-1}}{\beta_n} \qquad 0 < n \le m$$

$$\overset{*}{p}_n = \overset{*}{p}_0 \frac{\alpha_0 \cdots \alpha_{n-1}}{\beta_1 \cdots \beta_n} = \overset{*}{p}_0 \prod_{j=1}^{n} \frac{\alpha_{j-1}}{\beta_j}$$

The probability $\overset{*}{p}_0$ is determined by the condition that the sum of all probabilities $\overset{*}{p}_j$ for $0 \le j \le m$ is one.

$$\overset{*}{p}_0 + \sum_{n=1}^{m} \overset{*}{p}_j = 1$$

$$\overset{*}{p}_0 = \frac{1}{1 + C} \qquad C = \sum_{n=1}^{m} \prod_{j=1}^{n} \frac{\alpha_{j-1}}{\beta_j}$$

The stationary distribution depends only on the ratios α_{j-1} / β_j. These ratios are introduced as traffic densities ϱ_j :

$$\varrho_j := \alpha_{j-1} / \beta_j \qquad 0 < j \le m$$

$$C = \sum_{n=1}^{m} \prod_{j=1}^{n} \varrho_j$$

$$\overset{*}{p}_n = \frac{1}{1 + C} \prod_{j=1}^{n} \varrho_j \qquad 0 < n \le m$$

A model with an unrestricted waiting room has a stationary distribution only if C tends to a finite value for $m \to \infty$. The stationary distributions of the queue length are treated in the following for some typical queue models.

Queue model M / M / 1 / m : The queue model M / M / 1 / m consists of one service station with restricted waiting capacity. The arrival process and the service process are Markov processes. The rates of arrival and rates of service are independent of the length n of the queue.

$$\alpha_n = \alpha$$

$$\beta_n = \beta$$

$$\varrho_n = \varrho = \alpha / \beta$$

For $\varrho = 1$, the stationary distribution of the queue length is a uniform distribution :

$$C = \sum_{n=1}^{m} \varrho^n = m \qquad \varrho = 1$$

$$\overset{*}{p}_n = 1 / (1 + m) \qquad 0 \le n \le m$$

For $\varrho \neq 1$, the stationary distribution of the length n of the queue decreases monotonically with increasing n for $\varrho < 1$ and increases monotonically with increasing n for $\varrho > 1$.

$$C = \sum_{n=1}^{m} \varrho^n = \varrho \frac{1 - \varrho^m}{1 - \varrho} \qquad \varrho \neq 1$$

$$\overset{*}{p}_n = \varrho^n (1 - \varrho) / (1 - \varrho^{m+1}) \qquad 0 \le n \le m$$

The probability $\overset{*}{p}_m$ is the probability that the waiting capacity is exhausted. It is also the probability that an arriving customer is turned away.

Queue model M / M / 1 / ∞ : The queue model M / M / 1 / ∞ differs from the queue model M / M / 1 / m only in that the waiting capacity is unrestricted. The stationary distribution of the queue length is therefore obtained by taking the limit $m \to \infty$. It exists only for $\varrho < 1$ and is a geometric distribution.

$$C = \sum_{n=1}^{\infty} \varrho^n = \frac{\varrho}{1 - \varrho} \qquad \varrho < 1$$

$$\overset{*}{p}_n = \varrho^n (1 - \varrho) \qquad n \ge 0$$

The existence of a stationary distribution for $\varrho = \alpha / \beta < 1$ is plausible. The condition $\alpha < \beta$ means that on average fewer customers arrive than can be served. The queue length can therefore reach an equilibrium. This is not possible for $\alpha \ge \beta$. The average length $\overset{*}{\mu}$ of the queue is the mean of the geometric distribution.

$$\overset{*}{\mu} = \varrho / (1 - \varrho)$$

The waiting time W of a customer is the time between the arrival at the service station and the departure from the service station. At arrival, the customer finds a queue with n customers with probability $\overset{*}{p}_n$. In the case of n waiting customers, the waiting time is T_{n+1}. The average waiting time W the arriving customer should expect is given by

$$W = \sum_{n=0}^{\infty} \overset{*}{p}_n \, T_{n+1}$$

The waiting time T_{n+1} for serving n + 1 customers with the constant rate of service β has a gamma distribution with the density function $f_{n+1}(t)$. This leads to the following density function $f_W(t)$ for the waiting time W :

$$f_W(t) = \sum_{n=0}^{\infty} \overset{*}{p}_n \, f_{n+1}(t) = \sum_{n=0}^{\infty} (1-\varrho)\, \varrho^n \, \beta \, \frac{(\beta t)^n}{n!} \, e^{-\beta t}$$

$$f_W(t) = (1-\varrho)\, \beta \, e^{-\beta t} \sum_{n=0}^{\infty} \frac{(\varrho\beta t)^n}{n!} = (1-\varrho)\, \beta \, e^{-\beta t} \, e^{\varrho\beta t}$$

$$f_W(t) = (1-\varrho)\, \beta \, e^{-(1-\varrho)\beta t} = (\beta-\alpha)\, e^{-(\beta-\alpha)t}$$

The waiting time W of an arriving customer is exponentially distributed with the parameter $\beta-\alpha$. The average waiting time μ_W is the mean of the exponential distribution.

$$\mu_W = 1 \,/\, (\beta-\alpha)$$

Queue model M / M / s / ∞ : The queue model M / M / s / ∞ consists of $s \geq 1$ service stations and an unrestricted waiting capacity for a queue. The arrival process and the service process are Markov processes. The rates of arrival and the rates of service per service station are independent of the length n of the queue. Up to s customers can be served simultaneously at the s service stations.

$$\begin{array}{llllllll}
\alpha_n = \alpha & & & & & & \\
\beta_n = n\,\beta & \text{for} & 0 < n \leq s & \quad & \beta_n = s\,\beta & \text{for} & n \geq s \\
\varrho_n = \varrho / n & \text{for} & 0 < n \leq s & \quad & \varrho_n = \varrho / s & \text{for} & n \geq s \\
\varrho = \alpha / \beta & & & & & &
\end{array}$$

Assuming $\varrho / s < 1$, the following stationary distribution of the queue length is obtained :

$$\overset{*}{p}_n = \frac{1}{1+C} \frac{\varrho^n}{n!} \qquad 0 \leq n \leq s$$

$$\overset{*}{p}_n = \frac{1}{1+C} \frac{\varrho^s}{s!} \left(\frac{\varrho}{s}\right)^{n-s} \qquad n \geq s$$

$$C = \sum_{n=1}^{\infty} \prod_{j=1}^{n} \varrho_j = \sum_{n=1}^{s} \frac{\varrho^n}{n!} + \sum_{n=s+1}^{\infty} \frac{\varrho^s}{s!} \left(\frac{\varrho}{s}\right)^{n-s}$$

$$C = \sum_{n=1}^{s} \frac{\varrho^n}{n!} + \frac{\varrho^s}{s!} \sum_{j=1}^{\infty} \left(\frac{\varrho}{s}\right)^{j} = \sum_{n=1}^{s} \frac{\varrho^n}{n!} + \frac{\varrho^s}{s!} \frac{\varrho / s}{1-\varrho / s}$$

Queue model M / M / ∞ / ∞ : The queue model M / M / ∞ / ∞ differs from the queue model M / M / s / ∞ in that it possesses an infinite number of service stations. Each arriving customer is served immediately. The number of customers being served is the length of the queue. Taking the limit $s \rightarrow \infty$ for the service model M / M / s / ∞ yields :

$$\begin{array}{lll} \alpha_n &= \alpha & n \geq 0 \\ \beta_n &= n\,\beta & n > 0 \\ \varrho_n &= \varrho / n & n > 0 \\ \varrho &= \alpha / \beta & \end{array}$$

The stationary distribution of the queue length is a Poisson distribution.

$$C = \sum_{n=1}^{\infty} \prod_{j=1}^{n} \varrho_j = \sum_{n=1}^{\infty} \frac{\varrho^n}{n!} = e^{\varrho} - 1$$

$$\overset{*}{p}_n = \frac{1}{1+C} \frac{\varrho^n}{n!} = \frac{\varrho^n}{n!} e^{-\varrho}$$

The average length $\overset{*}{\mu}$ of the queue is the mean of the Poisson distribution.

$$\overset{*}{\mu} = \varrho$$

The queue model M / M / ∞ / ∞ corresponds to a queue model with only one service station if the arriving customers join the queue with a probability depending on the queue length. If the rate of arrival α_n is inversely proportional to the length $n+1$ of the queue, one obtains :

$$\begin{array}{lll} \alpha_n &= \alpha / (n+1) & n \geq 0 \\ \beta_n &= \beta & n > 0 \\ \varrho_n &= \alpha_{n-1} / \beta = \varrho / n & n > 0 \\ \varrho &= \alpha / \beta & \end{array}$$

This queue model possesses the same traffic densities ϱ_n as the queue model M / M / ∞ / ∞. The two queue models therefore have the same stationary distribution of the queue length.

Example 1 : Client and server

Let several clients be connected to a server in a computer network. The clients access the server 20 times per minute on average. Every access involves a request which on average takes 2.0 seconds to process. The computer configuration corresponds to the queue model M / M / 1 / ∞. From the average number of accesses and the time required to process a request, the rate of arrival α, the rate of service β and the traffic density ϱ are calculated as follows :

$$\begin{array}{llll} \text{rate of arrival} & \alpha = 20/60\ \mathrm{s}^{-1} & = 1/3\ \mathrm{s}^{-1} \\ \text{rate of service} & \beta = 1/2.0\ \mathrm{s}^{-1} & = 1/2\ \mathrm{s}^{-1} \\ \text{traffic density} & \varrho = \alpha/\beta & = 2/3 \end{array}$$

The stationary queue for the server accesses is geometrically distributed with the average length $\overset{*}{\mu}$:

$$\overset{*}{p}_n = \varrho^n (1-\varrho) = \left(\frac{2}{3}\right)^n \frac{1}{3} \qquad n \geq 0$$

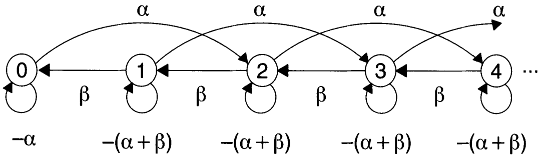
$$\overset{*}{\mu} = \varrho / (1-\varrho) = 2.0$$

The waiting time required until a request is processed is called the response time. It is exponentially distributed with the mean μ_W.

$$f_W(t) = (\beta-\alpha)\, e^{-(\beta-\alpha)t} \qquad t \geq 0$$

$$\mu_W = 1/(\beta-\alpha) = 6.0 \text{ s}$$

Queue model with arrival in groups : In the preceding queue models the customers are assumed to arrive at the service station one by one. The customers may, however, also arrive in groups. In the simplest queue model with arrival in groups, it is assumed that every customer group consists of k customers, that customer groups arrive at the service station with the rate of arrival α and that each customer is served with the rate of service β. This queueing process is a homogeneous Markov process in continuous time $t \geq 0$. It is represented in the following transition graph for the group size $k = 2$.



α rate of arrival

β rate of service

Queue model with service in phases : In the preceding queue models it is assumed that customers are served in a single phase. The service may, however, consist of several phases. In the simplest queue model with service in phases, it is assumed that the customers arrive at the service station one by one with the rate of arrival α, that each customer is served in k consecutive service phases and that each service phase is determined by the rate of service β. While the service time in the preceding queue models is exponentially distributed with the parameter β, in the case of service in phases the service time has a gamma distribution with the parameters k and β.

The queue model with arrival in groups is equivalent to the service model with service in phases. A customer served in k phases corresponds to a group of customers consisting of k customers. A customer m in the phase $1 \leq j \leq k$ in the queue model with service in phases corresponds to a customer $n = k(m-1) + j$ in the queue model with arrival in groups.

Stationary queue : The simple queue model with arrival in groups and an unrestricted waiting capacity has a unique equilibrium distribution for $\alpha / \beta < 1 / k$. It corresponds to the stationary distribution of the length of the queue. According to Section 10.5.3.2, the probabilities $\overset{*}{p}_n$ of the equilibrium distribution satisfy the following system of linear equations :

$$\begin{array}{lll} -\alpha \overset{*}{p}_0 + \beta \overset{*}{p}_1 & = 0 & \\ -(\alpha + \beta) \overset{*}{p}_n + \beta \overset{*}{p}_{n+1} & = 0 & 0 < n < k \\ \alpha \overset{*}{p}_{n-k} - (\alpha + \beta) \overset{*}{p}_n + \beta \overset{*}{p}_{n+1} & = 0 & k \le n \end{array}$$

With $\varrho = \alpha / \beta$, the probabilities $\overset{*}{p}_n$ may be calculated recursively starting from the probability $\overset{*}{p}_0$:

$$\begin{array}{lll} \overset{*}{p}_1 & = \varrho \overset{*}{p}_0 & \\ \overset{*}{p}_{n+1} & = (1 + \varrho) \overset{*}{p}_n & 0 < n < k \\ \overset{*}{p}_{n+1} & = (1 + \varrho) \overset{*}{p}_n - \varrho \overset{*}{p}_{n-k} & k \le n \end{array}$$

The probability $\overset{*}{p}_0$ is determined by the condition that the sum of the probabilities $\overset{*}{p}_n$ for $n \ge 0$ is one. This leads to the following formula :

$$\overset{*}{p}_0 = 1 - k\varrho$$

The probability $\overset{*}{p}_{j,m}$ is the probability that the j-th customer of the first customer group is being served and $m - 1$ customer groups with k customers each are waiting. It is equal to the probability that the queue has length $n = km - j + 1$.

$$\overset{*}{p}_{j,m} = \overset{*}{p}_{km-j+1} \qquad m = 1, 2, \ldots \qquad j = 1, \ldots, k$$

The probability $\overset{*}{p}_m$ is the probability that m customer groups are waiting or being served. It is calculated by summing the probabilities $\overset{*}{p}_{j,m}$ for $j = 1, \ldots, k$. The probability that no customer group is waiting or being served is $\overset{*}{p}_0$.

$$\overset{*}{p}_m = \sum_{j=1}^{k} \overset{*}{p}_{j,m} \qquad m = 1, 2, \ldots$$

The equilibrium distribution of the queue model with service in phases is calculated like the distribution for the queue model with arrival in groups. The probability $\overset{*}{p}_m$ is the probability that the queue has length m.

Example 2 : Service in phases

Consider a service model in which each customer is served in k = 3 phases. Let the rate of service in each phase be β. The service time for a customer has a gamma distribution with the density function $f_3(t)$ and the average service time μ_3.

$$f_3(t) = \beta \frac{(\beta t)^2}{2!} e^{-\beta t} \qquad\qquad \mu_3 = 3/\beta$$

For comparison, consider a service model in which every customer is served in one phase with the rate of service $\beta / 3$. Let the service time be exponentially distributed with the density function $f_1(t)$ and the average service time μ_1. The rate of service is chosen such that the average service times of the two service models coincide.

$$f_1(t) = \frac{\beta}{3} e^{-\beta t/3} \qquad\qquad \mu_1 = 3/\beta = \mu_3$$

The probability distributions of the stationary queue are calculated for the two queue models with $\alpha/\beta = 0.2$. The probabilities $\overset{*}{p}_m$ for the lengths m of the queue are compiled in the following table :

m	0	1	2	3	4	5	6	7	8
$\overset{*}{p}_m$ (k = 3)	0.400	0.291	0.158	0.078	0.038	0.018	0.009	0.004	0.002
$\overset{*}{p}_m$ (k = 1)	0.400	0.240	0.144	0.086	0.052	0.031	0.019	0.011	0.007

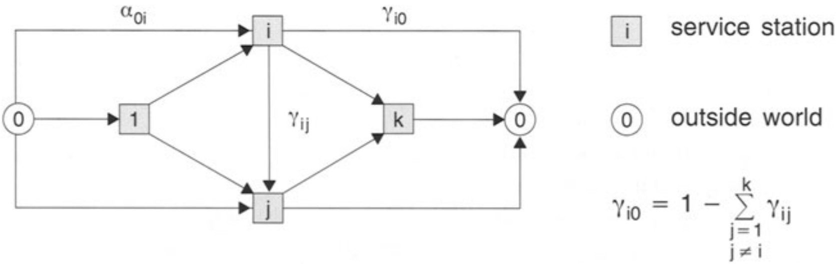
The model with k = 1 corresponds to a queue model M / M / 1 / ∞ with the traffic density $\varrho = \alpha/\beta = 0.6$. In the stationary case, it has a geometric distribution. Both models yield the same probability that the queue has length m = 0. The queue model with k = 3 phases exhibits higher probabilities for the lengths m = 1, 2 than the other model, and lower probabilities for $m \geq 3$. The average length of the queue is therefore also shorter than in the other model.

Generalization : Different rates of service for the individual phases are often considered in practical applications of the queue model with service in phases. In this case the service time for a customer is the sum of exponentially distributed service times for the different phases. Its distribution may be analytically determined from the conditions for first passage from the beginning of the first phase to the end of the last phase. The stationary distribution of the queue is determined from the equilibrium conditions.

10.5.3.5 Queue systems

Introduction : A queue system is a network of service stations, each with its own queue. Customers arrive at the various service stations from the outside. After being served at a service station, they go to another service station or leave the system with a certain probability. A queue forms at each service station. The behavior of such a network of service stations is treated as a multidimensional Markov process. The equilibrium state of a queue system is particularly important in practical applications. The fundamentals of simple queue systems with exponentially distributed service time are treated in the following.

Queue systems : A queue system consists of k service stations. Customers arrive at the service station i with the rate of arrival α_{0i}. After being served at the service station i, the customers go to another service station $j \neq i$ with probability γ_{ij} or leave the system with probability γ_{i0}. A queue system, including the outside world, is represented in a transition graph. A simple example of a queue system is shown in the following diagram.



Customers arriving at the service station i are served with the rate of service β_i. Each service station i has its own queue with unrestricted waiting capacity. The length of this queue at time t is a random variable $X_i(t)$ which takes natural numbers $n_i \in \mathbb{N}$. The state of the queue system at time t is described by a random vector $\mathbf{X}(t)$ containing the lengths of the queues at all service stations. The equilibrium state of the queue system is of particular interest; it corresponds to the limit $t \rightarrow \infty$.

Rates of arrival and departure : If a queue system is in equilibrium, then on average as many customers must leave a service station per unit of time as are arriving there. The rate of departure at a service station is thus equal to the rate of arrival. The rate of arrival α_i at a service station is the sum of the rate α_{0i} with which customers are arriving from the outside and the rates of arrival $\alpha_j \gamma_{ji}$ for the customers coming from the service stations $j \neq i$.

$$\alpha_i = \alpha_{0i} + \sum_{\substack{j=1 \\ j \neq i}}^{k} \alpha_j \gamma_{ji} \qquad\qquad 1 \le i \le k$$

The sum of the rates α_{0i} with which customers are arriving at the service stations from the outside is equal to the sum of the rates $\alpha_i \gamma_{i0}$ with which customers are departing from the service stations to leave the system.

$$\begin{aligned} \sum_{i=1}^{k} \alpha_{0i} &= \sum_{i=1}^{k} (\alpha_i - \sum_{\substack{j=1 \\ j \neq i}}^{k} \alpha_j \gamma_{ji}) = \sum_{i=1}^{k} \alpha_i - \sum_{j=1}^{k} \alpha_j \sum_{\substack{i=1 \\ i \neq j}}^{k} \gamma_{ji} = \\ &= \sum_{i=1}^{k} \alpha_i - \sum_{j=1}^{k} \alpha_j (1 - \gamma_{j0}) = \sum_{j=1}^{k} \alpha_j \gamma_{j0} \end{aligned}$$

The equations which determine the rates of arrival α_i are linear; they are conveniently formulated in vector and matrix notation. The rates of arrival α_i and α_{0i} are arranged in the vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}_0$. The transition probabilities γ_{ij} for $i \neq j$ are arranged in the matrix $\boldsymbol{\Gamma}$.

$$(\mathbf{I} - \boldsymbol{\Gamma}^T)\, \boldsymbol{\alpha} = \boldsymbol{\alpha}_0 \qquad\qquad \boldsymbol{\alpha} \ge \mathbf{0}$$

The system of linear equations has unique solutions α_i if and only if every service station can be reached from every other service station (possibly via the outside world). In the following, the rates of arrival α_i for all service stations of the queue system are assumed to be uniquely determined.

Equilibrium distribution : A state of the queue system is described by a state vector **n** containing the length $n_i \ge 0$ of the queue at each service station $1 \le i \le k$. If the rate of arrival α_i for each service station is less than the rate of service β_i, then an equilibrium distribution exists. The probability $\overset{*}{p}_{\mathbf{n}}$ for the state **n** of the queue system in equilibrium is :

$$\begin{aligned} \varrho_i &= \alpha_i / \beta_i & \qquad \varrho_i &< 1 \\ \overset{*}{p}_{n_i} &= (1 - \varrho_i)\, \varrho_i^{n_i} & \qquad n_i &\ge 0 \\ \overset{*}{p}_{\mathbf{n}} &= \prod_{i=1}^{k} \overset{*}{p}_{n_i} & \qquad \mathbf{n} &\ge \mathbf{0} \end{aligned}$$

The equilibrium distribution $\overset{*}{p}_{\mathbf{n}}$ of the queue system is the product of geometric distributions, each of which corresponds to the equilibrium distribution $\overset{*}{p}_{n_i}$ for a service station i with the traffic density ϱ_i according to the queue model M / M / 1 / ∞. Thus in equilibrium the queue system behaves like k independent queue models M / M / 1 / ∞ for the service stations.

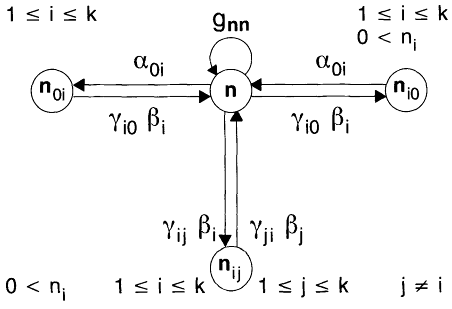
Proof : The matrix **G** of the queue system contains the transition rates for the transitions between the following states :

$$\begin{aligned}
\mathbf{n} &= (\ldots, n_i \quad, \ldots, n_j \quad, \ldots) \\
\mathbf{n}_{0i} &= (\ldots, n_i + 1, \ldots, n_j \quad, \ldots) \\
\mathbf{n}_{i0} &= (\ldots, n_i - 1, \ldots, n_j \quad, \ldots) \\
\mathbf{n}_{ij} &= (\ldots, n_i - 1, \ldots, n_j + 1, \ldots)
\end{aligned}$$

The transitions between the states have the following significance :

- The state **n** changes to the state $\mathbf{n}_{0i}$ if a customer comes to the station i from the outside. The transition rate is α_{0i}.
- The state **n** changes to the state $\mathbf{n}_{i0}$ if a customer leaves the station i after being served and departs. The transition rate is $\gamma_{i0}\,\beta_i$ for $n_i > 0$.
- The state **n** changes to the state $\mathbf{n}_{ij}$ if a customer goes to the station j after being served at the station $i \neq j$. The transition rate is $\gamma_{ij}\,\beta_i$ for $n_i > 0$.
- The state $\mathbf{n}_{i0}$ changes to the state **n** if a customer comes to the station i from the outside. The transition rate is α_{0i} for $n_i > 0$.
- The state $\mathbf{n}_{0i}$ changes to the state **n** if a customer leaves the station i after being served and departs. The transition rate is $\gamma_{i0}\,\beta_i$.
- The state $\mathbf{n}_{ij}$ changes to the state **n** if a customer goes to the station i after being served at the station $j \neq i$. The transition rate is $\gamma_{ji}\,\beta_j$ for $n_i > 0$.

The transitions between the state **n** and the neighboring states $\mathbf{n}_{0i}$, $\mathbf{n}_{i0}$, $\mathbf{n}_{ij}$ and the corresponding transition rates may be graphically represented as follows :



$$\begin{aligned}
g_{\mathbf{n}\,\mathbf{n}_{i0}} &= \gamma_{i0}\,\beta_i \\
g_{\mathbf{n}\,\mathbf{n}_{0i}} &= \alpha_{0i} \\
g_{\mathbf{n}\,\mathbf{n}_{ij}} &= \gamma_{ij}\,\beta_i \\
g_{\mathbf{n}_{i0}\,\mathbf{n}} &= \alpha_{0i} \\
g_{\mathbf{n}_{0i}\,\mathbf{n}} &= \gamma_{i0}\,\beta_i \\
g_{\mathbf{n}_{ij}\,\mathbf{n}} &= \gamma_{ji}\,\beta_j
\end{aligned}$$

The rate $g_{\mathbf{nn}}$ is the negative sum of the transition rates of all possible transitions from the state **n** to the neighboring states :

$$g_{\mathbf{nn}} = -\sum_{i=1}^{k} g_{\mathbf{nn}_{0i}} - \sum_{\substack{i=1 \\ n_i > 0}}^{k} \left(g_{\mathbf{nn}_{i0}} + \sum_{\substack{j=1 \\ j \neq i}}^{k} g_{\mathbf{nn}_{ij}} \right)$$

If the probabilities $\overset{*}{p}_{\mathbf{n}}$ for all possible states $\mathbf{n} \geq \mathbf{0}$ are arranged in the vector $\overset{*}{\mathbf{p}}$ and the transition rates for all possible transitions between states are arranged in the matrix $\mathbf{G}$, the equilibrium condition becomes $\mathbf{G}^T \overset{*}{\mathbf{p}} = \mathbf{0}$. For the state $\mathbf{n}$ this yields :

$$g_{\mathbf{nn}} \overset{*}{p}_{\mathbf{n}} + \sum_{i=1}^{k} g_{\mathbf{n}_{0i}\mathbf{n}} \overset{*}{p}_{\mathbf{n}_{0i}} + \sum_{\substack{i=1\\ n_i>0}}^{k} (g_{\mathbf{n}_{i0}\mathbf{n}} \overset{*}{p}_{\mathbf{n}_{i0}} + \sum_{\substack{j=1\\ j\neq i}}^{k} g_{\mathbf{n}_{ij}\mathbf{n}} \overset{*}{p}_{\mathbf{n}_{ij}}) = 0$$

The above theorem for the equilibrium distribution yields the following state probabilities :

$$\begin{aligned} \varrho_i &= \alpha_i / \beta_i \\ \overset{*}{p}_{\mathbf{n}_{0i}} &= \overset{*}{p}_{\mathbf{n}} \varrho_i \\ \overset{*}{p}_{\mathbf{n}_{i0}} &= \overset{*}{p}_{\mathbf{n}} / \varrho_i \\ \overset{*}{p}_{\mathbf{n}_{ij}} &= \overset{*}{p}_{\mathbf{n}} \varrho_j / \varrho_i \end{aligned}$$

The state probabilities and transition rates are substituted into the equilibrium condition :

$$\sum_{i=1}^{k} (\gamma_{i0} \beta_i \varrho_i - \alpha_{0i}) + \sum_{\substack{i=1\\ n_i>0}}^{k} (\alpha_{0i}/\varrho_i - \gamma_{i0} \beta_i + \sum_{\substack{j=1\\ j\neq i}}^{k} (\gamma_{ji} \beta_j \varrho_j / \varrho_i - \gamma_{ij} \beta_i)) = 0$$

With the relationships

$$\gamma_{i0} = 1 - \sum_{\substack{j=1\\ j\neq i}}^{k} \gamma_{ij} \qquad \alpha_i = \alpha_{0i} + \sum_{\substack{j=1\\ j\neq i}}^{k} \alpha_j \gamma_{ji}$$

$$\sum_{i=1}^{k} \alpha_{0i} = \sum_{i=1}^{k} \alpha_i \gamma_{i0} \qquad \varrho_i = \frac{\alpha_i}{\beta_i}$$

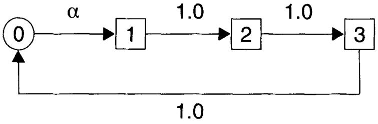
the equilibrium condition becomes

$$\begin{aligned} &\sum_{i=1}^{k} (\gamma_{i0} \alpha_i - \alpha_{0i}) + \sum_{\substack{i=1\\ n_i>0}}^{k} (\alpha_{0i}/\varrho_i - \beta_i + \sum_{\substack{j=1\\ j\neq i}}^{k} (\gamma_{ji} \ \alpha_j / \varrho_i)) = \\ &\sum_{i=1}^{k} (\gamma_{i0} \alpha_i - \alpha_{0i}) + \sum_{\substack{i=1\\ n_i>0}}^{k} (\alpha_i / \varrho_i - \beta_i) = \\ &\sum_{i=1}^{k} (\gamma_{i0} \alpha_i - \alpha_{0i}) + \sum_{\substack{i=1\\ n_i>0}}^{k} (\beta_i - \beta_i) = 0 \end{aligned}$$

Since the equilibrium condition is satisfied, the above equilibrium distribution $\overset{*}{p}_{\mathbf{n}}$ for the queue system is shown to be correct.

Example 1 : Serial queue system

Let a queue system with 3 service stations in a serial arrangement be given. On average, α customers arrive at the service station 1 from the outside per unit of time. The average service time for a customers at each service station is $1/\beta$.



The rates of arrival α_i at the service stations are obtained by solving the following system of linear equations :

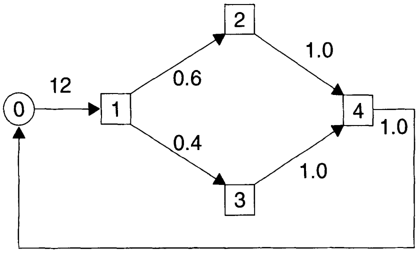
$$\begin{aligned} \alpha_1 &= \alpha \\ \alpha_2 &= \alpha_1 \\ \alpha_3 &= \alpha_2 \end{aligned}$$

The solution is $\alpha_1 = \alpha_2 = \alpha_3 = \alpha$. The rate of service for each service station i is β. The following equilibrium distribution is obtained in terms of the traffic density $\varrho = \alpha / \beta$:

$$\overset{*}{p}_{\mathbf{n}} = \prod_{i=1}^{3} (1-\varrho)\, \varrho^{n_i} = (1-\varrho)^3\, \varrho^{n_1+n_2+n_3}$$

Example 2 : Queue system with a branching

Let the illustrated queue system with 4 service stations be given. On average 12 customers arrive at station 1 from the outside per unit of time. After being served at station 1 they go to stations 2 and 3 with probabilities 0.6 and 0.4, respectively, and then leave the queue system via station 4.



α_0
12.0
0.0
0.0
0.0

Γ			
0.0	0.6	0.4	0.0
0.0	0.0	0.0	1.0
0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0

The rates of arrival from the outside at all stations are arranged in the vector $\boldsymbol{\alpha}_0$. The transition probabilities from every station to every other station are arranged in the matrix $\boldsymbol{\Gamma}$. The rates of arrival at all stations for the equilibrium state of the queue system are arranged in the vector $\boldsymbol{\alpha}$. They are calculated by solving a system of linear equations :

$$(\mathbf{I} - \boldsymbol{\Gamma}^{\mathrm{T}}) \quad * \quad \boldsymbol{\alpha} \quad = \quad \boldsymbol{\alpha}_0$$

$$\begin{array}{|c|c|c|c|}\hline 1.0 & & & \\ \hline -0.6 & 1.0 & & \\ \hline -0.4 & 0.0 & 1.0 & \\ \hline 0.0 & -1.0 & -1.0 & 1.0 \\ \hline \end{array} * \begin{array}{|c|}\hline 12.0 \\ \hline 7.2 \\ \hline 4.8 \\ \hline 12.0 \\ \hline \end{array} = \begin{array}{|c|}\hline 12.0 \\ \hline 0.0 \\ \hline 0.0 \\ \hline 0.0 \\ \hline \end{array}$$

Let the traffic density for all stations be $\varrho = 0.8$. The rates of service for the stations are determined as follows :

$$\begin{aligned} \beta_1 &= \alpha_1 / \varrho = 12.0 / 0.8 = 15.0 \\ \beta_2 &= \alpha_2 / \varrho = 7.2 / 0.8 = 9.0 \\ \beta_3 &= \alpha_3 / \varrho = 4.8 / 0.8 = 6.0 \\ \beta_4 &= \alpha_4 / \varrho = 12.0 / 0.8 = 15.0 \end{aligned}$$

In equilibrium, the stations of the queue system are occupied uniformly, since they all have the same traffic density. The probability that n_1, n_2, n_3, n_4 customers are at stations 1, 2, 3, 4 is given by

$$\overset{*}{p}_{\mathbf{n}} = \prod_{i=1}^{4} (1-\varrho)\, \varrho^{n_i} = (1-\varrho)^4\, \varrho^{n_1+n_2+n_3+n_4}$$

10.5.4 STATIONARY PROCESSES

10.5.4.1 Introduction

A random process is described by a random function of time. If the random function has certain time-independent stochastic characteristics, then the random process is stationary. For many applications it is sufficient to consider the moments of random functions. The mean and the variance of a stationary random process defined on the time domain from $-\infty$ to ∞ are constant over time. The states of the stationary random process at consecutive times are linearly dependent. This dependence is described by the covariance function or the correlation function. The basic definitions for stationary random processes rely on probability theory for random vectors; they are treated in Section 10.5.4.2.

The time domain of a stationary random process may be discrete or continuous. Accordingly, discrete and continuous processes are distinguished. They are related by an approximation. The uncorrelated process, the averaging process and the regression process are typical stationary random processes. The harmonic process is a particularly important process in continuous time. These processes are treated in Sections 10.5.4.3 and 10.5.4.4.

10.5.4.2 Probability distributions and moments

Introduction : A random process X(t) is said to be stationary if certain stochastic properties are independent of the time t. The conditions for stationary processes are formulated on the basis of probability distributions and moments. Probability theory for random vectors with several random variables is applied by considering the random process at different points in time. The basic definitions of stationary random processes with respect to their distribution functions and moments are treated in the following.

Distribution functions : A random process is described by a random function X(t), which takes real values as a function of the time t. For a given time t, the quantity X(t) is a random variable with a one-dimensional distribution function F(x ; t). A random process is stationary of order 1 if the one-dimensional distribution function is invariant with respect to a time shift t_0. An equivalent property is that the one-dimensional distributions are the same at all time points.

$$F(x;t) = F(x;t+t_0)$$

For two given time points t_1 and t_2 the quantities $X(t_1)$, $X(t_2)$ are random variables with a common two-dimensional distribution function $F(x_1, x_2\,;\,t_1, t_2)$. A random process is stationary of order 2 if every two-dimensional distribution function is invariant with respect to a time shift t_0. An equivalent property is that every two-dimensional distribution depends only on the time difference $\tau = t_2 - t_1$.

$$F(x_1, x_2\,;\,t_1, t_2) \;=\; F(x_1, x_2\,;\,t_1 + t_0, t_2 + t_0)$$

If a random process is stationary of order 2, then it is also stationary of order 1, since every one-dimensional distribution is a marginal distribution of a two-dimensional distribution, and hence also has the stationary property. In general form, stationary processes of order n may be defined by n-dimensional distributions which are invariant with respect to a time shift. However, for many practical applications it suffices to consider stationary processes of second order.

Moments : The stationary properties of a second-order process $X(t)$ lead to the stationary properties of its first and second moments. For a given time t, the mean of the random variable $X(t)$ is given by its first moment $E(X(t))$. This moment is invariant with respect to a time shift t_0 and is therefore constant.

$$E(X(t)) \;=\; E(X(t + t_0)) \;=\; \mu_X$$

For two given time points t_1 and t_2 the random variables $X(t_1)$, $X(t_2)$ possess the second central moment $D(X(t_1)\;X(t_2))$ with respect to their means. This moment is the variance for $t_1 = t_2$ and the covariance for $t_1 \neq t_2$. It is invariant with respect to a time shift t_0. Equivalently, the variance is constant and the covariance depends only on the time difference $\tau = t_2 - t_1$.

$$\begin{aligned}
D(X(t_1)\;X(t_1)) &= D(X(t_1 + t_0)\;X(t_1 + t_0)) = \sigma_X^2 & & \\
D(X(t_1)\;X(t_2)) &= D(X(t_1 + t_0)\;X(t_2 + t_0)) = \gamma_X(\tau) & \qquad \tau &= t_2 - t_1 \\
D(X(t_2)\;X(t_1)) &= D(X(t_2 + t_0)\;X(t_1 + t_0)) = \gamma_X(-\tau) & \qquad \tau &= t_2 - t_1
\end{aligned}$$

The second central moments are arranged in a covariance matrix $\mathbf{V}_X$, which is invariant with respect to a time shift and depends only on the time difference $\tau = t_2 - t_1$.

$$\mathbf{V}_X(\tau) \;=\; \begin{array}{|c|c|}\hline \sigma_X^2 & \gamma_X(\tau) \\ \hline \gamma_X(-\tau) & \sigma_X^2 \\ \hline \end{array}$$

Covariance function : The function $\gamma_X(\tau)$ for the second central moment of the random variables X(t) and $X(t+\tau)$ is called the covariance function. For $\tau = 0$ it is equal to the variance $\sigma_X^2 > 0$.

$$\gamma_X(0) = \sigma_X^2 > 0$$

The covariance matrix for two random variables is symmetric. The covariance function for stationary processes is therefore also symmetric.

$$\gamma_X(\tau) = \gamma_X(-\tau)$$

The covariance matrix for two random variables is positive semidefinite. Its determinant is non-negative. This implies the following inequality :

$$\gamma_X^2(\tau) \le \sigma_X^2$$

Correlation function : The function $\gamma_X(\tau)/\gamma_X(0)$ is called the correlation function and is designated by $\varrho_X(\tau)$. It is a measure of the linear dependence of the random variable $X(t+\tau)$ on the random variable X(t) and only takes values in the range from -1 to $+1$. The properties of the covariance function imply the following properties of the correlation function :

$$\varrho_X(\tau) = \gamma_X(\tau) / \gamma_X(0)$$

$$\varrho_X(0) = 1$$

$$\varrho_X(\tau) = \varrho_X(-\tau)$$

$$\varrho_X^2(\tau) \le 1$$

Gaussian process : A stationary process X(t) whose random variables $X(t_1), \ldots, X(t_n)$ at different times $t_1, \ldots, t_n$ have a multinormal distribution is called a Gaussian process. The multinormal distribution is completely described by the means $\mu_X(t_j)$, the variances $\sigma_X^2(t_j)$ and the covariances $\gamma_X(t_j, t_k)$ for j, k = 1,...,n. A Gaussian process is said to be strongly stationary if its multinormal distribution is invariant with respect to a time shift. This is the case if and only if the means $\mu_X(t_j)$ and the variances $\sigma_X^2(t_j)$ are constant and the covariances $\gamma_X(t_j, t_k)$ depend only on the time difference $\tau = t_k - t_j$. A strongly stationary Gaussian process is therefore completely described by the mean μ_X, the variance σ_X^2 and the correlation function $\varrho_X(\tau)$.

Ergodic process : A stationary process X(t) is said to be ergodic if the moments may be determined from a single realization x(t) of the process. For stationary processes in the continuous time domain $-\infty \le t \le \infty$, the mean μ_X and the covariance function $\gamma_X(\tau)$ are calculated by integrating over time :

$$\mu_X \quad = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(t)\, dt$$

$$\gamma_X(\tau) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} (x(t) - \mu_X)\,(x(t+\tau) - \mu_X)\, dt$$

The ergodicity of a stationary process is derived from the properties of the application or from statistical considerations. Often ergodicity is assumed if the absolute value of the correlation function $\varrho_X(\tau)$ decays sufficiently rapidly with increasing $\tau > 0$.

10.5.4.3 Stationary processes in discrete time

Introduction : A random process in discrete time is a sequence of random variables $X(t_n)$ at equidistant times t_n. Without loss of generality, the time points are assumed to be integers. The simplest stationary process is an uncorrelated process, for which the random variables $X(t_n)$ at different times t_n are stochastically independent. This process forms the basis for the construction of further processes. The averaging process and the regression processes of first and second order are typical examples. The rules of construction of these stationary processes are of fundamental importance for the computer-aided simulation of processes. The mean and the variance as well as the covariance function and the correlation function of the stationary process may be determined from the rule of construction.

Rules of calculation for moments : The rules of construction for stationary processes can often be reduced to linear combinations of random variables Z_j :

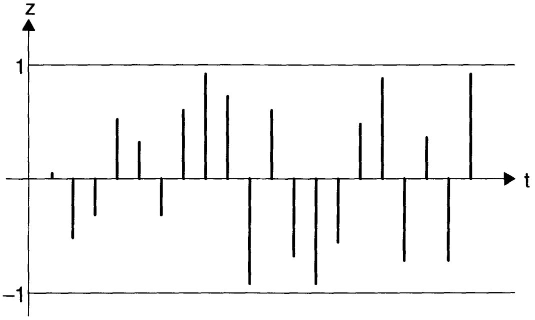
$$X = \sum_k a_k Z_k \qquad\qquad Y = \sum_j b_j Z_j$$

If the moments of the random variables Z_j are known, then the moments of the random variables X, Y may be obtained from them. Using the rules for linear combinations of random vectors in Section 10.4.4, the means are determined as first moments and the variances and covariances are determined as second central moments :

$$\begin{aligned}
E(X) &= E(\sum_k a_k Z_k) &&= \sum_k a_k E(Z_k) \\
D(X^2) &= D(\sum_k \sum_j a_k a_j Z_k Z_j) &&= \sum_k \sum_j a_k a_j D(Z_k Z_j) \\
D(X Y) &= D(\sum_k \sum_j a_k b_j Z_k Z_j) &&= \sum_k \sum_j a_k b_j D(Z_k Z_j)
\end{aligned}$$

In the following these rules of calculation are used to determine the mean, the variance and the covariance function of stationary processes.

Uncorrelated process : Let a sequence of random variables $Z(t_n)$ in the integer time domain $-\infty < t_n < \infty$ be given. Let the random variables be stochastically independent, and let them have identical distributions with the mean $\mu_Z = 0$ and the variance σ_Z^2. A realization of such a stationary process for a uniform distribution on the range $-1 \le z \le 1$ is shown below :



The first and second moments are given by :

$$\begin{aligned}
\mu_Z &= E(Z(t_n)) = 0 \\
\sigma_Z^2 &= D(Z^2(t_n)) = \gamma_Z(0) \\
\gamma_Z(\tau) &= D(Z(t_n)\, Z(t_{n+\tau})) = 0 \quad \text{for} \quad \tau \ne 0
\end{aligned}$$

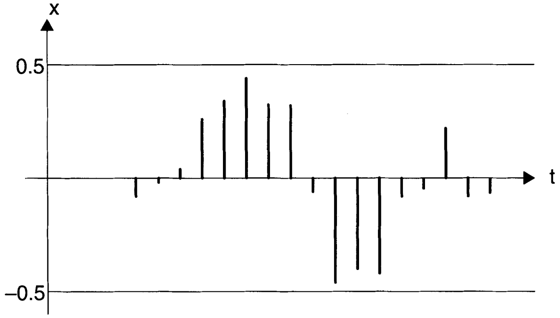
Due to the stochastic independence of the random variables $Z(t_n)$, the covariance function $\gamma_Z(\tau)$ and the correlation function $\varrho_Z(\tau)$ are zero for $\tau \ne 0$. Thus the stationary process is uncorrelated.

$$\begin{aligned}
\varrho_Z(\tau) &= \gamma_Z(\tau) / \gamma_Z(0) \\
\varrho_Z(0) &= 1 \\
\varrho_Z(\tau) &= 0 \quad \text{for} \quad \tau \ne 0
\end{aligned}$$

Averaging process : In an averaging process, the random variable $X(t_n)$ is formed by averaging over $r+1$ random variables $Z(t_{n-r})$ to $Z(t_n)$ of a stationary uncorrelated process with mean 0.

$$X(t_n) = \sum_{j=0}^{r} c_j Z(t_{n-j}) \qquad c_j = 1/(r+1)$$

A realization of the averaging process $X(t_n)$ with $r = 3$ for the process $Z(t_n)$ illustrated above is shown below :



Since the mean of the process $Z(t_n)$ is zero, the mean of the process $X(t_n)$ is also zero.

$$\mu_X = E(X(t_n)) = \sum_{j=0}^{r} c_j \, E(Z(t_{n-j})) = \sum_{j=0}^{r} c_j \, \mu_Z = 0$$

The covariance function is calculated as the second central moment for $\tau \geq 0$:

$$\gamma_X(\tau) = D(X(t_n)\,X(t_{n+\tau})) = \sum_{i=0}^{r} \sum_{k=0}^{r} c_j \, c_k \, D(Z(t_{n-j})\, Z(t_{n+\tau-k}))$$

$$\gamma_X(\tau) = \sum_{j=0}^{r} \sum_{k=0}^{r} c_j \, c_k \, \gamma_Z(\tau - k + j)$$

The covariance function $\gamma_Z(\tau)$ of the process $Z(t_n)$ is equal to the variance σ_Z^2 for $\tau = 0$ and zero for $\tau \neq 0$. This yields :

$$\gamma_X(\tau) = \sigma_Z^2 \sum_{k=\tau}^{r} c_{k-\tau} \, c_k = \sigma_Z^2 \, \frac{r + 1 - \tau}{(r + 1)^2} \qquad 0 \leq \tau \leq r + 1$$

The variance σ_X^2 is the value of the covariance function for $\tau = 0$:

$$\sigma_X^2 = \gamma_X(0) = \sigma_Z^2 \,/\, (r + 1)$$

The correlation function is obtained from the covariance function :

$$\varrho_X(\tau) = \gamma_X(\tau) \,/\, \gamma_X(0) \qquad \varrho_X(\tau) = \varrho_X(-\tau)$$

$$\varrho_X(\tau) = 1 - |\tau| \,/\, (r + 1) \qquad \text{for} \quad 0 \leq |\tau| \leq r + 1$$

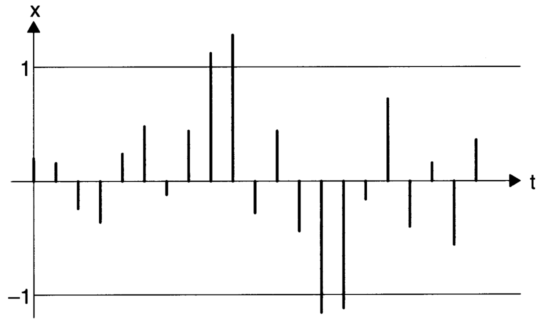
$$\varrho_X(\tau) = 0 \qquad \text{for} \quad |\tau| \geq r + 1$$

The averaging process $X(t_n)$ is a stationary process with mean 0 and variance σ_X^2. Its correlation function decays linearly from 1 to 0 for $0 \leq |\tau| \leq r + 1$.

Regression process of first order : In a regression process of first order, the random variable $X(t_n)$ depends linearly on the random variable $X(t_{n-1})$.

$$X(t_n) = \alpha X(t_{n-1}) + Z(t_n)$$

The additive term $Z(t_n)$ is a stationary uncorrelated process with mean 0. A realization of the regression process of first order with $\alpha = 0.5$ for the process $Z(t_n)$ illustrated above is shown below :



The random variable $X(t_n)$ is reduced to a linear combination of the random variables $Z(t_{n-j})$ for $j \geq 0$ by recursive substitution.

$$X(t_n) = \sum_{j=0}^{\infty} \alpha^j Z(t_{n-j})$$

Since the mean of the process $Z(t_n)$ is zero, the mean of the process $X(t_n)$ is also zero. The covariance function is calculated as a second central moment for $\tau \geq 0$:

$$\gamma_X(\tau) = D\,(X(t_n)\,X(t_{n+\tau})) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \alpha^{j+k} D\,(Z(t_{n-j})\,Z(t_{n+\tau-k}))$$

$$\gamma_X(\tau) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} \alpha^{j+k} \gamma_Z(\tau - k + j)$$

The covariance function $\gamma_X(\tau)$ of the process $Z(t_n)$ is equal to the variance σ_Z^2 for $\tau = 0$ and is zero for $\tau \neq 0$. The summation formula for a geometric series yields :

$$\gamma_X(\tau) = \sigma_Z^2 \sum_{j=0}^{\infty} \alpha^{2j+\tau} = \sigma_Z^2 \frac{\alpha^\tau}{1 - \alpha^2} \qquad \text{for } |\alpha| < 1 \text{ and } \tau \geq 0$$

The variance σ_X^2 is the value of the covariance function for $\tau = 0$:

$$\sigma_X^2 = \gamma_X(0) = \sigma_Z^2 / (1 - \alpha^2)$$

The correlation function is obtained from the covariance function :

$$\varrho_X(\tau) = \gamma_X(\tau) / \gamma_X(0) \qquad \varrho_X(\tau) = \varrho_X(-\tau)$$

$$\varrho_X(\tau) = \alpha^{|\tau|} \qquad |\alpha| < 1$$

For $|\alpha| < 1$, the regression process of first order is a stationary process with mean 0 and variance σ_X^2. The absolute value of its correlation function decays geometrically from 1 to 0 for $|\tau| \geq 0$. The correlation function alternates for $\alpha < 0$.

The correlation function may also be determined directly using the recursive rule for the random variable $X(t_n)$. The rule is multiplied by $X(t_{n-\tau})$ for $\tau > 0$. The second central moment is formed under the assumption that the process $X(t_n)$ is stationary.

$$D\,(X(t_{n-\tau})\,X(t_n)) \;=\; \alpha\,D\,(X(t_{n-\tau})\,X(t_{n-1})) \;+\; D\,(X(t_{n-\tau})\,Z(t_n))$$

The moment $D(X(t_{n-\tau})\ X(t_n))$ is the covariance function $\gamma_X(\tau)$. The moment $D\,(X(t_{n-\tau})\ X(t_{n-1}))$ is the covariance function $\gamma_X(\tau-1)$. The moment $D(X(t_{n-\tau})\ Z(t_n))$ is zero, since $X(t_{n-\tau})$ is independent of $Z(t_n)$. Dividing by $\gamma_X(0)$ yields the following recursion formula for the correlation function:

$$\varrho_X(\tau) \;=\; \alpha\,\varrho_X(\tau-1) \qquad\qquad \text{for} \quad \tau > 0$$

Since $\varrho_X(0) = 1$, the solution of this recursive equation is the geometric sequence α^τ for $\tau \geq 0$:

$$\varrho_X(\tau) \;=\; \alpha^\tau \qquad\qquad \text{for} \quad \tau \geq 0$$

Regression process of second order : In a regression process of second order, the random variable $X(t_n)$ depends linearly on the random variables $X(t_{n-1})$ and $X(t_{n-2})$:

$$X(t_n) \;=\; \alpha\,X(t_{n-1}) \;+\; \beta\,X(t_{n-2}) \;+\; Z(t_n)$$

The additive term $Z(t_n)$ is a stationary uncorrelated process with mean 0. As for the regression process of first order, the mean μ_X of the process $X(t_n)$ is zero. To determine the correlation function, the recursion formula for $X(t_n)$ is multiplied by $X(t_{n-\tau})$ for $\tau \geq 0$. The second central moment is formed under the assumption that the process is stationary.

$$\begin{aligned} D\,(X(t_{n-\tau})\,X(t_n)) \;&=\; \alpha\,D\,(X(t_{n-\tau})\,X(t_{n-1})) \;+\; \beta\,D\,(X(t_{n-\tau})\,X(t_{n-2})) \\ &+\; D\,(X(t_{n-\tau})\,Z(t_n)) \end{aligned}$$

The moment $D(X(t_{n-\tau})\ X(t_n))$ is the covariance function $\gamma_X(\tau)$. The moment $D\,(X(t_{n-\tau})\ X(t_{n-j}))$ is the covariance function $\gamma_X(\tau-j)$. The moment $D\,(X(t_{n-\tau})\ Z(t_n))$ is zero, since $X(t_{n-\tau})$ is independent of $Z(t_n)$. The equation for the central moments is divided by $\gamma_X(0)$ to obtain the following recursion formula for the correlation function for $\tau > 1$:

$$\varrho_X(\tau) \;=\; \alpha\,\varrho_X(\tau-1) \;+\; \beta\,\varrho_X(\tau-2) \qquad\qquad \tau > 1 \qquad\qquad (1)$$

For $\tau = 1$, the symmetry condition $\varrho_X(\tau) = \varrho_X(-\tau)$ implies :

$$\varrho_X(1) = \alpha\,\varrho_X(0) + \beta\,\varrho_X(1) \tag{2}$$

For $\tau = 0$, the correlation function is 1 :

$$\varrho_X(0) = 1 \tag{3}$$

The equations for the correlation function have the form of an initial value problem. Equation (1) is a linear difference equation of second order. Equations (2) and (3) specify the initial values.

$$\varrho_X(\tau) = \alpha\,\varrho_X(\tau-1) + \beta\,\varrho_X(\tau-2) \qquad \varrho_X(0) = 1 \qquad \varrho_X(1) = \alpha/(1-\beta)$$

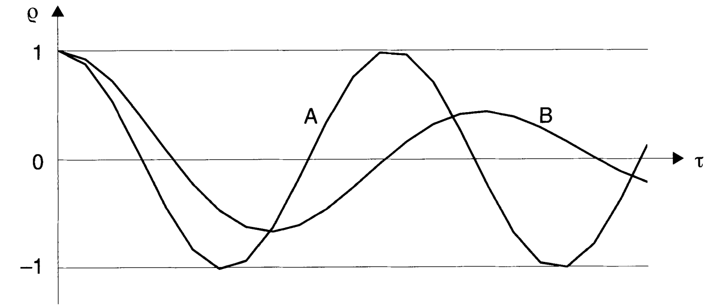
Substituting the trial function $\varrho_X(\tau) = c\,\lambda^\tau$ into the difference equation yields a quadratic equation for λ with the solutions λ_1 and λ_2 :

$$\lambda^2 - \alpha\,\lambda - \beta = 0 \qquad\qquad \lambda_{1,2} = \frac{\alpha}{2} \pm \sqrt{\frac{\alpha^2}{4} + \beta}$$

For $\lambda_1 \neq \lambda_2$, the linear combination $c_1\,\lambda_1^\tau + c_2\,\lambda_2^\tau$ is a solution of the difference equation. The constants c_1 and c_2 are determined such that the initial conditions for $\tau = 0$ and $\tau = 1$ are satisfied. This yields the following solution of the initial value problem for $\tau \geq 0$:

$$\varrho_X(\tau) = \frac{\lambda_2 - a}{\lambda_2 - \lambda_1}\,\lambda_1^\tau + \frac{a - \lambda_1}{\lambda_2 - \lambda_1}\,\lambda_2^\tau \qquad a = \frac{\alpha}{1-\beta} \qquad \lambda_1 \neq \lambda_2$$

The regression process of second order is stationary if the absolute values of λ_1 and λ_2 are less than 1. The cases in which λ_1 and λ_2 are conjugate complex numbers are especially important. In these cases, the correlation function corresponds to a harmonic oscillation. The following diagram shows typical examples; the individual values of the correlation function are joined by line segments.



A : $\alpha = 1.75$ $\beta = -1.0$ B : $\alpha = 1.75$ $\beta = -0.9$

Random oscillations : A regression process of second order describes stationary random oscillations of a one-mass oscillator in discrete time subjected to impulses. The displacements x(t) of an oscillator with the angular eigenfrequency ω and the damping constant δ under the load p(t) are described by the following linear differential equation in continuous time :

$$\ddot{x}(t) + 2\,\delta\,\omega\,\dot{x}(t) + \omega^2\,x(t) = p(t)$$

In discrete time with equidistant time points t_n at a distance $\Delta t = 1$ from each other, the differential equation is approximated by a difference equation. For $t = t_{n-1}$ one obtains :

$$x(t_n) - 2\,x(t_{n-1}) + x(t_{n-2}) + \delta\,\omega\,(x(t_n) - x(t_{n-2})) + \omega^2\,x(t_{n-1}) = p(t_{n-1})$$

The difference equation is solved for $x(t_n)$:

$$x(t_n) = \frac{2 - \omega^2}{1 + \delta\omega}\,x(t_{n-1}) - \frac{1 - \delta\omega}{1 + \delta\omega}\,x(t_{n-2}) + \frac{1}{1 + \delta\omega}\,p(t_{n-1})$$

The displacement $x(t_n)$ is a realization of the random variable $X(t_n)$ at time t_n. The load $p(t_{n-1})$ first takes effect at time t_n, so that the load term $p(t_{n-1})\,/\,(1 + \delta\,\omega)$ is considered as a realization of the random variable $Z(t_n)$. The following regression process of second order is obtained :

$$X(t_n) = \alpha\,X(t_{n-1}) + \beta\,X(t_{n-2}) + Z(t_n)$$

$$\alpha = (2 - \omega^2)\,/\,(1 + \delta\,\omega)$$

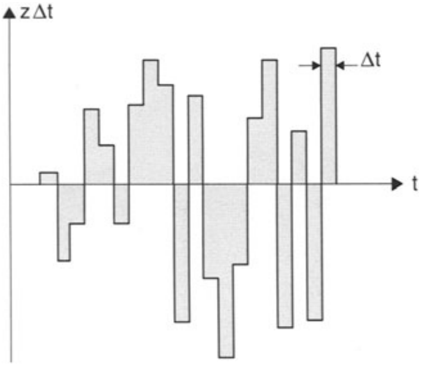
$$\beta = -(1 - \delta\,\omega)\,/\,(1 + \delta\,\omega)$$

For an undamped oscillation, $\delta = 0$. Then $\beta = -1$ and $\alpha = 2 - \omega^2$. The correlation function $\varrho(\tau)$ for $\tau \geq 0$ corresponds to an oscillation of the one-mass oscillator without load with the initial displacement $\varrho(0) = 1$ and the initial velocity $\dot{\varrho}(0) = 0$.

10.5.4.4 Stationary processes in continuous time

Introduction : The stationary random processes in discrete time may be transferred to continuous time. The process X(t) is considered at equidistant times t at a distance $\Delta t \rightarrow 0$ from each other. In analogy with the case of discrete time, the uncorrelated process and the averaging and regression processes derived from it are treated. Harmonic processes, which exhibit a sinusoidal time dependence, require a different approach. In engineering, harmonic processes are particularly important for describing random oscillations.

Uncorrelated process : Let a sequence of random rectangular impulses $Z(t)\Delta t$ of finite intensity in consecutive infinitesimal time intervals $\Delta t \rightarrow 0$ of the continuous time domain $-\infty < t < \infty$ be given. Let the impulses be stochastically independent, and let them have identical distributions with the mean $\mu_Z = 0$. A realization of such a stationary process can only be achieved approximately by replacing the infinitesimal time intervals by small finite intervals. A typical example for uniformly distributed impulses is shown below :



By virtue of the independence of the impulses at different time, the covariance function $\gamma_Z(\tau)$ is zero for $\tau \neq 0$. It is mathematically formulated using the delta function :

$$\gamma_Z(\tau) = \sigma_Z^2 \, \delta(\tau)$$

$$\delta(\tau) = 0 \quad \text{for} \quad \tau \neq 0 \qquad\qquad \int_{-\infty}^{\infty} \delta(\tau)\, d\tau = 1$$

This stationary process is uncorrelated. In mathematics it is sometimes called a delta-correlated process; in physics it is referred to as "white noise".

Averaging process : In the averaging process, the random variable Y(t, T) is formed by averaging a stationary uncorrelated process with mean 0 over the time interval from t – T to t :

$$Y(t,T) \;=\; \frac{1}{T}\int_0^T Z(t-s)\,ds$$

The mean μ_Y of the averaging process is zero, since the mean μ_Z of the stationary uncorrelated process is zero. The covariance function $\gamma_Y(\tau)$ is calculated as a second moment :

$$\gamma_Y(\tau,T) = \;D\,(Y(t,T)\;Y(t+\tau,T)) \;=\; \frac{1}{T^2}\int_0^T\int_\tau^{T+\tau} D\,(Z(t-s)\;Z\,(t-r))\;ds\,dr$$

$$\gamma_Y(\tau,T) = \;\frac{1}{T^2}\int_0^T\int_\tau^{T+\tau} \gamma_Z\,(r-s)\;ds\,dr$$

The correlation function $\gamma_Z(r-s) = \sigma_Z^2\;\delta(r-s)$ is non-zero only for $r = s$. With the integration rules for delta functions one obtains :

$$\gamma_Y(\tau,T) \;=\; \frac{1}{T^2}\int_\tau^T \sigma_Z^2\,ds \;=\; \sigma_Z^2\;\frac{T-\tau}{T^2} \qquad\qquad 0 \le \tau \le T$$

$$\gamma_Y(\tau,T) \;=\; 0 \qquad\qquad \tau \ge T$$

For $\tau = 0$ this yields the variance σ_Y^2 (T), which is inversely proportional to T :

$$\sigma_Y^2\;(T) \;=\; \sigma_Z^2\,/\,T$$

The correlation function $\varrho_Y(\tau,T)$ is determined from the covariance function :

$$\varrho_Y(\tau,T) \;=\; \gamma_X(\tau,T)\,/\,\gamma_Y(0,T) \qquad\qquad \varrho_Y(\tau,\,T) \;=\; \varrho_Y(-\tau,\,T)$$

$$\varrho_Y(\tau,T) \;=\; 1 \,-\, |\tau|\,/\,T \qquad\qquad |\tau| \;\le\; T$$

$$\varrho_Y(\tau,T) \;=\; 0 \qquad\qquad |\tau| \;\ge\; T$$

Like the averaging process in discrete time, the averaging process in continuous time is a stationary process with a linearly decaying correlation function.

Regression process of first order : In the regression process of first order, the random variable X(t) depends linearly on the random variable X(t – Δt) for Δt→0 :

$$X(t) = (1 - \alpha\,\Delta t)\; X(t - \Delta t) \;+\; \Delta t\; Y(t, \Delta t)$$

The quantity α is the regression parameter. The random variable Y(t, Δt) is the average of a stationary uncorrelated process in the interval Δt. It has the mean $\mu_Y = 0$ and the variance $\sigma_Y^2\,(\Delta t) = \sigma_Z^2\,/\,\Delta t$. In the following, the moments of the process X(t) are determined in the limit Δt→0 under the assumption that the process is stationary. The mean μ_X is zero, since with $\alpha \neq 0$ the first moments are given by

$$E(X(t)) = (1 - \alpha\,\Delta t)\; E(X(t - \Delta t)) \;+\; \Delta t\; E(Y(t, \Delta t))$$

$$\mu_X = (1 - \alpha\,\Delta t)\;\mu_X \;+\; \Delta t\;\mu_Y$$

$$\mu_X = \mu_Y\,/\,\alpha = 0$$

The random variable X(t) is a linear combination of the random variables X(t – Δt) and Y(t – Δt), which are independent. The rules for calculating the variances of linear combinations of independent random variables yield

$$\sigma_X^2 = (1 - \alpha\,\Delta t)^2\,\sigma_X^2 \quad + \quad \Delta t^2\,\sigma_Y^2\,(\Delta t)$$

$$\sigma_X^2 = \frac{\Delta t^2\,\sigma_Y^2\,(\Delta t)}{1 - (1 - \alpha\;\Delta t)^2} = \frac{\sigma_Z^2}{2\alpha - \alpha^2\;\Delta t}$$

$$\sigma_X^2 = \sigma_Z^2\,/\,2\alpha \qquad \text{for} \quad \Delta t \to 0$$

The covariance function $\gamma_X(\tau)$ for $\tau > 0$ is determined as a second central moment :

$$D(X(t - \tau)\;X(t)) = (1 - \alpha\,\Delta t)\;D(X(t - \tau)\;\;X(t - \Delta t)) \;+\; \Delta t\;D(X(t - \tau)\;Y(t, \Delta t))$$

Since the random variables X(t – τ) and Y(t, Δt) are independent, their covariance is zero. This implies :

$$\gamma_X(\tau) = (1 - \alpha\,\Delta t)\;\gamma_X(\tau - \Delta t)$$

$$\frac{\gamma_X(\tau) - \gamma_X(\tau - \Delta t)}{\Delta t} + \alpha\,\gamma_X(\tau - \Delta t) = 0$$

In the limit Δt→0 one obtains a first order linear differential equation for the covariance function with the following solution :

$$\dot{\gamma}_X(\tau) \;+\; \alpha\,\gamma_X(\tau) = 0$$

$$\gamma_X(\tau) = \gamma_X(0)\;e^{-\alpha\tau} = \sigma_X^2\;e^{-\alpha\tau} \qquad\qquad \tau \geq 0$$

The correlation function is determined from the covariance function :

$$\varrho_X(\tau) = \gamma_X(\tau)\,/\,\gamma_X(0) \qquad\qquad \varrho_X(\tau) = \varrho_X(-\tau)$$

$$\varrho_X(\tau) = e^{-\alpha\,|\tau|}$$

The regression process of first order is stationary for $\alpha > 0$. Its correlation function decays exponentially.

Simple harmonic process : In a simple harmonic random process, the random function X(t) is a harmonic oscillation with a given frequency ω, a random amplitude A and a random phase shift Φ.

$$X(t) = A\cos(\omega t + \Phi)$$

The amplitude A is described by a density function $f_A(a)$ with the mean μ_A and the variance σ_A^2. The phase shift Φ is described by a uniform distribution $f_\Phi(\phi) = 1/(2\pi)$ in the range $-\pi \le \phi \le \pi$. The amplitude A and the phase shift Φ are independent, so that the common density function is the product $f_A(a)\ f_\Phi(\phi)$. The moments of the random variable X(t) are obtained from the density functions using the rules for deterministic dependence. The mean μ_X is zero.

$$\mu_X = E(X(t)) = \int_{-\pi}^{\pi}\int_{-\infty}^{\infty} a\cos(\omega t + \phi)\ f_A(a)\ f_\Phi(\phi)\ da\ d\phi$$

$$\mu_X = \int_{-\infty}^{\infty} a\ f_A(a)\ da \int_{-\pi}^{\pi} \frac{1}{2\pi}\cos(\omega t + \phi)\ d\phi = \mu_A \cdot 0 = 0$$

The covariance function $\gamma_X(\tau)$ is obtained as a central moment :

$$\gamma_X(\tau) = D(X(t)\ X(t+\tau))$$

$$\gamma_X(\tau) = \int_{-\pi}^{\pi}\int_{-\infty}^{\infty} a^2\cos(\omega t + \phi)\cos(\omega(t+\tau) + \phi)\ f_A(a)\ f_\Phi(\phi)\ da\ d\phi$$

$$\gamma_X(\tau) = \int_{-\infty}^{\infty} a^2\ f_A(a)\ da \int_{-\pi}^{\pi} \frac{1}{2\pi}\cos(\omega t + \phi)\cos(\omega(t+\tau) + \phi)\ d\phi$$

$$\gamma_X(\tau) = E(A^2)\int_{-\pi}^{\pi} \frac{1}{2\pi}\cos(\omega t + \phi)\cos(\omega(t+\tau) + \phi)\ d\phi$$

With the theorems of trigonometry and the second moment $E(A^2) = \sigma_A^2 + \mu_A^2$ this becomes :

$$\gamma_X(\tau) = \frac{1}{2}(\sigma_A^2 + \mu_A^2)\cos\omega\tau$$

The variance σ_X^2 and the correlation function $\varrho_X(\tau)$ are obtained from the covariance function $\gamma_X(\tau)$:

$$\sigma_X^2 = \gamma_X(0) = (\sigma_A^2 + \mu_A^2)\ /\ 2$$

$$\varrho_X(\tau) = \gamma_X(\tau)\ /\ \gamma_X(0) = \cos\omega\tau$$

The harmonic process is a typical example of a stationary non-ergodic process. A realization x(t) of the process does not allow the probability distribution $f_A(a)$ for the amplitude A to be inferred. The correlation function of the process is periodic.

In contrast to the stationary processes in continuous time treated above, this process has a further special property. The amplitude a and the phase angle ϕ may be determined from two suitable samples $x(t_1)$ and $x(t_2)$ of a realization x(t). If a and ϕ are known the realization can be determined in the entire range $-\infty \leq t \leq \infty$.

General harmonic process : In a general harmonic random process, several harmonic oscillations with different given frequencies ω_j, random amplitudes A_j and random phase shifts Φ_j are superimposed :

$$X(t) \quad = \quad \sum_{j=1}^{n} A_j \cos(\omega_j t + \Phi_j)$$

If the amplitudes A_j and the phase shifts Φ_j are independent and the phase shifts are uniformly distributed, the mean μ_X and the correlation function $\gamma_X(\tau)$ are given by :

$$\mu_X \quad = \quad 0$$

$$\gamma_X(\tau) \quad = \quad \frac{1}{2} \sum_{j=1}^{n} S(\omega_j) \cos \omega_j \tau$$

$$S(\omega_j) \; = \; E(A_j^2) \; = \; \sigma_{Aj}^2 + \mu_{Aj}^2$$

The values $S(\omega_j)$ are called the spectral values of the general harmonic process. The discrete function $S(\omega_j)$ in the frequency space $0 < \omega_j < \infty$ is called a discrete spectral function. The general harmonic process with discrete frequencies forms the basis for a spectral analysis of stationary processes in the frequency domain. Although spectral analysis is important in the context of random oscillations, a detailed account is beyond the scope of the present treatment.

INDEX