Advanced Mathematical Tools for Automatic Control Engineers

Volume 1: Deterministic Techniques

Alexander S. Poznyak

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Elsevier
Linacre House, Jordan Hill, Oxford OX2 8DP, UK
Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2008

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978 0 08 044674 5

For information on all Elsevier publications
visit our web site at books.elsevier.com

Printed and bound in Hungary
08 09 10 10 9 8 7 6 5 4 3 2 1

Preface

Mathematics is playing an increasingly important role in the physical, biological and engineering sciences, provoking a blurring of boundaries between scientific disciplines and a resurgence of interest in the modern as well as the classical techniques of applied mathematics. Remarkable progress has been made recently in both theory and applications of all important areas of control theory.

Modern automatic control theory covers such topics as the algebraic theory of linear systems (including controllability, observability, feedback equivalence, minimality of realization, frequency domain analysis and synthesis etc.), Lyapunov stability theory, input–output method, optimal control (maximum principle and dynamic programing), observers and dynamic feedback, robust control (in Hardy and Lebesgue spaces), delay-systems control, the control of infinite-dimensional systems (governed by models in partial differential equations), conflict and game situations, stochastic processes and effects, and many others. Some elegant applications of control theory are presently being implemented in aerospace, biomedical and industrial engineering, robotics, economics, power systems etc.

The efficient implementation of the basic principles of control theory to different applications of special interest requires an interdisciplinary knowledge of advanced mathematical tools and such a combined expertise is hard to find. What is needed, therefore, is a textbook making these tools accessible to a wide variety of engineers, researchers and students.

Many suitable texts exist (practically there are no textbooks) that tackle some of the areas of investigation mentioned above. Each of these books includes one or several appendices containing the minimal mathematical background that is required of a reader to actively work with this material. Usually it is a working knowledge of some mathematical tools such as the elements of linear algebra, linear differential equations, fourier analysis and, perhaps, some results from optimization theory, as well. In fact, there are no textbooks containing all (or almost all) of the mathematical knowledge required for successful studying and research within the control engineering community. It is important to emphasize that the mathematical tools for automatic control engineers are specific and significantly differ from those needed by people involved in fluid mechanics, electrical engineering etc. To our knowledge, no similar books or publications exist. The material in these books partially overlaps with several other books. However, the material covered in each book cannot be found in a single book (dealing with deterministic or stochastic systems). Nevertheless, some books may be considered as partially competitive. For example:

- Guillemin (1949) *The Mathematics of Circuit Analysis: Extensions to the Mathematical Training of Electrical Engineers*, John Wiley and Sons, Inc., New York. In fact this

book is nicely written, but it is old, does not have anything on stochastic and is oriented only to electrical engineers.

- *Modern Mathematics for the Engineer* (1956), edited by E.F. Beckenbah, McGraw-Hill, New York. This is, in fact, a multi-author book containing chapters written by the best mathematicians of the second half of the last century such as N. Wiener, R. Bellman and others. There is no specific orientation to the automatic control community.
- *Systèmes Automatisés* (in French) (2001), Hermes-Science, Paris, five volumes. These are multi-authored books where each chapter is written by a specific author or authors.
- Hinrichsen & Pritchard (2005) *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*, Texts in Applied Mathematics, Springer. This excellent book is written by mathematicians for mathematicians working with mathematical aspects of control theory.

The wide community of automatic control engineers requires a book like this. The primary reason is that there exist no similar books, and, secondly, the mathematical tools are spread over many mathematical books written by mathematicians and the majority of them are unsuitable for the automatic control engineering community.

This book was conceived as a hybrid monograph/textbook. I have attempted to make the development didactic. Most of the material comes from reasonably current periodic literature and a fair amount of the material (especially in Volume 2) is my own work.

This book is practically self-contained since almost all lemmas and theorems within contain their detailed proofs. Here, it makes sense to remember the phrase of David Hilbert: "It is an error to believe that rigor in the proof is the enemy of simplicity. The very effort for rigor forces us to discover simpler methods of proof. . . ."

Intended audience

My teaching experience and developing research activities convinced me of the need for this sort of textbook. It should be useful for the average student yet also provide a depth and rigor challenging to the exceptional student and acceptable to the advanced scholar. It should comprise a basic course that is adequate for all students of automatic control engineering regardless of their ultimate speciality or research area. It is hoped that this book will provide enough incentive and motivation to new researchers, both from the "control community" and applied and computational mathematics, to work in the area. Generally speaking, this book is intended for students (undergraduate, postdoctoral, research) and practicing engineers as well as designers in different industries. It was written with two primary objectives in mind:

- to provide a list of references for researchers and engineers, helping them to find information required for their current scientific work, and
- to serve as a text in an advanced undergraduate or graduate level course in mathematics for automatic control engineering and related areas.

The particular courses for which this book might be used as a text, supplementary text or reference book are as follows:

Volume 1

- Introduction to automatic control theory,
- Linear and nonlinear control systems,
- Optimization,
- Control of robotic systems,
- Robust and adaptive control,
- Optimal control,
- Discrete-time and impulse systems,
- Sliding mode control,
- Theory of stability.

Volume 2

- Probability and stochastic processes in control theory,
- Signal and systems,
- Identification and parameters estimation,
- Adaptive stochastic control,
- Markov processes,
- Game theory,
- Machine learning,
- Intelligent systems,
- Design of manufacturing systems and operational research,
- Reliability,
- Signal processing (diagnosis, pattern recognition etc.).

This book can also be used in several departments: electrical engineering and electronics, systems engineering, electrical and computer engineering, computer science, information science and intelligent systems, electronics and communications engineering, control engineering, systems science and industrial engineering, cybernetics, aerospace engineering, econometrics, mathematical economics, finance, quality control, applied and computational mathematics, and applied statistics and operational research, quality management, chemical engineering, mechanical engineering etc.
The book is also ideal for self-study.

It seems to be more or less evident that any book on advanced mathematical methods is predetermined to be incomplete. It will also be evident that I have selected for inclusion in the book a set of methods based on my own preferences, reflected by my own experience, from among a wide spectrum of modern mathematical approaches. Nevertheless, my intention is to provide a solid package of materials, making the book valuable for postgraduate students in automatic control, mechanical and electrical engineering, as well as for all engineers dealing with advance mathematical tools in their daily practice. As I intended to write a textbook and not a handbook, the bibliography is by no means complete. It comprises only those publications which I actually used.

Acknowledgments

I have received invaluable assistance in many conversations with my friend and colleague Prof. Vladimir L. Kharitonov (CINVESTAV-IPN, Mexico). Special thanks are due to him for his direct contribution to the presentation of Chapter 20 concerning stability theory and for his thought-provoking criticism of many parts of the original manuscript. I am indebted to my friend Prof. Vadim I. Utkin (Ohio State University, USA) for being so efficient during my work with Chapters 9, 19 and 22 including the topics of theory of differential equations with discontinuous right-hand side and of sliding mode control. I also wish to thank my friend Prof. Kaddour Najim (Polytechnic Institute of Toulouse, France) for his initial impulse and suggestion to realize this project. Of course, I feel deep gratitude to my students over the years, without whose evident enjoyment and expressed appreciation the current work would not have been undertaken. Finally, I wish to acknowledge the editors at Elsevier for being so cooperative during the production process.

Alexander S. Poznyak
Mexico, 2007

Contents

PART I

Matrices and Related Topics

1 Determinants

Contents

The material presented in this chapter as well as in the next chapters is based on the following classical books dealing with matrix theory and linear algebra: Lancaster (1969), Lankaster & Tismenetsky (1985), Marcus & Minc (1992), Bellman (1960) and Gantmacher (1990). The numerical methods of linear algebra can be found in Datta (2004).

1.1 Basic definitions

1.1.1 Rectangular matrix

Definition 1.1. *An ordered array of elements a_{ij} $(i = 1, \ldots, m;\ j = 1, \ldots, n)$ taken from arbitrary field $\mathfrak{F}$ (here $\mathfrak{F}$ will always be the set of all real or all complex numbers, denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively) written in the form of the table*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} = [a_{ij}]_{i,j=1}^{m,n} \tag{1.1}$$

is said to be a ***rectangular*** *$m \times n$* ***matrix*** *where a_{ij} denotes the elements of this table lying on the intersection of the ith row and jth column.*

The set of all $m \times n$ matrices with real elements will be denoted by $\mathbb{R}^{m\times n}$ and with complex elements by $\mathbb{C}^{m\times n}$.

1.1.2 Permutations, number of inversions and diagonals

Definition 1.2. *If $j_1, j_2, \ldots, j_n$ are the numbers $1, 2, \ldots, n$ written in any order then $(j_1, j_2, \ldots, j_n)$ is said to be a* ***permutation*** *of $1, 2, \ldots, n$. A certain* ***number of inversions*** *associated with a given permutation $(j_1, j_2, \ldots, j_n)$ denoted briefly by $t(j_1, j_2, \ldots, j_n)$.*

Clearly, there exists exactly $n! = 1 \cdot 2 \cdots n$ permutations.

Example 1.1. $(1, 3, 2)$, $(3, 1, 2)$, $(3, 2, 1)$, $(1, 2, 3)$, $(2, 1, 3)$, $(2, 3, 1)$ *are the permutations of* $1, 2, 3$.

Example 1.2. $t\,(2, 4, 3, 1, 5) = 4$.

Definition 1.3. *A* ***diagonal*** *of an arbitrary square matrix* $A \in \mathbb{R}^{n\times n}$ *is a sequence of elements of this matrix containing one and only one element from each row and one and only one element from each column. Any diagonal of* A *is always assumed to be ordered according to the row indices; therefore it can be written in the form*

$$a_{1j_1}, a_{2j_2}, \ldots, a_{nj_n}$$

Any matrix $A \in \mathbb{R}^{n\times n}$ *has* $n!$ *different diagonals.*

Example 1.3. *If* $(j_1, j_2, \ldots, j_n) = (1, 2, \ldots, n)$ *we obtain the* ***main diagonal***

$$a_{11}, a_{22}, \ldots, a_{nn}$$

If $(j_1, j_2, \ldots, j_n) = (n, n-1, \ldots, 1)$ *we obtain the* ***secondary diagonal***

$$a_{1n}, a_{2(n-1)}, \ldots, a_{n1}$$

1.1.3 Determinants

Definition 1.4. *The* ***determinant*** $\det A$ *of a square matrix* $A \in \mathbb{R}^{n\times n}$ *is defined by*

$$\begin{aligned}\det A &:= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)}\, a_{1j_1} a_{2j_2} \cdots a_{nj_n} \\ &= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} \prod_{k=1}^{n} a_{kj_k}\end{aligned} \tag{1.2}$$

In other words, $\det A$ is a sum of $n!$ products involving n elements of A belonging to the same diagonal. This product is multiplied by $(+1)$ or (-1) according to whether $t\,(j_1, j_2, \ldots, j_n)$ is even or odd, respectively.

Example 1.4. *If* $A \in \mathbb{R}^{2\times 2}$ *then*

$$\det \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = a_{11}a_{22} - a_{12}a_{21}$$

Example 1.5. (Sarrius's rule) *If* $A \in \mathbb{R}^{3\times 3}$ *(see Fig. 1.1) then*

$$\det \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$
$$= a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{21}a_{13}a_{32}$$
$$- a_{31}a_{22}a_{13} - a_{32}a_{11}a_{23} - a_{21}a_{12}a_{33}$$

Example 1.6.

$$\det \begin{bmatrix} 0 & a_{12} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{24} & 0 \\ 0 & 0 & a_{33} & a_{34} & 0 \\ a_{41} & a_{42} & 0 & 0 & 0 \\ 0 & 0 & 0 & a_{54} & a_{55} \end{bmatrix} = (-1)^{t(2,4,3,1,5)} a_{12}a_{24}a_{33}a_{41}a_{55}$$
$$= (-1)^4 a_{12}a_{24}a_{33}a_{41}a_{55} = a_{12}a_{24}a_{33}a_{41}a_{55}$$

Example 1.7. *The determinant of a* ***low triangular matrix*** *is equal to the product of its diagonal elements, that is,*

$$\det \begin{bmatrix} a_{11} & \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{0} \\ a_{21} & a_{22} & \mathbf{0} & \cdot & \cdot & \mathbf{0} \\ \cdot & \cdot & \cdot & \mathbf{0} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \mathbf{0} & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \mathbf{0} \\ 0 & \cdot & \cdot & 0 & a_{n,n-1} & a_{nn} \end{bmatrix} = a_{11}a_{22}\cdots a_{nn} = \prod_{i=1}^{n} a_{ii}$$

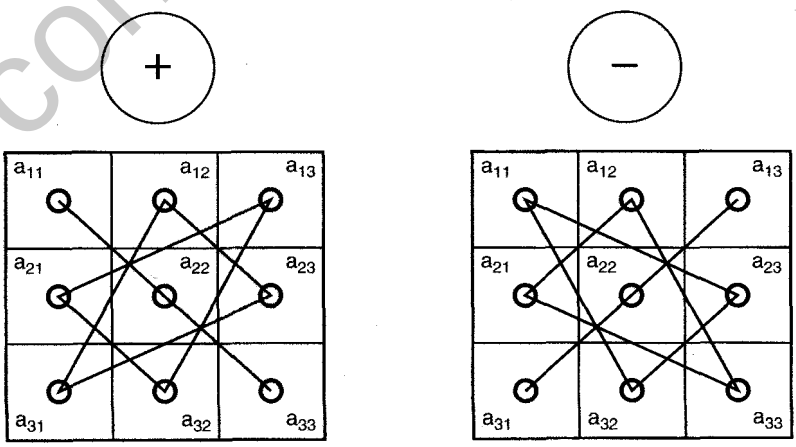

Fig. 1.1. Illustration of the Sarrius's rule.

Example 1.8. *The determinant of an* ***upper triangular matrix*** *is equal to the product of its diagonal elements, that is,*

$$\det \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1n} \\ \mathbf{0} & a_{22} & a_{23} & \cdot & \cdot & a_{2n} \\ \mathbf{0} & \mathbf{0} & a_{33} & a_{34} & \cdot & \cdot \\ \cdot & \cdot & \mathbf{0} & \cdot & \cdot & \cdot \\ \mathbf{0} & \cdot & \cdot & \mathbf{0} & \cdot & a_{n-1,n} \\ \mathbf{0} & \cdot & \cdot & \mathbf{0} & \mathbf{0} & a_{nn} \end{bmatrix} = a_{11}a_{22}\cdots a_{nn} = \prod_{i=1}^{n} a_{ii}$$

Example 1.9. *For the square matrix* $A \in \mathbb{R}^{n\times n}$ *having only zero elements above (or below) the secondary diagonal*

$$\det A = (-1)^{n(n-1)/2} a_{1n}a_{2,n-1}\cdots a_{n1}$$

Example 1.10. *The determinant of any matrix* $A \in \mathbb{R}^{n\times n}$ *containing a zero row (or column) is equal to zero.*

1.2 Properties of numerical determinants, minors and cofactors

1.2.1 Basic properties of determinants

Proposition 1.1. *If* $\tilde{A}$ *denotes a matrix obtained from a square matrix A by multiplying one of its rows (or columns) by a scalar k, then*

$$\det \tilde{A} = k \det A \tag{1.3}$$

Corollary 1.1. *The determinant of a square matrix is a* ***homogeneous*** *over field* $\mathfrak{F}$*, that is,*

$$\boxed{\det (kA) = \det [ka_{ij}]_{i,j=1}^{n,n} = k^n \det A}$$

Proposition 1.2.

$$\boxed{\det A = \det A^{\mathsf{T}}}$$

where A^{T} *is the* ***transpose*** *of the matrix A obtained by interchanging the rows and columns of A, that is,*

$$\boxed{A^{\mathsf{T}} = \begin{bmatrix} a_{11} & a_{21} & \cdot & \cdot & \cdot & a_{n1} \\ a_{12} & a_{22} & \cdot & \cdot & \cdot & a_{n2} \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & & & & \cdot \\ a_{1n} & a_{2n} & \cdot & \cdot & \cdot & a_{nn} \end{bmatrix} = [a_{ji}]_{i,j=1}^{n,n}}$$

Proof. It is not difficult to see that a diagonal $a_{1j_1}, a_{2j_2}, \ldots, a_{nj_n}$, ordered according to the row indices and corresponding to the permutation $(j_1, j_2, \ldots, j_n)$, is also a diagonal of $A^\top$ since the elements of A and $A^\top$ are the same. Consider now pairs of indices

$$(1, j_1), (2, j_2), \ldots, (n, j_n) \tag{1.4}$$

corresponding to a term of $\det A$ and pairs

$$(k_1, 1), (k_2, 2), \ldots, (j_n, n) \tag{1.5}$$

obtained from the previous pairs collection by a reordering according to the second term and corresponding to the term of $\det A^\top$ with the same elements. Observe that each interchange of pairs in (1.4) yields a simultaneous interchange of numbers in the permutations $(1, 2, \ldots, n)$, $(j_1, j_2, \ldots, j_n)$ and $(k_1, k_2, \ldots, k_n)$. Hence,

$$t(j_1, j_2, \ldots, j_n) = t(k_1, k_2, \ldots, k_n)$$

This completes the proof. □

Proposition 1.3. *If the matrix $B \in \mathbb{R}^{n\times n}$ is obtained by interchanging two rows (or columns) of $A \in \mathbb{R}^{n\times n}$ then*

$$\det A = -\det B \tag{1.6}$$

Proof. Observe that the terms of $\det A$ and $\det B$ consist of the same factors taking one and only one from each row and each column. It is sufficient to show that the signs of each elements are changed. Indeed, let the rows be in general position with rows r and s (for example, $r < s$). Then with $(s - r)$-interchanges of neighboring rows, the rows

$$r, r+1, \ldots, s-1, s$$

are brought into positions

$$r+1, r+2, \ldots, s, r$$

A further $(s - r - 1)$-interchanges of neighboring rows produces the required order

$$s, r+1, r+2, \ldots, s-1, r$$

Thus, a total $2(s - r) - 1$ interchanges is always odd that completes the proof. □

Corollary 1.2. *If the matrix $A \in \mathbb{R}^{n\times n}$ has two rows (or columns) alike, then*

$$\det A = 0$$

Proof. It follows directly from the previous proposition that since making the interchanging of these two rows (or columns) we have

$$\det A = -\det A$$

which implies the result. □

Corollary 1.3. *If a row (or column) is a multiple of another row (or column) of the same matrix A then*

$$\det A = 0$$

Proof. It follows from the previous propositions that

$$\det \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ \cdot & \cdot & \cdots & \cdot \\ a_{1j} & a_{2j} & \cdots & a_{nj} \\ ka_{1j} & ka_{2j} & \cdots & ka_{nj} \\ \cdot & \cdot & & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix}$$

$$= k \det \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ \cdot & \cdot & \cdots & \cdot \\ a_{1j} & a_{2j} & \cdots & a_{nj} \\ a_{1j} & a_{2j} & \cdots & a_{nj} \\ \cdot & \cdot & & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{nn} \end{bmatrix} = 0$$

The result is proven. □

Proposition 1.4. *Let B be the matrix obtained from A by adding the elements of the ith row (or column) to the corresponding elements of its kth ($k \neq i$) row (or column) multiplied by a scalar α. Then*

$$\det B = \det A \tag{1.7}$$

Proof. Taking into account that in the determinant representation (1.2) each term contains only one element from each row and only one from each column of the given matrix, we have

$$\begin{aligned} \det A &= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} b_{1j_1} b_{2j_2} \cdot b_{ij_i} \cdot b_{nj_n} \\ &= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} a_{1j_1} a_{2j_2} \cdot \left(a_{ij_i} + \alpha a_{kj_k}\right) \cdot a_{kj_k} \cdot a_{nj_n} \\ &= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} a_{1j_1} a_{2j_2} \cdot a_{ij_i} \cdot a_{kj_k} \cdot a_{nj_n} \\ &\quad + \alpha \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} a_{1j_1} a_{2j_2} \cdot a_{kj_k} \cdot a_{kj_k} \cdot a_{nj_n} \end{aligned}$$

The second determinant is equal to zero since it has two rows alike. This proves the result. □

Corollary 1.4. (Gauss's method of determinants evaluation) *When the operation described above is applied several times,* ***the evaluation of a determinant can be reduced to that of a triangular matrix.***

Example 1.11.

$$\det\begin{bmatrix} 1 & 2 & 1 & -1 \\ 1 & -2 & 1 & 0 \\ -1 & 1 & -2 & 1 \\ 0 & -1 & 1 & 2 \end{bmatrix} = \det\begin{bmatrix} 1 & 2 & 1 & -1 \\ 0 & -4 & 0 & 1 \\ -1 & 1 & -2 & 1 \\ 0 & -1 & 1 & 2 \end{bmatrix}$$

$$= \det\begin{bmatrix} 1 & 2 & 1 & -1 \\ 0 & -4 & 0 & 1 \\ 0 & 3 & -1 & 0 \\ 0 & -1 & 1 & 2 \end{bmatrix} = \det\begin{bmatrix} 1 & 3 & 1 & -1 \\ 0 & -4 & 0 & 1 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

$$= \det\begin{bmatrix} 1 & 3 & 1 & -1 \\ 0 & -4 & 0 & 1 \\ 0 & 0 & -1 & 0.5 \\ 0 & 0 & 1 & 2 \end{bmatrix} = \det\begin{bmatrix} 1 & 3 & 1 & -1 \\ 0 & -4 & 0 & 2 \\ 0 & 0 & -1 & 0.5 \\ 0 & 0 & 0 & 2.5 \end{bmatrix} = 10$$

Corollary 1.5. *If* $\bar{A}$ *denotes the complex conjugate of* $A \in \mathbb{C}^{n\times n}$, *then*

$$\boxed{\det \bar{A} = \overline{\det A}}$$

Proof. Transforming $\det \bar{A}$ to the determinant of a triangular matrix $[\text{triang } \bar{A}]$ and applying the rule

$$\overline{ab} = \bar{a}\bar{b}$$

valid within the field $\mathbb{C}$ of complex values, we get

$$\det \overline{A} = \det\left(\text{triang } \overline{A}\right) = \prod_{i=1}^{n} \left(\text{triang } \overline{A}\right)_{ii}$$

$$= \overline{\prod_{i=1}^{n} (\text{triang } A)_{ii}} = \overline{\det\left(\text{triang } A\right)} = \overline{\det A}$$

The result is proven. □

Corollary 1.6.

$$\det(A^*) = \overline{\det A} \tag{1.8}$$

Proposition 1.5. *Let us consider the, so-called,* $n \times n$ ***companion matrix***

$$\boxed{C_a := \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1 \\ -a_0 & -a_1 & \cdot & \cdot & \cdot & -a_{n-1} \end{bmatrix}}$$

associated with the vector $a = (a_0, \ldots, a_{n-1})^\intercal$. *Then*

$$\boxed{\det C_a = (-1)^n a_0}$$

Proof. Multiplying the ith row $(i = 1, \ldots, n-1)$ by a_i, adding it to the last one and moving the first column to the last right-hand side position, we obtain

$$\det \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1 \\ -a_0 & -a_1 & \cdot & \cdot & \cdot & -a_{n-1} \end{bmatrix} = \det \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1 \\ -a_0 & 0 & \cdot & \cdot & \cdot & 0 \end{bmatrix}$$

$$= (-1)^{n-1} \det \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & 1 & 0 & 0 & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & 1 & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & -a_0 \end{bmatrix} = (-1)^n a_0$$

The proposition is proven. □

1.2.2 Minors and cofactors

Definition 1.5. *A* ***minor*** M_{ij} *of a matrix* $A \in \mathbb{R}^{n \times n}$ *is the determinant of a submatrix of* A *obtained by striking out the* i*th row and* j*th column.*

Example 1.12.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad M_{23} = \det \begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix} = -6$$

Definition 1.6. *The* ***cofactor*** A_{ij} *(or* ij***-algebraic complement) of the element*** a_{ij} *of the matrix* $A \in \mathbb{R}^{n\times n}$ *is defined as*

$$\boxed{A_{ij} := (-1)^{i+j} M_{ij}} \tag{1.9}$$

Example 1.13.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}, \quad A_{23} = (-1)^{2+3} \det \begin{bmatrix} 1 & 2 \\ 7 & 8 \end{bmatrix} = 6$$

Lemma 1.1. (Cofactor expansion) *For any matrix* $A \in \mathbb{R}^{n\times n}$ *and any indices* $i, j = 1, \ldots, n$

$$\boxed{\det A = \sum_{j=1}^{n} a_{ij} A_{ij} = \sum_{i=1}^{n} a_{ij} A_{ij}} \tag{1.10}$$

Proof. Observing that each term in (1.2) contains an element of the ith row (or, analogously, of the jth column) and collecting together all terms containing a_{ij} we obtain

$$\begin{aligned} \det A &:= \sum_{j_1, j_2, \ldots, j_n} (-1)^{t(j_1, j_2, \ldots, j_n)} a_{1 j_1} a_{2 j_2} \cdots a_{n j_n} \\ &= \sum_{j_i=1}^{n} a_{i j_i} \left[\sum_{j_k, k\neq i} (-1)^{t(j_1, j_2, \ldots, j_n)} \prod_{k\neq i} a_{k j_k} \right] \end{aligned}$$

To fulfill the proof it is sufficient to show that

$$\tilde{A}_{ij_i} := \sum_{j_k, k\neq i} (-1)^{t(j_1, j_2, \ldots, j_n)} \prod_{k\neq i} a_{k j_k} = A_{ij_i}$$

In view of the relation

$$(-1)^{t(j_1, j_2, \ldots, j_n)} = (-1)^{i+j_i} (-1)^{t(j_1, j_2, \ldots, j_{i-1}, j_{j+1}, \ldots, j_n)}$$

it follows that

$$\begin{aligned} \tilde{A}_{ij_i} &= (-1)^{i+j_i} \left[\sum_{j_k, k\neq i} (-1)^{t(j_1, j_2, \ldots, j_{i-1}, j_{j+1}, \ldots, j_n)} \prod_{k\neq i} a_{k j_k} \right] \\ &= (-1)^{i+j_i} M_{ij_i} = A_{ij_i} \end{aligned}$$

which completes the proof. □

Example 1.14.

$$\det\begin{bmatrix}1&2&3\\4&5&6\\7&8&9\end{bmatrix} = 1\,(-1)^{1+1}\det\begin{bmatrix}5&6\\8&9\end{bmatrix} + 4\,(-1)^{2+1}\det\begin{bmatrix}2&3\\8&9\end{bmatrix} + 7\,(-1)^{3+1}\begin{bmatrix}2&3\\5&6\end{bmatrix}$$
$$= -3 - 4\,(-6) + 7\,(-3) = 0$$

Lemma 1.2. *For any matrix $A \in \mathbb{R}^{n\times n}$ and any indices $i \neq r$, $j \neq s$ $(i, j = 1, \ldots, n)$ it follows that*

$$\boxed{\sum_{j=1}^{n} a_{ij}A_{rj} = \sum_{i=1}^{n} a_{ij}A_{is} = 0}$$

Proof. The result follows directly if we consider the matrix obtained from A by replacing the row i (column j) by the row r (column s) and then use the property of a determinant with two rows (or columns) alike that says that it is equal to zero. □

Lemma 1.3. Vandermonde determinant

$$\boxed{V_{1,n} := \det\begin{bmatrix}1&1&\cdots&1\\x_1&x_2&\cdots&x_n\\x_1^2&x_2^2&\cdots&x_n^2\\\cdot&\cdot&\cdots&\cdot\\x_1^{n-1}&x_2^{n-1}&\cdots&x_n^{n-1}\end{bmatrix} = \prod_{j=1}^{n}\prod_{i>j}^{n}(x_i - x_j)}$$

Proof. Adding the ith row multiplied by $(-x_1)$ to the $(i+1)$ row $(i = n-1, n-2, \ldots, 1)$ and applying the iteration implies

$$V_n = \det\begin{bmatrix}1&1&\cdots&1\\0&x_2 - x_1&\cdots&x_n - x_1\\0&x_2^2 - x_1x_2&\cdots&x_n^2 - x_1x_n\\\cdot&\cdot&\cdots&\cdot\\0&x_2^{n-1} - x_1x_2^{n-2}&\cdots&x_n^{n-1} - x_1x_n^{n-2}\end{bmatrix}$$

$$= \det\begin{bmatrix}x_2 - x_1&\cdots&x_n - x_1\\(x_2 - x_1)\,x_2&\cdots&(x_n - x_1)\,x_n\\\cdot&\cdots&\cdot\\(x_2 - x_1)\,x_2^{n-2}&\cdots&(x_n - x_1)\,x_n^{n-2}\end{bmatrix}$$

$$= (x_2 - x_1) \cdots (x_n - x_1) \det \begin{bmatrix} 1 & \cdots & 1 \\ x_2 & \cdots & x_n \\ \cdot & \cdots & \cdot \\ x_2^{n-2} & \cdots & x_n^{n-2} \end{bmatrix}$$

$$= \prod_{i=2}^{n} (x_i - x_1)\, V_{2,n} = \cdots = \prod_{j=1}^{n} \prod_{i>j}^{n} (x_i - x_j) V_{n,n} = \prod_{j=1}^{n} \prod_{i>j}^{n} (x_i - x_j)$$

since $V_{n,n} = 1$. □

1.2.3 Laplace's theorem

Definition 1.7.

(a) If A is an $m \times n$ matrix, then the determinant of a $p \times p$ $(1 \leq p \leq \min(m, n))$ submatrix of A, obtained by striking out $(m - p)$ rows and $(n - p)$ columns, is called a ***minor of order p of A****. If rows and columns retained are given by subscripts*

$$1 \leq i_1 < i_2 < \cdots < i_p \leq m, \quad 1 \leq j_1 < j_2 < \cdots < j_p \leq n \tag{1.11}$$

respectively, then the corresponding minor is denoted by

$$A \begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} := \det [a_{i_k j_k}]_{k=1}^{p} \tag{1.12}$$

(b) The minors for which

$$i_k = j_k \quad (k = 1, 2, \ldots, p)$$

are called the ***principal minors****.*

(c) The minors for which

$$i_k = j_k = k \quad (k = 1, 2, \ldots, p)$$

are called the ***leading principal minors****.*

Definition 1.8. *The determinant of a square matrix A resulting from the deletion of the rows and columns listed in (1.11) is called the* ***complementary minor*** *and is denoted by*

$$A \begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix}^c$$

The ***complementary cofactor*** *to (1.12) is defined by*

$$A^c \begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} := (-1)^s A \begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix}^c$$

$$s = (i_1 + i_2 + \cdots + i_p) + (j_1 + j_2 + \cdots + j_p)$$

Example 1.15. *For* $A = [a_{ij}]_{i,j=1}^{5}$ *we have*

$$A\begin{pmatrix} 2 & 3 & 5 \\ 1 & 2 & 4 \end{pmatrix} = \begin{bmatrix} a_{21} & a_{22} & a_{24} \\ a_{31} & a_{32} & a_{34} \\ a_{51} & a_{52} & a_{54} \end{bmatrix}$$

$$A\begin{pmatrix} 2 & 3 & 5 \\ 1 & 2 & 4 \end{pmatrix}^{c} = A\begin{pmatrix} 1 & 4 \\ 3 & 5 \end{pmatrix} = \begin{bmatrix} a_{13} & a_{15} \\ a_{43} & a_{45} \end{bmatrix}$$

$$A^{c}\begin{pmatrix} 2 & 3 & 5 \\ 1 & 2 & 4 \end{pmatrix} = (-1)^{17}\begin{bmatrix} a_{13} & a_{15} \\ a_{43} & a_{45} \end{bmatrix} = -(a_{13}a_{45} - a_{15}a_{43})$$

Theorem 1.1. (Laplace's theorem) *Let* A *be an arbitrary* $n \times n$ *matrix and let any* p *rows (or columns) of* A *be chosen. Then*

$$\det A = \sum_{j_1, j_2, \cdots, j_p} A\begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} A^{c}\begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} \tag{1.13}$$

where the summation extends over all $C_n^p := \dfrac{n!}{p!\,(n-p)!}$ *distinct sets of column indices*

$$j_1, j_2, \cdots, j_p \quad (1 \le j_1 < j_2 < \cdots < j_p \le n)$$

Or, equivalently,

$$\det A = \sum_{i_1, j_2, \cdots, i_p} A\begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} A^{c}\begin{pmatrix} i_1 & i_2 \cdots i_p \\ j_1 & j_2 \cdots j_p \end{pmatrix} \tag{1.14}$$

where

$$1 \le i_1 < i_2 < \cdots < i_p \le n$$

Proof. It can be arranged similarly to that of the cofactor expansion formula (1.10). □

1.2.4 Binet–Cauchy formula

Theorem 1.2. (Binet–Cauchy formula) *Two matrices* $A \in \mathbb{R}^{p \times n}$ *and* $B \in \mathbb{R}^{n \times p}$ *are given, that is,*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ a_{p1} & a_{p2} & \cdot & a_{pn} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} & \cdot & b_{1p} \\ b_{21} & b_{22} & \cdot & b_{2p} \\ \cdot & \cdot & \cdot & \cdot \\ b_{n1} & b_{n2} & \cdot & b_{np} \end{bmatrix}$$

Multiplying the rows of A by the columns of B let us construct p^2 numbers

$$\boxed{c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \quad (i, j = 1, \ldots, p)}$$

and consider the determinant $D := |c_{ij}|_{i,j=1}^{p}$. *Then*

1. if $p \leq n$ *we have*

$$D = \sum_{1 \leq j_1 < j_2 < \cdots < j_p \leq n} A \begin{pmatrix} 1 & 2 & \cdots & p \\ j_1 & j_2 & \cdots & j_p \end{pmatrix} B \begin{pmatrix} j_1 & j_2 & \cdots & j_p \\ 1 & 2 & \cdots & p \end{pmatrix} \tag{1.15}$$

2. if $p > n$ *we have*

$$D = 0$$

Proof. It follows directly from Laplace's theorem. □

Example 1.16. *Let us prove that*

$$\begin{bmatrix} \sum_{k=1}^{n} a_k c_k & \sum_{k=1}^{n} a_k d_k \\ \sum_{k=1}^{n} b_k c_k & \sum_{k=1}^{n} b_k d_k \end{bmatrix} = \sum_{1 \leq j < k \leq n} \begin{bmatrix} a_j & a_k \\ b_j & b_k \end{bmatrix} \begin{bmatrix} c_j & c_k \\ d_j & d_k \end{bmatrix} \tag{1.16}$$

Indeed, considering two matrices

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & a_{2n} \end{bmatrix}, \quad B = \begin{bmatrix} c_{11} & d_{12} \\ c_{21} & d_{22} \\ \cdot & \cdot \\ c_{n1} & d_{n2} \end{bmatrix}$$

and applying (1.15) we have (1.16).

Example 1.17. (Cauchy identity) *The following identity holds*

$$\boxed{\left(\sum_{i=1}^{n} a_i c_i\right)\left(\sum_{i=1}^{n} b_i d_i\right) - \left(\sum_{i=1}^{n} a_i d_i\right)\left(\sum_{i=1}^{n} b_i c_i\right) = \sum_{1 \leq j < k \leq n} (a_j b_k - a_k b_j)(c_j d_k - c_k d_j)}$$

It is the direct result of (1.16).

1.3 Linear algebraic equations and the existence of solutions

1.3.1 Gauss's method

Let us consider the set of *m linear equations* (a system of linear equations)

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned} \tag{1.17}$$

in n unknowns $x_1, x_2, \ldots, x_n \in \mathbb{R}$ and $m \times n$ coefficients $a_{ij} \in \mathbb{R}$. An n-tuple $(x_1^*, x_2^*, \ldots, x_n^*)$ is said to be a solution of (1.17) if, upon substituting x_i^* instead of $x_i \cdot (i = 1, \ldots n)$ in (1.17), equalities are obtained. A system of linear equations (1.17) may have

- a unique solution;
- infinitely many solutions;
- no solutions (to be *inconsistent*).

Definition 1.9. *A system of linear equations*

$$\begin{aligned} \tilde{a}_{11}x_1 + \tilde{a}_{12}x_2 + \cdots + \tilde{a}_{1n}x_n &= \tilde{b}_1 \\ \tilde{a}_{21}x_1 + \tilde{a}_{22}x_2 + \cdots + \tilde{a}_{2n}x_n &= \tilde{b}_2 \\ &\cdots \\ \tilde{a}_{m1}x_1 + \tilde{a}_{m2}x_2 + \cdots + \tilde{a}_{mn}x_n &= \tilde{b}_m \end{aligned} \tag{1.18}$$

is said to be ***equivalent*** *to a system (1.17) if their sets of solutions coincide or they do not exist simultaneously.*

It is easy to see that the following *elementary operations transform* the given system of linear equations to an equivalent one:

- interchanging equations in the system;
- multiplying an equation in the given system by a nonzero constant;
- adding one equation, multiplied by a number, to another.

Proposition 1.6. (Gauss's rule) *Any system of m linear equations in n unknowns has an equivalent system in which the augmented matrix has a reduced* ***row-echelon form****, for example, for* $m < n$

$$\begin{aligned} &\tilde{a}_{11}x_1 + \tilde{a}_{12}x_2 + \cdots + \tilde{a}_{1n}x_n = \tilde{b}_1 \\ &\tilde{a}_{21}x_1 + \tilde{a}_{22}x_2 + \cdots + \tilde{a}_{2n}x_n = \tilde{b}_2 \\ &\qquad\qquad \cdots \\ &0 \cdot x_1 + \cdots + 0 \cdot x_{n-m-1} + \tilde{a}_{mn}x_{n-m} + \cdots + \tilde{a}_{mn}x_n = \tilde{b}_m \\ &\qquad\qquad \cdots \\ &0 \cdot x_1 + 0 \cdot x_2 + \cdots + 0 \cdot x_n = 0 \end{aligned}$$

Example 1.18.

$$\begin{array}{r} 2x_1 - x_2 - x_3 + 3x_4 = 1 \\ 4x_1 - 2_2 - x_3 + x_4 = 5 \\ 6x_1 - 3_2 - x_3 - x_4 = 9 \\ 2x_1 - x_2 + 2x_3 - 12x_4 = 10 \end{array} \sim \begin{array}{r} 2x_1 - x_2 - x_3 + 3x_4 = 1 \\ 0 \cdot x_1 + 0 \cdot x_2 + x_3 - 5x_4 = 3 \\ 0 \cdot x_1 + 0 \cdot x_2 + 2x_3 - 10x_4 = 6 \\ 0 \cdot x_1 + 0 \cdot x_2 + 3x_3 - 15x_4 = 9 \end{array}$$

$$\sim \begin{array}{r} 2x_1 - x_2 - x_3 + 3x_4 = 1 \\ 0 \cdot x_1 + 0 \cdot x_2 + x_3 - 5x_4 = 3 \\ 0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 = 0 \\ 0 \cdot x_1 + 0 \cdot x_2 + 0 \cdot x_3 + 0 \cdot x_4 = 0 \end{array}$$

Here the first elementary transform consists in multiplying the first row by $2, 3, 1$ *and adding (with minus) to the following rows, correspondingly. The second elementary transform consists in multiplying the second row by* $2, 3$ *and adding (with minus) to the following rows, correspondingly. Finally, one gets*

$$\begin{array}{l} 2x_1 - x_2 - x_3 + 3x_4 = 1 \\ x_3 - 5x_4 = 3 \end{array}$$

Taking x_2 *and* x_4 *as free variables it follows that*

$$\begin{array}{l} x_1 = \dfrac{1}{2}x_2 + x_4 + 2 \\ x_3 = 5x_4 + 3 \end{array}$$

1.3.2 Kronecker–Capelli criterion

Lemma 1.4. (Kronecker–Capelli) *A system of linear equations given in the form (1.17) has*

- *a unique solution if* $m = n$ *and*

$$\det A = \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & a_{2n} \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdot & a_{nn} \end{bmatrix} \neq 0$$

- *infinitely many solutions if the minimal number of linearly independent rows of* A *(denoted by* rank A*) is equal to one of the extended matrices (denoted by* rank $[A \mid b]$*), that is,*

$$\text{rank } A = \text{rank} \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} \\ a_{21} & a_{22} & \cdot & a_{2n} \\ \cdot & \cdot & & \cdot \\ a_{n1} & a_{n2} & \cdot & a_{nn} \end{bmatrix} = \text{rank } [A \mid b] = \text{rank} \begin{bmatrix} a_{11} & a_{12} & \cdot & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdot & a_{2n} & b_2 \\ \cdot & \cdot & & \cdot & \cdot \\ a_{n1} & a_{n2} & \cdot & a_{nn} & b_n \end{bmatrix}$$

- *no solutions (to be inconsistent) if*

$$\text{rank } A \neq \text{rank } [A \mid b]$$

The proof of this fact will be clarified in the next chapter where the inverse matrix will be introduced.

1.3.3 Cramer's rule

Proposition 1.7. (Cramer) *If* $m = n$ *and* $\det A \neq 0$ *the unique solution of (1.17) is given by*

$$x_i = \frac{1}{\det A} \begin{bmatrix} a_{11} & \cdot & a_{1,i-1} & b_1 & a_{1,i+1} & \cdot & a_{1n} \\ a_{21} & \cdot & a_{2,i-1} & b_2 & a_{2,i+1} & \cdot & a_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & \cdot & a_{n,i-1} & b_n & a_{n,i+1} & \cdot & a_{nn} \end{bmatrix} \quad (i = 1, \ldots, n)$$

The proof of this fact will be also done in the next chapter.

Example 1.19.

$$\left.\begin{matrix} x_1 - 2x_2 = 1 \\ 3x_1 - 4x_2 = 7 \end{matrix}\right\}, \quad \det A = \begin{vmatrix} 1 & -2 \\ 3 & -4 \end{vmatrix} = 2 \neq 0$$

$$x_1 = \frac{1}{2}\begin{vmatrix} 1 & -2 \\ 7 & -4 \end{vmatrix} = \frac{10}{2}, \qquad x_2 = \frac{1}{2}\begin{vmatrix} 1 & 1 \\ 3 & 7 \end{vmatrix} = \frac{4}{2} = 2$$

2 Matrices and Matrix Operations

Contents

2.1 Basic definitions

2.1.1 Basic operations over matrices

The definition of a matrix has been done in (1.1). Here the basic properties of matrices and the operations with them will be considered.

Three basic operations over matrices are defined: summation, multiplication and multiplication of a matrix by a scalar.

Definition 2.1.

1. *The* ***sum*** $A + B$ *of two matrices* $A = [a_{ij}]_{i,j=1}^{m,n}$ *and* $B = [b_{ij}]_{i,j=1}^{m,n}$ *of the same size is defined as*

$$\boxed{A + B := [a_{ij} + b_{ij}]_{i,j=1}^{m,n}}$$

2. *The* ***product*** C *of two matrices* $A = [a_{ij}]_{i,j=1}^{m,n}$ *and* $B = [b_{ij}]_{i,j=1}^{n,p}$ *may be of different sizes (but, as required, the number of columns of the first matrix coincides with the number of rows of the second one) and is defined as*

$$\boxed{C = [c_{ij}]_{i,j=1}^{m,p} = AB := \left[\sum_{k=1}^{n} a_{ik} b_{kj}\right]_{i,j=1}^{m,p}} \tag{2.1}$$

(If $m = p = 1$ *this is the definition of the scalar product of two vectors). In general,*

$$AB \neq BA$$

3. *The operation of* ***multiplication*** *of a matrix* $A \in \mathbb{R}^{m\times n}$ ***by a scalar*** $\alpha \in \mathbb{R}$ *is defined as follows*

$$\boxed{\alpha A = A\alpha := [\alpha a_{ij}]_{i,j=1}^{m,n}}$$

4. *The* ***difference*** $A - B$ *of two matrices* $A = [a_{ij}]_{i,j=1}^{m,n}$ *and* $B = [b_{ij}]_{i,j=1}^{m,n}$ *of the same size is called a matrix* X *satisfying*

$$X + B = A$$

Obviously,

$$\boxed{X = A - B := [a_{ij} - b_{ij}]_{i,j=1}^{m,n}}$$

2.1.2 Special forms of square matrices

Definition 2.2.

1. *A* ***diagonal matrix*** *is a particular case of a squared matrix* $(m = n)$ *for which all elements lying outside the main diagonal are equal to zero:*

$$\boxed{A = \begin{bmatrix} a_{11} & 0 & \cdot & 0 \\ 0 & a_{22} & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & a_{nn} \end{bmatrix} = \operatorname{diag}[a_{11}, a_{22}, \ldots, a_{nn}]}$$

If $a_{11} = a_{22} = \ldots = a_{nn} = 1$ *the matrix* A *becomes the unit (or identity) matrix*

$$\boxed{I_{n\times n} := \begin{bmatrix} 1 & 0 & \cdot & 0 \\ 0 & 1 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 1 \end{bmatrix}}$$

(usually, the subindex in the unit matrix definition is omitted). If $a_{11} = a_{22} = \ldots = a_{nn} = 0$ *the matrix* A *becomes a zero-square matrix:*

$$\mathrm{O}_{n\times n} := \begin{bmatrix} 0 & 0 & \cdot & 0 \\ 0 & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 0 \end{bmatrix}$$

2. *The matrix* $A^{\mathsf{T}} \in \mathbb{R}^{n\times m}$ *is said to be* ***transposed*** *to a matrix* $A \in \mathbb{R}^{m\times n}$ *if*

$$\boxed{A^{\mathsf{T}} = [a_{ji}]_{j,i=1}^{n,m}}$$

3. *The* ***adjoint*** *(or* ***adjudged****) of a square matrix* $A \in \mathbb{R}^{n\times n}$*, written* adj A*, is defined to be the transposed matrix of cofactors* A_{ji} *(1.9) of* A*, that is,*

$$\boxed{\text{adj}\, A := \left([A_{ji}]_{j,i=1}^{n}\right)^{\mathsf{T}}}$$

4. *A square matrix* $A \in \mathbb{R}^{n\times n}$ *is said to be* ***singular*** *or* ***nonsingular*** *according to whether* det A *is zero or nonzero.*
5. *A square matrix* $B \in \mathbb{R}^{n\times n}$ *is referred to as an* ***inverse*** *of the square matrix* $A \in \mathbb{R}^{n\times n}$ *if*

$$AB = BA = I_{n\times n} \tag{2.2}$$

and when this is the case, we write $B = A^{-1}$.
6. *A matrix* $A \in \mathbb{C}^{n\times n}$
   - *is* ***normal*** *if* $AA^* = A^*A$ *and* ***real normal*** *if* $A \in \mathbb{R}^{n\times n}$ *and* $AA^{\mathsf{T}} = A^{\mathsf{T}}A$;
   - *is* ***Hermitian*** *if* $A = A^*$ *and* ***symmetric*** *if* $A \in R^{n\times n}$ *and* $A = A^{\mathsf{T}}$;
   - *is* ***skew-Hermitian*** *if* $A^* = -A$ *and skew-symmetric if* $A \in R^{n\times n}$ *and* $A^{\mathsf{T}} = -A$.
7. *A matrix* $A \in \mathbb{R}^{n\times n}$ *is said to be* ***orthogonal*** *if* $A^{\mathsf{T}}A = AA^{\mathsf{T}} = I_{n\times n}$*, or, equivalently, if* $A^{\mathsf{T}} = A^{-1}$ *and* ***unitary*** *if* $A \in \mathbb{C}^{n\times n}$ *and* $A^*A = AA^* = I_{n\times n}$*, or, equivalently, if* $A^* = A^{-1}$.

2.2 Some matrix properties

The following matrix properties hold:

1. *Commutativity* of the summing operation, that is,

$$A + B = B + A$$

2. *Associativity* of the summing operation, that is,

$$(A + B) + C = A + (B + C)$$

3. *Associativity* of the multiplication operation, that is,

$$(AB)\,C = A\,(BC)$$

4. *Distributivity* of the multiplication operation with respect to the summation operation, that is,

$$(A + B)\,C = AC + BC, C\,(A + B) = CA + CB$$
$$AI = IA = A$$

5. For $A = [a_{ij}]_{i,j=1}^{m,n}$ and $B = [b_{ij}]_{i,j=1}^{n,p}$ it follows that

$$\boxed{AB = \sum_{k=1}^{n} a^{(k)} b^{(k)\mathsf{T}}} \tag{2.3}$$

where

$$a^{(k)} := \begin{pmatrix} a_{1k} \\ \vdots \\ a_{mk} \end{pmatrix}, \quad b^{(k)} := \begin{pmatrix} b_{k1} \\ \vdots \\ b_{kp} \end{pmatrix}$$

6. For the *power matrix* A^p (p is a nonnegative integer number) defined as

$$A^p = \underbrace{AA \cdots A}_{p}, \quad A^0 := I$$

the following *exponent laws* hold

$$\boxed{\begin{aligned} A^p A^q &= A^{p+q} \\ (A^p)^q &= A^{pq} \end{aligned}}$$

where p and q are any nonnegative integers.

7. If two matrices *commute*, that is,

$$AB = BA$$

then

$$(AB)^p = A^p B^p$$

and the formula of Newton's binom holds:

$$(A + B)^m = \sum_{i=0}^{m} C_m^i A^{m-i} B^i$$

where $C_m^i := \dfrac{m!}{i!\,(m-i)!}$.

8. A matrix $U \in \mathbb{C}^{n\times n}$ is unitary if and only if for any $x, y \in \mathbb{C}^n$

$$(Ux, Uy) := (Ux)^* Uy = (x, y)$$

Indeed, if $U^*U = I_{n\times n}$ then $(Ux, Uy) = (x, U^*Uy) = (x, y)$. Conversely, if $(Ux, Uy) = (x, y)$, then $\left([U^*U - I_{n\times n}]x, y\right) = 0$ for any $x, y \in \mathbb{C}^n$ that proves the result.

9. If A and B are unitary, then AB is unitary too.

10. If A and B are normal and $AB = BA$ (they commute), then AB is normal too.
11. If A_i are Hermitian (skew-Hermitian) and α_i are any real numbers, then the matrix $\sum_{i=1}^{m} \alpha_i A_i$ is Hermitian (skew-Hermitian) too.
12. Any matrix $A \in \mathbb{C}^{n\times n}$ can be represented as

$$\boxed{A = H + iK}$$

where

$$\boxed{H = \frac{A + A^*}{2}, \quad K = \frac{A - A^*}{2i}, \quad i^2 = -1}$$

are both Hermitian. If $A \in \mathbb{R}^{n\times n}$, then

$$\boxed{A = S + T}$$

where

$$\boxed{S = \frac{A + A^{\mathsf{T}}}{2}, \quad T = \frac{A - A^{\mathsf{T}}}{2}}$$

and S is symmetric and T is skew-symmetric.
13. For any two square matrices A and B the following *determinant rule* holds:

$$\boxed{\det(AB) = \det(A)\det(B)}$$

This fact directly follows from Binet–Cauchy formula (1.15).
14. For any $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$

$$(AB)^{\mathsf{T}} = B^{\mathsf{T}} A^{\mathsf{T}}$$

Indeed,

$$(AB)^{\mathsf{T}} = \left[\sum_{k=1}^{n} a_{jk} b_{ki}\right]_{i,j=1}^{m,p} = \left[\sum_{k=1}^{n} b_{ki} a_{jk}\right]_{i,j=1}^{m,p} = B^{\mathsf{T}} A^{\mathsf{T}}$$

15. For any $A \in \mathbb{R}^{n\times n}$

$$\boxed{\begin{array}{c} \operatorname{adj} A^{\mathsf{T}} = (\operatorname{adj} A)^{\mathsf{T}} \\ \operatorname{adj} I_{n\times n} = I_{n\times n} \\ \operatorname{adj}(\alpha A) = \alpha^{n-1} \operatorname{adj} A \text{ for any } \alpha \in \mathfrak{F} \end{array}}$$

16. For any $A \in \mathbb{C}^{n\times n}$

$$\operatorname{adj} A^* = (\operatorname{adj} A)^*$$

17. For any $A \in \mathbb{R}^{n\times n}$

$$\boxed{A\,(\text{adj}\,A) = (\text{adj}\,A)\,A = (\det A)\,I_{n\times n}} \tag{2.4}$$

This may be directly proven if we calculate the matrix product in the left-hand side using the row expansion formula (1.10) that leads to a matrix with the number det A in each position on its main diagonal and zeros elsewhere.

18. If $\det A \neq 0$, then

$$\boxed{A^{-1} = \frac{1}{\det A}\text{adj}\,A} \tag{2.5}$$

This may be easily checked if we substitute (2.5) into (2.2) and verify its validity, using (2.4). As a consequence of (2.5) we get

$$\boxed{\det\,(\text{adj}\,A) = (\det A)^{n-1}}$$

As a consequence, we have

$$\boxed{\det\left(A^{-1}\right) = \frac{1}{\det A}} \tag{2.6}$$

This follows from (2.5) and (2.4).

19. If $\det A \neq 0$, then

$$\boxed{\left(A^{-1}\right)^{\intercal} = \left(A^{\intercal}\right)^{-1}}$$

Indeed,

$$I_{n\times n} = AA^{-1} = \left(AA^{-1}\right)^{\intercal} = \left(A^{-1}\right)^{\intercal} A^{\intercal}$$

So, by definitions, $\left(A^{-1}\right)^{\intercal} = \left(A^{\intercal}\right)^{-1}$.

20. If A and B are invertible matrices of the same size, then

$$\boxed{(AB)^{-1} = B^{-1}A^{-1}}$$

As the result, the following fact holds: if $\det A = \det B$, then there exists a matrix C such that

$$\begin{aligned} A &= BC \\ \det C &= 1 \end{aligned}$$

Indeed, $C = B^{-1}A$ and

$$\det C = \det\left(B^{-1}A\right) = \det\left(B^{-1}\right)\det A = \frac{\det A}{\det\,(B)} = 1$$

21. It is easy to see that for any unitary matrix $A \in \mathbb{C}^{n\times n}$ is always invertible and the following properties hold:

$$\boxed{\begin{array}{c} A^{-1} = A^* \\ \det A = \pm 1 \end{array}}$$

Indeed, since $A^*A = AA^* = I_{n\times n}$, it follows that $A^{-1} = A^*$. Also in view of (1.8) we have

$$\det A^*A = \det I_{n\times n} = 1$$
$$\det A^*A = (\det A^*)\,(\det A) = \overline{(\det A)}\,(\det A) = |\det A| = 1$$

22. Let $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times r}$, $C \in \mathbb{R}^{r\times r}$ and $D \in \mathbb{R}^{r\times n}$

(a) If A^{-1} and $\left(I_{r\times r} + DA^{-1}BC\right)^{-1}$ exist, then

$$\boxed{\begin{array}{c} (A+BCD)^{-1} \\ = A^{-1} - A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}CDA^{-1} \\ = A^{-1} - A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}DA^{-1} \end{array}} \tag{2.7}$$

Indeed, the simple matrix multiplication implies

$$\begin{aligned} &\left[A^{-1} - A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}CDA^{-1}\right](A+BCD) \\ &= I_{n\times n} - A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}CD + A^{-1}BCD \\ &\quad -A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}\left[-C + C + CDA^{-1}BC\right]D \\ &= I_{n\times n} - A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}CD + A^{-1}BCD \\ &\quad -A^{-1}BC\left(C + CDA^{-1}BC\right)^{-1}CD - A^{-1}BCD = I_{n\times n} \end{aligned}$$

Analogously,

$$\begin{aligned} &\left[A^{-1} - A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}DA^{-1}\right](A+BCD) \\ &= I_{n\times n} - A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}D + A^{-1}BCD \\ &\quad - A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}\left[-I_{r\times r} + I_{r\times r} + DA^{-1}BC\right]D \\ &= I_{n\times n} - A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}D + A^{-1}BCD \\ &\quad + A^{-1}BC\left(I_{r\times r} + DA^{-1}BC\right)^{-1}D - A^{-1}BCD = I_{n\times n} \end{aligned}$$

(b) In the partial case when $r = 1$, $C = 1$, $B = u$, $D = v^\top$ and $v^\top A^{-1}u \neq -1$, we obtain the *Sherman–Morrison* formula:

$$\boxed{(A + uv^\top)^{-1} = A^{-1} - \frac{A^{-1}v^\top A^{-1}}{1 + v^\top A^{-1}u}} \tag{2.8}$$

The next statement seems to be important for understanding the internal relationship within different classes of matrices.

Claim 2.1.

1. ***Any complex unitary, Hermitian, skew-Hermitian and real orthogonal, symmetric and skew-symmetric matrix is normal****, that is, it satisfies the condition*

$$AA^* = A^*A \text{ for complex matrices}$$

and

$$AA^{\mathsf{T}} = A^{\mathsf{T}}A \text{ for real matrices}$$

2. ***A matrix*** A ***is normal if and only if the matrices*** A ***and*** A^* ***have the same eigenvectors.***

Both of these properties can be easily checked directly.

2.3 Kronecker product

Definition 2.3. *For two matrices* $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{p\times q}$ *the direct (tensor)* ***Kronecker product****, written* $A \otimes B$*, is defined to be the partitioned matrix*

$$A \otimes B := \begin{bmatrix} a_{11}B & a_{12}B & \cdot & a_{1n}B \\ a_{21}B & a_{22}B & \cdot & a_{2n}B \\ \cdot & \cdot & \cdot & \cdot \\ a_{m1}B & a_{m2}B & \cdot & a_{mn}B \end{bmatrix} = [a_{ij}B]_{i,j=1}^{m,n} \in \mathbb{R}^{mp\times nq} \tag{2.9}$$

Example 2.1. *If*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}, \quad B = \begin{bmatrix} b_{11} & b_{12} \end{bmatrix}$$

then

$$A \otimes B = \begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & a_{12}b_{11} & a_{12}b_{12} & a_{13}b_{11} & a_{13}b_{12} \\ a_{21}b_{11} & a_{21}b_{12} & a_{22}b_{11} & a_{22}b_{12} & a_{23}b_{11} & a_{23}b_{12} \end{bmatrix}$$

The following properties for the Kronecker product are fulfilled:

(a)

$$I_{n\times n} \otimes A = \begin{bmatrix} A & O & \cdot & \cdot & O \\ O & A & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & O & A & O \\ O & \cdot & \cdot & O & A \end{bmatrix} = \operatorname{diag}[A, A, \ldots, A] \tag{2.10}$$

(b) $I_{m\times n} \otimes I_{p\times q} = I_{mp\times nq}$

(c) for any $\alpha \in \mathfrak{F}$ it follows that

$$(\alpha A) \otimes B = A \otimes (\alpha B) = \alpha (A \otimes B)$$

(d) $(A + C) \otimes B = A \otimes B + C \otimes B$

(e) $A \otimes (B + C) = A \otimes B + A \otimes C$

(f) $(A \otimes B) \otimes C = A \otimes (B \otimes C)$

(g) $(A \otimes B)^{\mathsf{T}} = A^{\mathsf{T}} \otimes B^{\mathsf{T}}$

and for complex matrices

$$\overline{(A \otimes B)} = \bar{A} \otimes \bar{B}$$
$$(A \otimes B)^* = A^* \otimes B^*$$

Next, very useful properties are less obvious.

Proposition 2.1. *If* $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{p\times q}$, $C \in \mathbb{R}^{n\times k}$ *and* $D \in \mathbb{R}^{q\times r}$ *then*

$$\boxed{(A \otimes B)(C \otimes D) = (AC) \otimes (BD)} \tag{2.11}$$

Proof. It follows from the identity that

$$\sum_{j=1}^{n} (a_{ij}B)(c_{js}D) = \left(\sum_{j=1}^{n} a_{ij}c_{js}\right) BD$$

□

Corollary 2.1. *If* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$ *then*
1.

$$\boxed{A \otimes B = (A \otimes I_{n\times n})(I_{m\times m} \otimes B) = (I_{m\times m} \otimes B)(A \otimes I_{n\times n})}$$

(to prove this it is sufficient to take $C = I_{n\times n}$ *and* $D = I_{m\times m}$*).*
2.

$$\boxed{\begin{aligned}&(A_1 \otimes B_1)(A_2 \otimes B_2) \cdots (A_p \otimes B_p)\\ &= (A_1 A_2 \cdots A_p) \otimes (B_1 B_2 \cdots B_p)\end{aligned}}$$

for all matrices $A_i \in \mathbb{R}^{n\times n}$ *and* $B_i \in \mathbb{R}^{m\times m}$ $(i = 1, \ldots, p)$.

3.

$$\boxed{(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}}$$

provided that both A^{-1} *and* B^{-1} *exist. Indeed,*

$$\begin{aligned}(A \otimes B)(A \otimes B)^{-1} &= (A \otimes B)\left(A^{-1} \otimes B^{-1}\right) \\ &= \left(AA^{-1}\right) \otimes \left(BB^{-1}\right) = I_{n\times n} \otimes I_{m\times m} = I_{nm\times nm}\end{aligned}$$

Proposition 2.2. *If* $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times m}$ *then there exists a permutation* $P \in \mathbb{R}^{nm\times nm}$ *such that*

$$\boxed{P^{\mathsf{T}}(A \otimes B)P = B \otimes A}$$

Proof. It is easy to check that there exists a permutation matrix such that

$$\begin{aligned}P^{\mathsf{T}}(A \otimes I_{n\times n})P &= I_{n\times n} \otimes A \\ P^{\mathsf{T}}(I_{m\times m} \otimes B)P &= B \otimes I_{m\times m}\end{aligned} \tag{2.12}$$

Then, since for any permutation matrix $PP^{\mathsf{T}} = I_{nm\times nm}$, by (2.1) it follows that

$$\begin{aligned}P^{\mathsf{T}}(A \otimes B)P &= P^{\mathsf{T}}(A \otimes I_{n\times n})(I_{m\times m} \otimes B)P \\ &= P^{\mathsf{T}}(A \otimes I_{n\times n})PP^{\mathsf{T}}(I_{m\times m} \otimes B)P \\ &= (I_{n\times n} \otimes A)(B \otimes I_{m\times m}) = B \otimes A\end{aligned}$$

□

Corollary 2.2.

$$\boxed{\det(A \otimes B) = (\det A)^n (\det B)^m} \tag{2.13}$$

Indeed, by (2.1)

$$\det(A \otimes B) = [\det(A \otimes I_{n\times n})][\det(I_{m\times m} \otimes B)]$$

In view of (2.12) and (2.10) one has

$$\begin{aligned}\det(A \otimes I_{n\times n}) &= \det[P^{\mathsf{T}}(I_{n\times n} \otimes A)P] \\ &= \det[PP^{\mathsf{T}}(I_{n\times n} \otimes A)] = \det(I_{n\times n} \otimes A) \\ &= \det(\operatorname{diag}(A, A, \ldots, A)) = (\det A)^n\end{aligned}$$

Analogously,

$$\det(I_{m\times m} \otimes B) = (\det B)^m$$

which completes the proof.

2.4 Submatrices, partitioning of matrices and Schur's formulas

Given matrix $A = [a_{ij}]_{i,j=1}^{m,n}$, if a number of complete rows or columns of A are deleted, or if some complete rows or complete columns are deleted, the new matrix that is obtained is called a *submatrix* of A. A division of matrices into submatrices is referred to as a *partition* of the matrix.

Proposition 2.3. *If the matrices A and B are partitioned as follows*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

where the corresponding submatrices A_{ij} and B_{ij} have the same size, then by direct computation we get

$$\boxed{A \pm B = \begin{bmatrix} A_{11} \pm B_{11} & A_{12} \pm B_{12} \\ A_{21} \pm B_{21} & A_{22} \pm B_{22} \end{bmatrix}} \tag{2.14}$$

and

$$\boxed{AB = \begin{bmatrix} A_{11}B_{11} + A_{12}B_{21} & A_{11}B_{12} + A_{12}B_{22} \\ A_{21}B_{11} + A_{22}B_{21} & A_{21}B_{12} + A_{22}B_{22} \end{bmatrix}} \tag{2.15}$$

Proposition 2.4. *If A_{11} and A_{22} are square matrices then*

1. for

$$A = \begin{bmatrix} A_{11} & \mathrm{O} \\ A_{21} & A_{22} \end{bmatrix}$$

it follows that

$$\det A = (\det A_{11})\,(\det A_{22})$$

2. for

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & \mathrm{O} \end{bmatrix} \in \mathbb{R}^{n\times n}$$

$$A_{12} \in \mathbb{R}^{p\times p}, \quad A_{12} \in \mathbb{R}^{(n-p)\times(n-p)}$$

it follows that

$$\det A = (-1)^{(n+1)p} (\det A_{12}) (\det A_{21})$$

Proposition 2.5. *If A is block-diagonal, that is,*

$$A = \operatorname{diag}\left[A_{11}\ A_{22} \cdots A_{nn}\right]$$

then
1.

$$\boxed{\det A = \prod_{i=1}^{n} \det A_{ii}}$$

2.

$$\boxed{\operatorname{rank} \operatorname{diag}\left[A_{11}\ A_{22} \cdots A_{nn}\right] = \sum_{i=1}^{n} (\operatorname{rank} A_{ii})}$$

Lemma 2.1. (Schur's formulas) *For block-matrix*

$$\tilde{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

where A and D may be of different sizes, the following properties hold:

$$\boxed{\det \tilde{A} = \begin{cases} \det A \det\left(D - CA^{-1}B\right) & \text{if} \quad \det A \neq 0 \\ \det\left(A - BD^{-1}C\right) \det D & \text{if} \quad \det D \neq 0 \\ \det\left(AD - CB\right) & \text{if} \quad AC = CA \\ \det\left(AD - BC\right) & \text{if} \quad CD = DC \end{cases}} \tag{2.16}$$

Proof. Notice that for the case when $\det A \neq 0$, we have

$$\boxed{\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & \mathrm{O} \\ CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & \mathrm{O} \\ \mathrm{O} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ \mathrm{O} & I \end{bmatrix}}$$

which leads directly to the first formula. The second formula, when $\det D \neq 0$, may be proven by the analogous way taking into account the decomposition

$$\boxed{\begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} I & BD^{-1} \\ O & I \end{bmatrix} \begin{bmatrix} A - BD^{-1}C & O \\ O & D \end{bmatrix} \begin{bmatrix} I & O \\ D^{-1}C & I \end{bmatrix}}$$

For proofs of formulas three and four see Gantmacher (1990). □

Lemma 2.2. (on the inversion of a block-matrix) *If*

$$S := \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}, \quad S_{11} \in \mathbb{R}^{l\times l}, \quad S_{22} \in \mathbb{R}^{k\times k}$$

then

$$\boxed{\begin{aligned} S^{-1} &:= \begin{bmatrix} A & B \\ C & D \end{bmatrix} \\ A &= \left(S_{11} - S_{12}S_{22}^{-1}S_{21}\right)^{-1} \in \mathbb{R}^{l\times l} \\ B &= -S_{11}^{-1}S_{12}D \\ C &= -S_{22}^{-1}S_{21}A \\ D &= \left(S_{22} - S_{21}S_{11}^{-1}S_{12}\right)^{-1} \in \mathbb{R}^{k\times k} \end{aligned}}$$

provided by the condition that all inverse matrices exist.

Proof. In view of the identity

$$SS^{-1} = \begin{bmatrix} S_{11}A + S_{12}C & S_{11}B + S_{12}D \\ S_{21}A + S_{22}C & S_{21}B + S_{22}D \end{bmatrix} = I_{(l+k)\times(l+k)}$$

it is sufficient to check the equalities

$$\begin{aligned} S_{11}A + S_{12}C &= I_{l\times l} \\ S_{11}B + S_{12}D &= O_{l\times k} \\ S_{21}A + S_{22}C &= O_{k\times l} \\ S_{21}B + S_{22}D &= I_{k\times k} \end{aligned}$$

The second and third ones hold automatically. Then we have:

$$\begin{aligned} S_{11}A + S_{12}C &= S_{11}A - S_{12}S_{22}^{-1}S_{21}A \\ &= \left(S_{11} - S_{12}S_{22}^{-1}S_{21}\right)A = A^{-1}A = I_{l\times l} \end{aligned}$$

and

$$\begin{aligned} S_{21}B + S_{22}D &= -S_{21}S_{11}^{-1}S_{12}D + S_{22}D \\ &= \left(S_{22} - S_{21}S_{11}^{-1}S_{12}\right)D = D^{-1}D = I_{k\times k} \end{aligned}$$

□

2.5 Elementary transformations on matrices

The determinants' properties, mentioned above, indicate a very close relationship between the determinants of square matrix A and matrices obtained from A by the following *elementary operations*:

- **Operation 1:** interchanging two rows (or columns) in A;
- **Operation 2:** multiplying all elements of a row (or column) of A by a nonzero constant;
- **Operation 3:** adding to any row (or column) of A any other row (or column), multiplied by a nonzero number.

Proposition 2.6. *Each of the elementary operations described above can be achieved by pre- and post-multiplications (respectively) of A by appropriate matrices containing only* 0, 1 *elements and, for the 3rd operation, also one element equal to k, that is,*

- **Matrix form of operation 1:** *the elementary operation of i_1 and i_2 rows interchanging is equal to the following left-hand side multiplication $E^{(1)}A$ where*

$$E^{(1)}_{i_1,i_2} := \begin{array}{c} \\ \\ \\ (i_1) \\ \\ \\ \\ (i_2) \\ \\ \end{array} \begin{bmatrix} 1 & 0 & \cdot & & & & & & \cdot & 0 \\ 0 & \cdot & 0 & \cdot & & & & & & \cdot \\ \cdot & 0 & 1 & \cdot & & & & & & \\ & & 0 & \mathbf{0} & \cdot & \cdot & \cdot & \mathbf{1} & & \\ & & & \mathbf{0} & \mathbf{1} & \cdot & \cdot & \cdot & & \\ & & & \cdot & \cdot & \cdot & \cdot & \cdot & & \\ & & & \mathbf{0} & \cdot & \mathbf{0} & \mathbf{1} & \cdot & & \\ & & & \mathbf{1} & \mathbf{0} & \cdot & \mathbf{0} & \mathbf{0} & \cdot & \\ & & & & & & \cdot & 0 & 1 & \cdot & \cdot \\ \cdot & & & & & & & \cdot & 0 & \cdot & 0 \\ 0 & \cdot & & & & & & & \cdot & 0 & 1 \end{bmatrix}$$

- **Matrix form of operation 2:** *the elementary operation of multiplication of all elements of a row i of A by a nonzero constant α is equal to the following left-hand side multiplication $E^{(2)}_i(\alpha)\,A$ where*

$$E^{(2)}_i(\alpha) := \begin{array}{c} \\ (i) \\ \\ \\ \\ \\ \end{array} \begin{bmatrix} 1 & 0 & \cdot & & & & \cdot & 0 \\ 0 & \cdot & 0 & \cdot & & & & \\ \cdot & 0 & 1 & \cdot & & & & \\ & & 0 & \alpha & 0 & \cdot & & \\ & & & 0 & 1 & 0 & \cdot & \\ & & & \cdot & \cdot & \cdot & \cdot & \\ 0 & \cdot & & & \cdot & 0 & 1 & \end{bmatrix}$$

- **Matrix form of operation 3:** *the elementary operation of adding to a row i_1 of A any other row i_2, multiplied by a* ***nonzero*** *number α, is equal to the following left-hand*

side multiplication $E^{(3)}_{i_1,i_2}(\alpha)\,A$ *where*

$$E^{(3)}_{i_1,i_2}(\alpha) := \begin{matrix} \\ \\ (i_1) \\ \\ \\ \\ (i_2) \\ \\ \end{matrix} \begin{bmatrix} 1 & 0 & \cdot & & & & \cdot & 0 \\ 0 & \cdot & \cdot & & & & & \\ \cdot & 0 & \mathbf{1} & \mathbf{0} & \cdot & \boldsymbol{\alpha} & & \\ & & \mathbf{0} & \mathbf{1} & \mathbf{0} & \cdot & & \\ & & \cdot & \cdot & \cdot & \cdot & & \\ & & \mathbf{0} & \cdot & \mathbf{0} & \mathbf{1} & & \\ & & & & & \cdot & \cdot & \cdot & \cdot \\ \cdot & & & & & & \cdot & \cdot & 0 \\ 0 & \cdot & & & & & \cdot & 1 \end{bmatrix}$$

Similarly, the multiplication of A on the right-hand side by the appropriate matrices $E^{(1)}_{j_1,j_2}$, $E^{(2)}_{j}(\alpha)$ or $E^{(3)}_{j_1,j_2}(\alpha)$ leads to analogous changes in *columns*.

Example 2.2.

(a)

$$A = \begin{bmatrix} a_{11}\,a_{12}\,a_{13} \\ a_{21}\,a_{22}\,a_{23} \\ a_{31}\,a_{32}\,a_{33} \end{bmatrix}, \quad E^{(1)}_{1,3} = \begin{bmatrix} 0\,0\,1 \\ 0\,1\,0 \\ 1\,0\,0 \end{bmatrix}$$

$$E^{(1)}_{1,3}A = \begin{bmatrix} 0\,0\,1 \\ 0\,1\,0 \\ 1\,0\,0 \end{bmatrix} \begin{bmatrix} a_{11}\,a_{12}\,a_{13} \\ a_{21}\,a_{22}\,a_{23} \\ a_{31}\,a_{32}\,a_{33} \end{bmatrix} = \begin{bmatrix} a_{31}\,a_{32}\,a_{33} \\ a_{21}\,a_{22}\,a_{23} \\ a_{13}\,a_{12}\,a_{13} \end{bmatrix}$$

(b)

$$A = \begin{bmatrix} a_{11}\,a_{12}\,a_{13} \\ a_{21}\,a_{22}\,a_{23} \\ a_{31}\,a_{32}\,a_{33} \end{bmatrix}, \quad E^{(2)}_{2} = \begin{bmatrix} 1\,0\,0 \\ 0\,\alpha\,0 \\ 0\,0\,1 \end{bmatrix}$$

$$E^{(2)}_{2}A = \begin{bmatrix} 1\,0\,0 \\ 0\,\alpha\,0 \\ 0\,0\,1 \end{bmatrix} \begin{bmatrix} a_{11}\,a_{12}\,a_{13} \\ a_{21}\,a_{22}\,a_{23} \\ a_{31}\,a_{32}\,a_{33} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ \alpha a_{21} & \alpha a_{22} & \alpha a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

(c)

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}, \quad E^{(3)}_{1,3}(\alpha) = \begin{bmatrix} 1 & 0 & \alpha \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$E^{(3)}_{1,3}(\alpha) A = \begin{bmatrix} 1 & 0 & \alpha \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

$$= \begin{bmatrix} a_{11} + \alpha a_{31} & a_{12} + \alpha a_{32} & a_{13} + \alpha a_{33} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Proposition 2.7. *For any subindices*

$$\det E^{(1)}_{i_1,i_2} = -1, \quad \det E^{(2)}_{i}(\alpha) = \alpha, \quad \det E^{(3)}_{i_1,i_2}(\alpha) = 1 \tag{2.17}$$

and

$$\det E^{(1)}_{i_1,i_2} A = \det A E^{(1)}_{j_1,j_2} = \det A$$
$$\det E^{(2)}_{i}(\alpha) A = \det A E^{(2)}_{j}(\alpha) = \alpha \det A$$
$$\det E^{(3)}_{i_1,i_2}(\alpha) A = \det A E^{(3)}_{j_1,j_2}(\alpha) = \det A$$

Proof. The formulas above are the simple mathematical expressions of the determinants, properties (1.6), (1.3) and (1.7). □

Proposition 2.8. *For any $m \times n$ matrix A there exists a finite sequence of elementary matrices $E_1, E_2, \ldots, E_{k+s}$ such that*

$$E_k \cdots E_2 E_1 A E_{k+1} \cdots E_{k+s}$$

is one of the following matrices

1. $I_{n\times n}$ *for* $m = n$

2. $\begin{bmatrix} I_{m\times m} & \mathrm{O}_{m\times(n-m)} \end{bmatrix}$ *for* $m < n$

3. $\begin{bmatrix} I_{n\times n} \\ \mathrm{O}_{(m-n)\times n} \end{bmatrix}$ *for* $m > n$ (2.18)

4. $\begin{bmatrix} I_{r\times r} & \mathrm{O}_{r\times(n-r)} \\ \mathrm{O}_{(n-r)\times r} & \mathrm{O}_{(n-r)\times(n-r)} \end{bmatrix}$ $(r \leq \min(m, n))$

known as ***canonical*** *ones.*

Proof. It follows directly from the elementary operations definition and its relation to the canonical matrix forms. □

Example 2.3. *For*

$$A = \begin{bmatrix} 0 & 1 & 2 & -1 \\ -2 & 0 & 0 & 6 \\ 4 & -2 & -4 & -10 \end{bmatrix}$$

we get

$$A_1 := E_2^{(2)}(-1/2)\,A = \begin{bmatrix} 0 & 1 & 2 & -1 \\ 1 & 0 & 0 & -3 \\ 4 & -2 & -4 & -10 \end{bmatrix}$$

$$A_2 := E_{2,3}^{(3)}(-4)\,A_1 = \begin{bmatrix} 0 & 1 & 2 & -1 \\ 1 & 0 & 0 & -3 \\ 0 & -2 & -4 & 2 \end{bmatrix}$$

$$A_3 := E_{1,2}^{(1)} A_2 = \begin{bmatrix} 1 & 0 & 0 & -3 \\ 0 & 1 & 2 & -1 \\ 0 & -2 & -4 & 2 \end{bmatrix}$$

$$A_4 := E_{3,2}^{(3)}(2)\,A_3 = \begin{bmatrix} 1 & 0 & 0 & -3 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_5 := A_4 E_{4,1}^{(3)}(3) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_6 := A_5 E_{4,2}^{(3)}(1) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$A_7 := A_6 E_{3,2}^{(3)}(-2) = \begin{bmatrix} \mathbf{1} & \mathbf{0} & 0 & 0 \\ \mathbf{0} & \mathbf{1} & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

that is,

$$E_{3,2}^{(3)}(2)\,E_{1,2}^{(1)} E_{2,3}^{(3)}(-4)\,E_2^{(2)}(-1/2)\,A E_{4,1}^{(3)}(3)\,E_{4,2}^{(3)}(1)\,E_{3,2}^{(3)}(-2) = \begin{bmatrix} I_{2\times 2} & \mathrm{O} \\ \mathrm{O} & \mathrm{O} \end{bmatrix}$$

Corollary 2.3. *For any $m \times n$ matrix A there exist matrices P ($\det P \neq 0$) and Q ($\det Q \neq 0$) such that PAQ is equal to one of the canonical matrices (2.18).*

Proof. It follows from (2.7) and the fact that

$$\det\left(\prod_{i=1}^{k} E_i\right) = \prod_{i=1}^{k} \det(E_i) \neq 0$$

□

Definition 2.4. *Two matrices A and B are said to be* ***equivalent*** *(or belonging to the same equivalent class) if they may be transformed by the elementary operations application to the same canonical matrix form. This is written as* $A \sim B$.

Proposition 2.9. *An* $n \times n$ *matrix A is nonsingular if and only if* $A \sim I_{n\times n}$.

Proof. Since

$$\det(PAQ) = (\det A)(\det P)(\det Q) = \det S$$

where S is one of the canonical matrices (2.18), we have

$$\det A = \frac{\det S}{(\det P)(\det Q)} \neq 0$$

if and only if $S = I_{n\times n}$. □

Definition 2.5. *A square* $n \times n$ *matrix A is said to be* ***simple*** *if it is equivalent to a diagonal matrix D.*

These definitions will be used frequently below.

2.6 Rank of a matrix

Definition 2.6. *For a matrix* $A \in \mathbb{R}^{m\times n}$ *the size*

$$r \quad (1 \leq r \leq \min(m, n))$$

of the identity matrix in the canonical form for A is referred to as the ***rank*** *of A, written* $r = \operatorname{rank} A$. *If* $A = O_{m\times n}$ *then* $\operatorname{rank} A = 0$, *otherwise* $\operatorname{rank} A \geq 1$.

For each four canonical forms in (2.18) we have

$$\operatorname{rank} I_{n\times n} = n \text{ for } m = n$$

$$\operatorname{rank}\begin{bmatrix} I_{m\times m} & O_{m\times(n-m)} \end{bmatrix} = m \text{ for } m < n$$

$$\operatorname{rank}\begin{bmatrix} I_{n\times n} \\ O_{(m-n)\times n} \end{bmatrix} = n \text{ for } m > n$$

$$\begin{bmatrix} I_{r\times r} & O_{r\times(n-r)} \\ O_{(n-r)\times r} & O_{(n-r)\times(n-r)} \end{bmatrix} = r \text{ for } r \leq \min(m,n)$$

Proposition 2.10. *For a square matrix* $A \in \mathbb{R}^{n\times n}$ $\operatorname{rank} A = n$ *if and only if it is nonsingular.*

Proof. It follows straightforwardly from proposition (2.9). □

Corollary 2.4. *The rank of a matrix* $A \in \mathbb{R}^{m\times n}$ *is equal to the order of its largest nonzero minor.*

Several important properties of rank are listed below.

1. (**Frobenius inequality**) If A, B and C are rectangular matrices and the product ABC is well defined, then

$$\boxed{\operatorname{rank}(AB) + \operatorname{rank}(BC) \leq \operatorname{rank}(A) + \operatorname{rank}(ABC)} \tag{2.19}$$

$$\boxed{\operatorname{rank}(AB) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}} \tag{2.20}$$

Indeed, taking in (2.19) first A and C to be appropriate size we obtain (2.20).

2. For any complex matrix A

$$\boxed{\operatorname{rank}(A) = \operatorname{rank}(AA^*) = \operatorname{rank}(A^*A)}$$

3. If P and Q are nonsingular and A is square, then

$$\boxed{\operatorname{rank}(PAQ) = \operatorname{rank}(A)} \tag{2.21}$$

Indeed, by (2.20) it follows

$$\begin{aligned}
&\operatorname{rank}(PAQ) \leq \min\{\operatorname{rank}(P), \operatorname{rank}(AQ)\} \\
&= \min\{n, \operatorname{rank}(AQ)\} = \operatorname{rank}(AQ) \\
&\quad \leq \min\{\operatorname{rank}(A), \operatorname{rank}(Q)\} = \operatorname{rank}(A) \\
&= \operatorname{rank}\left(P^{-1}[PAQ]Q^{-1}\right) \\
&\quad \leq \min\left\{\operatorname{rank}\left(P^{-1}\right), \operatorname{rank}\left([PAQ]Q^{-1}\right)\right\} \\
&= \operatorname{rank}\left([PAQ]Q^{-1}\right) \\
&\quad \leq \min\left\{\operatorname{rank}(PAQ), \operatorname{rank}\left(Q^{-1}\right)\right\} = \operatorname{rank}(PAQ)
\end{aligned}$$

4.

$$\boxed{\operatorname{rank}(A) = \operatorname{rank}(A^\intercal) = \operatorname{rank}(A^*)} \tag{2.22}$$

5. For any $A, B \in \mathbb{R}^{m\times n}$

$$\boxed{\operatorname{rank}(A+B) \leq \operatorname{rank}(A) + \operatorname{rank}(B)} \tag{2.23}$$

6. **(Sylvester's rule)** For any $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{n\times p}$

$$\boxed{\begin{array}{c}\operatorname{rank}(A) + \operatorname{rank}(B) - n \le \operatorname{rank}(AB) \\ \le \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}\end{array}} \tag{2.24}$$

7. For any $A \in \mathbb{R}^{m\times m}$ and $B \in \mathbb{R}^{n\times m}$

$$\boxed{\operatorname{rank}(A \otimes B) = (\operatorname{rank} A)\,(\operatorname{rank} B)} \tag{2.25}$$

This follows from (2.11).

2.7 Trace of a quadratic matrix

Definition 2.7. *The trace of a matrix $A \in \mathbb{C}^{n\times n}$ (may be with complex elements), written* $\operatorname{tr} A$*, is defined as the sum of all elements lying on the main diagonal of A, that is,*

$$\boxed{\operatorname{tr} A := \sum_{i=1}^{n} a_{ii}} \tag{2.26}$$

Some evident properties of trace follow.

1. For any $A, B \in \mathbb{C}^{n\times n}$ and any $\alpha, \beta \in \mathbb{C}$

$$\boxed{\operatorname{tr}(\alpha A + \beta B) = \alpha \operatorname{tr} A + \beta \operatorname{tr} B} \tag{2.27}$$

2. For any $A \in \mathbb{C}^{m\times n}$ and any $B \in \mathbb{C}^{n\times m}$

$$\boxed{\operatorname{tr}(AB) = \operatorname{tr}(BA)} \tag{2.28}$$

Indeed,

$$\boxed{\begin{aligned}\operatorname{tr}(AB) &:= \sum_{i=1}^{m}\sum_{k=1}^{n} a_{ik}b_{ki} = \sum_{i=1}^{m}\sum_{k=1}^{n} b_{ki}a_{ik} \\ &= \sum_{k=1}^{n}\sum_{i=1}^{m} b_{ki}a_{ik} = \operatorname{tr}(BA)\end{aligned}}$$

3. For any $A \in \mathbb{C}^{n\times n}$

$$\boxed{\operatorname{tr}(AA^*) = \operatorname{tr}(A^*A) = \sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2} \tag{2.29}$$

(this follows directly from property 2).

4. If S^{-1} $\left(S \in \mathbb{C}^{n\times n}\right)$ exists then for any $A \in \mathbb{C}^{n\times n}$

$$\boxed{\operatorname{tr}\left(S^{-1}AS\right) = \operatorname{tr}\left(SS^{-1}A\right) = \operatorname{tr}(A)} \tag{2.30}$$

5. For any $A \in \mathbb{C}^{n\times n}$ and any $B \in \mathbb{C}^{p\times p}$

$$\boxed{\operatorname{tr}(A \otimes B) = \operatorname{tr}(A)\operatorname{tr}(B)} \tag{2.31}$$

Indeed,

$$\begin{aligned}\operatorname{tr}(A \otimes B) &:= \sum_{i=1}^{n}\left(a_{ii}\sum_{j=1}^{p} b_{ii}\right)\\ &= \left(\sum_{i=1}^{n} a_{ii}\right)\left(\sum_{j=1}^{p} b_{ii}\right) = \operatorname{tr}(A)\operatorname{tr}(B)\end{aligned}$$

3 Eigenvalues and Eigenvectors

Contents

3.1 Vectors and linear subspaces

Definition 3.1. *The matrix* $A \in \mathbb{C}^{n\times 1} := \mathbb{C}^n$, *written as*

$$a := \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix}$$

is called a ***vector*** $a \in \mathbb{C}^{n\times 1}$.

Definition 3.2. *The matrix product (2.1), written for* $a \in \mathbb{C}^{n\times 1}$ *and* $b \in \mathbb{C}^{1\times n}$, *is called a* ***scalar*** *(inner)* ***product*** *of two vectors* a *and* b *and denoted by*

$$\boxed{(a, b) := a^* b = \sum_{i=1}^{n} \bar{a}_i b_i} \tag{3.1}$$

which for real vectors $a, b \in \mathbb{R}^n$ *becomes*

$$\boxed{(a, b) := a^{\mathsf{T}} b = \sum_{i=1}^{n} a_i b_i} \tag{3.2}$$

Definition 3.3. *For the set of vectors* $x_1, x_2, \ldots, x_k \in \mathbb{C}^n$ *and elements* $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{C}$ *the following notions may be introduced:*

1. ***Linear combinations*** *of* $x_1, x_2, \ldots, x_{k\leq n}$ *over* $\mathbb{C}$ *are an element of the form*

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k$$

2. *The set of all linear combinations of* $x_1, x_2, \ldots, x_k$ *over* $\mathbb{C}$ *is called a* ***subspace*** *or the* ***span*** *of* $x_1, x_2, \ldots, x_k$, *denoted by*

$$\boxed{\begin{array}{c} \operatorname{span}\{x_1, x_2, \ldots, x_k\} := \\ \{x = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k : \alpha_i \in \mathbb{C},\ i = 1, \ldots, k\} \end{array}} \tag{3.3}$$

3. *Some vectors* $x_1, x_2, \ldots, x_k \in \mathbb{C}^n$ *are said to be* ***linearly dependent*** *over* $\mathbb{C}$ *if there exist* $\alpha_1, \alpha_2, \ldots, \alpha_k \in \mathbb{C}$ *not all zero* $\left(\sum_{i=1}^{n} |\alpha_i|^2 > 0\right)$ *such that*

$$\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k = 0$$

Otherwise, they are said to be ***linearly independent***.

4. *If* S *is a subspace of* $\mathbb{C}^n$, *then a set of vectors* $x_1, x_2, \ldots, x_k \in \mathbb{C}^n$ *is called a* ***basis*** *for* S *if*

- $x_1, x_2, \ldots, x_k$ *are linearly independent;*
- $S = \text{span}\{x_1, x_2, \ldots, x_k\}$.

Such basis for a subspace S *is not unique: all bases for* S *have the same number of elements which is called the* ***dimension*** *of* S, *denoted by* $\dim S$.

5. *Vectors* $x_1, x_2, \ldots, x_k \in \mathbb{C}^n$ *are* ***mutually orthogonal*** *if*

$$\left(x_i^*, x_j\right) = 0 \quad \textit{for all} \quad i \neq j$$

and are ***orthonormal*** *if for all* $i, j = 1, \ldots, n$

$$\left(x_i^*, x_j\right) = \delta_{ij}$$

where $x_i^* = (\bar{x}_i)^\mathsf{T}$ *and* $\delta_{ij} = \begin{cases} 1 & \text{if} \quad i = j \\ 0 & \text{if} \quad i \neq j \end{cases}$ *is the Kronecker (delta-function) symbol.*

6. *The* ***orthogonal completion*** $S^\perp$ *of a subspace* $S \subset \mathbb{C}^n$ *is defined by*

$$S^\perp = \text{span}\{x_{k+1}, x_{k+2}, \ldots, x_n\} \tag{3.4}$$

where the vectors x_{k+j} $(j = 1, \ldots, n-k)$ *are orthonormal.*

Any matrix $A \in \mathbb{C}^{m \times n}$ may be considered as **a linear transformation** from $\mathbb{C}^n$ to $\mathbb{C}^m$, i.e.,

$$A : \mathbb{C}^n \longmapsto \mathbb{C}^m$$

Definition 3.4.

(a) The ***kernel*** *(or* ***null space****) of the linear transformation* $A : \mathbb{C}^n \longmapsto \mathbb{C}^m$ *is defined by*

$$\boxed{\text{Ker}\, A = \mathcal{N}(A) := \{x \in \mathbb{C}^n : Ax = 0\}} \tag{3.5}$$

(b) The ***image*** *(or* ***range****) of the linear transformation* $A : \mathbb{C}^n \longmapsto \mathbb{C}^m$ *is*

$$\boxed{\text{Im}\, A = \mathcal{R}(A) := \{y \in \mathbb{C}^m : y = Ax, \ x \in \mathbb{C}^n\}} \tag{3.6}$$

(c) The dimension of the subspace $\text{Ker}\, A = \mathcal{N}(A)$ *is referred to as the* ***defect*** *of the transformation* A *and is denoted by* $\text{def}\, A$, *that is,*

$$\boxed{\text{def}\, A := \dim \text{Ker}\, A} \tag{3.7}$$

By the definitions above, it is clear that both Ker A and ImA are subspaces in $\mathbb{C}^n$ and $\mathcal{R}(A)$, respectively. Moreover, it can be easily seen that

$$\boxed{\begin{array}{c}\dim \operatorname{Ker} A + \dim \operatorname{Im} A = n \\ \dim \operatorname{Im} A = \dim(\operatorname{Ker} A)^{\perp} \\ \operatorname{rank} A \leq \dim \operatorname{Im} A \\ \operatorname{rank} A = \operatorname{rank} A^*\end{array}} \tag{3.8}$$

Proposition 3.1. *If* $A_1, A_2 : S_1 \subseteq \mathbb{C}^n \longmapsto S_2 \subseteq \mathbb{C}^m$ *then*

$$\boxed{\begin{array}{c}\operatorname{Im}(A_1 + A_2) \subset \operatorname{Im}(A_1) + \operatorname{Im}(A_2) \\ \operatorname{rank} A_1 - \operatorname{rank} A_2 \leq \operatorname{rank}(A_1 + A_2) \leq \operatorname{rank} A_1 + \operatorname{rank} A_2\end{array}}$$

It follows directly from the definitions above.

Proposition 3.2. *If* $A : S_1 \subseteq \mathbb{C}^n \longmapsto S_2 \subseteq \mathbb{C}^m$ *and* $B : S_3 \subseteq \mathbb{C}^m \mathbb{C}^m \longmapsto S_4 \subseteq \mathbb{C}^p$ *then*

$$\boxed{\begin{array}{c}\operatorname{Im}(BA) = B\operatorname{Im}(A) \\ \operatorname{rank}(BA) \leq \min\{\operatorname{rank} A, \operatorname{rank} B\}\end{array}}$$

This is the consequence of (3.8).

It is not difficult to verify that for any $A : S_1 \subseteq \mathbb{C}^n \longmapsto S_2 \subseteq \mathbb{C}^m$ and $B : S_3 \subseteq \mathbb{C}^m \longmapsto S_4 \subseteq \mathbb{C}^p$ one has

1.

$$\boxed{\operatorname{rank} A + \operatorname{def} A = \dim(S_1)} \tag{3.9}$$

2. for any $S \subset S_1$

$$\boxed{\dim A(S) = \dim S - \dim(S \cap \operatorname{Ker} A)}$$

3. for any $A, B : S_1 \subseteq \mathbb{C}^n \longmapsto S_2 \subseteq \mathbb{C}^m$

$$\boxed{\operatorname{Ker} A \cap \operatorname{Ker} B \subset \operatorname{Ker}(A + B)}$$

4. If $S = S_1$ then

$$\boxed{\dim A(S) \geq \dim S - \operatorname{def} T \geq \operatorname{def} AB}$$

5.

$$\boxed{\operatorname{def} AB \leq \operatorname{def} A + \operatorname{def} B}$$

6. If A is left invertible ($S \subset S_1$) if and only if

$$\boxed{\dim A(S) = \dim S}$$

7.

$$\boxed{\text{def } A = n - m}$$

3.2 Eigenvalues and eigenvectors

Definition 3.5. *Let $A \in \mathbb{C}^{n\times n}$ be a* ***squared*** *$n \times n$ matrix (may be with complex elements). Then*

(a) any nonzero vector $x \in \mathbb{C}^n$ is referred to as ***a right eigenvector*** *of the matrix A if it satisfies the equation*

$$\boxed{Ax = \lambda^{(r)}x} \tag{3.10}$$

for some (may be zero) complex value $\lambda^{(r)} \in \mathbb{C}$ called the eigenvalue of the matrix A which corresponds to this right eigenvector x;

(b) any nonzero vector $x \in \mathbb{C}^n$ is referred to as ***a left eigenvector*** *of the matrix A if it satisfies the equation*

$$\boxed{x^*A = \lambda^{(l)}x^*} \tag{3.11}$$

for some (may be zero) complex value $\lambda^{(l)} \in \mathbb{C}$ called the ***eigenvalue*** *of the matrix A which corresponds to this left eigenvector x.*

Remark 3.1. *If x is an eigenvalue, then for any nonzero $\alpha \in \mathbb{C}$ the vector αx is also the eigenvector. This means that for each $\lambda^{(r)}$ (the same for $\lambda^{(l)}$) there exists a single dimensional subspace*

$$S_\lambda = \{\alpha x : Ax = \lambda x,\ \alpha \in \mathbb{C}\}$$

of the corresponding eigenvectors αx.

Proposition 3.3. *For any matrix $A \in \mathbb{C}^{n\times n}$ any eigenvalue $\lambda^{(r)}$ (as well as $\lambda^{(l)}$) satisfies the, so-called,* ***characteristic equation***

$$p_A(\lambda) := \det(\lambda I_{n\times n} - A) = 0 \tag{3.12}$$

Proof. By (3.10) (or (3.11)) it follows that

$$\left(\lambda^{(r)}I_{n\times n} - A\right)x = 0, \qquad x^*\left(\lambda^{(l)}I_{n\times n} - A\right) = 0$$

Hence these equations have nonzero solutions if and only if (see Proposition 1.7) (3.12) holds. □

Evidently, the *characteristic polynomial*

$$p(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \tag{3.13}$$

has exactly n roots $\lambda_i (i = 1, \ldots, n)$. Some of these roots may coincide.

Definition 3.6. *The set of all roots of $p_A(\lambda)$ is called the* ***spectrum*** *of A and is denoted by*

$$\sigma(A) := \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$$

where λ_j satisfies

$$p_A(\lambda_j) = 0$$

The maximum modulus of the eigenvalues is called the ***spectral radius*** *of A, denoted by*

$$\rho(A) := \max_{1 \le j \le n} |\lambda_j| \tag{3.14}$$

The following remark seems to be evident.

Remark 3.2. *The spectrum of A contains all right eigenvalues as well as all left eigenvalues which implies that any $\lambda_j^{(r)}$ has an equal $\lambda_j^{(l)}$, that is, there exist two indices i and j such that*

$$\lambda_i^{(r)} = \lambda_j^{(l)}$$

Proposition 3.4. *If x is a right (left) eigenvector of a real matrix $A \in \mathbb{R}^{n \times n}$ with the corresponding eigenvalue λ, that is, $Ax = \lambda x$, then the complex conjugated vector $\bar{x}$ is also an eigenvector of A with the corresponding eigenvalue $\bar{\lambda}$.*

Proof. Let $x = u + iv$ and $\lambda = \alpha + i\beta$. Then we have

$$\begin{aligned} Ax &= A(u + iv) = Au + i(Av) \\ &= \lambda x = (\alpha + i\beta)(u + iv) = (\alpha u - \beta v) + i(\beta u + \alpha v) \end{aligned}$$

This implies

$$\begin{aligned} Au &= \alpha u - \beta v \\ Av &= \beta u + \alpha v \end{aligned}$$

or, equivalently,

$$\begin{aligned} Au &= \alpha u - (-\beta)(-v) \\ A(-v) &= (-\beta)u + \alpha(-v) \end{aligned}$$

which, after multiplication of the second equality by the complex unite i and summation of both equalities, leads to the following identity

$$A(u - iv) = A\bar{x} = (\alpha - i\beta)(u - iv) = \bar{\lambda}\bar{x}$$

For the left eigenvalues the proof is similar. □

Corollary 3.1. *The following presentation of the characteristic polynomial* $p(\lambda)$ *(3.13) takes place*

$$\boxed{p(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i) = \left(\prod_{i=1}^{s}[\lambda_i - u_i]\right)\left(\prod_{i=1}^{n-s}[(\lambda_i - u_i)^2 + v_i^2]\right)} \tag{3.15}$$

where s *is the number of pure real eigenvalues and*

$$u_i = Re\lambda_i, \quad v_i = Im\lambda_i$$

Example 3.1. *For*

$$A = \begin{bmatrix} 1 & -1 & 0 \\ 2 & 3 & 2 \\ 1 & 1 & 2 \end{bmatrix}$$

by Sarrius's rule we have

$$\begin{aligned} p_A(\lambda) = \det(\lambda I_{3\times 3} - A) &= \det \begin{bmatrix} \lambda - 1 & 1 & 0 \\ -2 & \lambda - 3 & -2 \\ -1 & -1 & \lambda - 2 \end{bmatrix} \\ &= (\lambda - 1)(\lambda - 3)(\lambda - 2) + 2 - 2(\lambda - 1) + 2(\lambda - 2) \\ &= (\lambda - 1)(\lambda - 2)(\lambda - 3) \end{aligned}$$

which implies

$$\lambda_1 = 1, \lambda_2 = 2, \lambda_3 = 3$$

One of the solutions of the equation $Ax^{(i)} = \lambda_i x^{(i)}$ $(i = 1, 2, 3)$ *is*

$$x^{(1)} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -2 \\ 2 \\ 1 \end{bmatrix}, \quad x^{(2)} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix}$$

In this example a nonsymmetric matrix has real eigenvalues and the corresponding eigenvectors. The next proposition shows when this occurs.

Proposition 3.5. *If a square* $n \times n$ *matrix* A *is* ***Hermitian*** *or* ***real symmetric****, that is, if*

$$A^* = A \quad or \quad A = A^\intercal$$

then obligatory all eigenvalues λ_j *are* ***real****. If* A *is real symmetric, then the corresponding eigenvectors* $x^{(j)}$ $(j = 1, \ldots, n)$ *are real too.*

Proof. Suppose that λ_j and $x^{(j)}$ are complex, i.e.,

$$\lambda_j = \alpha_j + i\beta_j$$
$$x^{(j)} = u^{(j)} + iv^{(j)}$$

Then

$$x^{(j)*}Ax^{(j)} = x^{(j)*}\left(\lambda_j x^{(j)}\right) = \lambda_j \left|x^{(j)}\right|^2$$

and calculating the complex conjugation plus transposition from both sides by symmetricity we get

$$\left(x^{(j)*}Ax^{(j)}\right)^* = x^{(j)*}A^*x^{(j)} = x^{(j)*}Ax^{(j)}x^{(j)*} = \bar{\lambda}_j \left|x^{(j)}\right|^2$$

which leads to the following relation

$$\lambda_j = \bar{\lambda}_j$$

This means that λ_j is real. Finally, by the Proposition 3.4 in the case of a real matrix we obtain that the solution of the linear uniform system $Ax^{(j)} = \lambda_j x^{(j)}$, containing only real elements, with respect to $x^{(j)}$ may give only a real solution. □

Proposition 3.6. *Eigenvectors corresponding to distinct eigenvalues are linearly independent.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_s$ $(s \leq n)$ be the distinct eigenvalues of a matrix A and x_1, $x_2, \ldots, x_s$ denote corresponding eigenvectors. Suppose that there exist numbers $\alpha_1, \alpha_2, \ldots, \alpha_s$ such that

$$\sum_{i=1}^{s} \alpha_i x_i = 0, \quad \sum_{i=1}^{s} |\alpha_i|^2 > 0 \tag{3.16}$$

Show that this is impossible. To prove that $\alpha_i = 0$ $(i = 1, \ldots, s)$ we first multiply both sides of (3.16) on the left by

$$(A - \lambda_{s-1}I_{n\times n})(A - \lambda_{s-2}I_{n\times n}) \cdots (A - \lambda_1 I_{n\times n})$$

and noting that $(A - \lambda_k I_{n\times n})x_k = 0$, we get

$$\begin{aligned}
&(A - \lambda_{s-1}I_{n\times n})(A - \lambda_{s-2}I_{n\times n}) \cdots (A - \lambda_1 I_{n\times n}) \sum_{i=1}^{s} \alpha_i x_i \\
&\quad = (A - \lambda_{s-1}I_{n\times n})(A - \lambda_{s-2}I_{n\times n}) \cdots (A - \lambda_2 I_{n\times n}) \sum_{i=2}^{s} (\lambda_i - \lambda_1)\alpha_i x_i \\
&\quad = (A - \lambda_{s-1}I_{n\times n})(A - \lambda_{s-2}I_{n\times n}) \cdots (A - \lambda_3 I_{n\times n}) \\
&\qquad \times \sum_{i=3}^{s} (\lambda_i - \lambda_1)(\lambda_i - \lambda_2)\alpha_i x_i = \cdots \\
&\quad = (\lambda_s - \lambda_1)(\lambda_s - \lambda_2) \cdots (\lambda_s - \lambda_{s-1})\alpha_s x_s = 0
\end{aligned}$$

which implies that $\alpha_s = 0$. Analogously, we may prove that

$$\alpha_1 = \alpha_2 = \cdots = \alpha_{s-1} = 0$$

So, all $\alpha_i = 0$ which contradicts (3.16). □

Proposition 3.7. *Eigenvectors of an* ***Hermitian matrix*** $(A = A^*)$*, corresponding to distinct eigenvalues, are* ***orthogonal****, that is,*

$$x^{(i)*}x^{(j)} = 0 \tag{3.17}$$

for any indices i and j such that $\lambda_i \neq \lambda_j$.

Proof. Pre-multiplying both equations below

$$x^{(j)*} : Ax^{(i)} = \lambda_i x^{(i)}$$
$$x^{(i)*} : Ax^{(j)} = \lambda_j x^{(j)}$$

by $x^{(j)*}$ and $x^{(j)}$, we derive

$$x^{(j)*}Ax^{(i)} = \left(Ax^{(j)}\right)^* x^{(i)} = \lambda_i x^{(j)*}x^{(i)}$$
$$\left(x^{(i)*}Ax^{(j)}\right)^* = \left(Ax^{(j)}\right)^* x^{(i)} = \lambda_j \left(x^{(i)*}x^{(j)}\right)^* = \lambda_j x^{(j)*}x^{(i)}$$

Multiplying the second equation by (-1) and summing both equalities we obtain

$$0 = (\lambda_i - \lambda_j)x^{(j)*}x^{(i)}$$

Since $\lambda_i \neq \lambda_j$ the result follows. □

Example 3.2. *The matrix $A = uv^\top$ $(u, v \in \mathbb{R}^{n\times 1})$ has one eigenvalue λ_1 equal to $v^\top u$ (with the corresponding eigenvector $x^{(1)} = u$) and all other eigenvalues $\lambda_{i\neq 1}$ equal to 0 (with the corresponding eigenvectors $x^{(i\neq 1)} = w^{(i\neq 1)} \perp v$). Indeed,*

$$Au = u(v^\top u) = (v^\top u)u$$
$$Av^{(i\neq 1)} = u(v^\top w^{(i\neq 1)}) = \mathrm{O}_{n\times 1} = 0 \cdot w^{(i\neq 1)} \ (i = 2, \ldots, n)$$

Proposition 3.8. *If $B = T^{-1}AT$ and $p_A(\lambda)$ is the characteristic polynomial of A, then*

$$\boxed{p_A(\lambda) = p_B(\lambda)}$$

that is, equivalent matrices have the same characteristic polynomials.

Proof. Indeed,

$$p_A(\lambda) = \det(\lambda I_{n\times n} - A) = \det T \cdot \det(\lambda I_{n\times n} - A) \cdot \det T^{-1}$$
$$= \det(\lambda T I_{n\times n} T^{-1} - TAT^{-1}) = \det(\lambda I_{n\times n} - B) = p_B(\lambda)$$

□

Corollary 3.2. *The eigenvector $x^{(j)}(B)$ of the matrix $B = T^{-1}AT$, corresponding to the eigenvalue λ_i, is as follows*

$$x^{(j)}(B) = T^{-1}x^{(j)}(A)$$

where $x^{(j)}(A)$ is the eigenvector of the matrix A.

Proof. We have

$$Ax^{(j)}(A) = \lambda_j x^{(j)}(A)$$
$$(T^{-1}AT)T^{-1}x^{(j)}(A) = Bx^{(j)}(B) = \lambda_j T^{-1}x^{(j)}(A) = \lambda_j x^{(j)}(B)$$

□

Proposition 3.9. *For any $A, B \in \mathbb{C}^{n\times n}$, the matrices AB and BA have the same characteristic polynomial and hence the same eigenvalues, that is,*

$$\boxed{\begin{aligned} p_{AB}(\lambda) &= p_{BA}(\lambda) \\ \sigma(AB) &= \sigma(BA) \end{aligned}} \tag{3.18}$$

Proof. Let us select $\mu \in \mathbb{C}$ such that $A - \mu I_{n\times n}$ is nonsingular ($[A - \mu I_{n\times n}]^{-1}$ exists). Then we have

$$\begin{aligned} &\det\left(\lambda I_{n\times n} - [A - \mu I_{n\times n}]\, B\right) \\ &\quad = \det\left([A - \mu I_{n\times n}]\right) \det\left(\lambda [A - \mu I_{n\times n}]^{-1} - B\right) \\ &\quad = \det\left(\lambda [A - \mu I_{n\times n}]^{-1} - B\right) \det\left([A - \mu I_{n\times n}]\right) \\ &\quad = \det\left(\lambda I_{n\times n} - B\,[A - \mu I_{n\times n}]\right) \end{aligned}$$

which for $\tilde{\lambda} := \lambda + \mu$ implies

$$p_{AB}\left(\tilde{\lambda}\right) = \det\left(\tilde{\lambda} I_{n\times n} - AB\right) = \det\left(\tilde{\lambda} I_{n\times n} - BA\right) = p_{BA}\left(\tilde{\lambda}\right)$$

□

Proposition 3.10. *Let us show that the $n \times n$ companion matrix*

$$C_a := \begin{bmatrix} 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & 1 & 0 & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & & 0 & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & 0 & 1 \\ -a_n & -a_{n-2} & \cdot & \cdot & \cdot & -a_1 \end{bmatrix}$$

has the characteristic polynomial $p_{C_a}(\lambda)$ equal to

$$\boxed{p_{C_a}(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n}$$

Proof. Indeed,

$$
\begin{aligned}
p_{C_a}(\lambda)O = \det &\begin{bmatrix} \lambda & -1 & 0 & \cdot & \cdot & 0 \\ 0 & \lambda & -1 & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & -1 \\ a_n & a_{n-1} & \cdot & \cdot & \cdot & \lambda + a_1 \end{bmatrix} \\
= (\lambda + a_1) \det &\begin{bmatrix} \lambda & -1 & 0 & \cdot & \cdot \\ 0 & \lambda & -1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & -1 \\ 0 & 0 & \cdot & \cdot & \lambda \end{bmatrix} \\
- a_2 \det &\begin{bmatrix} \lambda & -1 & 0 & \cdot & 0 \\ 0 & \lambda & -1 & 0 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \lambda & 0 \\ 0 & 0 & \cdot & \cdot & -1 \end{bmatrix} \\
+ \cdots - a_{n-1} \det &\begin{bmatrix} \lambda & 0 & \cdot & \cdot & 0 \\ 0 & -1 & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & 0 \\ 0 & \cdot & \cdot & \cdot & -1 \end{bmatrix} \\
+ a_n \det &\begin{bmatrix} -1 & 0 & \cdot & \cdot & 0 \\ \lambda & -1 & 0 & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & & \cdot & \cdot & 0 \\ 0 & \cdot & \cdot & \lambda & -1 \end{bmatrix} \\
= (\lambda + a_1)\lambda^{n-1} &+ a_2\lambda^{n-2} + \cdots + a_{n-1}\lambda + a_n
\end{aligned}
$$

□

Proposition 3.11. *For any square $n \times n$ matrix A*

$$
\boxed{\begin{aligned} a_n &= \det A = \prod_{i=1}^{n} \lambda_i \\ a_{n-1} &= \operatorname{tr} A = \sum_{i=1}^{n} \lambda_i \end{aligned}} \tag{3.19}
$$

Proof. The first formula follows directly from (3.15) if $\lambda = 0$. To prove that $a_{n-1} = \sum_{i=1}^{n} \lambda_i$ it is sufficient to open parentheses in

$$
p_A(\lambda) = \prod_{i=1}^{n} (\lambda - \lambda_i) = \lambda^n + \sum_{i=1}^{n} a_i \lambda^{n-i} \tag{3.20}
$$

and to calculate the coefficient corresponding λ in the right-hand side. The second identity $\operatorname{tr} A = \sum_{i=1}^{n} \lambda_i$ may be easily proven using the so-called diagonal form transformation which can be done below. □

Corollary 3.3.

$$\boxed{\operatorname{tr} A = \frac{d}{d\lambda} \det(\lambda I_{n\times n} - A)\,|_{\lambda=0}} \tag{3.21}$$

Proof. It follows directly from (3.20). □

Proposition 3.12. *The **spectrum of a unitary matrix** U lies on the unit circle.*

Proof. If $Ux = \lambda x$ and $x^*x = 1$, then

$$(Ux)^*Ux = x^*U^*Ux = x^*x = 1$$

On the other hand

$$(Ux)^*Ux = (\lambda x)^*(\lambda x) = x^*\lambda^*\lambda x = |\lambda|^2 x^*x = |\lambda|^2$$

The comparison yields $|\lambda|^2 = 1$. □

Proposition 3.13. *If x is a complex nonzero vector in $\mathbb{C}^n$, then the **Householder matrix** defined as*

$$H = I_{n\times n} - 2\frac{xx^*}{x^*x} \tag{3.22}$$

is unitary.

Proof. We should show that

$$HH^* = H^*H = I_{n\times n}$$

One has

$$HH^* = H^*H = \left(I_{n\times n} - 2\frac{xx^*}{x^*x}\right)\left(I_{n\times n} - 2\frac{xx^*}{x^*x}\right)$$

$$= I_{n\times n} - 4\frac{xx^*}{x^*x} + 4\frac{x(x^*x)x^*}{(x^*x)^2} = I_{n\times n}$$

□

Proposition 3.14. *For any square $n \times n$ matrix A with the eigenvectors $x^{(i)}(A)$ corresponding to the eigenvalue $\lambda_i(A)$ $(i = 1, \ldots n)$ it follows*

1.

$$\boxed{\begin{array}{c} x^{(i)}(A) = x^{(i)}(I_{n\times n} - A) \\ \lambda_i(I_{n\times n} - A) = 1 - \lambda_i(A) \end{array}}$$

Indeed,

$$\begin{aligned}(I_{n\times n} - A)x^{(i)}(A) &= x^{(i)}(A) - Ax^{(i)}(A) \\ &= x^{(i)}(A) - \lambda_i(A)x^{(i)}(A) \\ &= [1 - \lambda_i(A)]x^{(i)}(A)\end{aligned}$$

2.

$$\boxed{\begin{array}{c} x^{(i)}(A^p) = x^{(i)}(A), \quad p = 2, 3, \ldots \\ \lambda_i(A^p) = \lambda_i^p(A) \end{array}}$$

since

$$\begin{aligned}A^p x^{(i)}(A) &= A^{(p-1)}Ax^{(i)}(A) = \lambda_i(A)A^{(p-1)}x^{(i)}(A) \\ &= \lambda_i^2(A)A^{(p-2)}x^{(i)}(A) = \cdots = \lambda_i^p(A)x^{(i)}(A)\end{aligned}$$

*3. If, in addition, A is real orthogonal $(A^\top = A^{-1})$ and $(I_{n\times n} + A)$ is nonsingular $(\lambda_i(I_{n\times n} + A) \neq 0$ for all $i = 1, \ldots, n)$, then A can be represented as (**Cayley transformation**)*

$$\boxed{A = (I_{n\times n} - S)(I_{n\times n} + S)^{-1}} \tag{3.23}$$

where S is a real skew-matrix (2.6), i.e. $S^\top = -S$. This result follows directly based on the construction of S: if (3.23) holds then

$$\begin{aligned}&A(I_{n\times n} + S) = (I_{n\times n} - S) \\ &AS + S = I_{n\times n} - A \\ &S = (I_{n\times n} - A)(I_{n\times n} + A)^{-1}\end{aligned}$$

and

$$\begin{aligned}S^\top &= \left[(I_{n\times n} + A)^{-1}\right]^\top (I_{n\times n} - A)^\top \\ &= (I_{n\times n} + A^\top)^{-1}(I_{n\times n} - A^\top) \\ &= (I_{n\times n} + A^{-1})^{-1}(I_{n\times n} - A^{-1}) \\ &= \left[A^{-1}(A + I_{n\times n})\right]^{-1}\left[A^{-1}(A - I_{n\times n})\right] \\ &= \left[(A + I_{n\times n})\right]^{-1}\left[AA^{-1}\right](A - I_{n\times n}) = -S\end{aligned}$$

which means that S is a skew matrix.

3.3 The Cayley–Hamilton theorem

The theorem discussed in this subsection plays a key role in matrix theory and has important applications to Classical Control Theory.

Theorem 3.1. (Cayley–Hamilton theorem) *If*

$$\boxed{p_A(\lambda) = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0} \tag{3.24}$$

is the characteristic polynomial of a squared matrix A, then

$$\boxed{p_A(A) := A^n + a_1A^{n-1} + \cdots + a_{n-1}A + a_nI_{n\times n} = 0} \tag{3.25}$$

that is, the matrix A satisfies its characteristic equation.

Proof. By (2.5)

$$A^{-1} = \frac{1}{\det A}\operatorname{adj} A$$

we have

$$A \operatorname{adj} A = (\det A)I_{n\times n}$$

which leads to the following identity

$$\begin{aligned}&(\lambda I_{n\times n} - A)\operatorname{adj}(\lambda I_{n\times n} - A)\\ &\quad = (\det(\lambda I_{n\times n} - A))I_{n\times n} = p_A(A)I_{n\times n}\end{aligned} \tag{3.26}$$

It is clear that $\operatorname{adj}(\lambda I_{n\times n} - A)$ is matrix λ^{n-1} as the maximal order, i.e.,

$$\operatorname{adj}(\lambda I_{n\times n} - A) = B_{n-1}\lambda^{n-1} + B_{n-2}\lambda^{n-2} + \cdots + B_1\lambda + B_0$$

Then (3.26) becomes

$$\begin{aligned}&(\lambda I_{n\times n} - A)(B_{n-1}\lambda^{n-1} + B_{n-2}\lambda^{n-2} + \cdots + B_1\lambda + B_0)\\ &\quad = (\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n)I_{n\times n}\end{aligned}$$

Comparing coefficients, we obtain

$$\begin{aligned}B_{n-1} &= I_{n\times n}\\ B_{n-2} - AB_{n-1} &= a_1I_{n\times n}\\ &\vdots\\ B_0 - AB_1 &= a_{n-1}I_{n\times n}\\ -AB_0 &= a_nI_{n\times n}\end{aligned}$$

Multiplying the first of these equalities by A^n, the second one by A^{n-1}, the jth by A^{n-j+1} and adding them together yields

$$0 = A^n + a_1 A^{n-1} + \cdots + a_{n-1} A + a_n I_{n\times n}$$

which is exactly (3.25). □

Corollary 3.4. *If A^{-1} exists (A is nonsingular), then*

$$\boxed{A^{-1} = -\frac{1}{a_n}(A^{n-1} + a_1 A^{n-2} + \cdots + a_{n-2} A + a_{n-1} I_{n\times n})}$$

Proof. Since A is nonsingular (A has no zero eigenvalues) we have that $a_n \neq 0$. Then the result follows from the identity

$$A^{-1} p_A(A) = A^{-1}(A^n + a_1 A^{n-1} + \cdots + a_{n-1} A + a_n I_{n\times n}) = 0$$

□

3.4 The multiplicities and generalized eigenvectors

3.4.1 Algebraic and geometric multiplicities

For any $n \times n$ matrix A it may happen that some of its eigenvalues are equal, that is, the corresponding characteristic polynomial $p_A(\lambda)$ may have the following structure

$$\boxed{p_A(\lambda) = \prod_{i=1}^{K} (\lambda - \lambda_i)^{\mu_i}, \quad \sum_{i=1}^{K} \mu_i = n} \tag{3.27}$$

where μ_i is the number of times the factor $(\lambda - \lambda_i)$ appears in (3.27).

Definition 3.7.

*(a) the number μ_i is called the **algebraic multiplicity** of the eigenvalue λ_i of the matrix A;*

(b) the number

$$\boxed{\varkappa_i := \dim \operatorname{Ker}(\lambda_i I_{n\times n} - A)} \tag{3.28}$$

*is called the **geometric multiplicity** of the eigenvalue λ_i of the matrix A.*

Example 3.3. *For*

$$A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$$

we have

$$p_A(\lambda) = \det \begin{bmatrix} \lambda & -1 \\ 0 & \lambda \end{bmatrix} = \lambda^2$$

$$\lambda_1 = \lambda_2 = 0$$

which implies

$$\mu_1 = 2$$

and

$$\varkappa_1 = \dim \operatorname{Ker}((\lambda_1 I_{n\times n} - A)) = \dim \operatorname{Ker}(A) = 1$$

Here we see that in this example $\varkappa_1 < \mu_1$.

In general, the following property holds.

Lemma 3.1. *The geometric multiplicity of an eigenvalue does not exceed its algebraic multiplicity, that is, for any* $i = 1, \ldots, K$

$$\boxed{\varkappa_i \leq \mu_i} \tag{3.29}$$

Proof. If r is the rank of $(\lambda_i I_{n\times n} - A)$, then by (3.9) $r = n - \mu_i$ and all minors of $(\lambda_i I_{n\times n} - A)$ greater than $(n - \mu_i)$ are equal to zero. Hence, in

$$p_A(\lambda) = \prod_{i=1}^{n}(\lambda - \lambda_i) = \lambda^n + \sum_{i=1}^{n} a_{n-i}\lambda^i$$

for all $i \leq \mu_i - 1$ we have $a_i = 0$, which leads to the following

$$p_A(\lambda) = \prod_{i=\mu_i}^{n}(\lambda - \lambda_i)$$

This means that the last polynomial has a zero of multiplicity greater than or equal to μ_i which implies the result. □

Corollary 3.5.

$$\boxed{\sum_{i=1}^{K} \varkappa_i \leq \sum_{i=1}^{K} \mu_i = n} \tag{3.30}$$

3.4.2 Generalized eigenvectors

Definition 3.8. *Let an eigenvalue λ_i of a square $n \times n$ matrix A have the algebraic multiplicity μ_i. Then the vectors $x^{(i,k)}(A)$ satisfying the equations*

$$\boxed{[A - \lambda_i I_{n\times n}]\, x^{(i,k)}(A) = x^{(i,k-1)}(A), \quad k = 2, 3, \ldots, r \leq \mu_i} \tag{3.31}$$

are called the ***generalized eigenvectors*** *of A. Evidently,*

$$x^{(i,1)}(A) = x^{(i)}(A)$$

is the corresponding eigenvector of A (by the definition $x^{(i,0)}(A) \equiv 0$). The sequence of vectors $x^{(i,1)}(A)$, $x^{(i,2)}(A)$, $\ldots$, $x^{(i,r)}(A)$ is called a ***Jordan chain*** *of length $r \leq \mu_i$.*

If the eigenvector $x^{(i,1)}(A) = x^{(i)}(A)$ is selected, then the next vectors $x^{(i,2)}(A)$, $\ldots$, $x^{(i,r)}(A)$ are generated successively as far as the nonhomogeneous equation (3.31) has a solution.

Proposition 3.15. *$x^{(i,k)}(A)$ $(k \geq 2)$ is a generalized eigenvector of A if and only if the vector $[A - \lambda_i I_{n\times n}]^{k-1} x^{(i,k)}(A)$ is the eigenvector of A, or equivalently, if and only if*

$$[A - \lambda_i I_{n\times n}]^k x^{(i,k)}(A) = 0$$

Proof.

(a) *Necessity*. Let us prove this fact by induction. For $k = 2$ we have

$$[A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A) = x^{(i,1)}$$

and, hence, pre-multiplying by A implies

$$A\left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right) = Ax^{(i,1)}$$
$$= \lambda_i x^{(i,1)} = \lambda_i \left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right)$$

This means that $\left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right)$ is the eigenvector of A. Notice that the last identity may be rewritten as

$$0 = A\left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right) - \lambda_i \left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right)$$
$$= [A - \lambda_i I_{n\times n}]\left([A - \lambda_i I_{n\times n}]\, x^{(i,2)}(A)\right) = [A - \lambda_i I_{n\times n}]^2 x^{(i,2)}(A)$$

So, for $k = 2$ the proposition is true. Suppose that it is valid for some k. Show that it will be valid for $k + 1$ too. By this supposition we have

$$[A - \lambda_i I_{n\times n}]^k x^{(i,k)}(A) = 0$$

Then

$$[A - \lambda_i I_{n\times n}]\, x^{(i,k+1)}(A) = x^{(i,k)}$$

and pre-multiplying by $[A - \lambda_i I_{n\times n}]^k$ implies

$$[A - \lambda_i I_{n\times n}]^k \left([A - \lambda_i I_{n\times n}]\, x^{(i,k+1)}(A)\right) = [A - \lambda_i I_{n\times n}]^k\, x^{(i,k)} = 0$$

or, equivalently,

$$[A - \lambda_i I_{n\times n}]^{k+1}\, x^{(i,k+1)}(A) = 0$$

(b) *Sufficiency*. It follows directly from the definition of a generalized vector. □

Proposition 3.16. *Any Jordan chain consists of* ***linearly independent*** *elements.*

Proof. Suppose that there exist $\alpha_s (s = 1, \ldots, r)$ such that

$$\sum_{s=1}^{r} \alpha_i x^{(i,s)}(A) = 0$$

Applying the transformation $[A - \lambda_i I_{n\times n}]^{r-1}$ to both sides, we get

$$0 = \sum_{s=1}^{r} \alpha_i\, [A - \lambda_i I_{n\times n}]^{r-1-s}\, [A - \lambda_i I_{n\times n}]^s\, x^{(i,s)}(A) = \alpha_r$$

Then applying again the transformation $[A - \lambda_i I_{n\times n}]^{r-2}$, in view of the result before, we obtain that $\alpha_{r-1} = 0$. Repeating this procedure, we obtain the contradiction, and hence, the result is established. □

4 Matrix Transformations

Contents

4.1 Spectral theorem for Hermitian matrices

4.1.1 Eigenvectors of a multiple eigenvalue for Hermitian matrices

Proposition 4.1. *Let $x^{(i,1)}, x^{(i,2)}, \ldots, x^{(i,\mu_i)}$ be the eigenvectors of an $n \times n$ Hermitian matrix $A \in \mathbb{C}^{n\times n}$ corresponding to the eigenvalue λ_i (which by (3.5) is always real) with the algebraic multiplicity μ_i. Then* ***these vectors may be supposed to be linearly independent.***

Proof. Indeed, if $\operatorname{rank}(\lambda_i I_{n\times n} - A) = n - \mu_i$, then, selecting the last components $z^{(i,2)} := \left[x^{(i)}_{\mu_i+1}, x^{(i)}_{\mu_i+2}, \ldots, x^{(i)}_n\right]^{\mathsf{T}}$ as free variables and solving the linear systems

$$(\lambda_i I_{n\times n} - A)\, x^{(i)} = 0$$

with respect to $z^{(i,1)} := \left(x^{(i)}_1, x^{(i)}_2, \ldots, x^{(i)}_{n-\mu_i}\right)^{\mathsf{T}}$, we get

$$x^{(i)} := \begin{pmatrix} z^{(i,1)} \\ z^{(i,2)} \end{pmatrix}, \quad A := \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad \operatorname{rank} A_{11} = n - \mu_i$$

$$\begin{bmatrix} \left(\lambda_i I_{(n-\mu_i)\times(n-\mu_i)} - A_{11}\right) & -A_{12} \\ -A_{21} & \left(\lambda_i I_{\mu_i\times\mu_i} - A_{22}\right) \end{bmatrix} \begin{pmatrix} z^{(i,1)} \\ z^{(i,2)} \end{pmatrix} = 0$$

and, hence,

$$\left(\lambda_i I_{(n-\mu_i)\times(n-\mu_i)} - A_{11}\right) z^{(i,1)} = A_{12} z^{(i,2)}$$

or, equivalently,

$$z^{(i,1)} = \left[\left(\lambda_i I_{(n-\mu_i)\times(n-\mu_i)} - A_{11}\right)^{-1} A_{12}\right] z^{(i,2)} \tag{4.1}$$

Taking then the free components as $x_s^{(i,r)} = \delta_{s,r}$ $(s, r = \mu_i + 1, \ldots, n)$, we may define the following linearly independent (in $\mathbb{C}^{n-\mu_i}$) vectors

$$z^{(i,2,r)} = \begin{pmatrix} x_{\mu_i+1}^{(i,r)} \\ x_{\mu_i+2}^{(i,r)} \\ \vdots \\ x_n^{(i,r)} \end{pmatrix}, \quad r = 1, \ldots, n - \mu_i$$

Evidently, in spite of the fact that $z^{(i,1,r)}$ are linearly dependent on $z^{(i,2,r)}$ by (4.1), the joint vectors $\begin{pmatrix} z^{(i,1,r)} \\ z^{(i,2,r)} \end{pmatrix} = x^{(i,r)}$ remain linearly independent. □

4.1.2 Gram–Schmidt orthogonalization

Proposition 4.2. *Eigenvectors of a Hermitian matrix are linearly independent; some of them even correspond to the same eigenvalues.*

Proof. This result immediately follows from the previous proposition and from (3.17). □

Let $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ be the set of linearly independent eigenvectors of an $n \times n$ Hermitian matrix $A \in \mathbb{C}^{n\times n}$.

Lemma 4.1. (Gram–Schmidt orthogonalization process) *The set* $\{\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(n)}\}$ *of vectors obtained from* $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$ *by the procedure*

$$\boxed{\begin{aligned} \tilde{x}^{(1)} &= x^{(1)} \\ \tilde{x}^{(r)} &= \left[I_{n\times n} - \sum_{s=1}^{r-1} \frac{\tilde{x}^{(s)} \tilde{x}^{(s)\mathsf{T}}}{\left\| \tilde{x}^{(s)} \right\|^2} \right] x^{(r)}, \quad (r = 2, \ldots, n) \end{aligned}} \tag{4.2}$$

is ***orthogonal****, that is,*

$$\left(\tilde{x}^{(r)}, \tilde{x}^{(s)}\right) = 0 \quad (r \neq s = 1, \ldots, n) \tag{4.3}$$

Proof. Let us do it by induction. For $r = 2$ we have

$$\tilde{x}^{(2)} = \left[I_{n\times n} - \frac{\tilde{x}^{(1)} \tilde{x}^{(1)\mathsf{T}}}{\left\| \tilde{x}^{(1)} \right\|^2} \right] x^{(2)} = x^{(2)} - \frac{\tilde{x}^{(1)}}{\left\| \tilde{x}^{(1)} \right\|^2} \left(\tilde{x}^{(1)}, x^{(2)}\right)$$

and, as the result, it follows that

$$\left(\tilde{x}^{(1)}, \tilde{x}^{(2)}\right) = \left(\tilde{x}^{(1)}, x^{(2)}\right) - \left(\tilde{x}^{(1)}, x^{(2)}\right) = 0$$

Supposing that the vectors $\tilde{x}^{(1)}, \tilde{x}^{(2)}, \ldots, \tilde{x}^{(r-1)}$ are orthogonal, we get

$$\begin{aligned}\left(\tilde{x}^{(r-1)}, \tilde{x}^{(r)}\right) &= \left(\tilde{x}^{(r-1)}, x^{(r)}\right) - \sum_{s=1}^{r-1} \frac{\left(\tilde{x}^{(r-1)}, \tilde{x}^{(s)}\right)}{\left\|\tilde{x}^{(s)}\right\|^2} \left(\tilde{x}^{(s)}, x^{(r)}\right) \\ &= \left(\tilde{x}^{(r-1)}, x^{(r)}\right) - \frac{\left(\tilde{x}^{(r-1)}, \tilde{x}^{(r-1)}\right)}{\left\|\tilde{x}^{(r-1)}\right\|^2} \left(\tilde{x}^{(r-1)}, x^{(r)}\right) = 0\end{aligned}$$

□

Remark 4.1. *The set of the vectors* $\left\{ \frac{\tilde{x}^{(1)}}{\left\|\tilde{x}^{(1)}\right\|}, \ldots, \frac{\tilde{x}^{(n)}}{\left\|\tilde{x}^{(n)}\right\|} \right\}$ *may be considered as an orthonormal basis in* $\mathbb{C}^n$.

4.1.3 Spectral theorem

Theorem 4.1. (Spectral theorem) *If* $\left\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\right\}$ *is the set of linearly independent eigenvectors of an* $n \times n$ *Hermitian matrix* $A \in \mathbb{C}^{n \times n}$ *corresponding to the eigenvalues* $\lambda_1, \ldots, \lambda_n$ *(maybe multiple), then the following representation holds*

$$\boxed{A = X \Lambda X^{-1} = X \Lambda X^* = \sum_{i=1}^{n} \lambda_i x^{(i)} x^{(i)*}} \tag{4.4}$$

where $\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$ *and* $X := \left[x^{(1)}\, x^{(2)} \cdots x^{(n)}\right]$ *is unitary matrix, i.e.,* $X^* = X^{-1}$.

Proof. Notice that the relations

$$Ax^{(i)} = \lambda_i x^{(i)}, \quad i = 1, \ldots, n$$

may be rewritten as

$$AX = X\Lambda \tag{4.5}$$

Since $x^{(1)}, x^{(2)}, \ldots, x^{(n)}$ are linearly independent it follows that X^{-1} exists and, hence,

$$\begin{aligned} A &= X \Lambda X^{-1} \\ A^* = A &= \left(X^{-1}\right)^* \Lambda X^* \end{aligned}$$

and

$$A\left(X^{-1}\right)^* = \left(X^{-1}\right)^* \Lambda \tag{4.6}$$

The comparison of (4.5) and (4.6) implies

$$X = \left(X^{-1}\right)^*, \quad X^* = X^{-1}$$

and, hence, by (2.3) it follows that

$$A = X\Lambda X^{-1} = X\Lambda X^* = X \begin{bmatrix} \lambda_1 x^{(1)*} \\ \vdots \\ \lambda_n x^{(n)*} \end{bmatrix} = \sum_{i=1}^{n} \lambda_i x^{(i)} x^{(i)*}$$

□

4.2 Matrix transformation to the Jordan form

This subsection deals with the transformation of nonobligatory Hermitian matrices to triplet form analogous to one given before.

4.2.1 The Jordan block

Definition 4.1. *The matrix*

$$J_i := \begin{bmatrix} \lambda_i & 1 & 0 & \cdot & 0 \\ 0 & \lambda_i & 1 & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & 0 \\ \cdot & \cdot & \cdot & \lambda_i & 1 \\ 0 & \cdot & \cdot & 0 & \lambda_i \end{bmatrix} \tag{4.7}$$

*is referred to as a **Jordan block** (or **cell**) of order μ_i corresponding to the eigenvalue λ_i.*

4.2.2 The Jordan matrix form

Theorem 4.2. (The Jordan normal canonical representation) *For any square complex matrix $A \in \mathbb{C}^{n\times n}$ there exists a nonsingular matrix T such that*

$$A = TJT^{-1} \tag{4.8}$$

where

$$J = \operatorname{diag}(J_1, J_2, \ldots, J_K)$$
$$J_i \in \mathbb{C}^{\mu_i \times \mu_i}$$

with

$$\sum_{i=1}^{K} \mu_i = n$$

and with λ_i $(i = 1, \ldots, K)$ as the distinct eigenvalues of A with the multiplicity μ_i. The transformation T has the following form

$$T = \left[T_1\ T_2 \cdots T_K\right]$$
$$T_i = \left[x^{(i,1)}\ x^{(i,2)} \cdots x^{(i,\mu_i)}\right] \tag{4.9}$$

where $x^{(i,1)}$ *are the eigenvectors of* A *corresponding to the eigenvalue* λ_i *and* $x^{(i,s)}$ $(s = 2, \ldots, \mu_i)$ *are the generalized vectors of* A *generated by (3.31).*

Proof. Taking sufficiently large space, nevertheless it may be realized by the direct verification of the identity $AT = TJ$. □

Example 4.1.

$$A = \begin{bmatrix} 6 & 2 & 2 \\ -2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

We have

$$\det(A - \lambda I_{3\times 3}) = (2 - \lambda)(4 - \lambda)^2$$

So,

$$\lambda_1 = 2, \lambda_2 = 4 \quad with \quad \mu_2 = 2$$

and

$$x^{(1)} = \begin{bmatrix} 0 & -1 & 1 \end{bmatrix}^{\mathsf{T}},\ x^{(2,1)} = \begin{bmatrix} 2 & -2 & 0 \end{bmatrix}^{\mathsf{T}}$$
$$(A - \lambda I_{3\times 3})\, x^{(2,2)} = x^{(2,1)} \Longrightarrow x^{(2,2)} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}^{\mathsf{T}}$$

As the result we have

$$A = T \begin{bmatrix} 2 & 0 & 0 \\ 0 & \mathbf{4} & \mathbf{1} \\ 0 & \mathbf{0} & \mathbf{4} \end{bmatrix} T^{-1}$$
$$T = \left[x^{(1)}\ \ x^{(2,1)}\ \ x^{(2,2)}\right] = \begin{bmatrix} 0 & 2 & 1 \\ -1 & -2 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

4.3 Polar and singular-value decompositions

4.3.1 Polar decomposition

Proposition 4.3. (Polar factorization) *For any square complex matrix* $A \in \mathbb{C}^{n\times n}$ *there exist unique positive semidefinite* ($\lambda_i \geq 0$ *for all* $i = 1, \ldots, n$) *Hermitian matrices* H, K *and unitary matrices* U, V *all in* $\mathbb{C}^{n\times n}$ *such that*

$$\boxed{A = UH = KV} \tag{4.10}$$

Proof. First, notice the matrix A^*A is Hermitian and, hence, by (3.5) it has all eigenvalues real. Moreover, all of them are nonnegative, since by

$$x^{(i)*} : A^*Ax^{(i)} = \lambda_i x^{(i)}$$

one has

$$x^{(i)*}A^*Ax^{(i)} = \left\|Ax^{(i)}\right\|^2 = \lambda_i x^{(i)*}x^{(i)} = \lambda_i \left\|x^{(i)}\right\|^2$$

and, hence, for all $i = 1, \ldots, n$

$$\lambda_i = \frac{\left\|Ax^{(i)}\right\|^2}{\left\|x^{(i)}\right\|^2} \geq 0 \tag{4.11}$$

Define $r_1^2 := \lambda_1, r_2^2 := \lambda_2, \ldots, r_n^2 := \lambda_n$ such that

$$r_i > 0 \text{ for } i = 1, \ldots, k$$

and

$$r_i = 0 \text{ for } i = k+1, \ldots, n$$

Then, for $i, j = 1, \ldots, k$ and for the corresponding orthonormal eigenvectors $x^{(i)}$, $x^{(j)}$ $\left(\left(x^{(i)}, x^{(j)}\right) = \delta_{ij}\right)$, we have

$$\left(\frac{Ax^{(i)}}{r_i}, \frac{Ax^{(j)}}{r_j}\right) = \left(\frac{A^*Ax^{(i)}, x^{(j)}}{r_i r_j}\right) = \delta_{ij}\frac{r_i^2}{r_i r_j}$$

and thus the vectors

$$z^{(i)} := \frac{Ax^{(i)}}{r_i}, \quad i = 1, \ldots, k \tag{4.12}$$

are orthonormal. Define also two unitary matrices

$$\begin{aligned} X &:= \left[x^{(1)} \ldots x^{(k)} \ldots x^{(n)}\right] \\ Z &:= \left[z^{(1)} \ldots z^{(k)}\right] \end{aligned} \tag{4.13}$$

Then by (4.12) we have

$$Ax^{(i)} = z^{(i)}r_i$$

or, with $R := \text{diag}\{r_1, \ldots, r_n\}$,

$$AX = ZR$$

which, by post-multiplying by X^*, implies

$$AXX^* = A = ZRX^* \tag{4.14}$$

Now, let

$$\boxed{U = ZX^*, \quad H = XRX^*} \tag{4.15}$$

Clearly, U is unitary since X and Z are unitary. H evidently is Hermitian and has all eigenvalues nonnegative, or in other words, it is positive semidefinite. Moreover, by (4.14)

$$UH = ZX^*XRX^* = ZRX^* = A$$

Applying the above result to A^* we obtain $A = KV$. □

Corollary 4.1. *If $A \in \mathbb{C}^{n\times n}$ is nonsingular (or, equivalently all eigenvalues $\lambda_i = r_i^2$ of A^*A are strictly positive $(k = n)$), then*

$$H^\intercal = H, \quad K^\intercal = K, \quad U^\intercal = U^{-1}, \quad V^\intercal = V^{-1}$$

and

$$A^\intercal A = H^2, \quad AA^\intercal = V^2$$

Example 4.2. *For*

$$A = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}$$

it follows that

$$A^\intercal A = \begin{bmatrix} 5 & -2 \\ -2 & 1 \end{bmatrix}$$

$$\lambda_1 (A^\intercal A) = 3 - 2\sqrt{2}, \ \tilde{x}^{(1)} = \begin{bmatrix} \sqrt{2} - 1 \\ 1 \end{bmatrix}$$

$$\lambda_2 (A^\intercal A) = 3 + 2\sqrt{2}, \ \tilde{x}^{(2)} = \begin{bmatrix} -\sqrt{2} - 1 \\ 1 \end{bmatrix}$$

Notice that $(\tilde{x}^{(1)}, \tilde{x}^{(2)}) = 0$. The normalized eigenvectors are

$$x^{(1)} = \frac{1}{\sqrt{2\left(2 - \sqrt{2}\right)}} \begin{bmatrix} \sqrt{2} - 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 0.38268 \\ 0.92388 \end{bmatrix}$$

$$x^{(2)} = \frac{1}{\sqrt{2\left(2 + \sqrt{2}\right)}} \begin{bmatrix} -\sqrt{2} - 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -0.92388 \\ 0.38268 \end{bmatrix}$$

According to (4.13), we construct

$$X = \begin{bmatrix} 0.38268 & -0.92388 \\ 0.92388 & 0.38268 \end{bmatrix}$$

$$Z = \begin{bmatrix} 0.92387 & -0.38268 \\ -0.3827 & -0.92388 \end{bmatrix}$$

$$R = \begin{bmatrix} 0.41421 & 0.0 \\ 0.0 & 2.4142 \end{bmatrix}$$

Then by (4.15) we, finally, obtain

$$U = ZX^{\intercal} = \begin{bmatrix} 0.70710 & 0.7071 \\ 0.7071 & -0.70712 \end{bmatrix}$$

$$H = XRX^{\intercal} = \begin{bmatrix} 2.1213 & -0.70710 \\ -0.70710 & 0.70710 \end{bmatrix}$$

To check the calculation just made above, we compute

$$A = UH = \begin{bmatrix} 0.70710 & 0.7071 \\ 0.7071 & -0.70712 \end{bmatrix} \times \begin{bmatrix} 2.1213 & -0.70710 \\ -0.70710 & 0.70710 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}$$

4.3.2 Singular-value decomposition

Let us consider a matrix $A \in \mathbb{C}^{m\times n}$. Evidently, all roots of the matrices $A^*A \in \mathbb{C}^{n\times n}$ and $AA^* \in \mathbb{C}^{m\times m}$ are real and nonnegative. Indeed, if $\lambda_i(A^*A)$, $\lambda_i(AA^*)$ are some eigenvalues and $x^{(i)}(A^*A)$ and $x^{(i)}(AA^*)$ are the corresponding eigenvectors, then

$$x^{(i)*}(A^*A) : A^*Ax^{(i)}(A^*A) = \lambda_i(A^*A)x^{(i)}(A^*A)$$
$$x^{(i)*}(AA^*) : AA^*x^{(i)}(AA^*) = \lambda_i(AA^*)x^{(i)}(AA^*)$$

and, thus,

$$x^{(i)*}(A^*A)\,A^*Ax^{(i)}(A^*A) = \left\|Ax^{(i)}(A^*A)\right\|^2 = \lambda_i(A^*A)x^{(i)*}(A^*A)\,x^{(i)}(A^*A) = \lambda_i(A^*A)\left\|x^{(i)}(A^*A)\right\|^2$$

$$x^{(i)*}(AA^*)\,AA^*x^{(i)}(AA^*) = \left\|A^*x^{(i)}(AA^*)\right\|^2 = \lambda_i(AA^*)x^{(i)*}(AA^*)\,x^{(i)}(AA^*) = \lambda_i(AA^*)\left\|x^{(i)}(AA^*)\right\|^2$$

or, equivalently,

$$\boxed{\begin{aligned} \lambda_i(A^*A) &= \frac{\left\|Ax^{(i)}(A^*A)\right\|^2}{\left\|x^{(i)}(A^*A)\right\|^2} \geq 0 \\ \lambda_i(AA^*) &= \frac{\left\|Ax^{(i)}(AA^*)\right\|^2}{\left\|x^{(i)}(AA^*)\right\|^2} \geq 0 \end{aligned}} \tag{4.16}$$

It also follows (if $m = n$ this is the result of (3.18)) that the spectrums of A^*A and AA^* coincide, that is,

$$\sigma(A^*A) = \sigma(AA^*)$$

and eigenvalues have the same algebraic multiplicity (the geometric multiplicity may be different because of zero eigenvalues if they exist). More exactly,

Proposition 4.4.

$$\boxed{\sigma_i(A^*A) = \sigma_i(AA^*)}$$

Proof. Indeed, if

$$[A^*A]\, x^{(i)}(A^*A) = \lambda_i(A^*A)\, x^{(i)}(A^*A)$$

then

$$AA^*\left[Ax^{(i)}(A^*A)\right] = \lambda_i(A^*A)\left[Ax^{(i)}(A^*A)\right]$$

Thus, $Ax^{(i)}(A^*A)$ is the eigenvector of AA^* which corresponds to the same eigenvalue $\lambda_i(A^*A)$. □

In view of this property we may introduce the following definition.

Definition 4.2. *The number*

$$\boxed{\sigma_i(A) := \sqrt{\lambda_i(A^*A)} = \sqrt{\lambda_i(AA^*)}} \tag{4.17}$$

is called the ith ***singular value*** *of* $A \in \mathbb{C}^{n\times n}$.

Remark 4.2. *If a square matrix* $A \in \mathbb{C}^{n\times n}$ *is* ***normal****, that is, satisfies the relation*

$$AA^* = A^*A$$

then

$$\sigma_i(A) = \sqrt{\lambda_i(A^*A)} = |\lambda_i(A)|$$

Proposition 4.5. *The singular values of a squared matrix are invariant under unitary transformation, that is, if* $U \in \mathbb{C}^{n\times n}$ *satisfies* $U^*U = UU^* = I_{n\times n}$, *then for any* $A \in \mathbb{C}^{n\times n}$ *we have*

$$\boxed{\sigma_i(UA) = \sigma_i(AU) = \sigma_i(A)} \tag{4.18}$$

for all $i = 1, \ldots, n$.

Proof. Indeed,

$$\sigma_i(UA) = \sqrt{\lambda_i(A^*U^*UA)} = \sqrt{\lambda_i(A^*A)} = \sigma_i(A)$$

and

$$\begin{aligned}\sigma_i(AU) &= \sqrt{\lambda_i(U^*A^*AU)} = \sqrt{\lambda_i(U^*A^*AU)^*} \\ &= \sqrt{\lambda_i(AUU^*A^*)} = \sqrt{\lambda_i(AA^*)} = \sigma_i(A)\end{aligned}$$

□

The next theorem represents the main result of this subsection.

Theorem 4.3. (Singular-value decomposition) *Let $A \in \mathbb{C}^{m\times n}$ and $\sigma_i(A)$ $(i = 1, \ldots, r \leq \min\{m, n\})$ be the nonzero singular values of A. Then A can be represented in the triplet form*

$$\boxed{A = UDV^*} \tag{4.19}$$

where $U \in \mathbb{C}^{m\times m}$ and $V \in \mathbb{C}^{n\times n}$ are unitary (i.e. satisfy $U^*U = UU^* = I_{m\times m}$ and $V^*V = VV^* = I_{n\times n}$) and $D \in \mathbb{C}^{m\times n}$ has $\sigma_i(A)$ in the (i, i)th position $(i = 1, \ldots, r)$ and zero elsewhere.

Proof. Following (4.12) we have that

$$\begin{aligned}Ax^{(i)} &= \sigma_i(A)\, z^{(i)},\ i = 1, \ldots, r \\ Ax^{(i)} &= 0,\ i = r+1, \ldots, n\end{aligned} \tag{4.20}$$

where $x^{(i)}$ are orthogonal eigenvectors of A^*A and $z^{(i)}$ are orthogonal eigenvectors of AA^*. Constructing the matrices

$$\begin{aligned}V &= \left[x^{(1)}\, x^{(2)} \cdots x^{(n)}\right] \\ U &= \left[z^{(1)}\, z^{(2)} \cdots z^{(m)}\right]\end{aligned}$$

we may note that in view of (4.15) they are unitary by the construction. Then (4.20) implies

$$AV = \left[\sigma_1(A)\, z^{(1)}\, \sigma_2(A)\, z^{(2)} \cdots \sigma_r(A)\, z^{(r)}\, 0 \cdots 0\right] = UD$$

or, equivalently,

$$A = UDV^{-1} = UDV^*$$

The result is established. □

Example 4.3. *For*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

we have

$$A^*A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}, \quad AA^* = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

and

$$\sigma_1(A) = \sqrt{2}, \quad x^{(1)}(A^*A) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad z^{(1)}(AA^*) = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\sigma_2(A) = 1, \quad x^{(2)}(A^*A) = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad z^{(2)}(AA^*) = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

$$z^{(3)}(AA^*) = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

So,

$$A = \begin{bmatrix} 1/\sqrt{2} & 0 & -1/\sqrt{2} \\ 0 & 1 & 0 \\ 1/\sqrt{2} & 0 & 1/\sqrt{2} \end{bmatrix} \begin{bmatrix} \sqrt{2} & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Proposition 4.6. *Two matrices* $A, B \in \mathbb{C}^{m\times n}$ *are* ***unitary equivalent****, that is,*

$$A = UBV^* \tag{4.21}$$

where $U \in \mathbb{C}^{m\times m}$ *and* $V \in \mathbb{C}^{n\times n}$ *are unitary, if and only if they have the same singular values.*

Proof. Necessity. Assuming $A = UBV^*$ one has

$$A = U_A D_A V_A^* = UBV^*, \qquad B = U_B D_B V_B^*$$

$$U_A D_A V_A^* = U U_B D_B V_B^* V^*$$

$$D_A = \left(U_A^* U U_B\right) D_B \left(V_B^* V^* V_A\right) = \underbrace{\left(U_A^* U U_B\right)}_{\tilde{U}} D_B \underbrace{\left(V_A^* V V_B\right)^*}_{\tilde{V}}$$

$$D_A = \tilde{U} D_B \tilde{V}^*$$

Viewing the last relation as the singular-value decomposition for D_A and noting that D_B is uniquely defined, we conclude that $D_A = D_B$.

Sufficiency. If $D_A = D_B$, then

$$A = U_A D_A V_A^* = U_A D_B V_A^* = U_A \left(U_B^* U_B D_B V_B^* V_B\right) V_A^*$$

$$= U_A U_B^* \left(U_B D_B V_B^*\right) V_B V_A^* = \underbrace{\left(U_A U_B^*\right)}_{U} B \underbrace{\left(V_B V_A^*\right)}_{V^*} = UBV^*$$

□

4.4 Congruent matrices and the inertia of a matrix

4.4.1 Congruent matrices

Definition 4.3. *Two square matrices $A, B \in \mathbb{C}^{n\times n}$ are said to be* ***congruent*** *if there exists a nonsingular matrix $P \in \mathbb{C}^{n\times n}$ such that*

$$\boxed{A = PBP^*} \tag{4.22}$$

It is clear that for a unitary P the matrix A is unitary equivalent to B.

Theorem 4.4. *Any Hermitian matrix $H \in \mathbb{C}^{n\times n}$ is congruent to the matrix*

$$\boxed{\Lambda_0 := \begin{bmatrix} I_{s\times s} & 0 & 0 \\ 0 & I_{(r-s)\times(r-s)} & 0 \\ 0 & 0 & 0_{(n-r)\times(n-r)} \end{bmatrix}} \tag{4.23}$$

where $r = \operatorname{rank} H$, and s is the number of positive eigenvalues of H counted according to multiplicity.

Proof. By the spectral theorem (4.4) any H can be represented as

$$H = X\Lambda X^* \tag{4.24}$$

where Λ is a diagonal matrix of eigenvalues of H and X is unitary. Ordering the eigenvalues so that the first s scalars $\lambda_1, \ldots, \lambda_s$ on the main diagonal of Λ are positive and the next $(r-s)$ numbers $\lambda_{s+1}, \ldots, \lambda_{r-s}$ are negative, one may write

$$\Lambda = U\Lambda_1\Lambda_0\Lambda_1U^* \tag{4.25}$$

where Λ_0 is as in (4.23) and

$$\Lambda_1 = \operatorname{diag}\left(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_s}, \sqrt{|\lambda_{s+1}|}, \ldots, \sqrt{|\lambda_{r-s}|}, 0, \ldots, 0\right)$$

The matrix U is a permutation (and therefore a unitary) matrix. So, substituting (4.25) into (3.22) gives

$$H = X\Lambda X^* = (XU\Lambda_1)\,\Lambda_0\,(\Lambda_1U^*X^*) = P\Lambda_0P^*$$

with $P = XU\Lambda_1$. Theorem is proven. □

4.4.2 Inertia of a square matrix

Definition 4.4. *The* ***inertia of a square matrix*** *$A \in \mathbb{R}^{n\times n}$, written as* $\operatorname{In} A$, *is the triple of integers*

$$\boxed{\operatorname{In} A := \{\pi(A), \quad \nu(A), \quad \delta(A)\}} \tag{4.26}$$

where

- $\pi(A)$ *denotes the number of eigenvalues of* A*, counted with their algebraic multiplicities, lying in the open right half-plane of* $\mathbb{C}$*;*
- $\nu(A)$ *denotes the number of eigenvalues of* A*, counted with their algebraic multiplicities, lying in the open left half-plane of* $\mathbb{C}$*;*
- $\delta(A)$ *is the number of eigenvalues of* A*, counted with their algebraic multiplicities, lying on the imaginary axis.*

Notice that

$$\boxed{\pi(A) + \nu(A) + \delta(A) = n} \tag{4.27}$$

Remark 4.3. *In the particular case of Hermitian matrices* $\pi(H)$ *and* $\nu(H)$ *merely denote the number of positive and, negative eigenvalues of* H *respectively. Notice that for Hermitian matrices*

$$\pi(H) + \nu(H) = \operatorname{rank} H$$

The difference

$$\boxed{\operatorname{sig} H := \pi(H) - \nu(H)} \tag{4.28}$$

is referred to as the ***signature*** *of* H*.*

Theorem 4.5. *Let* $A, B \in \mathbb{R}^{n\times n}$ *be Hermitian matrices of the same rank* r *and*

$$A = MBM^*$$

for some matrix M *(not obligatory nonsingular). Then*

$$\operatorname{In} A = \operatorname{In} B$$

Proof. By Theorem 4.4 there exist nonsingular matrices P and Q such that

$$\begin{aligned} &PAP^* = \operatorname{diag}[I_t, -I_{r-t}, 0] := \Lambda_0(A) \\ &Q^{-1}B\left(Q^{-1}\right)^* = \operatorname{diag}[I_s, -I_{r-s}, 0] := \Lambda_0(B) \\ &t = \pi(A),\ s = \pi(B) \end{aligned}$$

To prove the theorem it is sufficient to show that $t = s$. Suppose that $s < t$ and let us seek a contradiction. Notice that since $A = MBM^*$ we have

$$\begin{aligned} \Lambda_0(A) = PAP^* &= PMBM^*P^* \\ &= (PMQ)\,\Lambda_0(B)\,(Q^*M^*P^*) = R\Lambda_0(B)\,R^* \end{aligned} \tag{4.29}$$
$$R := PMQ$$

Let $x \in \mathbb{C}^n$ as

$$x = \begin{bmatrix} \tilde{x} \\ 0 \end{bmatrix},\ \tilde{x} \in \mathbb{C}^t, \quad \|\tilde{x}\| > 0$$

Then

$$x^* \Lambda_0(A)\, x = \sum_{i=1}^{t} |x_i|^2 = \|\tilde{x}\|^2 > 0 \tag{4.30}$$

Partitioning of R in the form

$$R^* := \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}, \quad R_{11} \in \mathbb{C}^{s \times t}$$

implies that $\tilde{x}$ can be chosen such that

$$R_{11}\tilde{x} = 0$$

keeping $\tilde{x} \neq 0$. Define now $y = R_{21}\tilde{x} \in \mathbb{C}^{n-s}$ which leads to the following identity

$$R^* x = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

Then by (4.29)

$$x^* \Lambda_0(A)\, x = x^* R \Lambda_0(B)\, R^* x = y^* \Lambda_0(B)\, y = -\sum_{j=1}^{r-s} |y_j|^2 = -\|y\|^2 \leq 0$$

which contradicts (4.30). Similarly, interchanging the roles of $\Lambda_0(A)$ and $\Lambda_0(B)$, one can find that $t < s$ is impossible. Hence, $s = t$. Theorem is proven. □

Corollary 4.2. (Sylvester's law of inertia) *Congruent Hermitian matrices have the same inertia characteristics.*

Proof. Since $A = PBP^*$ and P is nonsingular, then rank A = rank B and the result follows. □

Example 4.4. *Consider the quadratic form*

$$f_A(x) = (x, Ax) = 2x_1x_2 + 2x_2x_3 + x_3^2$$

which corresponds to the following matrix

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

The transformation

$$x = Tz, \; T = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & -1 & -1 \end{bmatrix}$$

implies

$$f_A(x) = (x, Ax) = \left(z\,[,T^{\mathsf{T}}AT]\,z\right)$$

$$= \left(z, \begin{bmatrix} 1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix} z\right) = z_1^2 - z_2^2 + z_3^2$$

So,

$$r = 3, \quad \pi(A) = 2, \quad \nu(A) = 1, \quad \delta(A) = 0$$

4.5 Cholesky factorization

In this subsection we follow Highan (1996).

4.5.1 Upper triangular factorization

Theorem 4.6. (Cholesky factorization) *Let $A = A^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ be a real symmetric $n \times n$ matrix with positive definite eigenvalues $\lambda_i(A) > 0$ $(i = 1, \ldots, n)$. Then there is a unique upper triangular matrix $R \in \mathbb{R}^{n\times n}$ with positive diagonal elements such that*

$$\boxed{A = R^{\mathsf{T}}R} \tag{4.31}$$

Proof. It may be done by induction. For $n = 1$ the result is clear. Assume that it is true for $(n-1)$. Let us consider $A_n = A_n^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ which can be represented in the following block form

$$A_n = \begin{bmatrix} A_{n-1} & c \\ c^{\mathsf{T}} & \alpha \end{bmatrix}, \quad c \in \mathbb{R}^{1\times(n-1)}, \quad \alpha \in \mathbb{R} \tag{4.32}$$

where $A_{n-1} = A_{n-1}^{\mathsf{T}} \in \mathbb{R}^{(n-1)\times(n-1)}$ by the assumption of the induction method has a unique Cholesky factorization $A_{n-1} = R_{n-1}^{\mathsf{T}}R_{n-1}$. Then (4.32) may be rewritten as

$$A_n = \begin{bmatrix} A_{n-1} & c \\ c^{\mathsf{T}} & \alpha \end{bmatrix} = \begin{bmatrix} R_{n-1}^{\mathsf{T}} & 0 \\ r^{\mathsf{T}} & \beta \end{bmatrix} \begin{bmatrix} R_{n-1} & r \\ 0 & \beta \end{bmatrix} := R_n^{\mathsf{T}}R_n \tag{4.33}$$

if

$$R_{n-1}^{\mathsf{T}}r = c \tag{4.34}$$

$$r^{\mathsf{T}}r + \beta^2 = \alpha \tag{4.35}$$

Notice that (4.34) has a unique solution since R_{n-1}^{T} is nonsingular. Then (4.35) gives

$$\beta^2 = \alpha - r^{\mathsf{T}} r = \alpha - \left[\left(R_{n-1}^{\mathsf{T}}\right)^{-1} c\right]^{\mathsf{T}} \left[\left(R_{n-1}^{\mathsf{T}}\right)^{-1} c\right]$$
$$= \alpha - c^{\mathsf{T}} \left(R_{n-1}\right)^{-1} \left(R_{n-1}^{\mathsf{T}}\right)^{-1} c = \alpha - c^{\mathsf{T}} A_{n-1}^{-1} c$$

It remains to check that there exists a unique real positive β satisfying this equation, that is, we need to show that

$$\alpha - c^{\mathsf{T}} A_{n-1}^{-1} c > 0$$

One has

$$0 < \det A_n = \det(R_n^{\mathsf{T}} R_n) = \det(R_n^{\mathsf{T}}) \det(R_n)$$
$$= \left[\det(R_{n-1}^{\mathsf{T}})\beta\right] \left[\det(R_{n-1})\beta\right] = \left[\det(R_{n-1})\right]^2 \beta^2$$

which implies

$$\beta^2 = \frac{\det A_n}{\left[\det(R_{n-1})\right]^2} > 0$$

Hence there is a unique $\beta > 0$. So, (4.33) is valid. □

Corollary 4.3. *Given the Cholesky factorization*

$$A^{\mathsf{T}} A = R^{\mathsf{T}} R$$

the system of linear equations

$$Ax = b$$

or, equivalently,

$$\left(A^{\mathsf{T}} A\right) x = A^{\mathsf{T}} b := \tilde{b}$$

can be solved via the two triangular linear systems

$$R^{\mathsf{T}} y = \tilde{b}$$
$$Rx = y$$

which can be resolved by the simple Gaussian elimination procedure (1.6).

Corollary 4.4. *Let r_{ij} be the elements of R and $D := \operatorname{diag}\left(r_{11}^2, \ldots r_{nn}^2\right) > 0$. The Cholesky factorization $A = R^{\mathsf{T}} R$ (4.31) may be represented as*

$$A = LDL^{\mathsf{T}}$$

where

$$L = R^{\mathsf{T}} \operatorname{diag}\left(r_{11}^{-1}, \ldots, r_{nn}^{-1}\right)$$

4.5.2 Numerical realization

The following procedure is the direct algorithm for computation of matrix R:

for $j = 1 : n$

  for $i = 1 : j - 1$

$$r_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right) / r_{ii}$$

  end

$$r_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{i-1} r_{kj}^2}$$

end

Example 4.5.

$$A = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sqrt{3} & 0 & 0 \\ \frac{1}{\sqrt{3}} & \frac{1}{3}\sqrt{15} & 0 \\ \frac{1}{\sqrt{3}} & -\frac{1}{15}\sqrt{15} & \frac{1}{5}\sqrt{15} \end{bmatrix} \begin{bmatrix} \sqrt{3} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{1}{3}\sqrt{15} & -\frac{1}{15}\sqrt{15} \\ 0 & 0 & \frac{1}{5}\sqrt{15} \end{bmatrix}$$

5 Matrix Functions

Contents

5.1 Projectors

Definition 5.1. *A Hermitian $n \times n$ matrix P is said to be a* ***projector*** *or an* ***idempotent matrix*** *if it satisfies the condition*

$$P^*P = PP^* = P^2 = P = P^* \tag{5.1}$$

Proposition 5.1. *If $P \in \mathbb{C}^{n\times n}$ is a projector, then*

(a) the matrix

$$Q := I_{n\times n} - P \tag{5.2}$$

is a projector too and named the ***complementary projector*** *to P;*

(b)

$$\mathrm{Im}(I_{n\times n} - P) = \mathrm{Ker}\, P \tag{5.3}$$

(c)

$$\mathrm{Ker}(I_{n\times n} - P) = \mathrm{Im} P \tag{5.4}$$

Proof.

(a) To prove that Q is a projector too, note that

$$\begin{aligned}(I_{n\times n} - P)^2 &= I_{n\times n} - P - P + P^2 \\ &= I_{n\times n} - P - P + P = I_{n\times n} - P\end{aligned}$$

(b) If $y \in \mathrm{Im}(I_{n\times n} - P)$, then $y = (I_{n\times n} - P)\, x$ for some $x \in \mathbb{C}^n$. Thus,

$$Py = P\,(I_{n\times n} - P)\, x = \left(P - P^2\right) x = 0$$

It means that $y \in \mathrm{Ker}\, P$.

(c) By a similar argument, if $y \in \operatorname{Im} P$, then $y = Px$ for some $x \in \mathbb{C}^n$ and, hence,

$$(I_{n\times n} - P)\, y = (I_{n\times n} - P)\, Px = \left(P - P^2\right) x = 0$$

which exactly means (5.4).

□

Corollary 5.1. *If $P \in \mathbb{C}^{n\times n}$ is a projector, then*

$$\operatorname{Ker} P + \operatorname{Im} P = \mathbb{C}^n$$

Proof. Evidently, any $x \in \mathbb{C}^n$ may be represented as

$$x = (I_{n\times n} - P)\, x + Px = x^{(1)} + x^{(2)}$$
$$x^{(1)} := (I_{n\times n} - P)\, x \in \operatorname{Ker} P$$
$$x^{(2)} := Px \in \operatorname{Ker}(I_{n\times n} - P) = \operatorname{Im} P$$

□

Corollary 5.2. *Any $x^{(1)} \in \operatorname{Ker} P$ and $x^{(2)} \in \operatorname{Im} P$ are orthogonal, that is*

$$\left(x^{(1)}, x^{(2)}\right) = 0$$

Proof. By the previous corollary, we have

$$x^{(1)*} x^{(2)} = x^* (I_{n\times n} - P)^*\, Px$$
$$= x^* (I_{n\times n} - P)\, Px = x^* \left(P - P^2\right) x = 0$$

□

The property given above in (5.2) exactly justifies the name projector for P.

Theorem 5.1. *If $P \in \mathbb{C}^{n\times n}$ is a projector, then*

1. *its eigenvalues $\lambda_i(P)$ are either equal to 1 or 0;*
2. *it is a simple matrix, that is, it is equivalent to a diagonal matrix with the diagonal elements equal to 1 or 0;*
3. *it may be represented as*

$$\boxed{P = \sum_{i=1}^{r} x^{(i)} x^{(i)*}} \tag{5.5}$$

where $r = \operatorname{rank} P$ and $\left\{x^{(1)}, \ldots, x^{(r)}\right\}$ is the system of the eigenvectors of P corresponding to $\lambda_i(P) = 1$.

Proof.

1. If $x^{(i)}(P)$ is an eigenvector of P corresponding to an eigenvalue $\lambda_i(P)$, then

$$Px^{(i)}(P) = \lambda_i(P)\,x^{(i)}(P)$$

and pre-multiplication of this identity by $(I_{n\times n} - P)$ implies

$$\begin{aligned} 0 &= (I_{n\times n} - P)\,Px^{(i)}(P) = \lambda_i(P)\,(I_{n\times n} - P)\,x^{(i)}(P) \\ &= \lambda_i(P)\left[x^{(i)}(P) - Px^{(i)}(P)\right] = \lambda_i(P)\left[x^{(i)}(P) - \lambda_i(P)\,x^{(i)}(P)\right] \\ &= \lambda_i(P)\left[1 - \lambda_i(P)\right]x^{(i)}(P) \end{aligned}$$

which proves the first assertion.
(2) and (3) result from (1) and the spectral theorem (4.4).

□

5.2 Functions of a matrix

5.2.1 Main definition

Definition 5.2. *Let $A \in \mathbb{C}^{n\times n}$ be any square complex matrix and $T \in \mathbb{C}^{n\times n}$ is the nonsingular matrix T, defined by (4.9), transforming A to the Jordan canonical form, that is,*

$$\begin{aligned} A &= TJT^{-1} \\ J &= \operatorname{diag}(J_1, J_2, \ldots, J_K) \end{aligned}$$

where $J_i \in \mathbb{C}^{\mu_i\times\mu_i}$ is the ith Jordan block defined by (4.7). Also let

$$f(\lambda) : \mathbb{C} \longrightarrow \mathbb{C}$$

be a function which is $(l-1)$-times differentiable in a neighborhood of each $\lambda_i \in \sigma(A)$ where

$$l := \max_{i=1,\ldots,K} \mu_i, \quad \sum_{i=1}^{K} \mu_i = n$$

(μ_i is the multiplicity of λ_i). Then

$$\boxed{f(A) := T\operatorname{diag}(f(J_1), f(J_2), \ldots, f(J_K))\,T^{-1}} \tag{5.6}$$

where

$$\boxed{f(J_i) := \begin{bmatrix} f(\lambda_i) & \frac{1}{1!}f^{(1)}(\lambda_i) & \cdots & \frac{1}{(\mu_i-1)!}f^{(\mu_i-1)}(\lambda_i) \\ 0 & f(\lambda_i) & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{1!}f^{(1)}(\lambda_i) \\ 0 & \cdots & 0 & f(\lambda_i) \end{bmatrix}} \tag{5.7}$$

and $f^{(k)}(\lambda)$ is the kth derivative of $f(\lambda)$ in the point λ.

Example 5.1. *Let*

$$A = \begin{bmatrix} 6 & 2 & 2 \\ -2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

Then the eigenvalues, the corresponding eigenvectors and the generalized eigenvectors are as follows

$$\lambda_1 = 2, \quad x^{(1)} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix}, \quad \mu_1 = 1$$

$$\lambda_2 = 4, \quad x^{(2,1)} = \begin{bmatrix} -2 \\ 2 \\ 0 \end{bmatrix}, \quad \mu_2 = 2$$

From equation (3.31)

$$(A - 4I_{3\times 3})\, x^{(2,2)} = x^{(2,1)}$$

we obtain

$$x^{(2,2)} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

and, hence

$$T = \begin{bmatrix} 0 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix}, \quad T^{-1} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1/2 & 1/2 \\ 1 & 1 & 1 \end{bmatrix}$$

So, for example,

(a)

$$\ln(A) = T \begin{bmatrix} \ln 2 & 0 & 0 \\ 0 & \ln 4 & 1/4 \\ 0 & 0 & \ln 4 \end{bmatrix} T^{-1}$$

$$= \begin{bmatrix} 0 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \ln 2 & 0 & 0 \\ 0 & \ln 4 & 1/4 \\ 0 & 0 & \ln 4 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1/2 & 1/2 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.88629 & -0.5 & -0.5 \\ 0.5 & 1.8863 & 1.1931 \\ 0.0 & 0.0 & 0.69315 \end{bmatrix}$$

(b)

$$\sin(A) = T \begin{bmatrix} \sin(2) & 0 & 0 \\ 0 & \sin(4) & \cos(4) \\ 0 & 0 & \sin(4) \end{bmatrix} T^{-1}$$

$$= \begin{bmatrix} 0 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sin(2) & 0 & 0 \\ 0 & \sin(4) & \cos(4) \\ 0 & 0 & \sin(4) \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1/2 & 1/2 \\ 1 & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.55048 & 1.3073 & 1.3073 \\ -1.3073 & -2.0641 & -2.9734 \\ 0.0 & 0.0 & 0.90930 \end{bmatrix}$$

5.2.2 *Matrix exponent*

There exist *two definitions* of the matrix exponent e^A of an arbitrary *square* matrix $A \in \mathbb{C}^{n\times n}$.

1. The *first* definition may be done according to the general rule (5.6):

$$\boxed{\begin{aligned} e^A &:= T\,\mathrm{diag}\left(e^{J_1}, e^{J_2}, \ldots, e^{J_K}\right) T^{-1} \\ f(J_i) &:= \begin{bmatrix} e^{\lambda_i} & \frac{1}{1!}e^{\lambda_i} & \cdots & \frac{1}{(\mu_i-1)!}e^{\lambda_i} \\ 0 & e^{\lambda_i} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{1!}e^{\lambda_i} \\ 0 & \cdots & 0 & e^{\lambda_i} \end{bmatrix} \\ &= e^{\lambda_i} J_I(\mu_i) \end{aligned}} \tag{5.8}$$

where

$$J_I(\mu_i) := \begin{bmatrix} 1 & \frac{1}{1!} & \cdots & \frac{1}{(\mu_i-1)!} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \frac{1}{1!} \\ 0 & \cdots & 0 & 1 \end{bmatrix}$$

2. The *second* definition is as follows:

$$\boxed{e^A := \sum_{k=0}^{\infty} \frac{1}{k!} A^k} \tag{5.9}$$

Notice that the series in (5.9) always converges since the series $\sum_{k=0}^{\infty} \frac{1}{k!} \left(A^k\right)_{ij}$ always converges for any $A \in \mathbb{C}^{n\times n}$ and any $i, j = 1, \ldots, n$.

Lemma 5.1. *Definitions (5.8) and (5.9) coincide.*

Proof. For normal matrices this claim is evident since in this case the Jordan blocks are diagonal. In general cases this result can be checked by the simple evaluation (by induction) of A^k using both definitions (5.8) and (5.9). □

A rather surprising formula holds.

Lemma 5.2.

$$\boxed{\det\left(e^A\right) = \exp\left(\mathrm{tr}A\right)} \tag{5.10}$$

Proof. Using (5.8) and (3.19) one has

$$\begin{aligned}
\det\left(e^A\right) &= \det\left(T\,\mathrm{diag}\left(e^{J_1}, e^{J_2}, \ldots, e^{J_K}\right) T^{-1}\right) \\
&= \left[\det T\right]\left[\det \mathrm{diag}\left(e^{J_1}, e^{J_2}, \ldots, e^{J_K}\right)\right]\left[\det T^{-1}\right] \\
&= \left[\det \mathrm{diag}\left(e^{J_1}, e^{J_2}, \ldots, e^{J_K}\right)\right] = \prod_{i=1}^{K} \det e^{J_i} \\
&= \prod_{i=1}^{K} \left[\exp(\lambda_i)\right]^{\mu_i} = \prod_{i=1}^{K} \exp(\mu_i \lambda_i) = \exp\left(\sum_{i=1}^{K} \mu_i \lambda_i\right) \\
&= \exp\left(\sum_{s=1}^{n} \lambda_s\right) = \exp(\mathrm{tr}A)
\end{aligned}$$

□

Example 5.2. *For the matrix A from the previous example (5.1) we have*

$$\begin{aligned}
&\exp\left(\begin{bmatrix} 6 & 2 & 2 \\ -2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}\right) \\
&= \begin{bmatrix} 0 & -2 & 1 \\ -1 & 2 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} e^2 & 0 & 0 \\ 0 & e^4 & e^4 \\ 0 & 0 & e^4 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1/2 & 1/2 \\ 1 & 1 & 1 \end{bmatrix} \\
&= \begin{bmatrix} -54.598 & -109.20 & -109.20 \\ 109.20 & 163.79 & 156.41 \\ 0.0 & 0.0 & 7.3891 \end{bmatrix}
\end{aligned}$$

and by (5.10)

$$\det\left[\exp\left(\begin{bmatrix} 6 & 2 & 2 \\ -2 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}\right)\right]$$

$$= \exp(\mathrm{tr}A) = \exp(10)$$

The next result is very important in various matrix applications and especially in the theory of ordinary differential equations.

Proposition 5.2. *The identity*

$$e^{(A+B)t} = e^{At}e^{Bt}$$

is valid for all t (including $t = 1$) if and only if the matrices A and B commute, that is, when

$$AB = BA$$

Proof. This statement is sufficient for the class of normal matrices when the definition (5.9) is applied. Since

$$e^{(A+B)t} = I_{n\times n} + t(A+B) + \frac{t^2}{2}(A+B)^2 + \cdots$$

and

$$e^{At}e^{Bt} = \left(I_{n\times n} + tA + \frac{t^2}{2}A^2 + \cdots\right)\left(I_{n\times n} + tB + \frac{t^2}{2}B^2 + \cdots\right)$$

$$= I_{n\times n} + t(A+B) + \frac{t^2}{2}\left(A^2 + B^2 + 2AB\right) + \cdots$$

we obtain

$$e^{(A+B)t} - e^{At}e^{Bt} = (BA - AB)\frac{t^2}{2} + \cdots$$

which proves the proposition. □

Corollary 5.3. *For any $s, t \in \mathbb{C}$*

$$\boxed{e^{A(s+t)} = e^{As}e^{At}} \tag{5.11}$$

Corollary 5.4. *The matrix exponent e^{At} is **always nonsingular** and its inverse matrix is e^{-At}.*

Proof. Indeed, taking in (5.11) $s = -t$ we get

$$e^{A\cdot 0} = I_{n\times n} = e^{-At}e^{At}$$

which implies the result. □

5.2.3 Square root of a positive semidefinite matrix

In this subsection we will discuss the construction of the function which satisfies the condition

$$A = A^{1/2}A^{1/2} \tag{5.12}$$

The formal implementation of the definition (5.6) demands to consider only the matrices with *nonnegative spectrum of eigenvalues*, that is, when $\lambda_i(A) \geq 0$ for all $i = 1, \ldots, n$. But to fulfill (5.12) this is not sufficient. Indeed, if there exists at least one complete Jordan block this property never holds. Thus, we need to ask whether the Jordan block would not be presented which is true only for *Hermitian* (in the case of real matrices, *symmetric*) matrices. So, now we are ready to formulate the following proposition.

Proposition 5.3. *The matrix $A^{1/2}$ is well defined for a positive semidefinite Hermitian matrix and, moreover, it is positive semidefinite Hermitian itself.*

Proof. For Hermitian matrices the transforming matrix T is always unitary, that is, $T^{-1} = T^*$ and all eigenvalues are real. Thus,

$$A^{1/2} = T\,\mathrm{diag}\left((J_1)^{1/2}, (J_2)^{1/2}, \ldots, (J_K)^{1/2}\right) T^{-1}$$

and, hence,

$$\begin{aligned}\left(A^{1/2}\right)^* &= \left(T^{-1}\right)^* \mathrm{diag}\left((J_1)^{1/2}, (J_2)^{1/2}, \ldots, (J_K)^{1/2}\right) T^* \\ &= T\,\mathrm{diag}\left((J_1)^{1/2}, (J_2)^{1/2}, \ldots, (J_K)^{1/2}\right) T^{-1} = A^{1/2}\end{aligned}$$

□

Example 5.3.

$$\begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}^{1/2} = \begin{bmatrix} \sqrt{2}+1 & -\sqrt{2}+1 \\ 1 & 1 \end{bmatrix} \times \begin{bmatrix} \sqrt{\sqrt{2}+2} & 0 \\ 0 & \sqrt{2-\sqrt{2}} \end{bmatrix} \begin{bmatrix} \dfrac{\sqrt{2}+2}{4\sqrt{2}+4} & \dfrac{\sqrt{2}}{4\sqrt{2}+4} \\ -\dfrac{1}{4}\sqrt{2} & \dfrac{1}{4}\sqrt{2}+\dfrac{1}{2} \end{bmatrix}$$

$$= \begin{bmatrix} 1.6892 & 0.38268 \\ 0.38268 & 0.92388 \end{bmatrix}$$

5.3 The resolvent for a matrix

The complex function theory provides a third approach to the definition of $f(A)$ applicable when $f(\lambda)$ is an analytical function of a complex variable λ. Sure, this approach is consistent with general definition (5.6) of $f(A)$ valid for multiplying differentiable functions defined on the spectrum of A.

Let us consider a matrix $A(\lambda) : \mathbb{C} \to \mathbb{C}^{m\times n}$ whose elements are functions of a complex variable λ and define also

- *the derivatives*

$$\frac{d^r}{d\lambda^r} A(\lambda) = A^{(r)}(\lambda), \quad r = 0, 1, 2, \ldots \tag{5.13}$$

of the matrix $A(\lambda)$ to be the matrix obtained by differentiating each element of $A(\lambda)$;

- *the integral*

$$\int_L A(\lambda)\, d\lambda \tag{5.14}$$

of the matrix $A(\lambda)$ to be the matrix obtained by integrating each element of $A(\lambda)$ in the positive direction along a path L in a complex plane, which will be assumed to be a finite system of simple piecewise smooth closed contours without points of intersections.

Example 5.4. *For a normal matrix $A \in \mathbb{C}^{n\times n}$ using the series representation (5.9) the following properties may be proven:*

1.

$$\boxed{\frac{d}{dt}\left(e^{At}\right) = Ae^{At} = e^{At}A} \tag{5.15}$$

2.

$$\boxed{\frac{d}{dt}(A(t))^2 = \left(\frac{d}{dt}A(t)\right)A + A\left(\frac{d}{dt}A(t)\right)} \tag{5.16}$$

Notice that in general

$$\frac{d}{dt}(A(t))^2 \neq 2A\left(\frac{d}{dt}A(t)\right)$$

3. if all derivatives exist and $p = 1, 2, \ldots$ then

$$\boxed{\frac{d}{dt}(A(t))^p = \sum_{i=1}^{p} A^{i-1}\left(\frac{d}{dt}A(t)\right)A^{p-i}} \tag{5.17}$$

4.

$$\boxed{\frac{d}{dt}(A(t))^{-p} = -A^{-p}\left[\sum_{i=1}^{p} A^{i-1}\left(\frac{d}{dt}A(t)\right)A^{p-i}\right]A^{-p}} \tag{5.18}$$

This relation may be easily proven by simply differentiating the identity

$$A^{-p}A^{p} = I_{n\times n}$$

which implies

$$\left[\frac{d}{dt}(A(t))^{-p}\right]A^{p} + A^{-p}\left[\frac{d}{dt}(A(t))^{p}\right] = O_{n\times n}$$

and, thus,

$$\left[\frac{d}{dt}(A(t))^{-p}\right] = -A^{-p}\left[\frac{d}{dt}(A(t))^{p}\right]A^{-p}$$

Definition 5.3. *The matrix*

$$\boxed{R_{\lambda}(A) := (\lambda I_{n\times n} - A)^{-1}} \tag{5.19}$$

defined for all $\lambda \in \mathbb{C}$ which do not belong to the spectrum of $A \in \mathbb{C}^{n\times n}$ is known as the ***resolvent*** *of A.*

The following properties of $R_{\lambda}(A)$ seem to be important for the considerations below.

Proposition 5.4. *For all $\lambda \notin \sigma(A)$*

1.

$$R_{\lambda}(A) - R_{\mu}(A) = (\mu - \lambda)R_{\lambda}(A)R_{\mu}(A) \tag{5.20}$$

2.

$$\frac{d}{d\lambda}R_{\lambda}(A) = -R_{\lambda}^{2}(A) \tag{5.21}$$

3.

$$\frac{d^{r}}{d\lambda^{r}}R_{\lambda}(A) = (-1)^{r}r!R_{\lambda}^{r+1}(A) \tag{5.22}$$

Proof. Formula (5.22) may be proven by induction taking into account (5.21). But (5.21) results from (5.20). To prove (5.20) notice that

$$\begin{aligned} & R_\lambda^{-1}(A)\left[R_\lambda(A) - R_\mu(A)\right] R_\mu^{-1}(A) \\ & = R_\lambda^{-1}(A)\left[(\lambda I_{n\times n} - A)^{-1} - (\mu I_{n\times n} - A)^{-1}\right] R_\mu^{-1}(A) \\ & = R_\mu^{-1}(A) - R_\lambda^{-1}(A) = (\mu - \lambda) I_{n\times n} \end{aligned}$$

which implies (5.20). □

Theorem 5.2. *The resolvent $R_\lambda(A)$ of $A \in \mathbb{C}^{n\times n}$ is a rational function of λ with poles at the points of the spectrum of A and $R_\infty(A) = 0$. Moreover, each $\lambda_k \in \sigma(A)$ is a pole of $R_\lambda(A)$ of order μ_k where μ_k is the multiplicity of the eigenvalue λ_k, that is,*

$$R_\lambda(A) = \frac{1}{m(\lambda)} \sum_{j=1}^{K-1} \left(\sum_{i=1}^{K-1} \gamma_{ij} \lambda^i \right) A^j \tag{5.23}$$

where $m(\lambda) = \prod_{s=1}^{K} (\lambda - \lambda_s)^{\mu_s}, \quad \sum_{s=1}^{K} \mu_s = n.$

Proof. Evidently

$$\frac{m(\lambda) - m(\mu)}{\lambda - \mu} = \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} \gamma_{ij} \lambda^i \mu^j$$

for some numbers γ_{ij}. Using the matrix polynomial definition (as in the Cayley–Hamilton theorem) for $\lambda_k \notin \sigma(A)$ the last relation (after formal substituting A for μ) implies

$$[m(\lambda) I_{n\times n} - m(A)] R_\lambda(A) = \sum_{i=1}^{K-1} \sum_{j=1}^{K-1} \gamma_{ij} \lambda^i A^j$$

Since by the Cayley–Hamilton theorem $m(A) = 0$, we obtain (5.23). □

Thus, using the terminology of complex analysis, the spectrum of a matrix A can be described in terms of its resolvent. The next theorem establishes this relation exactly.

Theorem 5.3. (Cauchy integral theorem for matrices) *Let $f(\lambda)$ be a function of the complex variable λ* ***analytical in*** *an open set $D \in \mathbb{C}$, that is, $f(\lambda)$ has a convergent Taylor series expansion about each point of D. If $A \in \mathbb{C}^{n\times n}$ has distinct eigenvalues*

$\lambda_1, \ldots, \lambda_{s \leq n}$, the path L is a closed contour having $\lambda_1, \ldots, \lambda_s$ as its interior, and $f(\lambda)$ is continuous in and analytic within L, then

$$\boxed{\begin{aligned} f(A) &= \frac{1}{2\pi i} \int_L f(\lambda) \left[(\lambda I_{n\times n} - A)^{-1}\right] d\lambda \\ &= \frac{1}{2\pi i} \int_L f(\lambda) R_\lambda(A)\, d\lambda \end{aligned}} \tag{5.24}$$

Proof. This result may be established using (5.23) and the Cauchy theorem which asserts that

$$f^{(r)}(\lambda_0) = \frac{r!}{2\pi i} \int_L \frac{f(\lambda_0)}{(\lambda - \lambda_0)^{r+1}} d\lambda$$

for any $\lambda_0 \in D$. □

Corollary 5.5. *The following identities are valid for any $A \in \mathbb{C}^{n\times n}$:*

$$\boxed{\begin{aligned} I_{n\times n} &= \frac{1}{2\pi i} \int_L R_\lambda(A)\, d\lambda \\ A &= \frac{1}{2\pi i} \int_L \lambda R_\lambda(A)\, d\lambda \end{aligned}} \tag{5.25}$$

5.4 Matrix norms

5.4.1 Norms in linear spaces and in $\mathbb{C}^n$

Definition 5.4. *A real-valued function $\|x\| : \mathcal{L} \to \mathbb{R}$ defined on all elements x of a linear space $\mathcal{L}$ of complex or real numbers, is called a* ***norm*** *(on $\mathcal{L}$), if it satisfies the following axioms:*

1.

$$\boxed{\|x\| \geq 0}$$

for all $x \in \mathcal{L}$ and $\|x\| = 0$ if and only if $x = 0$;

2.

$$\boxed{\|\alpha x\| = |\alpha|\, \|x\|}$$

for all $x \in \mathcal{L}$ and all $\alpha \in \mathbb{C}$;

3. the triangle inequality holds, that is,

$$\boxed{\|x + y\| \leq \|x\| + \|y\|}$$

for all $x, y \in \mathcal{L}$.

A linear space $\mathcal{L}$ together with a norm defined on it is called a ***normed linear space****.*

Consider the following examples of norms in $\mathbb{C}^n$.

Example 5.5. *Let $x = (x_1, x_2, \ldots, x_n)^\top$ be a typical vector in $\mathbb{C}^n$ (or, in particular, in $\mathbb{R}^n$). Then the following functions are norms in a finite-dimensional space $\mathbb{C}^n$ (or $\mathbb{R}^n$):*

1. Modul-sum norm

$$\boxed{\|x\|_1 := \max_{1 \le i \le n} |x_i|} \tag{5.26}$$

2. Euclidean norm

$$\boxed{\|x\|_2 := \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}} \tag{5.27}$$

3. Hölder norm

$$\boxed{\|x\|_p := \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}, \quad p \ge 1} \tag{5.28}$$

4. Chebyshev norm

$$\boxed{\|x\|_\infty := \max_{1 \le i \le n} |x_i|} \tag{5.29}$$

5. Weighted norm

$$\boxed{\|x\|_H := \sqrt{(Hx, x)} = \sqrt{x^* H x}} \tag{5.30}$$

where H is a Hermitian (or symmetric) matrix with all positive definite eigenvalues.

It is not so difficult to check that functions (5.26)–(5.30) satisfy all three norm axioms.

Definition 5.5. *Two norms $\|x\|'$ and $\|x\|''$ are said to be* ***equivalent*** *in $\mathcal{L}$, if there exist positive numbers $r_1, r_2 \in (0, \infty)$ such that for any $x \in \mathcal{L}$*

$$\|x\|' \ge r_2 \|x\|'', \quad \|x\|'' \ge r_1 \|x\|'$$

Proposition 5.5. *Any two norms in a finite-dimensional linear space are equivalent.*

Proof. It is clear that a norm in a finite-dimensional linear space is a continuous function since the inequality

$$\|x + z\| \leq \|x\| + \|z\|$$

leads to the following relations

$$\|x + z\| - \|x\| \leq \|z\|$$

and

$$\|y\| - \|x\| \leq \|y - x\|$$
$$|\|y\| - \|x\|| \leq \|y - x\|$$

if $z = y - x$. The last inequality corresponds exactly to the continuity definition. Let us consider two sets

$$X_1 := \left\{x \in \mathcal{L} : \|x\|' = 1\right\}$$
$$X_2 := \left\{x \in \mathcal{L} : \|x\|'' = 1\right\}$$

By the continuity property there exist two elements $x_{01} \in X_1$ and $x_{02} \in X_2$ such that

$$0 < \gamma_1 := \inf_{x \in X_1} \|x\|'' = \|x_{01}\|'$$
$$0 < \gamma_2 := \inf_{x \in X_2} \|x\|' = \|x_{02}\|''$$

Thus, for any nonzero element $x \in \mathcal{L}$ and the second norm axiom it follows that

$$0 < \gamma_1 := \inf_{x \in X_1} \|x\|'' \leq \left\| \frac{x}{\|x\|'} \right\|'' = \frac{\|x\|''}{\|x\|'}$$
$$0 < \gamma_2 := \inf_{x \in X_2} \|x\|' \leq \left\| \frac{x}{\|x\|''} \right\|' = \frac{\|x\|'}{\|x\|''}$$

which for $r_1 = \gamma_2$ and $r_2 = \gamma_1$ corresponds to the desired result. □

5.4.2 *Matrix norms*

Here we will pay attention to norms on the linear space $\mathbb{C}^{n \times n}$, or in other words, to norms in a space of squared matrices. Sure, all properties of norms discussed before should be valid for the matrix case. However, some additional axiom (or axioms) are required because of the possibility of multiplying any two matrices that give rise to the question regarding the relation of $\|AB\|$ and $\|A\|$, $\|B\|$ for any two matrices $A, B \in \mathbb{C}^{n \times n}$.

Definition 5.6. (Axiom 4 for matrix norms) *The function $\|A\|$ defined for any $A \in \mathbb{C}^{n\times n}$ is said to be a **matrix** or **submultiplicative norm** (in contrast to a standard norm in $\mathbb{C}^{n\times n}$) if the following axiom holds*

$$\boxed{\|AB\| \leq \|A\|\,\|B\|} \tag{5.31}$$

Example 5.6. *It is not difficult to check that the function*

$$\max_{1\leq i,j\leq n} |a_{ij}|$$

where a_{ij} is an element of $A \in \mathbb{C}^{n\times n}$, is a norm on $\mathbb{C}^{n\times n}$, but it is not a matrix norm.

Proposition 5.6. *The following functions are the matrix norms for the matrix $A = [a_{ij}]_{1\leq i,j\leq n}$:*

1. *Frobenius (Euclidean) norm*

$$\boxed{\|A\|_F := \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}} \tag{5.32}$$

2. *Hölder norm*

$$\boxed{\|A\|_p := \left(\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^p\right)^{1/p}} \tag{5.33}$$

is a matrix norm if and only if

$$\boxed{1 \leq p \leq 2}$$

3. *Weighted Chebyshev norm*

$$\boxed{\|A\|_p := n \max_{1\leq i,j\leq n} |a_{ij}|} \tag{5.34}$$

4. *Trace norm*

$$\boxed{\|A\|_{tr} := \sqrt{\mathrm{tr}\,(A^*A)} = \sqrt{\mathrm{tr}\,(AA^*)}} \tag{5.35}$$

5. Maximal singular-value norm

$$\boxed{\begin{aligned}\|A\|_\sigma &:= \sqrt{\max_{1\le i\le n}\sigma_i(A)} \\ &= \sqrt{\max_{1\le i\le n}\lambda_i(A^*A)} = \sqrt{\max_{1\le i\le n}\lambda_i(AA^*)}\end{aligned}} \tag{5.36}$$

6. S-norm

$$\boxed{\|A\|_S := \left\|SAS^{-1}\right\|} \tag{5.37}$$

where S is any nonsingular matrix and $\|\cdot\|$ is any matrix norm.

We leave the proof of this proposition for readers as an exercise. The next statement also seems to be evident.

Proposition 5.7. *For any matrix norm $\|\cdot\|$*

$$\boxed{\begin{aligned}\|I_{n\times n}\| &\ge 1 \\ \left\|A^k\right\| &\le \|A\|^k, \quad k = 2, 3, \dots \\ \left\|A^{-1}\right\| &\ge \frac{1}{\|A\|}\end{aligned}}$$

There exists an estimate of any matrix norm related to the spectral radius (3.14).

Lemma 5.3. *For any matrix $A \in \mathbb{C}^{n\times n}$ with the spectral radius $\rho(A) = \max\limits_{1\le i\le n}|\lambda_i|$ and any matrix norm $\|\cdot\|$ the following estimate holds*

$$\boxed{\|A\| \ge \rho(A)} \tag{5.38}$$

Proof. Let λ be the eigenvalue of A with the maximal module, i.e., $\rho(A) = |\lambda|$. Then there exists the corresponding eigenvector $x \ne 0$ such that $Ax = \lambda x$. Define an $n \times n$ matrix

$$B := \begin{bmatrix} x & 0 & 0 \cdots 0 \end{bmatrix}$$

and observe that

$$AB = \lambda B$$

Then by the second and fourth norm axioms we deduce that

$$\|\lambda B\| = |\lambda| \ \|B\| = \rho(A)\|B\| \leq \|A\| \ \|B\|$$

Since $B \neq 0$ it follows that $\|B\| > 0$ which proves the desired result (5.38). □

The following comments are very important for practical applications.

Remark 5.1.

1. *The spectral radius $\rho(A)$ itself cannot be considered as a matrix norm (and as any norm in general) since it does not satisfy the first norm axiom, that is, if $\rho(A) = 0$, we cannot conclude that $A = O_{n\times n}$ (one can consider the matrix $A = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$ as an example).*
2. *The inequality (5.38) may be considered as an upper estimate for the spectral radius $\rho(A)$.*

5.4.3 Compatible norms

Definition 5.7. *The vector norm $\|\cdot\|_v$ and matrix norm $\|\cdot\|$ are said to be **compatible** if the inequality*

$$\boxed{\|Ax\|_v \leq \|A\| \ \|x\|_v} \tag{5.39}$$

is valid for any $x \in \mathbb{C}^n$ and any $A \in \mathbb{C}^{n\times n}$.

It is not difficult to check that

- the Frobenius matrix (5.32) and Euclidean vector (5.27) norm are compatible;
- the weighted Chebyshev norm (5.34) is compatible with Hölder norms (5.28) in C^n for $p = 1, 2, \infty$.

5.4.4 Induced matrix norm

Proposition 5.8. *The quotient*

$$\boxed{f(A) := \sup_{x\in C^n, x\neq 0} \frac{\|Ax\|_v}{\|x\|_v}} \tag{5.40}$$

*can be considered as **a matrix norm induced by the vector norm** $\|\cdot\|_v$. In particular, the matrix norm induced by the Euclidean vector norm is known as the **spectral matrix norm**. For calculus purposes it may be calculated as*

$$\boxed{f(A) := \max_{x\in C^n : \|x\|_v = 1} \|Ax\|_v} \tag{5.41}$$

Proof.

(a) First, let us prove (5.41). Notice that (5.40) can be represented as

$$f(A) := \sup_{x \in C^n, x \neq 0} \frac{\|Ax\|_v}{\|x\|_v}$$

$$= \sup_{x \in C^n, x \neq 0} \left\| A \frac{x}{\|x\|_v} \right\|_v = \sup_{x \in C^n : \|x\|_v = 1} \|Ax\|_v$$

Since any vector norm is a continuous function, there exists a vector $x_0 : \|x_0\|_v = 1$ such that $f(A) = \|Ax_0\|_v$, which means that sup is reachable, or in other words (5.41) holds. Now we are ready to prove that $f(A)$ defined by (5.41) is a vector norm.

(b) To check axiom 1 notice that $f(A) \geq 0$ and if $A \neq 0$ it follows that $Ax \neq 0$ ($\|x\|_v = 1$) and, hence, $\|Ax\|_v > 0$. So, the first axiom is established. The second axiom follows from the identity

$$f(\lambda A) = \max_{x \in C^n : \|x\|_v = 1} \|\lambda Ax\|_v$$

$$= |\lambda| \max_{x \in C^n : \|x\|_v = 1} \|Ax\|_v = |\lambda| f(A)$$

The third one results from triangle inequality for vectors, i.e.,

$$\|(A + B)x\|_v \leq \|Ax\|_v + \|Bx\|_v$$

which implies

$$\|(A + B)x\|_v \leq \max_{x \in C^n : \|x\|_v = 1} \|Ax\|_v + \max_{x \in C^n : \|x\|_v = 1} \|Bx\|_v$$

and

$$\max_{x \in C^n : \|x\|_v = 1} \|(A + B)x\|_v \leq \max_{x \in C^n : \|x\|_v = 1} \|Ax\|_v + \max_{x \in C^n : \|x\|_v = 1} \|Bx\|_v$$

So,

$$f(A + B) \leq f(A) + f(B)$$

The fourth axiom follows from the next inequalities

$$f(AB) = \max_{x \in C^n : \|x\|_v = 1} \|(AB)x\|_v = \|(AB)x_0\|_v = \|A(Bx_0)\|_v$$

$$\leq f(A) \|Bx_0\|_v \leq f(A) f(B) \|x_0\|_v = f(A) f(B)$$

□

Several properties of the induced norm given below turn out to be important in many practical implementations.

Proposition 5.9.

1. For any induced norm

$$f(I_{n\times n}) = 1 \tag{5.42}$$

2. If A is unitary

$$f(A) = 1$$

3. If $f(A)$ is the spectral norm then

$$f(A) = \sqrt{\max_{1\le i\le n} \lambda_i (A^*A)} = \sqrt{\max_{1\le i\le n} \lambda_i (AA^*)}$$

4. If U is a unitary matrix then

$$f(AU) = f(UA) = f(A)$$

5. For the vector norm (5.26) the corresponding induced norm is

$$f_1(A) = \max_{1\le j\le n} \sum_{i=1}^{n} |a_{ij}|$$

The proof of this proposition results in simple vector calculations and these therefore are omitted here.

Example 5.7.

$$f_1\left(\begin{bmatrix} 2 & -1 \\ 2 & 3 \end{bmatrix}\right) = \max\{3; 5\} = 5$$

6 Moore–Penrose Pseudoinverse

Contents

In this subsection we follow Albert (1972).

6.1 Classical least squares problem

Lemma 6.1. *Let x be a vector and $\mathcal{L}$ is a linear manifold in $\mathbb{R}^n$ (that is, if $x, y \in \mathcal{L}$, then $\alpha x + \beta y \in \mathcal{L}$ for any scalars α, β). Then if*

$$x = \hat{x} + \tilde{x} \tag{6.1}$$

where $\hat{x} \in \mathcal{L}$ and $\tilde{x} \perp \mathcal{L}$, then $\hat{x}$ is "nearest" to x, or, in other words, it is the ***projection*** *of x* ***to the manifold*** *$\mathcal{L}$.*

Proof. For any $y \in \mathcal{L}$ we have

$$\begin{aligned}\|x - y\|^2 &= \|\hat{x} + \tilde{x} - y\|^2 = \|(\hat{x} - y) + \tilde{x}\|^2 \\ &= \|\hat{x} - y\|^2 + 2(\hat{x} - y, \tilde{x}) + \|\tilde{x}\|^2 = \|\hat{x} - y\|^2 + \|\tilde{x}\|^2 \\ &\geq \|\tilde{x}\|^2 = \|x - \hat{x}\|^2\end{aligned}$$

with strict inequality holding unless $\|y - \hat{x}\|^2 = 0$. □

Theorem 6.1. *Let z be an n-dimensional real vector and $H \in \mathbb{R}^{n \times m}$.*

1. *There is always a vector, in fact a unique vector $\hat{x}$ of minimal (Euclidean) norm, which minimizes*

$$\|z - Hx\|^2$$

2. *The vector $\hat{x}$ is the unique vector in the range*

$$\mathcal{R}(H^\intercal) := \{x : x = H^\intercal z,\ z \in \mathbb{R}^n\}$$

which satisfies the equation

$$H\hat{x} = \hat{z}$$

where $\hat{z}$ is the projection of z on $\mathcal{R}(H)$.

Proof. By (6.1) we can write

$$z = \hat{z} + \tilde{z}$$

where $\hat{z}$ is the projection of z on the kernel (the null space)

$$\mathcal{N}(H^\intercal) := \{z \in \mathbb{R}^n : 0 = H^\intercal z\}$$

Since $Hx \in \mathcal{R}(H)$ for any $x \in \mathbb{R}^m$, it follows that

$$\hat{z} - Hx \in \mathcal{R}(H)$$

and, since $\tilde{z} \in \mathcal{R}^\perp(H)$,

$$\tilde{z} \perp \hat{z} - Hx$$

Therefore,

$$\begin{aligned}\|z - Hx\|^2 &= \|(\hat{z} - Hx) + \tilde{z}\|^2 \\ &= \|\hat{z} - Hx\|^2 + \|\tilde{z}\|^2 \geq \|\tilde{z}\|^2 = \|z - \hat{z}\|^2\end{aligned}$$

This low bound is attainable since $\hat{z}$, being the range of H, is the afterimage of some x^*, that is, $\hat{z} = Hx^*$.

1. Let us show that x^* has a minimal norm. Since x^* may be decomposed into two orthogonal vectors

$$x^* = \hat{x}^* + \tilde{x}^*$$

where $\hat{x}^* \in \mathcal{R}\left(H^\perp\right)$ and $\tilde{x}^* \in \mathcal{N}(H)$. Thus $Hx^* = H\hat{x}$ we have

$$\|z - Hx^*\|^2 = \|z - H\hat{x}\|^2$$

and

$$\|x^*\|^2 = \|\hat{x}^*\|^2 + \|\tilde{x}^*\|^2 \geq \|\hat{x}^*\|^2$$

with strict inequality unless $x^* = \hat{x}^*$. So, x^* may be selected equal to $\hat{x}^*$.

2. Show now that $x^* = \hat{x}^*$ is unique. Suppose that $Hx^* = Hx^{**} = \hat{z}$. Then

$$(x^* - x^{**}) \in \mathcal{R}(H)$$

But $H(x^* - x^{**}) = 0$, which implies,

$$(x^* - x^{**}) \in \mathcal{N}(H) = \mathcal{R}^{\perp}\left(H^{\perp}\right)$$

Thus $(x^* - x^{**})$ is orthogonal to itself, which means that $\|x^* - x^{**}\|^2 = 0$, or equivalently, $x^* = x^{**}$. □

Corollary 6.1. *$\|z - Hx\|^2$ is minimized by x_0 if and only if $Hx_0 = \hat{z}$ where $\hat{z}$ is the projection of z on $\mathcal{R}(H)$.*

Corollary 6.2. *There is always an n-dimensional vector y such that*

$$\|z - HH^{\mathsf{T}}y\|^2 = \inf_x \|z - Hx\|^2$$

and if

$$\|z - Hx_0\|^2 = \inf_x \|z - Hx\|^2$$

then

$$\|x_0\|^2 \geq \|H^{\mathsf{T}}y\|^2$$

with strict inequality unless $x_0 = H^{\mathsf{T}}y$. The vector y satisfies the equation

$$HH^{\mathsf{T}}y = \hat{z}$$

Theorem 6.2. (on the system of normal equations) *Among those vectors x, which minimize $\|z - Hx\|^2$, $\hat{x}$, the one having minimal norm, is the unique vector of the form*

$$\boxed{\hat{x} = H^{\mathsf{T}}y} \tag{6.2}$$

satisfying

$$\boxed{H^{\mathsf{T}}H\hat{x} = H^{\mathsf{T}}z} \tag{6.3}$$

Proof. By direct differentiation we have

$$\frac{\partial}{\partial x}\|z - Hx\|^2 = 2H^{\mathsf{T}}(z - Hx) = 0$$

which gives (6.3). The representation (6.2) follows from the previous corollary. □

6.2 Pseudoinverse characterization

We are now in the position to exhibit an explicit representation for the minimum norm solution to a least square problem.

Lemma 6.2. *For any real symmetric matrix $A \in \mathbb{R}^{n\times n}$ the limit*

$$\boxed{P_A := \lim_{\delta\to 0} (A + \delta I_{n\times n})^{-1} A} \tag{6.4}$$

always exists. For any vector $z \in \mathbb{R}^n$

$$\hat{z} = P_A z$$

is the ***projection*** *of x on $\mathcal{R}(A)$.*

Proof. By symmetricity of A for all $\delta > 0$ such that $0 < |\delta| < \min_{j:\lambda_j(A)\neq 0} |\lambda_j(A)|$ the matrix $(A + \delta I_{n\times n})^{-1}$ exists. Any $z \in \mathbb{R}^n$ may be represented as

$$z = \hat{z} + \tilde{z}$$

where $\hat{z} \in \mathcal{R}(A)$, $\tilde{z} \in \mathcal{N}(A)$ and $Az = A\hat{z}$. There exists x_0 such that $\hat{z} = Ax_0$, so

$$(A + \delta I_{n\times n})^{-1} Az = (A + \delta I_{n\times n})^{-1} A\hat{z} = (A + \delta I_{n\times n})^{-1} A (Ax_0)$$

By the spectral theorem (4.4) for symmetric matrices it follows that

$$A = T\Lambda T^{\intercal}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ and $T^{\intercal} = T^{-1}$. Thus

$$\begin{aligned}(A + \delta I_{n\times n})^{-1} Az &= (A + \delta I_{n\times n})^{-1} A^2 x_0 \\ &= (T\Lambda T^{\intercal} + \delta T T^{\intercal})^{-1} T\Lambda^2 T^{\intercal} x_0 \\ &= \left(T\left[\Lambda + \delta I_{n\times n}\right] T^{\intercal}\right)^{-1} T\Lambda^2 T^{\intercal} x_0 \\ &= T\left(\left[\Lambda + \delta I_{n\times n}\right]^{-1} \Lambda^2\right) T^{\intercal} x_0\end{aligned}$$

It is plain to see that

$$\begin{aligned}\lim_{\delta\to 0} \left[\Lambda + \delta I_{n\times n}\right]^{-1} \Lambda^2 &= \lim_{\delta\to 0} \left[\Lambda + \delta I_{n\times n}\right]^{-1} \left[\Lambda + \delta I_{n\times n} - \delta I_{n\times n}\right] \Lambda \\ &= \lim_{\delta\to 0} \left[I_{n\times n} - \delta \left[\Lambda + \delta I_{n\times n}\right]^{-1}\right] \Lambda \\ &= \left[\lim_{\delta\to 0} \operatorname{diag}\left(1 - \frac{\delta}{\lambda_1 + \delta}, \ldots, 1 - \frac{\delta}{\lambda_n + \delta}\right)\right] \Lambda = \Lambda\end{aligned}$$

since

$$1 - \frac{\delta}{\lambda_i + \delta} = \begin{cases} 0 & \text{if} \quad \lambda_i = 0 \\ \to 1 & \text{if} \quad \lambda_i \neq 0 \end{cases}$$

This implies

$$\lim_{\delta \to 0} (A + \delta I_{n\times n})^{-1} Az = T \Lambda T^{\intercal} x_0 = Ax_0 = \hat{z}$$

□

Theorem 6.3. *For any real* $(n \times m)$*-matrix* H *the limit*

$$\boxed{\begin{aligned} H^{+} &:= \lim_{\delta \to 0} \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} \\ &= \lim_{\delta \to 0} H^{\intercal} \left(H H^{\intercal} + \delta^2 I_{n\times n}\right)^{-1} \end{aligned}} \tag{6.5}$$

always exists. For any vector $z \in \mathbb{R}^n$

$$\hat{x} = H^{+} z$$

is the vector of minimal norm among those which minimize

$$\left\| z - Hx \right\|^2$$

Proof. It is clear that the right sides in (6.5) are equal, if either exists, since

$$\begin{aligned} H^{\intercal} H H^{\intercal} + \delta^2 H^{\intercal} &= \left(H^{\intercal} H + \delta^2 I_{m\times m}\right) H^{\intercal} \\ &= H^{\intercal} \left(H H^{\intercal} + \delta^2 I_{n\times n}\right) \end{aligned}$$

and the matrices $\left(H^{\intercal} H + \delta^2 I_{m\times m}\right)$ and $\left(H H^{\intercal} + \delta^2 I_{n\times n}\right)$ are inverse for any $\delta^2 > 0$. By the composition

$$z = \hat{z} + \tilde{z}$$

where $\hat{z} \in \mathcal{R}(H^{\intercal})$, $\tilde{z} \in \mathcal{N}(H^{\intercal})$ and $H^{\intercal} z = H^{\intercal} \hat{z}$, there exists x_0 such that $\hat{z} = Hx_0$. So,

$$\begin{aligned} \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} z &= \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} \hat{z} \\ &= \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} H x_0 \end{aligned}$$

By the previous Lemma there exists the limit

$$\lim_{\delta \to 0} \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} H = P_{H^{\intercal} H}$$

which gives

$$\lim_{\delta \to 0} \left(H^{\intercal} H + \delta^2 I_{m\times m}\right)^{-1} H^{\intercal} H x_0 = \left(P_{H^{\intercal} H}\right) x_0 := \hat{x}_0$$

where $\hat{x}_0$ is the projection on $\mathcal{R}(H^\top H) = \mathcal{R}(H^\top)$. Thus we conclude that

$$\begin{aligned}\hat{x}_0 &= \lim_{\delta\to 0}\left(H^\top H + \delta^2 I_{m\times m}\right)^{-1} H^\top \hat{z} \\ &= \lim_{\delta\to 0}\left(H^\top H + \delta^2 I_{m\times m}\right)^{-1} H^\top z\end{aligned}$$

always exists and is an element of $\mathcal{R}(H^\top)$ satisfying $\hat{z} = H\hat{x}_0$. □

Definition 6.1. *The matrix limit H^+ (6.5) is called the* ***pseudoinverse*** *(the generalized inverse) of H in the* ***Moore–Penrose sense.***

Remark 6.1. *It follows that*

- *(HH^+z) is the projection of z on $\mathcal{R}(H)$;*
- *(H^+Hx) is the projection of x on $\mathcal{R}(H^\top)$;*
- *$(I_{n\times n} - HH^+)z$ is the projection of z on $\mathcal{N}(H^\top)$;*
- *$(I_{n\times n} - H^+H)x$ is the projection of x on $\mathcal{N}(H)$.*

The following properties can be proven by the direct application of (6.5).

Corollary 6.3. *For any real $n \times m$ matrix H*

1.

$$\boxed{H^+ = (H^\top H)^+ H^\top} \tag{6.6}$$

2.

$$\boxed{(H^\top)^+ = \left(H^+\right)^\top} \tag{6.7}$$

3.

$$\boxed{H^+ = H^\top (HH^\top)^+} \tag{6.8}$$

4.

$$\boxed{H^+ = H^{-1}} \tag{6.9}$$

if H is square and nonsingular.

6.3 Criterion for pseudoinverse checking

The next theorem represents the **criterion** for a matrix B, to be the pseudoinverse H^+ of H.

Theorem 6.4. *For any real $n \times m$ matrix H the matrix $B = H^+$ if and only if*

1.

$$\boxed{HB \text{ and } BH \text{ are symmetric}} \tag{6.10}$$

2.

$$\boxed{HBH = H} \tag{6.11}$$

3.

$$\boxed{BHB = B} \tag{6.12}$$

Proof.

1. *Necessity.* Let $B = H^+$.
   (a) Since

$$HH^+ = \lim_{\delta \to 0} HH^{\mathsf{T}} \left(HH^{\mathsf{T}} + \delta^2 I_{n\times n}\right)^{-1}$$

$$\left(HH^+\right)^{\mathsf{T}} = \left(H \lim_{\delta \to 0} \left(H^{\mathsf{T}}H + \delta^2 I_{m\times m}\right)^{-1} H^{\mathsf{T}}\right)^{\mathsf{T}}$$

$$= H\left[\lim_{\delta \to 0} \left(H^{\mathsf{T}}H + \delta^2 I_{m\times m}\right)^{-1}\right] H^{\mathsf{T}} = HH^+$$

   and

$$H^+H = \left[\lim_{\delta \to 0} H^{\mathsf{T}} \left(HH^{\mathsf{T}} + \delta^2 I_{n\times n}\right)^{-1}\right] H$$

$$\left(H^+H\right)^{\mathsf{T}} = \left(\left[\lim_{\delta \to 0} H^{\mathsf{T}} \left(HH^{\mathsf{T}} + \delta^2 I_{n\times n}\right)^{-1}\right] H\right)^{\mathsf{T}}$$

$$H^{\mathsf{T}} \lim_{\delta \to 0} \left(H^{\mathsf{T}}H + \delta^2 I_{m\times m}\right)^{-1} H = H^+H$$

   the symmetricity (6.10) takes place.

   (b) Since by (6.1) HH^+ is a projector on $\mathcal{R}(H)$ and the projection of any vector from $\mathcal{R}(H)$ coincides with the same vector, one has for any $z \in \mathbb{R}^n$

$$HH^+(Hz) = Hz$$

   which gives (6.11). By (6.6)

$$H^+H = (H^{\mathsf{T}}H)^+ (H^{\mathsf{T}}H)$$

which, in view of (6.11) and the symmetricity property (6.10) of HH^+, implies

$$\begin{aligned} H^+ &= (H^\intercal H)^+ H^\intercal = (H^\intercal H)^+ \left(HH^+H\right)^\intercal \\ &= (H^\intercal H)^+ H^\intercal \left(HH^+\right)^\intercal = (H^\intercal H)^+ H^\intercal \left(HH^+\right) \\ &= (H^\intercal H)^+ (H^\intercal H)\, H^+ = H^+HH^+ \end{aligned}$$

So, (6.12) is proven.

2. *Sufficiency*. Suppose B satisfies (6.10), (6.11) and (6.12). Since

$$BH = (BH)^\intercal, \quad H = HBH$$

then

$$H = HBH = H\,(BH)^\intercal$$

Using this representation and since $HH^+H = H$, we derive

$$H^+H = H^+H\,(BH)^\intercal = \left[HH^+H\right]^\intercal B^\intercal = H^\intercal B^\intercal = BH \tag{6.13}$$

Analogously, since $B = BHB$ and HB is symmetric, we have

$$B^\intercal = HBB^\intercal$$

Pre-multiplying this identity by HH^+, we obtain

$$HH^+B^\intercal = HH^+HBB^\intercal = HBB^\intercal = B^\intercal$$

Taking transposes and in view of (6.13) we get

$$B = B\left(HH^+\right)^\intercal = B\left(HH^+\right) = BHH^+ = H^+HH^+ = H^+$$

□

The theorem above is extremely useful as a method for proving identities. If one thinks that a certain expression coincides with the pseudoinverse of a certain matrix H, a good way to decide is to run the expressions through conditions (6.10), (6.11), (6.12) and observe whether or not they hold.

6.4 Some identities for pseudoinverse matrices

Lemma 6.3. $b \in \mathcal{R}(A) := \mathrm{Im}(A) \subseteq \mathbb{R}^n$ *if and only if*

$$\boxed{AA^+b = b} \tag{6.14}$$

Proof.

(a) *Necessity*. If $b \in \mathcal{R}(A)$, then there exists a vector $d \in \mathbb{R}^n$ such that $b = Ad$, and, therefore,

$$AA^+b = AA^+(Ad) = \left(AA^+A\right)d = Ad = b$$

(b) *Sufficiency*. Suppose that (6.14) is true. Any vector b can be represented as $b = Ad + b^{\perp}$ with $b^{\perp} \perp Ad$, namely, $b^{\perp} = \left(I - AA^{+}\right) v$. Then

$$AA^{+}\left(Ad + b^{\perp}\right) = Ad + b^{\perp}$$

which implies $AA^{+}b^{\perp} = b^{\perp}$, and, hence,

$$AA^{+}\left(I - AA^{+}\right) v = 0 = \left(I - AA^{+}\right) v = b^{\perp}$$

So, $b^{\perp} = 0$, and, hence, $b = Ad$, or, equivalently, $b \in \mathcal{R}(A)$. Lemma is proven. □

The following identities can be proven more easily by simple verification of (6.10), (6.11), (6.12).

Claim 6.1.

1.

$$\boxed{(\mathrm{O}_{m\times n})^{+} = \mathrm{O}_{n\times m}} \tag{6.15}$$

2. For any $x \in \mathbb{R}^n$ $(x \neq 0)$

$$\boxed{x^{+} = \frac{x^{\mathsf{T}}}{\|x\|^2}} \tag{6.16}$$

3.

$$\boxed{\left(H^{+}\right)^{+} = H} \tag{6.17}$$

4. In general,

$$\boxed{(AB)^{+} \neq B^{+}A^{+}} \tag{6.18}$$

The identity takes place if

$$\boxed{\begin{array}{r} A^{\mathsf{T}}A = I \quad \textit{or} \\ BB^{\mathsf{T}} = I \quad \textit{or} \\ B = A^{\mathsf{T}} \quad \textit{or} \\ B = A^{+} \quad \textit{or} \\ \textit{both } A \textit{ and } B \textit{ are of full rank, or} \\ \operatorname{rank} A = \operatorname{rank} B \end{array}}$$

The identity in (6.18) holds ***if and only if***

$$\boxed{\mathcal{R}(BB^{\mathsf{T}}A^{\mathsf{T}}) \subseteq \mathcal{R}(A^{\mathsf{T}})} \tag{6.19}$$

and

$$\boxed{\mathcal{R}(A^{\intercal}AB) \subseteq \mathcal{R}(B)} \tag{6.20}$$

5.

$$\boxed{(AB)^{+} = B_1^{+}A_1^{+}} \tag{6.21}$$

where

$$\boxed{\begin{aligned} B_1 &= A^{+}AB \\ A_1 &= AB_1B_1^{+} \end{aligned}}$$

6.

$$\boxed{(H^{\intercal}H)^{+} = H^{+}(H^{\intercal})^{+}, (HH^{\intercal})^{+} = (H^{\intercal})^{+}H^{+}} \tag{6.22}$$

7. *If A is symmetric and $\alpha > 0$, then*

$$\boxed{\begin{aligned} (A^{\alpha})^{+} &= (A^{+})^{\alpha} \\ A^{\alpha}(A^{\alpha})^{+} &= (A^{\alpha})^{+}A^{\alpha} = AA^{+} \\ A^{+}A^{\alpha} &= A^{\alpha}A^{+} \end{aligned}} \tag{6.23}$$

8. *If $A = U\Lambda V^{\intercal}$ where U, V are orthogonal and Λ is a diagonal matrix, then*

$$\boxed{A^{+} = V\Lambda^{+}U^{\intercal}} \tag{6.24}$$

9. *Greville's formula (Greville 1960): if $C_{m+1} = \left[C_m \vdots c_{m+1}\right]$ then*

$$\boxed{C_{m+1}^{+} = \begin{bmatrix} C_m^{+}\left[I - c_{m+1}k_{m+1}^{\intercal}\right] \\ \cdots\cdots\cdots\cdots \\ k_{m+1}^{\intercal} \end{bmatrix}} \tag{6.25}$$

where

$$k_{m+1} = \begin{cases} \dfrac{\left[I - C_mC_m^{+}\right]c_{m+1}}{\left\|\left[I - C_mC_m^{+}\right]c_{m+1}\right\|^2} & \text{if} \quad \left[I - C_mC_m^{+}\right]c_{m+1} \neq 0 \\ \dfrac{\left(C_m^{+}\right)^{\intercal}C_m^{+}c_{m+1}}{1 + \left\|C_m^{+}c_{m+1}\right\|^2} & \text{otherwise} \end{cases}$$

10. *If H is rectangular and S is symmetric and nonsingular then*

$$\boxed{(SH)^{+} = H^{+}S^{-1}\left[I - \left(QS^{-1}\right)^{+}\left(QS^{-1}\right)\right]} \tag{6.26}$$

where

$$Q = I - H^{+}H$$

Example 6.1. *Simple verification of (6.10), (6.11), (6.12) shows that*

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix}^{+} = \begin{bmatrix} -1.333 & -0.333 & 0.667 \\ 1.8033 & 0.333 & -0.4167 \end{bmatrix}$$

6.5 Solution of least squares problem using pseudoinverse

Theorem 6.5.

(a) The vector x_0 minimizes $\|z - Hx\|^2$ if and only if it is of the form

$$\boxed{x_0 = H^{+}z + (I - H^{+}H)\, y} \tag{6.27}$$

for some vector y.

(b) Among all solutions x_0 (6.27) the vector

$$\boxed{\bar{x}_0 = H^{+}z} \tag{6.28}$$

has the minimal Euclidean norm.

Proof. By theorem (6.3) we know that $H^{+}z$ minimizes $\|z - Hx\|^2$ and by (6.1), any x_0 minimizes $\|z - Hx\|^2$ if and only if $Hx_0 = \hat{z}$ where $\hat{z}$ is the projection of z on $\mathcal{R}(H)$. In view of that

$$Hx_0 = H\left(H^{+}z\right)$$

This means that $x_0 - H^{+}z$ is a null vector of H that is true if and only if

$$x_0 - H^{+}z = (I - H^{+}H)\, y$$

for some y. So, (a) (6.27) is proven. To prove (b) (6.28) it is sufficient to notice that

$$\begin{aligned} \|x_0\|^2 &= \|H^{+}z + (I - H^{+}H)\, y\|^2 = \|H^{+}z\|^2 \\ &\quad + (H^{+}z, (I - H^{+}H)\, y) + \|(I - H^{+}H)\, y\|^2 = \|H^{+}z\|^2 \\ &\quad + ((I - H^{+}H)^{\mathsf{T}} H^{+}z, y) + \|(I - H^{+}H)\, y\|^2 = \|H^{+}z\|^2 \\ &\quad + ((I - H^{+}H)\, H^{+}z, y) + \|(I - H^{+}H)\, y\|^2 \\ &= \|H^{+}z\|^2 + \|(I - H^{+}H)\, y\|^2 \geq \|H^{+}z\|^2 = \|\bar{x}_0\|^2 \end{aligned}$$

□

Corollary 6.4. (LS problem with constraints) *Suppose the set*

$$\mathcal{J} = \{x : Gx = u\}$$

is not empty. Then the vector x_0 *minimizes* $\|z - Hx\|^2$ *over* $\mathcal{J}$ *if and only if*

$$\boxed{\begin{array}{c} x_0 = G^+u + \bar{H}^+z + \left(I - G^+G\right)\left(I - \bar{H}^+\bar{H}\right)y \\ \bar{H} := H\left(I - G^+G\right) \end{array}} \tag{6.29}$$

and among all solutions

$$\boxed{\bar{x}_0 = G^+u + \bar{H}^+z} \tag{6.30}$$

has the minimal Euclidean norm.

Proof. Notice that by the Lagrange multipliers method x_0 solves the problem if it minimizes the Lagrange function

$$\|z - Hx\|^2 + (\lambda, Gx - u)$$

for some λ. This λ and x_0 satisfy the equation

$$\frac{\partial}{\partial x}\left[\|z - Hx\|^2 + (\lambda, Gx - u)\right] = -2H^\intercal(z - Hx_0) + G^\intercal\lambda = 0$$

or, equivalently,

$$H^\intercal Hx_0 = \left[H^\intercal z - \frac{1}{2}G^\intercal\lambda\right] := \tilde{z}$$

which in view of (6.27) implies

$$\begin{aligned} x_0 &= (H^\intercal H)^+\tilde{z} + \left[I - (H^\intercal H)^+(H^\intercal H)\right]y \\ &= (H^\intercal H)^+\left[H^\intercal z - \frac{1}{2}G^\intercal\lambda\right] + \left[I - (H^\intercal H)^+(H^\intercal H)\right]y \end{aligned} \tag{6.31}$$

But this x_0 should satisfy $Gx_0 = u$ which leads to the following equality

$$Gx_0 = G\left[(H^\intercal H)^+\left(H^\intercal z - \frac{1}{2}G^\intercal\lambda\right) + \left[I - (H^\intercal H)^+(H^\intercal H)\right]y\right] = u$$

or, equivalently,

$$\begin{aligned} \left[G(H^\intercal H)^+\right]\left(\frac{1}{2}G^\intercal\lambda\right) = {} & G(H^\intercal H)^+H^\intercal z \\ & + G\left[I - (H^\intercal H)^+(H^\intercal H)\right]y - u \end{aligned}$$

or

$$\begin{aligned}\frac{1}{2}G^{\mathsf{T}}\lambda = & \left[G\left(H^{\mathsf{T}}H\right)^{+}\right]^{+}\left[G\left(H^{\mathsf{T}}H\right)^{+}H^{\mathsf{T}}z\right.\\ & + G\left[I-\left(H^{\mathsf{T}}H\right)^{+}\left(H^{\mathsf{T}}H\right)\right]y-u\Big]\\ & \left.+\left[I-\left[G\left(H^{\mathsf{T}}H\right)^{+}\right]^{+}\left[G\left(H^{\mathsf{T}}H\right)^{+}\right]\right]\tilde{y}\end{aligned} \tag{6.32}$$

Substitution of (6.32) into (6.31) and using the properties of the pseudoinverse implies (6.29). The statement (6.30) is evident. □

6.6 Cline's formulas

In fact, the direct verification leads to the following identities (see Cline (1964, 1965)).

Claim 6.2. (Pseudoinverse of a partitioned matrix)

$$\boxed{\left[U\,\vdots\,V\right]^{+}=\begin{bmatrix}U^{+}-U^{+}VJ\\ \cdots\cdots\cdots\\ J\end{bmatrix}} \tag{6.33}$$

where

$$\boxed{\begin{aligned}J&=C^{+}+\left(I-C^{+}C\right)KV^{\mathsf{T}}\left(U^{+}\right)^{\mathsf{T}}U^{+}\left(I-VC^{+}\right)\\ C&=\left(I-UU^{+}\right)V\\ K&=\left(I+\left[U^{+}V\left(I-C^{+}C\right)\right]^{\mathsf{T}}\left[U^{+}V\left(I-C^{+}C\right)\right]\right)^{-1}\end{aligned}} \tag{6.34}$$

Claim 6.3. (Pseudoinverse of sums of matrices)

$$\boxed{\begin{aligned}&\left(UU^{\mathsf{T}}+VV^{\mathsf{T}}\right)^{+}=\left(CC^{\mathsf{T}}\right)^{+}+\left[I-\left(VC^{+}\right)^{\mathsf{T}}\right]\\ &\times\left[\left(UU^{\mathsf{T}}\right)^{+}-\left(UU^{\mathsf{T}}\right)^{+}V\left(I-C^{+}C\right)KV^{\mathsf{T}}\left(UU^{\mathsf{T}}\right)^{+}\right]\\ &\times\left[I-\left(VC^{+}\right)^{\mathsf{T}}\right]\end{aligned}} \tag{6.35}$$

where C and K are defined in (6.34).

6.7 Pseudo-ellipsoids

6.7.1 Definition and basic properties

Definition 6.2. *We say that the set $\varepsilon(\mathring{x},A)\in\mathbb{R}^n$ is* ***the pseudo-ellipsoid*** *(or* ***elliptic cylinder****) in $\mathbb{R}^n$ with the center at the point $\mathring{x}\in\mathbb{R}^n$ and with the matrix $0\le A=A^{\mathsf{T}}\in\mathbb{R}^{n\times n}$ if it is defined by*

$$\boxed{\varepsilon(\mathring{x},A):=\left\{x\in\mathbb{R}^n\mid\|x-\mathring{x}\|_A^2=\left(x-\mathring{x},A\left(x-\mathring{x}\right)\right)\le 1\right\}} \tag{6.36}$$

If $A > 0$ *the set* $\varepsilon(\mathring{x}, A)$ *is an* ***ordinary ellipsoid*** *with the semi-axis equal to* $\lambda_i^{-1}(A)$ $(i = 1, \ldots, n)$.

Remark 6.2. *If*

(a) $A > 0$, *then* $\varepsilon(\mathring{x}, A)$ *is a bounded set;*
(b) $A \geq 0$, *then* $\varepsilon(\mathring{x}, A)$ *is an unbounded set.*

Lemma 6.4. *If* $0 < A = A^{\intercal} \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ *and* $\alpha < 1 - \|b\|_{A^{-1}}^2$, *then the set given by*

$$(x, Ax) - 2(b, x) + \alpha \leq 1$$

is the ellipsoid $\varepsilon\left(A^{-1}b, \dfrac{1}{1 - \alpha + \|b\|_{A^{-1}}^2} A\right)$.

Proof. It follows from the identity

$$\left\|x - A^{-1}b\right\|_A^2 - \|b\|_{A^{-1}}^2 = \|x\|_A^2 - 2(b, x)$$

□

Lemma 6.5. *If* $0 \leq A = A^{\intercal} \in \mathbb{R}^{n \times n}$, $b \in \mathcal{R}(A) \subseteq \mathbb{R}^n$ *and* $\alpha < 1 - \|b\|_{A^+}^2$, *then the set given by*

$$(x, Ax) - 2(b, x) + \alpha \leq 1$$

is the pseudo-ellipsoid $\varepsilon\left(A^{+}b, \dfrac{1}{1 - \alpha + \|b\|_{A^+}^2} A\right)$.

Proof. It follows from the identity

$$\left\|x - A^{+}b\right\|_A^2 - \|b\|_{A^+}^2 = \|x\|_A^2 - 2(b, x)$$

□

Lemma 6.6.

$$\boxed{\varepsilon\left(A^{+}A\mathring{x}, A\right) = \varepsilon(\mathring{x}, A)} \tag{6.37}$$

Proof. Indeed,

$$\begin{aligned}
&\left(x - A^{+}A\mathring{x}, A\left(x - A^{+}A\mathring{x}\right)\right) = \left(x - A^{+}A\mathring{x}, Ax - A\mathring{x}\right) \\
&(x, Ax) - \left(A^{+}A\mathring{x}, Ax\right) - \left(x, A\mathring{x}\right) + \left(A^{+}A\mathring{x}, A\mathring{x}\right) \\
&\quad = (x, Ax) - 2(x, A\mathring{x}) + (\mathring{x}, A\mathring{x}) = (x - \mathring{x}, A(x - \mathring{x}))
\end{aligned}$$

□

6.7.2 Support function

Definition 6.3. *The function* $f_S : \mathbb{R}^n \to \mathbb{R}$ *defined by*

$$\boxed{f_S(y) := \max_{x \in S} (y, x)} \tag{6.38}$$

is called the ***support*** *(or Legendre)* ***function*** *(SF) of the convex closed set* $S \subseteq \mathbb{R}^n$.

Lemma 6.7. *If* S *is the pseudo-ellipsoid* $\varepsilon(\mathring{x}, A)$ *(6.36), that is,*

$$S = \varepsilon(\mathring{x}, A) = \left\{x \in \mathbb{R}^n \mid \|x - \mathring{x}\|_A^2 = (x - \mathring{x}, A(x - \mathring{x})) \le 1\right\}$$

then

$$\boxed{f_S(y) = y^{\mathsf{T}} \mathring{x} + \sqrt{y^{\mathsf{T}} A^+ y}} \tag{6.39}$$

Proof. Using the Lagrange principle (see Theorem 21.12), for any $y \in \mathbb{R}^n$ we have

$$\arg\max_{x \in S} (y, x) = \arg\min_{\lambda \ge 0} \max_{x \in \mathbb{R}^n} L(x, \lambda \mid y)$$
$$L(x, \lambda \mid y) := (y, x) + \lambda \left[(x - \mathring{x}, A(x - \mathring{x})) - 1\right]$$

and, therefore, the extremal point (x^*, λ^*) satisfies

$$0 = \frac{\partial}{\partial x} L(x^*, \lambda^* \mid y) = y + \lambda^* A(x^* - \mathring{x})$$
$$\lambda^* \left[(x^* - \mathring{x}, A(x^* - \mathring{x})) - 1\right] = 0$$

The last identity is referred to as the *complementary slackness condition*. The x satisfying the first equation can be represented as follows

$$\arg_{x \in \mathbb{R}^n} \{y + \lambda A(x - \mathring{x}) = 0\} = \arg\min_{x \in \mathbb{R}^n} \|y + \lambda A(x - \mathring{x})\|^2$$

If $\lambda = 0$, it follows that $y = 0$. But $L(x, \lambda \mid y)$ is defined for any $y \in \mathbb{R}^n$. So, $\lambda > 0$, and hence, by (6.27),

$$x^* - \mathring{x} = \frac{1}{\lambda^*} A^+ y + \left(I - A^+ A\right) v, \quad v \in \mathbb{R}^n$$

Substitution of this expression in the complementary slackness condition and taking into account that $A^+ = \left(A^+\right)^{\mathsf{T}}$ implies

$$1 = \left(\frac{1}{\lambda^*} A^+ y + \left(I - A^+ A\right) v, \frac{1}{\lambda^*} A A^+ y\right)$$
$$= \frac{1}{(\lambda^*)^2} \left(y, A^+ A A^+ y\right) + \left(A^+ A \left(I - A^+ A\right) v, \frac{1}{\lambda^*} y\right)$$
$$= \frac{1}{(\lambda^*)^2} \left(y, A^+ y\right)$$

or, equivalently, $\lambda^* = \sqrt{(y, A^+ y)}$, which finally gives

$$\begin{aligned} f_S(y) &= \max_{x \in S} (y, x) = L(x^*, \lambda^* \mid y) = (y, x^*) \\ &= \left(y, \mathring{x} + \frac{1}{\lambda^*} A^+ y + \left(I - A^+ A\right) v \right) \\ &= (y, \mathring{x}) + \frac{1}{\lambda^*} \left(y, A^+ y\right) + \left(y, \left(I - A^+ A\right) v\right) \\ &= (y, \mathring{x}) + \frac{\left(y, A^+ y\right)}{\sqrt{(y, A^+ y)}} - \left(\lambda A (x - \mathring{x}), \left(I - A^+ A\right) v\right) \\ &= (y, \mathring{x}) + \frac{\left(y, A^+ y\right)}{\sqrt{(y, A^+ y)}} - \left(\lambda (x - \mathring{x}), A \left(I - A^+ A\right) v\right) \\ &= (y, \mathring{x}) + \frac{\left(y, A^+ y\right)}{\sqrt{(y, A^+ y)}} \end{aligned}$$

Lemma is proven. □

6.7.3 *Pseudo-ellipsoids containing vector sum of two pseudo-ellipsoids*

The support function $f_S(y)$ (6.38) is particularly useful since the vector sum of convex closed sets and a linear transformation $\bar{A}$ of S have their simple counterparts in the support function description (see Appendix in Schlaepfer & Schweppe (1972)).

Lemma 6.8. (on SF for the vector sum of convex sets) *Let*

$$S_1 \oplus S_2 := \{x \in \mathbb{R}^n \mid x = x_1 + x_2, x_1 \in S_1, x_2 \in S_2\}$$

where S_1, S_2 are convex closed sets. Then

$$\boxed{f_{S_1 \oplus S_2}(y) = f_{S_1}(y) + f_{S_2}(y)} \tag{6.40}$$

Proof. By (6.38), it follows

$$\begin{aligned} f_{S_1 \oplus S_2}(y) &= \max_{x \in S} (y, x) = \max_{x_1 \in S_1, x_2 \in S_2} ((y, x_1) + (y, x_2)) \\ &= \max_{x_1 \in S_1} (y, x_1) + \max_{x_2 \in S_2} (y, x_2) = f_{S_1}(y) + f_{S_2}(y) \end{aligned}$$

which completes the proof. □

Lemma 6.9. (on SF for a linear transformation) *Let*

$$\mathcal{B}_S := \{x \in \mathbb{R}^n \mid x = Bz, z \in S\}$$

where $B \in \mathbb{R}^{n \times n}$ is an $(n \times n)$ matrix. Then

$$\boxed{f_{\mathcal{B}_S}(y) = f_S(B^\top y)} \tag{6.41}$$

Proof. By (6.38), we have

$$f_{\mathcal{B}_S}(y) = \max_{x \in \mathcal{B}_S} (y, x) = \max_{z \in S} (y, Bz) = \max_{z \in S} (B^{\mathsf{T}} y, z) = f_S(B^{\mathsf{T}} y)$$

which proves the lemma. □

Lemma 6.10. (on SF for closed sets) *If two convex closed sets are related as $S_1 \supseteq S_2$, then for all $y \in S$*

$$\boxed{f_{S_1}(y) \geq f_{S_2}(y)} \tag{6.42}$$

Proof. It follows directly from the definition (6.38):

$$f_{S_1}(y) = \max_{x \in S_1} (y, x) \geq \max_{x \in S_2} (y, x) = f_{S_2}(y)$$

□

Lemma 6.11. (on SF for the vector sum of two ellipsoids) *Let S_1 and S_2 be two pseudo-ellipsoids, that is, $S_i = \varepsilon(\mathring{x}_i, A_i)$ $(i = 1, 2)$. Then*

$$\boxed{f_{S_1 \oplus S_2}(y) = f_S(y) = y^{\mathsf{T}} (\mathring{x}_1 + \mathring{x}_2) + \sqrt{y^{\mathsf{T}} A_1^{+} y} + \sqrt{y^{\mathsf{T}} A_2^{+} y}} \tag{6.43}$$

Proof. It results from (6.39) and (6.40). □

To bound $S_1 \oplus S_2$ by some pseudo-ellipsoid $S^*_{S_1 \oplus S_2}$ means to find $(\mathring{x}^*, A^*)$ such that (see Lemma 6.11) for all $y \in \mathbb{R}^n$

$$\boxed{y^{\mathsf{T}} (\mathring{x}_1 + \mathring{x}_2) + \sqrt{y^{\mathsf{T}} A_1^{+} y} + \sqrt{y^{\mathsf{T}} A_2^{+} y} \leq y^{\mathsf{T}} \mathring{x}^* + \sqrt{y^{\mathsf{T}} (A^*)^{+} y}} \tag{6.44}$$

Lemma 6.12. *The choice*

$$\boxed{\begin{array}{l} \mathring{x}^* = \mathring{x}_1 + \mathring{x}_2 \\ A^* = \left(\gamma^{-1} A_1^{+} + (1 - \gamma)^{-1} A_2^{+}\right)^{+}, \quad \gamma \in (0, 1) \end{array}} \tag{6.45}$$

is sufficient to satisfy (6.44).

Proof. Taking $\mathring{x}^* = \mathring{x}_1 + \mathring{x}_2$, we should to prove that

$$\sqrt{y^{\mathsf{T}} A_1^{+} y} + \sqrt{y^{\mathsf{T}} A_2^{+} y} \leq \sqrt{y^{\mathsf{T}} (A^*)^{+} y}$$

or, equivalently,

$$\left(\sqrt{y^{\mathsf{T}} A_1^{+} y} + \sqrt{y^{\mathsf{T}} A_2^{+} y}\right)^2 \leq y^{\mathsf{T}} (A^*)^{+} y$$

for some A^*. Applying the inequality (12.2), for any $\varepsilon > 0$ we have

$$\left(\sqrt{y^\intercal A_1^+ y} + \sqrt{y^\intercal A_2^+ y}\right)^2 \le (1+\varepsilon)\, y^\intercal A_1^+ y + \left(1+\varepsilon^{-1}\right) y^\intercal A_2^+ y$$
$$= y^\intercal \left[(1+\varepsilon)\, A_1^+ + \left(1+\varepsilon^{-1}\right) A_2^+\right] y$$

Denoting $\gamma^{-1} := (1+\varepsilon)$ and taking into account the identity (6.17) we get (6.45). □

6.7.4 *Pseudo-ellipsoids containing intersection of two pseudo-ellipsoids*

If $S_i = \varepsilon\,(\mathring{x}_i, A_i)$ $(i = 1, 2)$ are two pseudo-ellipsoids, then $S_1 \cap S_2$ is not a pseudo-ellipsoid. Sure, there exists a lot of pseudo-ellipsoids $S^*_{S_1 \cap S_2}$ (in fact, a set) containing $S_1 \cap S_2$. To bound $S_1 \cap S_2$ by $S^*_{S_1 \cap S_2}$ means to find $(\mathring{x}^*, A^*)$ such that

$$S_1 \cap S_2 \subseteq S^*_{S_1 \cap S_2}$$

where

$$\begin{aligned} &S_1 \cap S_2 := \{x \in \mathbb{R}^n \mid (x - \mathring{x}_1, A_1\,(x - \mathring{x}_1)) \le 1 \\ &\text{and} \quad (x - \mathring{x}_2, A_2\,(x - \mathring{x}_2)) \le 1\} \\ &S^*_{S_1 \cap S_2} := \{x \in \mathbb{R}^n \mid (x - \mathring{x}^*, A^*\,(x - \mathring{x}^*)) \le 1\} \end{aligned} \tag{6.46}$$

Lemma 6.13. *Let $S_1 \cap S_2 \neq \varnothing$. Then $(\mathring{x}^*, A^*)$ can be selected as follows*

$$\boxed{\begin{aligned} \mathring{x}^* &= A_\gamma^+ b_\gamma \\ b_\gamma &= \gamma A_1 \mathring{x}_1 + (1-\gamma) A_2 \mathring{x}_2 \\ A_\gamma &= \gamma A_1 + (1-\gamma) A_2, \ \gamma \in (0,1) \end{aligned}} \tag{6.47}$$

and

$$\boxed{\begin{aligned} A^* &= \frac{1}{\beta_\gamma} A_\gamma \\ \beta_\gamma &= 1 - \alpha_\gamma + \|b_\gamma\|^2_{A_\gamma^+} \\ \alpha_\gamma &= \gamma\,(\mathring{x}_1, A_1 \mathring{x}_1) + (1-\gamma)\,(\mathring{x}_2, A_2 \mathring{x}_2) \end{aligned}} \tag{6.48}$$

Proof. Notice that $S^*_{S_1 \cap S_2}$ can be selected as

$$\begin{aligned} S_1 \cap S_2 := \{x \in \mathbb{R}^n \mid &\gamma\,(x - \mathring{x}_1, A_1\,(x - \mathring{x}_1)) \\ &+ (1-\gamma)\,(x - \mathring{x}_2, A_2\,(x - \mathring{x}_2)) \le 1\}, \ \gamma \in (0,1) \end{aligned}$$

Straightforward calculations imply

$$\gamma\,(x - \mathring{x}_1, A_1\,(x - \mathring{x}_1)) + (1-\gamma)\,(x - \mathring{x}_2, A_2\,(x - \mathring{x}_2))$$
$$= \left(x, A_\gamma x\right) - 2\left(b_\gamma, x\right) + \alpha_\gamma \le 1$$

Applying Lemma 6.5 we get (6.48). Lemma is proven. □

7 Hermitian and Quadratic Forms

Contents

7.1 Definitions

Let $A \in \mathbb{C}^{n\times n}$ be a *Hermitian* matrix $\left(A = A^* := \left(\bar{A}\right)^{\mathsf{T}}\right)$ and $x \in \mathbb{C}^n$.

Definition 7.1. *The function*

$$\boxed{f_A(x) := (Ax, x)} \tag{7.1}$$

is called

- *the* ***Hermitian form***
- *converting to the* ***quadratic form*** *if* $A \in \mathbb{R}^{n\times n}$ *is a symmetric matrix* $(A = A^{\mathsf{T}})$ *and* $x \in \mathbb{R}^n$.

If $\mathcal{E} = \left\{x^{(1)}, \ldots, x^{(n)}\right\}$ is a basic in $\mathbb{C}^n$ such that

$$x = \sum_{i=1}^{n} \alpha_i x^{(i)}$$

then $f_A(x)$ may be represented as

$$\begin{aligned} f_A(x) := (Ax, x) &= \left(A\sum_{i=1}^{n}\alpha_i x^{(i)}, \sum_{j=1}^{n}\alpha_j x^{(j)}\right) \\ &= \left(\sum_{i=1}^{n}\alpha_i A x^{(i)}, \sum_{j=1}^{n}\alpha_j x^{(j)}\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_{ij}\alpha_i\bar{\alpha}_j \end{aligned} \tag{7.2}$$

where

$$\gamma_{ij} = \left(Ax^{(i)}, x^{(j)}\right), \quad i, j = 1, \ldots, n$$

If the basis $\mathcal{E}$ is the *standard basis* (orthonormal in the sense of the standard inner product, i.e., $\left(x^{(i)}, x^{(j)}\right) = \delta_{ij}$), then (7.2) becomes

$$f_A(x) := (Ax, x) = \sum_{i=1}^{n} \gamma_{ii} |\alpha_i|^2 \tag{7.3}$$

Conjecture 7.1. (Sylvester's law of inertia for quadratic forms) *The positive $\pi(A)$ and negative $\nu(A)$ squares in (7.3) are invariants of the Hermitian form $f_A(x)$ independent of an orthogonal basis in $\mathbb{C}^n$, namely,*

$$f_A(x) = \sum_{i=1}^{\pi(A)} |\alpha_i'|^2 - \sum_{j=1}^{\nu(A)} |\alpha_j'|^2 = \sum_{i=1}^{\pi(A)} |\alpha_i''|^2 - \sum_{j=1}^{\nu(A)} |\alpha_j''|^2$$

where

$$x = \sum_{i=1}^{n} \alpha_i' x^{(i)'}, \quad x = \sum_{i=1}^{n} \alpha_i'' x^{(i)''}$$

and $\left\{x^{(i)'}\right\}$, $\left\{x^{(i)''}\right\}$ are the orthogonal bases in $\mathbb{C}^n$.

Proof. Suppose that T is a unitary matrix $\left(T^* = T^{-1}\right)$ transforming an orthogonal basis $\mathcal{E}$ to another orthogonal one $\mathcal{E}'$, i.e.,

$$\left(x^{(1)'}, \ldots, x^{(n)'}\right) = T\left(x^{(1)}, \ldots, x^{(n)}\right)$$
$$\left(x^{(i)'}, x^{(j)'}\right) = \left(Tx^{(i)}, Tx^{(j)}\right) = \left(T^*Tx^{(i)}, x^{(j)}\right) = \left(x^{(i)}, x^{(j)}\right) = \delta_{ij}$$

Then by (7.3)

$$f_A(x) := (Ax, x) = \sum_{i=1}^{n} \gamma_{ii}' |\alpha_i'|^2$$
$$\gamma_{ij} = \left(Ax^{(i)'}, x^{(j)'}\right) = \left(ATx^{(i)}, Tx^{(j)}\right) = \left(T^*ATx^{(i)}, x^{(j)}\right)$$

Then by theorem on congruent Hermitian matrices there always exists a nonsingular matrix P such that

$$PAP^* = \operatorname{diag}[I_t, -I_{r-t}, 0] := \Lambda_0(A)$$

and (7.3) under the transformation $x = Px'$ becomes

$$f_A(x) := (Ax, x) = \left(P^*APx^{(i)}, x^{(j)}\right)$$

$$= \sum_{i=1}^{n} \gamma_{ii} |\alpha_i'|^2 = \sum_{i=1}^{\pi(A)} |\alpha_i'|^2 - \sum_{j=1}^{\nu(A)} |\alpha_j'|^2$$

For any other basis the unitary transformation U provides

$$f_A(x) := (Ax, x) = \left(U^* P^* A P U x'', x''\right)$$

which by (7.3) does not change the invariant indices In A. Theorem is proven. □

Claim 7.1. *If $A = A^\intercal$ is real and $x = u + iv$, then*

$$\boxed{f_A(x) := (Ax, x) = f_A(u) + f_A(v)} \tag{7.4}$$

Corollary 7.1. (on the extension)

- *Any real quadratic $f_A(u)$ can be uniquely extended up to the corresponding Hermitian form $f_A(x)$, using formula (7.4).*
- *It is very convenient to realize this extension by changing the product $u_i u_j$ with $\operatorname{Re} x_i x_j^*$ (and, hence, $|u_i|^2$ with $|x_i|^2$).*
- *If*

$$\boxed{f_A(u) = |a^\intercal u|^2 + (b^\intercal u)(c^\intercal u)} \tag{7.5}$$

then

$$\boxed{f_A(x) = |a^* x|^2 + \operatorname{Re}(b^* u)(c^* u)} \tag{7.6}$$

Corollary 7.2. *Evidently, by (7.4), $f_A(x) > 0$ ($f_A(x) \geq 0$) for any $x \in \mathbb{C}^n$ if and only if $f_A(u) > 0$ ($f_A(u) \geq 0$) for all $u \in \mathbb{R}^n$.*

Claim 7.2. *If $x = \begin{pmatrix} x^{(1)} \\ x^{(2)} \end{pmatrix}$ and $A = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^* & A_{22} \end{bmatrix}$, then*

$$\boxed{\begin{aligned} f_A(x) &:= (Ax, x) \\ &= \left(A_{11} x^{(1)}, x^{(1)}\right) + 2 \operatorname{Re}\left(A_{12} x^{(2)}, x^{(1)}\right) + \left(A_{22} x^{(2)}, x^{(2)}\right) \end{aligned}} \tag{7.7}$$

7.2 Nonnegative definite matrices

7.2.1 Nonnegative definiteness

Definition 7.2. *A symmetric matrix $S \in \mathbb{R}^{n \times n}$ is said to be* ***nonnegative definite*** *if*

$$\boxed{x^\intercal S x \geq 0} \tag{7.8}$$

for all $x \in \mathbb{R}^n$.

The next simple lemma holds.

Lemma 7.1. *(Bellman 1970) The following statements are equivalent:*

1. *S is nonnegative definite;*
2. *S may be represented as*

$$\boxed{S = HH^{\mathsf{T}}} \tag{7.9}$$

for some matrix H;

3. *the eigenvalues of S are nonnegative, that is, for all* $i = 1, \ldots, n$

$$\boxed{\lambda_i(S) \geq 0} \tag{7.10}$$

4. *there is a symmetric matrix* $R \in \mathbb{R}^{n \times n}$ *such that*

$$\boxed{S = R^2} \tag{7.11}$$

R is called the square root of S, and is denoted by the symbol $S^{1/2} := R$.

Definition 7.3. *If S is nonnegative and nonsingular, it is said to be* ***positive definite****.*

Remark 7.1. *In the case when S is positive definite,* $S^{1/2}$ *is also positive definite and for all* $x \neq 0$

$$x^{\mathsf{T}} S x > 0$$

The statement "S is nonnegative definite" is abbreviated

$$S \geq 0$$

and, similarly,

$$S > 0$$

means "S is positive definite".

Remark 7.2. *The abbreviation*

$$\boxed{A \geq B \ (or\ A > B)} \tag{7.12}$$

applied to two symmetric matrices of the same size, means that

$$A - B \geq 0 \ (or\ A - B > 0)$$

Remark 7.3. *Evidently, if* $A > 0$ *then for any quadratic nonsingular* B $(\det B \neq 0)$ *it follows that*

$$BAB^{\mathsf{T}} > 0$$

and, inversely, if $BAB^{\mathsf{T}} > 0$ *for some nonsingular matrix* B*, then* $A > 0$*.*

Remark 7.4. *If* $A \geq B$ *(or* $A > B$*), then for any quadratic nonsingular* T $(\det T \neq 0)$

$$TAT^{\mathsf{T}} \geq TBT^{\mathsf{T}} \text{ (or } TAT^{\mathsf{T}} > TBT^{\mathsf{T}})$$

and, inversely, if $TAT^{\mathsf{T}} \geq TBT^{\mathsf{T}}$ *(or* $TAT^{\mathsf{T}} > TBT^{\mathsf{T}}$*) for some nonsingular* T*, then* $A \geq B$ *(or* $A > B$*).*

Proposition 7.1. *If*

$$\boxed{A > B > 0}$$

then

$$\boxed{B^{-1} > A^{-1} > 0}$$

Proof. Let T_A be an orthogonal transformation which transforms A to a diagonal matrix $\Lambda_A := \operatorname{diag}(\lambda_1(A), \ldots, \lambda_n(A))$ and

$$\Lambda_A = \Lambda_A^{1/2}\Lambda_A^{1/2}, \quad \Lambda_A^{1/2} = \operatorname{diag}\left(\sqrt{\lambda_1(A)}, \ldots, \sqrt{\lambda_n(A)}\right) > 0$$

Then, by the previous remark,

$$T_A A T_A^{\mathsf{T}} = \Lambda_A > T_A B T_A^{\mathsf{T}}$$
$$I_{n\times n} > \Lambda_A^{-1/2} T_A B T_A^{\mathsf{T}} \Lambda_A^{-1/2}$$

Denoting by T an orthogonal transformation which transforms the right-hand side of the last inequality to a diagonal matrix Λ, we obtain

$$I_{n\times n} = TT^{\mathsf{T}} > T\left(\Lambda_A^{-1/2} T_A B T_A^{\mathsf{T}} \Lambda_A^{-1/2}\right) T^{\mathsf{T}} = \Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$$

Inverting this inequality by components, we have

$$I_{n\times n} < \Lambda^{-1} = \left[T\left(\Lambda_A^{-1/2} T_A B T_A^{\mathsf{T}} \Lambda_A^{-1/2}\right) T^{\mathsf{T}}\right]^{-1} = T\left(\Lambda_A^{1/2} T_A B^{-1} T_A^{\mathsf{T}} \Lambda_A^{1/2}\right) T^{\mathsf{T}}$$

which implies

$$I_{n\times n} < \Lambda_A^{1/2} T_A B^{-1} T_A^{\mathsf{T}} \Lambda_A^{1/2}$$

and

$$\Lambda_A^{-1} < T_A B^{-1} T_A^{\mathsf{T}}$$

Hence,

$$T_A^{\mathsf{T}} \Lambda_A^{-1} T_A = A^{-1} < B^{-1}$$

Proposition is proven. □

Proposition 7.2. *If $S \geq 0$ and $T \geq 0$, then*

$$S + T \geq 0$$

with strict inequality holding if and only if

$$\mathcal{N}(S) \cap \mathcal{N}(T) = \varnothing$$

The proof of these statements is evident.

7.2.2 Nonnegative (positive) definiteness of a partitioned matrix

Theorem 7.1. *(Albert 1972) Let S be a square matrix partitioned as*

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^{\mathsf{T}} & S_{22} \end{bmatrix}$$

where S_{11} is a symmetric $n \times n$ matrix and S_{22} is a symmetric $m \times m$ matrix. Then

(a) $S \geq 0$ if and only if

$$\boxed{\begin{aligned} &S_{11} \geq 0 \\ &S_{11} S_{11}^{+} S_{12} = S_{12} \\ &S_{22} - S_{12}^{\mathsf{T}} S_{11}^{+} S_{12} \geq 0 \end{aligned}} \tag{7.13}$$

*(b) $S > 0$ if and only if (**Schur's complement**)*

$$\boxed{\begin{aligned} &S_{11} > 0 \\ &S_{22} > 0 \\ &S_{11} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}} > 0 \\ &S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12} > 0 \end{aligned}} \tag{7.14}$$

Proof.

(a) *Necessity.* Suppose that $S \geq 0$. Then there exists a matrix H with $(n + m)$ rows such that $S = HH^{\mathsf{T}}$. Let us write H as a partitioned matrix

$$H = \begin{bmatrix} X \\ Y \end{bmatrix}, \quad X \in \mathbb{R}^{n\times n}, \quad Y \in \mathbb{R}^{m\times m}$$

Then

$$S = HH^{\mathsf{T}} = \begin{bmatrix} XX^{\mathsf{T}} & XY^{\mathsf{T}} \\ YX^{\mathsf{T}} & YY^{\mathsf{T}} \end{bmatrix}$$

so that

$$S_{11} = XX^{\mathsf{T}} \geq 0, \quad S_{12} = XY^{\mathsf{T}}$$

By (6.8)

$$S_{11}S_{11}^{+} = (XX^{\mathsf{T}})(XX^{\mathsf{T}})^{+} = X\left[X^{\mathsf{T}}(XX^{\mathsf{T}})^{+}\right] = XX^{+}$$

so that

$$S_{11}S_{11}^{+}S_{12} = XX^{+}(XY^{\mathsf{T}}) = \left(XX^{+}X\right)Y^{\mathsf{T}} = XY^{\mathsf{T}} = S_{12}$$

Finally, if we let

$$U := Y - S_{12}^{\mathsf{T}}S_{11}^{+}X$$

then

$$0 \leq UU^{\mathsf{T}} = S_{22} - S_{12}^{\mathsf{T}}S_{11}^{+}S_{12}$$

Sufficiency. Let (7.13) hold. Define

$$U := \left[I_{n\times n} \vdots O_{n\times m}\right], \quad V := \left[O_{m\times n} \vdots I_{m\times m}\right]$$

$$X := S_{11}^{1/2}U$$

$$Y := S_{12}^{\mathsf{T}}S_{11}^{+}S_{11}^{1/2}U + \left(S_{22} - S_{12}^{\mathsf{T}}S_{11}^{+}S_{12}\right)^{1/2}V$$

Since

$$UV^{\mathsf{T}} = O_{n\times m}$$

we can see that

$$0 \leq \begin{bmatrix} X \\ \cdots \\ Y \end{bmatrix}\begin{bmatrix} X \\ \cdots \\ Y \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} XX^{\mathsf{T}} & XY^{\mathsf{T}} \\ YX^{\mathsf{T}} & YY^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^{\mathsf{T}} & S_{22} \end{bmatrix} = S$$

(b) *Necessity*. Suppose that $S > 0$. Then by part (a), $S_{11} \geq 0$. Assuming that S_{11} has a zero eigenvalue with the corresponding eigenvector $\bar{x} \neq 0$, we can see that for the nonzero $\begin{pmatrix} \bar{x} \\ 0 \end{pmatrix}$ we have

$$\begin{pmatrix} \bar{x} \\ 0 \end{pmatrix}^{\mathsf{T}} S \begin{pmatrix} \bar{x} \\ 0 \end{pmatrix} = \bar{x}^{\mathsf{T}} S_{11} \bar{x} = 0$$

which contradicts the fact that $S > 0$. So, S_{11} is nonsingular and $S_{11} > 0$. Similarly, by (a)

$$S_{22} - S_{12}^{\mathsf{T}} S_{11}^{+} S_{12} = S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12} \geq 0$$

and by the same argument (by contradiction) $S_{22} > 0$.

The eigenvalues of S^{-1} are reciprocals of S's and so, $S^{-1} > 0$ if $S > 0$. Therefore

$$S^{-1} = \begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix} > 0,$$

with $0 < A \in \mathbb{R}^{n \times n}$, $0 < C \in \mathbb{R}^{m \times m}$, so that $A^{-1} > 0$ and $C^{-1} > 0$. The condition

$$SS^{-1} = I_{(n+m)\times(n+m)}$$

dictates that

$$\begin{aligned} \left(S_{11} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}}\right) A &= I_{n \times n} \\ \left(S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right) C &= I_{m \times m} \end{aligned}$$

or, equivalently,

$$\begin{aligned} \left(S_{11} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}}\right) &= A^{-1} > 0 \\ \left(S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right) &= C^{-1} > 0 \end{aligned}$$

This proves the necessity of (7.14).

Sufficiency. Suppose that (7.14) holds. By part (a), $S \geq 0$. Define

$$\begin{aligned} A &:= \left(S_{11} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}}\right)^{-1} \\ C &:= \left(S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right)^{-1} \\ B &:= -S_{11}^{-1} S_{12} C \end{aligned}$$

It is easy to show that

$$\begin{aligned} B &= -S_{11}^{-1} S_{12} C = -A \left[A^{-1} S_{11}^{-1} S_{12} C\right] \\ &= -A \left[\left(S_{11} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}}\right) S_{11}^{-1} S_{12} \left(S_{22} - S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right)^{-1}\right] \\ &= -A \left[\left(S_{11} S_{11}^{-1} S_{12} - S_{12} S_{22}^{-1} S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right) \left(I_{m \times m} - S_{22}^{-1} S_{12}^{\mathsf{T}} S_{11}^{-1} S_{12}\right)^{-1} S_{22}^{-1}\right] \end{aligned}$$

$$= -A\left[S_{12}\left(I_{m\times m} - S_{22}^{-1}S_{12}^{\mathsf{T}}S_{11}^{-1}S_{12}\right)\left(I_{m\times m} - S_{22}^{-1}S_{12}^{\mathsf{T}}S_{11}^{-1}S_{12}\right)^{-1}S_{22}^{-1}\right]$$

$$= -AS_{12}S_{22}^{-1}$$

Then routine calculations verify that

$$S\begin{bmatrix} A & B \\ B^{\mathsf{T}} & C \end{bmatrix} = I_{(n+m)\times(n+m)}$$

So, S is nonsingular.

□

Corollary 7.3. *Suppose that in the previous theorem $m=1$, that is, the following representation holds*

$$\boxed{\begin{aligned} S_{n+1} &= \begin{bmatrix} S_n & s_n \\ s_n^{\mathsf{T}} & \sigma_{n+1} \end{bmatrix} \\ s_n &\in \mathbb{R}^n,\ \sigma_{n+1} \in \mathbb{R} \end{aligned}} \tag{7.15}$$

where $0 \le S_n \in \mathbb{R}^{n\times n}$. Let

$$t_n := S_n^{+}s_n, \quad \alpha_n := \sigma_{n+1} - s_n^{\mathsf{T}}S_n^{+}s_n$$
$$\beta_n := 1 + \|t_k\|^2, \quad T_n - t_nt_n^{\mathsf{T}}/\beta_n$$

Then

(a) $S_{n+1} \ge 0$ if and only if

$$\boxed{S_nt_n = s_n \ \text{ and } \ \alpha_n \ge 0} \tag{7.16}$$

and

$$\boxed{S_{n+1}^{+} = \begin{cases} \begin{bmatrix} S_n^{+} + t_nt_n^{\mathsf{T}}\alpha_n^{-1} & -t_n\alpha_n^{-1} \\ -t_n^{\mathsf{T}}\alpha_n^{-1} & \alpha_n^{-1} \end{bmatrix} & \text{if } \ \alpha_n > 0 \\ \begin{bmatrix} T_nS_n^{+}T_n & T_nS_n^{+}t_n\beta_n^{-1} \\ \left(T_nS_n^{+}t_n\beta_n^{-1}\right)^{\mathsf{T}} & \left(t_n^{\mathsf{T}}S_n^{+}t_n\right)\beta_n^{-2} \end{bmatrix} & \text{if } \ \alpha_n = 0 \end{cases}} \tag{7.17}$$

(b) $S_{n+1} > 0$ if and only if

$$\boxed{\alpha_n = \sigma_{n+1} - s_n^{\mathsf{T}}S_n^{-1}s_n > 0} \tag{7.18}$$

and

$$\boxed{S_{n+1}^{-1} = \begin{bmatrix} S_n^{-1} + \left[S_n^{-1}s_ns_n^{\mathsf{T}}S_n^{-1}\right]\alpha_n^{-1} & -\left(S_n^{-1}s_n\right)\alpha_n^{-1} \\ -\left(S_n^{-1}s_n\right)^{\mathsf{T}}\alpha_n^{-1} & \alpha_n^{-1} \end{bmatrix}} \tag{7.19}$$

The proof of this corollary follows directly from the previous theorem and the application of Cline's formula (6.33).

7.3 Sylvester criterion

Here we present a simple proof of the known criterion which gives a power instrument for the numerical test of positive definiteness.

Theorem 7.2. (Sylvester criterion) *A symmetric matrix $S \in \mathbb{R}^{n\times n}$ is positive definite if and only if* ***all leading principal minors*** *(1.12) are* ***strictly positive****, that is, for all $p = 1, 2, \ldots, n$*

$$\boxed{A\begin{pmatrix} 1 & 2 & \cdots & p \\ 1 & 2 & \cdots & p \end{pmatrix} > 0} \tag{7.20}$$

Proof. Let us prove this result by the induction method. For $n = 2$ the result is evident. Indeed, for

$$S = \begin{bmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{bmatrix}$$

under the assumption that $a_{11} \neq 0$, we have

$$\begin{aligned} x^{\mathsf{T}} S x &= a_{11}x_1^2 + 2a_{12}x_1x_2 + a_{22}x_2^2 \\ &= a_{11}\left(x_1 + \frac{a_{12}}{a_{11}}\right)^2 + \left(a_{22} - \frac{a_{12}^2}{a_{11}}\right)x_2^2 \end{aligned}$$

from which it follows that $x^{\mathsf{T}} S x > 0$ $(x \neq 0)$, or equivalently, $S > 0$ if and only if

$$a_{11} > 0, \quad a_{22} - \frac{a_{12}^2}{a_{11}} = \det S > 0$$

Let us represent $S \in \mathbb{R}^{n\times n}$ in the form (7.15)

$$S_n = \begin{bmatrix} S_{n-1} & s_{n-1} \\ s_{n-1}^{\mathsf{T}} & \sigma_n \end{bmatrix}$$
$$s_{n-1} \in \mathbb{R}^{n-1}, \quad \sigma_n \in \mathbb{R}$$

and suppose that $S_{n-1} > 0$. This implies that $\det S_{n-1} > 0$. Then by (7.3) $S_n > 0$ if and only if the condition (7.18) holds, that is, when

$$\alpha_{n-1} = \sigma_n - s_{n-1}^{\mathsf{T}} S_{n-1}^{-1} s_{n-1} > 0$$

But by the Schur's formula

$$\begin{aligned} A\begin{pmatrix} 1 & 2 & \cdots & n \\ 1 & 2 & \cdots & n \end{pmatrix} &= \det S = \det\left(\sigma_n - s_{n-1}^{\mathsf{T}} S_{n-1}^{-1} s_{n-1}\right)(\det S_{n-1}) \\ &= \left(\sigma_n - s_{n-1}^{\mathsf{T}} S_{n-1}^{-1} s_{n-1}\right)(\det S_{n-1}) = \alpha_{n-1}(\det S_{n-1}) > 0 \end{aligned}$$

if and only if (7.18) holds, which proves the result. □

7.4 The simultaneous transformation of a pair of quadratic forms

7.4.1 The case when one quadratic form is strictly positive

Theorem 7.3. *For any two quadratic forms*

$$\boxed{\begin{aligned} &f_A(x) = (x, Ax), \quad f_B(x) = (x, Bx) \\ &A = A^{\mathsf{T}} > 0, \quad B = B^{\mathsf{T}} \end{aligned}}$$

when one quadratic form is strictly positive, i.e. $(x, Ax) > 0$ *for any* $x \neq 0$, *there exists a nonsingular transformation* T *such that in new variables* z *is defined as*

$$\boxed{z = T^{-1}x, \quad x = Tz}$$

and the given quadratic forms are

$$\boxed{\begin{aligned} &f_A(x) = (x, Ax) = (z, T^{\mathsf{T}}ATz) = (z, z) = \sum_{i=1}^{n} z_i^2 \\ &f_B(x) = (x, Bx) = (z, T^{\mathsf{T}}BTz) = \sum_{i=1}^{n} \beta_i z_i^2 \\ &T^{\mathsf{T}}AT = I_{n\times n}, \quad T^{\mathsf{T}}BT = \operatorname{diag}(\beta_1, \beta_2, \ldots, \beta_n) \end{aligned}} \tag{7.21}$$

Proof. Let T_A transform A to the diagonal forms, namely,

$$T_A^{\mathsf{T}} A T_A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n) := \Lambda_A$$

with $\alpha_1 > 0$ $(i = 1, \ldots, n)$. Notice that this transformation exists by the spectral theorem and is unitary, i.e. $T_A^{\mathsf{T}} = T_A^{-1}$. Then, defining $\Lambda_A^{1/2}$ such that

$$\Lambda_A = \Lambda_A^{1/2}\Lambda_A^{1/2}, \qquad \Lambda_A^{1/2} = \operatorname{diag}\left(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \ldots, \sqrt{\alpha_n}\right)$$

we have

$$\left[\Lambda_A^{-1/2} T_1^{\mathsf{T}}\right] A \left[T_1 \Lambda_A^{-1/2}\right] = I_{n\times n}$$

Hence,

$$\tilde{B} := \left[\Lambda_A^{-1/2} T_1^{\mathsf{T}}\right] B \left[T_1 \Lambda_A^{-1/2}\right]$$

is a symmetric matrix, i.e. $\tilde{B} = \tilde{B}^{\mathsf{T}}$. Let $T_{\tilde{B}}$ be a unitary matrix transforming $\tilde{B}$ to the diagonal form, that is,

$$T_{\tilde{B}}^{\mathsf{T}} \tilde{B} T_{\tilde{B}} = \operatorname{diag}\left(\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_n\right) := \Lambda_{\tilde{B}}$$

Then the transformation T defined by

$$T := \left[T_1 \Lambda_A^{-1/2}\right] T_{\tilde{B}}$$

exactly realizes (7.21) since

$$T^{\mathsf{T}} A T = T_{\tilde{B}}^{\mathsf{T}} \left(\left[\Lambda_A^{-1/2} T_1^{\mathsf{T}}\right] A \left[T_1 \Lambda_A^{-1/2}\right]\right) T_{\tilde{B}} = T_{\tilde{B}}^{\mathsf{T}} T_{\tilde{B}} = I_{n\times n}$$

□

Corollary 7.4.

$$\boxed{T = \left[T_1 \Lambda_A^{-1/2}\right] T_{\tilde{B}}} \tag{7.22}$$

where the matrices T_1, $\Lambda_A^{1/2}$ *and* $T_{\tilde{B}}$ *realize the following transformations*

$$\boxed{\begin{aligned} &T_A^{\mathsf{T}} A T_A = \operatorname{diag}(\alpha_1, \alpha_2, \ldots, \alpha_n) := \Lambda_A \\ &\Lambda_A^{1/2} = \operatorname{diag}\left(\sqrt{\alpha_1}, \sqrt{\alpha_2}, \ldots, \sqrt{\alpha_n}\right) \\ &T_{\tilde{B}}^{\mathsf{T}} \left(\left[\Lambda_A^{-1/2} T_1^{\mathsf{T}}\right] A \left[T_1 \Lambda_A^{-1/2}\right]\right) T_{\tilde{B}} = \operatorname{diag}\left(\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_n\right) := \Lambda_{\tilde{B}} \end{aligned}}$$

7.4.2 The case when both quadratic forms are nonnegative

Theorem 7.4. *Let two quadratic forms*

$$\boxed{f_A(x) = (x, Ax), \quad f_B(x) = (x, Bx)}$$

be nonnegative, that is,

$$\boxed{A = A^{\mathsf{T}} \geq 0, \quad B = B^{\mathsf{T}} \geq 0}$$

Then there exists a nonsingular matrix T *such that*

$$\boxed{\begin{aligned} TAT^{\mathsf{T}} &= \begin{bmatrix} \sum_1 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\ \left(T^{-1}\right)^{\mathsf{T}} B T^{-1} &= \begin{bmatrix} \sum_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sum_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{aligned}} \tag{7.23}$$

with $\sum_i$ $(i = 1, 2)$ *diagonal and positive definite, that is,*

$$\boxed{\sum_i = \operatorname{diag}\left(\sigma_1^{(i)}, \ldots, \sigma_{n_i}^{(i)}\right), \quad \sigma_s^{(i)} > 0\,(s = 1, \ldots, n_i)}$$

Proof. Since A is positive and semidefinite, then there exists a unitary matrix T_1 such that $T_1 A T_1^{\mathsf{T}} = \begin{bmatrix} I & 0 \\ 0 & 0 \end{bmatrix}$. Then, let it be $\left(T_1^{\mathsf{T}}\right)^{-1} B T_1^{-1} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$. Again, there exists a unitary matrix U_1 such that $U_1 B_{11} U_1^{\mathsf{T}} = \begin{bmatrix} \left(\sum_1\right)^2 & 0 \\ 0 & 0 \end{bmatrix}$ with $\sum_1 > 0$. Define the unitary matrix T_2^{T} by $\left(T_2^{\mathsf{T}}\right)^{-1} = \begin{bmatrix} U_1 & 0 \\ 0 & I \end{bmatrix}$. Then we have

$$\left(T_2^{\mathsf{T}}\right)^{-1}\left(T_1^{\mathsf{T}}\right)^{-1} B T_1^{-1}\left(T_2\right)^{-1} = \begin{bmatrix} \left(\sum_1\right)^2 & 0 & Q_{121} \\ 0 & 0 & Q_{122} \\ Q_{121}^{\mathsf{T}} & Q_{122}^{\mathsf{T}} & Q_{22} \end{bmatrix}$$

with $Q_{122} = 0$ since $B \geq 0$. Define $\left(T_3^{\mathsf{T}}\right)^{-1} = \begin{bmatrix} I & 0 & 0 \\ 0 & I & 0 \\ -Q_{121}^{\mathsf{T}}\left(\sum_1\right)^2 & 0 & I \end{bmatrix}$. Then

$$\left(T_3^{\mathsf{T}}\right)^{-1}\left(T_2^{\mathsf{T}}\right)^{-1}\left(T_1^{\mathsf{T}}\right)^{-1} B T_1^{-1}\left(T_2\right)^{-1}\left(T_3\right)^{-1} = \begin{bmatrix} \left(\sum_1\right)^2 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \left[Q_{22} - Q_{121}^{\mathsf{T}}\left(\sum_1\right)^2 Q_{121}\right] \end{bmatrix}$$

Next, define the unitary matrix U_2 such that

$$U_2\left[Q_{22} - Q_{121}^{\mathsf{T}}\left(\sum_1\right)^2 Q_{121}\right] U_2^{\mathsf{T}} = \begin{bmatrix} \sum_2 & 0 \\ 0 & 0 \end{bmatrix}$$

with $\sum_2 > 0$, and define also the unitary matrix T_4 such that $\left(T_4^{\mathsf{T}}\right)^{-1} = \begin{bmatrix} \left(\sum_1\right)^{-1/2} & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & U_2 \end{bmatrix}$. Then it is easy to check that for

$$T = T_4 T_3 T_2 T_1 \tag{7.24}$$

we get (7.23). □

Corollary 7.5. *The product of two nonnegative matrices is similar to a nonnegative matrix, that is, for T defined by (7.24) it follows that*

$$\boxed{T(AB)T^{-1} = \begin{bmatrix} \left(\sum_1\right)^2 & 0 \\ 0 & 0 \end{bmatrix}} \tag{7.25}$$

Proof. Indeed,

$$T(AB)T^{-1} = [TAT^{\intercal}]\left[\left(T^{-1}\right)^{\intercal} BT^{-1}\right]$$

$$= \begin{bmatrix} \sum_1 & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sum_1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \sum_2 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} \left(\sum_1\right)^2 & 0 \\ 0 & 0 \end{bmatrix}$$

which completes the proof. □

7.5 Simultaneous reduction of more than two quadratic forms

For the case of two quadratic forms define

$$T_2 = T$$

where T is given by (7.22). Let us apply the induction method, namely, suppose that the transformation T_{k-1} transforms simultaneously one strictly positive definite form with a matrix $A = A^{\intercal} > 0$ and $(k-2)$ another quadratic form with the matrices $B_i = B_i^{\intercal}$ $(i = 1, \ldots, k-1)$ to the sum of pure positive quadratics and the rest to the sum of quadratic elements (maybe with zero coefficients). Then the matrix

$$\tilde{B}_k := T_{k-1} B_k T_{k-1}$$

is a symmetric one. Hence, by the spectral theorem, there exists a unitary transformation $T_{\tilde{B}_n}$ such that

$$T_{\tilde{B}_k}^{\intercal} \tilde{B}_k T_{\tilde{B}_k} = \Lambda_{\tilde{B}_k} := \operatorname{diag}\left(\tilde{\beta}_1^{(k)}, \ldots, \tilde{\beta}_n^{(k)}\right)$$

$$T_{\tilde{B}_k}^{\intercal} T_{\tilde{B}_k} = I_{n \times n}$$

Then the transformation

$$T_k := T_{k-1} T_{\tilde{B}_k}$$

will keep all previous quadratic forms in the same presentation and will transform the last one to a diagonal form.

7.6 A related maximum–minimum problem

7.6.1 Rayleigh quotient

Definition 7.4. *The function* $f_H(x) : \mathbb{C}^n \longrightarrow \mathbb{R}$, *defined by*

$$\boxed{f_H(x) := \frac{(x, Hx)}{(x, x)}, \quad x \neq 0} \tag{7.26}$$

for any Hermitian matrix H, *is known as the* ***Rayleigh quotient.***

Evidently, $f_H(x)$ may be represented in the normalized form $F_H(e)$ as

$$\boxed{\begin{array}{l} f_H(x) = F_H(e) := (e, He), \quad \|e\| = 1 \\ e = \dfrac{x}{\|x\|}, \quad x \neq 0 \end{array}} \tag{7.27}$$

Below we will present the main properties of the Rayleigh quotient in the normalized form $F_H(e)$.

7.6.2 Main properties of the Rayleigh quotient

Theorem 7.5. *The normalized Rayleigh quotient* $F_H(e)$ *(7.27) is invariant to a unitary transformation of the argument as well as to unitary similarity transformation of the matrix* H, *namely, for any unitary matrix* $U \in \mathbb{C}^{n\times n}$

$$F_{UHU^*}(e) = F_H(\tilde{e})$$
$$\tilde{e} = U^* e$$

keeping the property

$$\|\tilde{e}\| = 1$$

Proof. Since U is unitary then $UU^* = I_{n\times n}$ and hence

$$F_{UHU^*}(e) = (e, UHU^*e) = (U^*e, HU^*e) = (\tilde{e}, H\tilde{e}) = F_H(\tilde{e})$$

and

$$\|\tilde{e}\| = \sqrt{(\tilde{e}, \tilde{e})} = \sqrt{(U^*e, U^*e)} = \sqrt{(e, UU^*e)} = \sqrt{(e, e)} = \|e\| = 1$$

Theorem is proven. □

Define the set $\mathcal{F}_H$ as

$$\boxed{\mathcal{F}_H := \{f \in \mathbb{R} \mid f = (e, He), \ \|e\| = 1\}} \tag{7.28}$$

that is, $\mathcal{F}_H$ is the set of all possible values of the normalized Rayleigh quotient $F_H(e)$ (7.27).

Lemma 7.2. *$\mathcal{F}_H$ contains the spectrum $\sigma(H)$ of all eigenvalues of H, i.e.,*

$$\boxed{\sigma(H) \subset \mathcal{F}_H} \tag{7.29}$$

Proof. If $\lambda \in \sigma(H)$, then there exists an eigenvector e of H which corresponds to this λ, that is,

$$He = \lambda e$$

and hence,

$$(e, He) = (e, \lambda e) = \lambda (e, e) = \lambda$$

So, $\lambda \in \mathcal{F}_H$. Thus, $\sigma(H) \subset \mathcal{F}_H$. □

Let now $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues of H.

Theorem 7.6. *$\mathcal{F}_H$ coincides with the convex hull $\overline{\text{co}}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ of the eigenvalues of H, namely,*

$$\mathcal{F}_H \equiv \overline{\text{co}}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} \tag{7.30}$$

where

$$\overline{\text{co}}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} := \left\{ \lambda \mid \lambda = \sum_{i=1}^{n} \alpha_i \lambda_i, \quad \sum_{i=1}^{n} \alpha_i = 1, \quad \alpha_i \geq 0 \ (i = 1, \ldots, n) \right\} \tag{7.31}$$

Proof. Since H is Hermitian, there exists a unitary matrix U, such that

$$UHU^* = \Lambda = \text{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$$

Hence by theorem (7.5), one has $\mathcal{F}_H = \mathcal{F}_\Lambda$. So, it is sufficient to show that the field of the eigenvalues of the diagonal matrix Λ coincides with $\overline{\text{co}}\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$. Indeed, by (7.31)

$$\lambda_j = \sum_{i=1}^{n} \alpha_i \lambda_i \quad \text{when} \quad \alpha_i = \delta_{ij}$$

□

Corollary 7.6.

$$\boxed{\mathcal{F}_H \equiv \overline{\text{co}}\{\lambda_1, \lambda_2, \ldots, \lambda_n\} = \left[\lambda_1 := \min_{i=1,\ldots,n} \lambda_i, \ \lambda_n := \max_{i=1,\ldots,n} \lambda_i\right]} \tag{7.32}$$

Corollary 7.7.

$$\boxed{\begin{aligned}\min_{e:\|e\|=1}(e, He) &= \min_{x\neq 0}\frac{(x, Hx)}{(x, x)} = \lambda_1 := \min_{i=1,\ldots,n}\lambda_i = \lambda_{\min}(H)\\ \max_{e:\|e\|=1}(e, He) &= \max_{x\neq 0}\frac{(x, Hx)}{(x, x)} = \lambda_n := \max_{i=1,\ldots,n}\lambda_i = \lambda_{\max}(H)\end{aligned}} \tag{7.33}$$

Corollary 7.8. *If* $H = \|h_{ij}\|_{i,j=1,n}$, *then*

$$\boxed{\begin{aligned}&\lambda_1 \leq h_{ij} \leq \lambda_n\\ &n\lambda_1 \leq \mathrm{tr}H \leq n\lambda_n\end{aligned}} \tag{7.34}$$

Corollary 7.9. (Stationary property) *If* λ_i *is an eigenvalue of a symmetric matrix* $A = A^{\mathsf{T}}$ *with the corresponding eigenvector* $x^{(i)}$, *then for* $f_A(x) = \dfrac{(x, Ax)}{(x, x)}$ $(x \neq 0)$ *it follows that*

$$\boxed{f_A(x^{(i)}) = \lambda_i} \tag{7.35}$$

and the stationary property holds, that is, for any $i = 1, \ldots, n$ *one has*

$$\boxed{\frac{\partial}{\partial x} f_A(x)\,|_{x=x^{(i)}} = 0} \tag{7.36}$$

Proof. The identity (7.35) follows directly from the simple calculation of $f_A(x^{(i)})$. As for (7.36) it is sufficient to notice that

$$\frac{\partial}{\partial x} f_A(x) = \frac{\partial}{\partial x}\frac{(x, Ax)}{(x, x)} = \frac{2Ax\,(x, x) - 2x\,(x, Ax)}{(x, x)^2}$$

and hence

$$\begin{aligned}\frac{\partial}{\partial x} f_A(x)\,|_{x=x^{(i)}} &= \frac{2Ax^{(i)}\left(x^{(i)}, x^{(i)}\right) - 2x^{(i)}\left(x^{(i)}, Ax^{(i)}\right)}{\left(x^{(i)}, x^{(i)}\right)^2}\\ &= 2\frac{\lambda_i x^{(i)}\left(x^{(i)}, x^{(i)}\right) - x^{(i)}\left(x^{(i)}, \lambda_i x^{(i)}\right)}{\left(x^{(i)}, x^{(i)}\right)^2}\\ &= 2\lambda_i\left(\frac{x^{(i)}}{\left(x^{(i)}, x^{(i)}\right)} - \frac{x^{(i)}}{\left(x^{(i)}, x^{(i)}\right)}\right) = 0\end{aligned}$$

Corollary is proven. □

7.7 The ratio of two quadratic forms

Consider the ratio $r(x)$ of two quadratic forms, i.e.,

$$\boxed{\begin{aligned} &r(x) = \frac{(x, Hx)}{(x, Gx)}, \quad (x, Gx) > 0 \\ &H = H^{\mathsf{T}}, \quad G = G^{\mathsf{T}} > 0 \end{aligned}} \tag{7.37}$$

Theorem 7.7.

$$\boxed{\min_{x:(x,Gx)>0} r(x) = \lambda_{\min}\left(G^{-1/2} H G^{-1/2}\right)} \tag{7.38}$$

and

$$\boxed{\max_{x:(x,Gx)>0} r(x) = \lambda_{\max}\left(G^{-1/2} H G^{-1/2}\right)} \tag{7.39}$$

Proof. Using the presentation

$$G = G^{1/2} G^{1/2}$$

valid for any symmetric nonnegative matrix, for $z = G^{1/2}x$ we have

$$r(x) = \frac{(x, Hx)}{(x, Gx)} = r(x) = \frac{(x, Hx)}{\left(x, G^{1/2}G^{1/2}x\right)} = \frac{(x, Hx)}{\left(G^{1/2}x, G^{1/2}x\right)}$$

$$= \frac{\left(G^{-1/2}z, HG^{-1/2}z\right)}{(z, z)} = \frac{\left(z, \left[G^{-1/2}HG^{-1/2}\right]z\right)}{(z, z)}$$

The result follows from corollary (7.7). □

8 Linear Matrix Equations

Contents

8.1 General type of linear matrix equation

8.1.1 General linear matrix equation

Here we consider the *general linear matrix equation*

$$\boxed{A_1 X B_1 + A_2 X B_2 + \cdots + A_p X B_p = C} \tag{8.1}$$

where $A_j \in \mathbb{C}^{m\times m}$, $B_j \in \mathbb{C}^{n\times n}$ $(j = 1, \ldots, p)$ are the given matrices and $X \in \mathbb{C}^{m\times n}$ is the unknown matrix to be found.

8.1.2 Spreading operator and Kronecker product

Together with the *Kronecker matrix product* definition

$$\boxed{\begin{array}{c} A \otimes B := \|a_{ij} B\| \in \mathbb{R}^{n^2\times n^2} \\ A \in \mathbb{C}^{n\times n}, \quad B \in \mathbb{C}^{n\times n} \end{array}} \tag{8.2}$$

given before let us introduce the *spreading operator* col $\{\cdot\}$ for some matrix $A \in \mathbb{C}^{m\times n}$ as

$$\boxed{\operatorname{col} A := \left(a_{1,1}, \ldots, a_{1,n}, a_{2,1}, \ldots, a_{2,n}, \ldots, a_{m,1}, \ldots, a_{m,n}\right)^{\mathsf{T}}} \tag{8.3}$$

that is, .

$$\boxed{\operatorname{col} A := \begin{bmatrix} A_{*1} \\ A_{*2} \\ \vdots \\ A_{*n} \end{bmatrix} \in \mathbb{C}^{mn}, \quad A = \left[A_{*1}\, A_{*2} \cdots A_{*n}\right], \quad A_{*j} \in \mathbb{C}^{m}}$$

Let $\mathrm{col}^{-1}A$ be the operator inverse to $\mathrm{col}\,A$. Evidently, for any $A, B \in \mathbb{C}^{m\times n}$ and $\alpha, \beta \in \mathbb{C}$

$$\boxed{\mathrm{col}\{\alpha A + \beta B\} = \alpha\,\mathrm{col}\{A\} + \beta\,\mathrm{col}\{B\}} \tag{8.4}$$

8.1.3 Relation between the spreading operator and the Kronecker product

Lemma 8.1. *For any matrices $A \in \mathbb{C}^{m\times m}$, $B \in \mathbb{C}^{n\times n}$ and $X \in \mathbb{C}^{m\times n}$ the following properties hold*

1.

$$\boxed{\mathrm{col}\{AX\} = (I_{n\times n} \otimes A)\,\mathrm{col}X} \tag{8.5}$$

2.

$$\boxed{\mathrm{col}\{XB\} = (B^{\mathsf{T}} \otimes I_{m\times m})\,\mathrm{col}X} \tag{8.6}$$

3.

$$\boxed{\mathrm{col}\{AX + XB\} = [(I_{n\times n} \otimes A) + (B^{\mathsf{T}} \otimes I_{m\times m})]\,\mathrm{col}X} \tag{8.7}$$

4.

$$\boxed{\mathrm{col}\{AXB\} = (B^{\mathsf{T}} \otimes A)\,\mathrm{col}\{X\}} \tag{8.8}$$

Proof. Properties 1–3 (8.5)–(8.7) follow directly from property 4 (8.8) and (8.4). So, let us prove (8.8). By definition (8.3) the jth column $(AXB)_{*,j}$ of the matrix AXB can be expressed as

$$(AXB)_{*,j} = AX\,(B)_{*,j} = \sum_{s=1}^{n} b_{s,j}\,(AX)_{*,s} = \sum_{s=1}^{n} \left(b_{s,j}A\right) X_{*,j}$$

$$= \left[b_{1,j}A\ b_{2,j}A \cdots b_{n,j}A\right]\mathrm{col}X$$

which corresponds to (8.8). □

Lemma 8.2. *The eigenvalues of $(A \otimes B)$ are*

$$\boxed{\lambda_i\mu_j\ (i = 1, \ldots, m;\ j = 1, \ldots, n)} \tag{8.9}$$

where λ_i are the eigenvalues of $A \in \mathbb{C}^{m\times m}$ and μ_j are the eigenvalues of $B \in \mathbb{C}^{n\times n}$. They correspond to the following eigenvectors

$$\boxed{\bar{e}_{ij} := \bar{x}_i \otimes \bar{y}_j} \tag{8.10}$$

where $\bar{x}_i$ and $\bar{y}_j$ are the eigenvectors of A and B, that is,

$$A\bar{x}_i = \lambda_i\bar{x}_i, \quad B\bar{y}_j = \mu_j\bar{y}_j$$

Proof. We have

$$(A \otimes B)\,\bar{e}_{ij} = \begin{Vmatrix} a_{11}B \vdots a_{12}B \vdots \cdots \vdots a_{1n}B \\ \cdots \\ a_{n1}B \vdots a_{n2}B \vdots \cdots \vdots a_{nn}B \end{Vmatrix} \begin{pmatrix} x_{1i}\bar{y}_j \\ \cdots \\ x_{ni}\bar{y}_j \end{pmatrix}$$

$$= \begin{Vmatrix} a_{11}x_{1i}B\bar{y}_j + \cdots + a_{1n}x_{ni}B\bar{y}_j \\ \cdots \\ a_{n1}x_{1i}B\bar{y}_j + \cdots + a_{nn}x_{ni}B\bar{y}_j \end{Vmatrix}$$

$$= \begin{Vmatrix} a_{11}x_{1i}\left(\mu_j\bar{y}_j\right) + \cdots + a_{1n}x_{ni}\left(\mu_j\bar{y}_j\right) \\ \cdots \\ a_{n1}x_{1i}\left(\mu_j\bar{y}_j\right) + \cdots + a_{nn}x_{ni}\left(\mu_j\bar{y}_j\right) \end{Vmatrix} = A\bar{x}_i \otimes \mu_j\bar{y}_j$$

$$= \lambda_i\bar{x}_i \otimes \mu_j\bar{y}_j = \lambda_i\mu_j\left(\bar{x}_i \otimes \bar{y}_j\right) = \lambda_i\mu_j\bar{e}_{ij}$$

which proves the lemma. □

Corollary 8.1. *The eigenvalues of the matrix (the* ***Kronecker sum****)*

$$\boxed{[(I_{n\times n} \otimes A) + (B \otimes I_{m\times m})]} \tag{8.11}$$

are as follows

$$\boxed{\lambda_i + \mu_j \ (i = 1, \ldots, m;\ j = 1, \ldots, n)} \tag{8.12}$$

with the corresponding eigenvector

$$\boxed{\tilde{e}_{ij} := \bar{y}_j \otimes \bar{x}_i} \tag{8.13}$$

Proof. Let us check that the vector (8.13) is the eigenvector of the matrix (8.11) with the eigenvalue $(\lambda_i + \mu_j)$. By (8.9) and taking into account that for the unitary matrix $I_{n\times n}$ any vector is an eigenvector with the eigenvalue equal to 1 and the relation

$$(A \otimes B)\left(\bar{x}_i \otimes \bar{y}_j\right) = (A\bar{x}_i) \otimes \left(B\bar{y}_j\right)$$

we get

$$[(I_{n\times n} \otimes A) + (B \otimes I_{m\times m})]\,\tilde{e}_{ij} = (I_{n\times n} \otimes A)\,\tilde{e}_{ij} + (B \otimes I_{m\times m})\,\tilde{e}_{ij}$$

$$= (I_{n\times n} \otimes A)\left(\bar{y}_j \otimes \bar{x}_i\right) + (B \otimes I_{m\times m})\left(\bar{y}_j \otimes \bar{x}_i\right)$$

$$= \left(I_{n\times n}\bar{y}_j\right) \otimes (A\bar{x}_i) + \left(B\bar{y}_j\right) \otimes (I_{m\times m}\bar{x}_i)$$

$$= (1 \cdot \lambda_i)\left(\bar{x}_i \otimes \bar{y}_j\right) + (\mu_i \cdot 1)\left(\bar{x}_i \otimes \bar{y}_j\right) = \left(\lambda_i + \mu_j\right)\bar{e}_{ij}$$

□

Corollary 8.2. *Since the spectrum of eigenvalues for a transposed matrix coincides with the spectrum of eigenvalues for the original matrix then*

$$\lambda_i + \mu_j \, (i = 1, \ldots, m; \; j = 1, \ldots, n)$$

will be eigenvalues for the following matrices

$$[(I_{n\times n} \otimes A) + (B^{\mathsf{T}} \otimes I_{m\times m})]$$

$$[(I_{n\times n} \otimes A^{\mathsf{T}}) + (B \otimes I_{m\times m})]$$

$$[(I_{n\times n} \otimes A^{\mathsf{T}}) + (B^{\mathsf{T}} \otimes I_{m\times m})]$$

8.1.4 Solution of a general linear matrix equation

Theorem 8.1. *The general linear matrix equation (8.1) has the solution* $X \in \mathbb{C}^{m\times n}$ *if and only if the vector* $x = \mathrm{col}X$ *is the solution of the vector equation*

$$\boxed{Gx = c} \tag{8.14}$$

where

$$\boxed{\begin{array}{c} G := \sum_{i=1}^{p} \left(B_i^{\mathsf{T}} \otimes A_i\right) \\ c := \mathrm{col}C \end{array}} \tag{8.15}$$

Proof. By the property (8.8) and since the operator *col* is linear, applying this operator to both sides of (8.1), we have

$$c = \mathrm{col}C = \mathrm{col}\left\{\sum_{i=1}^{p} (A_i X B_i)\right\} = \sum_{i=1}^{p} \mathrm{col}\{A_i X B_i\}$$

$$= \sum_{i=1}^{p} \left(B_i^{\mathsf{T}} \otimes A_i\right) \mathrm{col}X = Gx$$

□

Corollary 8.3. *The general linear matrix equation (8.1) has the* ***unique solution*** $X \in \mathbb{C}^{m\times n}$ *given by*

$$\boxed{X = \mathrm{col}^{-1}\left\{G^{-1}c\right\}} \tag{8.16}$$

if and only if

$$\boxed{\det G \neq 0} \tag{8.17}$$

8.2 Sylvester matrix equation

Now we will consider an important particular case of (8.1).

Lemma 8.3. *The* ***Sylvester matrix equation***

$$\boxed{\begin{array}{c} AX + XB = -Q \\ A \in \mathbb{C}^{m\times m}, \quad B \in \mathbb{C}^{n\times n} \quad \textit{and} \quad X, C \in \mathbb{C}^{m\times n} \end{array}} \tag{8.18}$$

has the unique solution

$$\boxed{X = \mathrm{col}^{-1}\left\{[(B^{\mathsf{T}} \otimes A)]^{-1}\,\mathrm{col}C\right\}} \tag{8.19}$$

if and only if

$$\boxed{\lambda_i + \mu_j \neq 0 \textit{ for any } i, j = 1, \ldots, n} \tag{8.20}$$

where λ_i are the eigenvalues of A and μ_j are the eigenvalues of B.

Proof. This follows directly from (8.16). □

8.3 Lyapunov matrix equation

Lemma 8.4. *Lyapunov (1892) The* ***Lyapunov matrix equation***

$$\boxed{\begin{array}{c} AP + PA^{\mathsf{T}} = -Q \\ A, P, Q = Q^{\mathsf{T}} \in \mathbb{R}^{n\times n} \end{array}} \tag{8.21}$$

has the unique symmetric solution $P = P^{\mathsf{T}}$ if and only if the matrix A has no neutral eigenvalues lying at the imaginary axis, i.e.,

$$\boxed{\mathrm{Re}\,\lambda_i \neq 0 \ (i = 1, \ldots, n)} \tag{8.22}$$

Proof. Equation (8.21) is a particular case of the Sylvester equation (8.18) with $B = A^{\mathsf{T}}$, which by (8.20) implies the uniqueness of the solution providing for all $(i, j = 1, \ldots, n)$

$$\lambda_i + \lambda_j \neq 0$$

This condition obviously is fulfilled if and only if (8.22) holds. □

9 Stable Matrices and Polynomials

Contents

9.1 Basic definitions

Definition 9.1. *A real valued $n \times n$ matrix is said to be* ***stable*** *if all its eigenvalues belong to the left open complex semi-plane*

$$\mathbb{C}^- := \{z \in \mathbb{C} \mid \operatorname{Re} z < 0\} \tag{9.1}$$

that is,

$$\lambda_i (A) < 0 \quad \textit{for any} \quad i = 1, \ldots, n \tag{9.2}$$

Denote the **characteristic polynomial** of a matrix $A \in \mathbb{R}^{n\times n}$ by $p_A(\lambda)$, i.e.,

$$p_A(\lambda) := \det \|A - \lambda I_{n\times n}\| = \lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n \tag{9.3}$$

Notice that $p_A(\lambda)$ is a *monic* polynomial whose leading coefficient (the coefficient of the highest power) is 1. It is clear from (9.3) that the stability property of a matrix $A \in \mathbb{R}^{n\times n}$ is definitely related to the values of the coefficients a_i $(i = 1, \ldots, n)$ in (9.3) since $\lambda_i(A)$ are the roots of the polynomial equation

$$\boxed{p_A(\lambda) = \prod_{i=1}^{n} [\lambda - \lambda_i (A)] = 0} \tag{9.4}$$

Below we will present several results providing the verification of the matrix stability property based only on the coefficients of the characteristic polynomial.

9.2 Lyapunov stability

9.2.1 Lyapunov matrix equation for stable matrices

Lemma 9.1. *Lyapunov (1892)*

1. *If the Lyapunov's matrix equation (8.21)*

$$\boxed{\begin{array}{c} AP + PA^{\mathsf{T}} = -Q \\ A, P, Q = Q^{\mathsf{T}} \in \mathbb{R}^{n\times n} \end{array}} \tag{9.5}$$

holds for some positive definite

$$Q = Q^{\mathsf{T}} > 0$$

and

$$P = P^{\mathsf{T}} > 0$$

then A is stable.

2. *Equation (8.21) has a positive definite solution*

$$P = P^{\mathsf{T}} = \int_{t=0}^{\infty} e^{At} Q e^{A^{\mathsf{T}} t}\, dt > 0 \tag{9.6}$$

if and only if matrix A is stable (Hurwitz) and
(a) or

$$Q = Q^{\mathsf{T}} > 0$$

(if $Q = Q^{\mathsf{T}} \geq 0$, then $P = P^{\mathsf{T}} \geq 0$),

(b) or Q has the structure as

$$Q = BB^{\mathsf{T}}$$

such that the pair (A, B) is ***controllable****, that is,*

$$\operatorname{rank}\left[B \vdots AB \vdots A^2B \vdots \cdots \vdots A^{n-1}B\right] = n \tag{9.7}$$

Proof.

1(a) Claim 1 of this lemma follows directly from the previous lemma 8.4 if let in (8.21)

$$B = A, \quad X = P$$

taking into account that the inequality (8.20) always fulfilled for different (nonconjugated) eigenvalues and for complex conjugated eigenvalues

$$\begin{aligned} \mu_j &= \bar{\lambda}_i = u_i - iv_i \\ \lambda_i &= u_i + iv_i \end{aligned}$$

The inequality (8.20) implies the existence of a solution P. The symmetry of P follows from the following fact: applying the transposition procedure to both sides of (8.21) we get

$$P^{\mathsf{T}}A^{\mathsf{T}} + AP^{\mathsf{T}} = -Q^{\mathsf{T}} = -Q$$

which coincides with (8.21). But this equation has a unique solution, hence $P = P^{\mathsf{T}}$.

1(b) Let λ_i be an eigenvalue of A^{T}, that is,

$$A^{\mathsf{T}}x_i = \lambda_i x_i, \quad x_i \neq 0$$

Then we also have

$$x_i^* A = \bar{\lambda}_i x_i^*$$

(here $A^* := \overline{(A^{\mathsf{T}})}$, i.e., the transposition together with the complex conjugation). Multiplying the left-hand side of (8.21) by x_i^* and the right-hand side by x_i, it follows that

$$\begin{aligned} x_i^*(AP + PA^{\mathsf{T}})x_i &= \bar{\lambda}_i x_i^* P x_i + x_i^* P \lambda_i x_i \\ &= (\bar{\lambda}_i + \lambda_i)\, x_i^* P x_i = -x_i^* Q x_i < 0 \end{aligned}$$

and, since, by the supposition, $x_i^* P x_i > 0$, we obtain $(\bar{\lambda}_i + \lambda_i) = 2 \operatorname{Re} \lambda_i < 0$, which means that A is stable.

2. *Sufficiency*. Let A be stable. Defining the matrices

$$H(t) := e^{At}Q, \quad U(t) := e^{A^{\mathsf{T}}t}$$

it follows that

$$dH(t) := Ae^{At}Q\,dt, \quad dU(t) := e^{A^{\mathsf{T}}t}A^{\mathsf{T}}dt$$

Then we have

$$\begin{aligned} \int_{t=0}^{T} d[H(t)\,U(t)] &= H(T)\,U(T) - H(0)\,U(0) \\ &= e^{AT}Qe^{A^{\mathsf{T}}T} - Q \\ &= \int_{t=0}^{T} H(t)\,dU(t) + \int_{t=0}^{T} dH(t)\,U(t) \\ &= \int_{t=0}^{T} e^{At}Qe^{A^{\mathsf{T}}t}A^{\mathsf{T}}\,dt + \int_{t=0}^{T} Ae^{At}Qe^{A^{\mathsf{T}}t}\,dt \\ &= \left[\int_{t=0}^{T} e^{At}Qe^{A^{\mathsf{T}}t}\,dt\right]A^{\mathsf{T}} + A\left[\int_{t=0}^{T} e^{At}Qe^{A^{\mathsf{T}}t}\,dt\right] \end{aligned} \tag{9.8}$$

The stability of A implies

$$e^{AT} R e^{A^\top T} \underset{T \longmapsto \infty}{\longmapsto} 0$$

and, moreover, the integral

$$P := \lim_{T \longmapsto \infty} \left[\int_{t=0}^{T} e^{At} Q e^{A^\top t} \, dt \right]$$

exists, since

$$\begin{aligned} \left\| \int_{t=0}^{T} e^{At} Q e^{A^\top t} \, dt \right\| &\leq \int_{t=0}^{T} \left\| e^{At} Q e^{A^\top t} \right\| dt \\ &\leq \|Q\| \int_{t=0}^{T} \left\| e^{At} \right\|^2 \, dt \leq \|Q\| \int_{t=0}^{T} e^{2\lambda_{\max} At} \, dt \\ &\leq \|Q\| \int_{t=0}^{T} e^{2\alpha t} \, dt \leq \|Q\| \int_{t=0}^{\infty} e^{-2|\alpha| t} \, dt = \frac{1}{2|\alpha|} \|Q\| < \infty \end{aligned}$$

where

$$\lambda_{\max}(A) \leq -\min_i \operatorname{Re} |\lambda_i| := \alpha < 0$$

So, taking $T \longmapsto \infty$ in (9.8), we obtain (9.5), which means that (9.6) is the solution of (9.5).

(a) If $Q > 0$, then

$$P = \int_{t=0}^{\infty} e^{At} Q e^{A^\top t} \, dt \geq \lambda_{\min}(Q) \int_{t=0}^{\infty} e^{(A+A^\top)t} \, dt > 0$$

(b) If $Q = BB^\top$, then for any $x \in \mathbb{R}^n$

$$x^\top P x = \int_{t=0}^{\infty} x^\top e^{At} B B^\top e^{A^\top t} x \, dt = \int_{t=0}^{\infty} \left\| B^\top e^{A^\top t} x \right\|^2 dt \tag{9.9}$$

Suppose that there exist $x \neq 0$ and the interval (t_0, t_1) $(t_0 < t_1)$ such that

$$\left\| x^\top e^{At} B \right\|^2 = 0 \quad \text{for all} \quad t \in (t_0, t_1) \, (t_0 < t_1)$$

and, hence,

$$x^\top e^{At} B = 0 \tag{9.10}$$

Then the sequent differentiation of (9.10) by t gives

$$x^\top e^{At} AB = 0, \quad x^\top e^{At} A^2 B = 0, \cdots, \quad x^\top e^{At} A^{(n-1)} B = 0$$

which may be rewritten in the matrix form as follows

$$x^{\mathsf{T}} e^{At} \left[B \vdots AB \vdots A^2 B \vdots \cdots \vdots A^{(n-1)} B \right] = 0$$

for all $t \in (t_0, t_1)$ $(t_0 < t_1)$. It means that

$$\operatorname{rank} \left[B \vdots AB \vdots A^2 B \vdots \cdots \vdots A^{n-1} B \right] < n$$

which is in contradiction to (9.7). So,

$$\left\| x^{\mathsf{T}} e^{At} B \right\|^2 > 0$$

at least at one interval (t_0, t_1) and, hence, by (9.9)

$$x^{\mathsf{T}} P x = \int_{t=0}^{\infty} \left\| B^{\mathsf{T}} e^{A^{\mathsf{T}} t} x \right\|^2 dt \geq \int_{t=t_0}^{t_1} \left\| B^{\mathsf{T}} e^{A^{\mathsf{T}} t} x \right\|^2 dt > 0$$

for all $x \neq 0$. It means that $P > 0$.

Necessity. Suppose that there exists a positive solution $P > 0$ given by (9.6). Then this integral exists only if A is stable.

(a) But P may be positive only if $Q > 0$ (this is easily seen by contradiction).

(b) Let $x^{*i} \neq 0$ be an unstable mode (a left eigenvector of A corresponding to an unstable eigenvalue λ_i), that is,

$$x^{*i} A = \lambda_i x^{*i}, \quad \operatorname{Re} \lambda_i \geq 0$$

By the relation

$$\begin{aligned} 0 < x^{*i} P x^i &= \int_{t=0}^{\infty} \left\| x^{*i} e^{At} B \right\|^2 dt = \int_{t=0}^{\infty} \left\| x^{*i} \left[\sum_{l=0}^{\infty} \frac{1}{l!} (At)^l \right] B \right\|^2 dt \\ &= \int_{t=0}^{\infty} \left\| \left[\sum_{l=0}^{\infty} \frac{1}{l!} x^{*i} A^l t^l \right] B \right\|^2 dt = \int_{t=0}^{\infty} \left\| \left[x^{*i} \sum_{l=0}^{\infty} \frac{1}{l!} \lambda_i^l t^l \right] B \right\|^2 dt \\ &= \int_{t=0}^{\infty} \left\| x^{*i} e^{\lambda t} B \right\|^2 dt = \int_{t=0}^{\infty} e^{2\lambda t} \left\| x^{*i} B \right\|^2 dt \end{aligned}$$

it follows that it should be

$$x^{*i} B \neq 0$$

because if not, we get $x^{*i} P x^i = 0$. But this means that the pair (A, B) is controllable (see PBH-test below). Lemma is proven. □

Remark 9.1. *Notice that for* $Q = qI$ *the matrix* P *as the solution of (9.5) can be represented as*

$$P = P^{\mathsf{T}} = q \int_{t=0}^{\infty} e^{At} e^{A^{\mathsf{T}} t} dt > 0$$

but never as $q \int_{t=0}^{\infty} e^{(A+A^{\mathsf{T}})t} dt$, *that is,*

$$P \neq q \int_{t=0}^{\infty} e^{(A+A^{\mathsf{T}})t} dt$$

since

$$e^{At} e^{A^{\mathsf{T}} t} \neq e^{(A+A^{\mathsf{T}})t}$$

that may be verified by the use of the Taylor series expansion for the matrix exponent (see Proposition 5.2).

9.3 Necessary condition of the matrix stability

Here we will present only a *necessary condition* of a matrix stability that gives the simple rule how quickly can we detect if a matrix is unstable.

Let $\{\lambda_i(A)\}_{i=1}^{m}$ be the set of zeros (roots) of the characteristic polynomial (9.3)

$$p_A(\lambda) = \lambda^n + a_1 \lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n$$

or, in another representation,

$$p_A(\lambda) = (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_m) \tag{9.11}$$

Theorem 9.1. (Stodola's rule) *If a matrix* A *is stable (or, equivalently, its characteristic polynomial* $p_A(\lambda)$ *(9.3) is Hurwitz) then all coefficients* a_i *in (9.3) are strictly positive, that is,*

$$\boxed{a_i > 0 \, (i = 1, \ldots, n)} \tag{9.12}$$

Proof. Since the roots $\{\lambda_i(A)\}_{i=1}^{m}$ of $p_A(\lambda)$ (9.11) in general are complex values, one can represent (9.11) as follows

$$p_A(\lambda) = \prod_{j=1}^{n_r} (\lambda - \lambda_j) \cdot \prod_{k=1}^{n_c/2} (\lambda - \lambda_k)(\lambda - \bar{\lambda}_k) \tag{9.13}$$

where

$$\lambda_j = -u_j, \quad j = 1, \ldots, n_r$$
$$\lambda_k = -u_k + iv_k, \quad k = 1, \ldots, n_c/2$$

So, the first n_r roots are purely real and the rest of them are complex. By the stability property of A all real parts are strictly positive, i.e., $u_j > 0$, $j = 1, \ldots, n_r$, and $u_k > 0$, $k = 1, \ldots, n_c/2$. Hence

$$p_A(\lambda) = \prod_{j=1}^{n_r} (\lambda + u_j) \prod_{k=1}^{n_c/2} (\lambda^2 + 2u_k\lambda + u_k^2 + v_k^2)$$

The right-hand side is a polynomial on λ with only positive coefficients which proves the theorem. □

The next useful conclusion follows immediately.

Corollary 9.1. *If the polynomial $p_A(\lambda)$ has coefficients of different signs (or some of them are absent ($a_i = 0$ for at least one i)) the corresponding matrix A is* ***unstable****.*

Example 9.1. *The polynomials*

$$p_A(\lambda) = \lambda^5 + 3\lambda^4 - \lambda^3 + \lambda^2 + \lambda + 1$$
$$p_A(\lambda) = \lambda^5 + \lambda^3 + \lambda^2 + \lambda + 1$$

and, hence, the corresponding matrices A, are unstable. Indeed, in the first polynomial $a_2 = -1 < 0$ and in the second one $a_1 = 0$.

9.4 The Routh–Hurwitz criterion

In this section we will present the *necessary and sufficient conditions* (or, in another words, *the criterion*) of a matrix stability.

Let us define the, so-called, *Hurwitz* matrix H^A as follows:

$$\mathrm{H}^A := \begin{bmatrix} a_1 & a_3 & a_5 & \cdot & \cdot & & 0 \\ 1 & a_2 & a_4 & & & & \cdot \\ 0 & a_1 & a_3 & & & & \cdot \\ \cdot & 1 & a_2 & \cdot & & \cdot & \cdot \\ \cdot & 0 & a_1 & & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & & & a_n & \cdot \\ \cdot & \cdot & \cdot & & & a_{n-1} & 0 \\ 0 & 0 & 0 & \cdot & \cdot & a_{n-2} & a_n \end{bmatrix} \tag{9.14}$$

Here in the main diagonal the coefficients are assigned starting from a_1. Each column has the aligned coefficients in increasing order. Denote also by H_i^A $(i = 1, \ldots, n)$ the leading principal minors of H, that is,

$$H_i^A := \mathrm{H}^A \begin{pmatrix} 1 & 2 & \cdots & i \\ 1 & 2 & \cdots & i \end{pmatrix} \tag{9.15}$$

such that

$$H_n^A = \det \mathrm{H}^A$$

Lemma 9.2. (Orlando's formula) *Let* $\lambda_i\,(i = 1, \ldots, n)$ *be zeros of the polynomial* $p_A(\lambda)$. *Then*

$$\boxed{H_{n-1}^A = (-1)^{n(n-1)/2} \prod_{k=1}^{n} \prod_{i=1}^{k-1} (\lambda_i + \lambda_k)} \tag{9.16}$$

Proof. The proof may be done by the induction by n. For $n = 2$, evidently,

$$H_{n-1} = H_1 = -(\lambda_1 + \lambda_2)$$

Suppose that (9.16) is valid for the polynomial of the order n. So, we need to prove that (9.16) is true for any polynomial of the order $(n+1)$. To do that let us introduce the polynomial $f_{A,h}(\lambda)$ according to the following formula:

$$\begin{aligned} f_{A,h}(\lambda) &= (\lambda + h)\, p_A(\lambda) \\ &= \lambda^{n+1} + a_1\lambda^n + \cdots + a_{n-1}\lambda^2 + a_n\lambda \\ &\quad + h\lambda^n + ha_1\lambda^{n-1} + \cdots + ha_{n-1}\lambda + ha_n \\ &= \lambda^{n+1} + (a_1 + h)\lambda^n + (ha_1 + a_2)\lambda^{n-1} \\ &\quad + \cdots + (ha_{n-1} + a_n)\lambda + ha_n \end{aligned}$$

This polynomial has the roots

$$\lambda_i\,(i = 1, \ldots, n), \quad \lambda_{n+1} = -h$$

Constructing the corresponding Hurwitz matrix $\mathrm{H}^{A,h}$ for $f_{A,h}(\lambda)$ we get

$$H^{A,h} := \begin{bmatrix} (a_1+h) & (ha_2+a_3) & (ha_4+a_5) & \cdot & \cdot & \cdot & 0 \\ 1 & ha_1+a_2 & ha_3+a_4 & & & \cdot & \cdot \\ 0 & (a_1+h) & ha_2+a_3 & & & \cdot & \cdot \\ \cdot & 1 & ha_1+a_2 & \cdot & & \cdot & \cdot \\ \cdot & 0 & (a_1+h) & & & \cdot & \cdot \\ \cdot & \cdot & \cdot & & & ha_{n-1}+a_n & 0 \\ 0 & 0 & 0 & \cdot & \cdot & ha_{n-3}+a_{n-2} & ha_n \end{bmatrix}$$

and

$$H_n^{A,h} = \det \begin{bmatrix} (a_1+h) & (ha_2+a_3) & (ha_4+a_5) & \cdot & \cdot & \cdot & 0 \\ 1 & ha_1+a_2 & ha_3+a_4 & & & & \cdot \\ 0 & (a_1+h) & ha_2+a_3 & \cdot & & & \cdot \\ \cdot & 1 & ha_1+a_2 & \cdot & & & \cdot \\ \cdot & 0 & (a_1+h) & \cdot & \cdot & & 0 \\ \cdot & \cdot & 1 & \cdot & & \cdot & ha_n \\ \cdot & \cdot & \cdot & & & \cdot & ha_{n-1}+a_n \end{bmatrix}$$

Introduce the, so-called, "bordering" determinant:

$$D = \det \begin{bmatrix} (a_1+h) & (ha_2+a_3) & (ha_4+a_5) & \cdot & \cdot & 0 & 0 \\ 1 & ha_1+a_2 & ha_3+a_4 & & & \cdot & \cdot \\ 0 & (a_1+h) & ha_2+a_3 & \cdot & & \cdot & \cdot \\ \cdot & 1 & ha_1+a_2 & \cdot & & \cdot & \cdot \\ \cdot & 0 & (a_1+h) & \cdot & & 0 & \cdot \\ \cdot & \cdot & \cdot & & \cdot & ha_{n-1}+a_n & a_{n-1} \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & (-1)^n \end{bmatrix} \tag{9.17}$$

Obviously,

$$D = (-1)^n H_n^{A,h}$$

Let us now apply some simple transformation to the determinant (9.17) which does not change its value. First, adding $(-h)$ times the column $(r+1)$ to the column r for all $r = 1, 2, \ldots, n$ leads to

$$D = \det \begin{bmatrix} a_1 & a_3 & a_5 & \cdot & \cdot & \cdot & a_{2n-1} & h^n \\ 1 & a_2 & a_4 & & & & \cdot & -h^{n-1} \\ 0 & a_1 & a_3 & \cdot & & & \cdot & \cdot \\ 0 & 1 & a_2 & \cdot & & & \cdot & \cdot \\ \cdot & 0 & a_1 & \cdot & \cdot & & 0 & \cdot \\ \cdot & \cdot & 1 & \cdot & & \cdot & 0 & (-1)^{n-2} h^2 \\ \cdot & \cdot & \cdot & & & \cdot & a_n & (-1)^{n-1} h \\ 0 & 0 & \cdot & \cdot & \cdot & a_{n-3} & a_{n-1} & (-1)^n \end{bmatrix}$$

Then replacing the first row by

$$\text{1st row} - a_1\,(\text{2nd row}) + a_2\,(\text{3rd row}) - \cdots + (-1)^n a_n\,\big([n+1]\,\text{st row}\big),$$

one can see that the last term of this row is exactly $p_A(h)$ and all others are zeros which implies

$$D = (-1)^n p_A(h)\, H_{n-1}^A \tag{9.18}$$

Comparing (9.17) with (9.18) and using (9.16) and (9.4) for $h = -\lambda_{n+1}$, we get

$$\begin{aligned} H_n^{A,h} &= p_A(h)\,|_{h=-\lambda_{n+1}}\, H_{n-1}^A \\ &= \prod_{i=1}^{n} [-\lambda_{n+1} - \lambda_i(A)]\,(-1)^{n(n-1)/2} \prod_{k=1}^{n} \prod_{i=1}^{k-1} (\lambda_i + \lambda_k) \\ &= (-1)^n\,(-1)^{n(n-1)/2} \prod_{i=1}^{n} [\lambda_{n+1} + \lambda_i(A)] \prod_{k=1}^{n} \prod_{i=1}^{k-1} (\lambda_i + \lambda_k) \\ &= (-1)^{(n+1)n/2} \prod_{k=1}^{n+1} \prod_{i=1}^{k-1} (\lambda_i + \lambda_k) \end{aligned}$$

Lemma is proven. □

Lemma 9.3. *Defining* $H_0^A := 1$, *for the leading minors*

$$\boxed{v_j := V\begin{pmatrix} 1 & 2 & \cdots & j \\ 1 & 2 & \cdots & j \end{pmatrix}} \tag{9.19}$$

of the $n \times n$ *matrix*

$$\boxed{V := \begin{bmatrix} a_0a_1 & 0 & a_0a_3 & \cdots \\ 0 & -a_0a_3 + a_1a_2 & 0 & \cdots \\ a_0a_3 & 0 & a_0a_5 - a_1a_4 + a_2a_3 & \cdots \\ \vdots & \vdots & \vdots & \vdots \end{bmatrix}} \tag{9.20}$$

with the elements v_{ij} *given by*

$$\boxed{v_{ij} := \left\{ \begin{array}{lll} \sum_{k=1}^{i} (-1)^{k+i} a_{k-1}a_{i+j-k} & \text{if} & j \geq i \\ v_{ji} & \text{if} & j < i \\ 0 & & \end{array} \right\} \begin{array}{l} \text{for} \quad (i+j) \ \text{even} \\ \\ \text{for} \quad (i+j) \ \text{odd} \end{array}}$$

for all $j = 1, \ldots, n$ *the following property holds*

$$\boxed{v_j = H_j^A H_{j-1}^A} \tag{9.21}$$

where H_j^A *is the* j*th leading minor (9.15) of the Hurwitz matrix* H^A *(9.14).*

Proof. Permute the rows and columns of V symmetrically, bring odd numbered columns and rows into the leading positions and even numbered rows and columns into the last positions which is achieved with the permutation matrix

$$P := \left[e_1 e_3 e_5 \cdots \mid \quad e_2 e_4 e_6 \cdots \right]$$

and leads to the resulting matrix

$$D_{n\times n} := P^{\mathsf{T}} V P$$

$$= \left\{ \begin{array}{lll} D_{2m\times 2m} = \begin{bmatrix} E_{m\times m} & 0 \\ 0 & F_{m\times m} \end{bmatrix} & \text{for} & n = 2m \\ D_{(2m+1)\times(2m+1)} = \begin{bmatrix} E_{(m+1)\times(m+1)} & 0 \\ 0 & F_{m\times m} \end{bmatrix} & \text{for} & n = 2m+1 \end{array} \right.$$

(the subscript denotes the order of the square matrices). Thus we get

$$\det V = \begin{cases} v_{2m} & \text{for} \quad n = 2m \\ v_{2m+1} & \text{for} \quad n = 2m+1 \end{cases}$$

$$= \begin{cases} \det D_{2m\times 2m} = \det E_{m\times m} \det F_{m\times m} & \text{for} \quad n = 2m \\ \det D_{(2m+1)\times(2m+1)} = \det E_{(m+1)\times(m+1)} \det F_{m\times m} & \text{for} \quad n = 2m+1 \end{cases} \tag{9.22}$$

Then define the matrices

$$K_{2m\times m} := \begin{bmatrix} 0 & \cdots & 0 & -a_0 \\ \vdots & & 0 & a_1 \\ \cdot & \cdot & -a_0 & a_2 \\ 0 & & & a_3 \\ -a_0 & & & \\ a_1 & a_3 & \cdots & a_{2m-1} \end{bmatrix}$$

$$K_{(2m+1)\times(m+1)} := \begin{bmatrix} 0 & \cdots & 0 & 0 & a_0 \\ \vdots & & 0 & 0 & -a_1 \\ \cdot & \cdot & 0 & a_0 & a_2 \\ 0 & \cdot & 0 & -a_1 & -a_3 \\ 0 & 0 & a_0 & a_2 & a_4 \\ & & & & \vdots \\ a_1 & a_3 & \cdots & & a_{2m-1} \end{bmatrix}$$

which satisfy the identities

$$\mathrm{H}^A K_{2m\times m} = \begin{bmatrix} 0 \\ F_{m\times m} \end{bmatrix} \quad \text{for} \quad n = 2m$$

$$\mathrm{H}^A K_{(2m+1)\times(m+1)} = \begin{bmatrix} 0 \\ E_{(m+1)\times(m+1)} \end{bmatrix} \quad \text{for} \quad n = 2m+1$$

Therefore it is easily deduced that by (9.22)

$$\det \mathrm{H}^A = H^A_{2m} = \det F_{m\times m} \qquad \text{for} \quad n = 2m$$
$$\det \mathrm{H}^A = H^A_{2m+1} = \det E_{(m+1)\times(m+1)} \qquad \text{for} \quad n = 2m$$

and

$$v_{2m} = \det E_{m\times m} \det F_{m\times m} = H^A_{2m-1} H^A_{2m} \qquad \text{for} \quad n = 2m$$
$$v_{2m+1} = \det E_{(m+1)\times(m+1)} \det F_{m\times m} \qquad \text{for} \quad n = 2m+1$$

Hence, in any case

$$v_j = H^A_j H^A_{j-1}$$

Lemma is proven. □

Corollary 9.2. *H_j^A $(j = 1, \ldots, n)$ are positive if and only if v_j $(j = 1, \ldots, n)$ are positive too, or, equivalently by the Sylvester criterion, H_j^A $(j = 1, \ldots, n)$* ***are positive if and only if the matrix V (9.20) is positive definite.***

Let us now introduce the *companion matrix* A_{com} defined by the coefficients of $p_A(\lambda)$ (9.3):

$$A_{\text{com}} := \begin{bmatrix} -a_1 & -a_2 & \cdot & \cdot & \cdot & -a_n \\ 1 & 0 & \cdot & \cdot & 0 & 0 \\ 0 & 1 & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & 0 & 1 & 0 \end{bmatrix} \tag{9.23}$$

Note that A_{com} is a stable matrix if and only if all zeros of $p_A(\lambda)$ have negative real parts. Indeed,

$$\gamma_n := \det(\lambda I_{n\times n} - A) = \det \begin{bmatrix} \lambda + a_1 & a_2 & \cdot & \cdot & a_n \\ -1 & \lambda & \cdot & \cdot & 0 \\ 0 & -1 & \lambda & \cdot & 0 \\ 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & -1 & \lambda \end{bmatrix}$$

$$= \lambda \det \begin{bmatrix} \lambda + a_1 & a_2 & \cdot & \cdot & a_{n-1} \\ -1 & \lambda & \cdot & \cdot & 0 \\ 0 & -1 & \lambda & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -1 & \lambda \end{bmatrix}$$

$$+ (-1)^{1+n} a_n \begin{bmatrix} -1 & \lambda & \cdot & \cdot & 0 \\ 0 & -1 & \lambda & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \lambda \\ 0 & 0 & \cdot & 0 & -1 \end{bmatrix}$$

$$= \lambda \gamma_{n-1} + (-1)^{1+n} a_n (-1)^{n-1}$$

$$= \lambda \gamma_{n-1} + a_n = \lambda [\lambda \gamma_{n-2} + a_{n-1}] + a_n = \lambda^2 \gamma_{n-2} + \lambda a_{n-1} + a_n$$

$$= \cdots = p_A(\lambda)$$

The claim is true.

Lemma 9.4. *The matrices A_{com} (9.23) and V (9.20) are related as*

$$\boxed{A_{\text{com}}^{\mathsf{T}} V + V A_{\text{com}} = -W} \tag{9.24}$$

where W is a nonnegative definite matrix equal to

$$W = 2\begin{bmatrix} a_1^2 & 0 & a_1a_3 & 0 & a_1a_5 & \cdots \\ 0 & 0 & 0 & \cdot & 0 & 0 \\ a_1a_3 & 0 & a_3^2 & \cdot & a_3a_5 & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \tag{9.25}$$

Proof. The direct computation of the left hand-side of (9.24) suffices for the verification of this identity. To prove that $W \geq 0$ it is sufficient to observe that for any $x \in \mathbb{C}^n$

$$x^*Wx = 2\,|a_1x_1 + a_2x_2 + \cdots + a_nx_n|^2 \tag{9.26}$$

Lemma is proven. □

Theorem 9.2. (The Routh–Hurwitz criterion) *A matrix A is stable if and only if $H_i^A > 0$ $(i = 1, 2, \ldots, n)$ where H_i^A are defined in (9.15).*

Proof.

(a) *Necessity.* By lemma (9.1) it follows that if A is stable and W (9.25) is nonnegative definite then $P = P^\top > 0$ or, equivalently, $v_j \geq 0$ $(j = 1, 2, \ldots, n)$. Then by (9.2) it follows that $H_j^A \geq 0$ $(j = 1, \ldots, n)$. Notice that if any $H_j^A = 0$, then $H_n^A = 0$. But this fact leads to contradiction and so deduce that $H_j^A > 0$ $(j = 1, \ldots, n)$. The stability of A implies that there are no zero roots and so $a_n \neq 0$. But $H_n^A = a_nH_{n-1}^A$ and so $H_n^A = 0$ leads to $H_{n-1}^A = 0$. In this case from the Orlando's formula (9.16) we deduce that there are a pair of roots λ_i, λ_k such that $\lambda_i + \lambda_k = 0$. But this contradicts the hypothesis that both λ_i, λ_k have negative real parts which completes the proof of the necessity part.

(b) *Sufficiency.* Suppose now that $H_j^A > 0$ $(j = 1, \ldots, n)$. We need to prove that A is stable. By (9.2) it follows that V (22.175) is positive definite. Then by the first part of Lemma 9.1 we have only to prove that $a^{(i)*}Wa^{(i)} > 0$ for all right eigenvectors $a^{(i)}$ associated with any given eigenvalue $\lambda_i(A)$. Notice that we may take as $a^{(i)}$ the following vector

$$a^{(i)\top} = \left(\lambda_i^{n-1}(A), \lambda_i^{n-2}(A), \ldots, \lambda_i(A), 1\right)$$

and by (9.26) we may conclude that $a^{(i)*}Wa^{(i)} = 0$ if and only if

$$a_1\lambda_i^{n-1}(A) + a_2\lambda_i^{n-2}(A) + \cdots + a_n = 0$$

Now $H_n^A > 0$ implies $a_n \neq 0$ and hence $\lambda_i(A) \neq 0$. But since $p_A(\lambda_i(A)) = 0$ we have

- for odd n

$$\lambda_i^n(A) + a_1\lambda_i^{n-2}(A) + \cdots + a_{n-1}\lambda_i(A) = 0$$
$$a_1\lambda_i^{n-1}(A) + a_3\lambda_i^{n-3}(A) + \cdots + a_n = 0$$

Hence

$$\begin{bmatrix} 1 & a_2 & a_4 & \cdots & a_{n-1} \\ a_1 & a_3 & a_5 & \cdots & a_n \end{bmatrix} \begin{bmatrix} \lambda_i^{n-1}(A) \\ \vdots \\ \lambda_i^2(A) \\ 1 \end{bmatrix} = 0$$

- for even n

$$\lambda_i^n(A) + a_2\lambda_i^{n-2}(A) + \cdots + a_n = 0$$
$$a_1\lambda_i^{n-1}(A) + a_3\lambda_i^{n-3}(A) + \cdots + a_{n-1}\lambda_i(A) = 0$$

Hence

$$\begin{bmatrix} 1 & a_2 & \cdots & a_{n-2} & a_n \\ a_1 & a_3 & \cdots & a_{n-1} & 0 \end{bmatrix} \begin{bmatrix} \lambda_i^n(A) \\ \vdots \\ \lambda_i^2(A) \\ 1 \end{bmatrix} = 0$$

Thus in both cases for odd or even n we obtain

$$a_1a_2 - a_3 = 0,\ a_1a_4 - a_5 = 0, \ldots$$

and the second row and column of V are zero which contradicts our deduction that V is positive definite. Hence $a^{(i)*}Wa^{(i)} \neq 0$ and thus A is stable. Theorem is proven. □

Example 9.2. *The polynomial*

$$p(\lambda) = \lambda^4 + 2\lambda^3 + 3\lambda^2 + 2\lambda + 1 - \alpha\beta \tag{9.27}$$

has the following Hurwitz matrix (9.14)

$$\mathrm{H}^A = \begin{bmatrix} 2 & 2 & 0 & 0 \\ 1 & 3 & 1-\alpha\beta & 0 \\ 0 & 2 & 2 & 0 \\ 0 & 1 & 3 & 1-\alpha\beta \end{bmatrix}$$

So, by the Routh–Hurwitz criterion the corresponding matrix A is stable if and only if all principal minors H_i^A $(i = 1, \ldots, 4)$ *are strictly positive, that is,*

$$H_1^A = 2 > 0, \quad H_2^A = \det \begin{bmatrix} 2 & 2 \\ 1 & 3 \end{bmatrix} = 4 > 0$$

$$H_3^A = \det \begin{bmatrix} 2 & 2 & 0 \\ 1 & 3 & 1-\alpha\beta \\ 0 & 2 & 2 \end{bmatrix} = 12 - 4(1-\alpha\beta) - 4 = 4(1+\alpha\beta) > 0 \tag{9.28}$$

$$H_4^A = H^A = \det \begin{bmatrix} 2 & 2 & 0 & 0 \\ 1 & 3 & 1-\alpha\beta & 0 \\ 0 & 2 & 2 & 0 \\ 0 & 1 & 3 & 1-\alpha\beta \end{bmatrix} = (1-\alpha\beta) H_3^A > 0$$

which give the following necessary and sufficient condition of stability:

$$1 + \alpha\beta > 0, \quad (1 - \alpha\beta) > 0$$

or, equivalently,

$$\boxed{|\alpha\beta| < 1} \tag{9.29}$$

9.5 The Liénard–Chipart criterion

The Routh–Hurwitz criterion may be represented in a form that requires at least twice the corresponding numerical calculation less compared with the original one. This simplified form is known as the *Liénard–Chipart criterion.*

Theorem 9.3. (The Liénard–Chipart criterion) *A matrix A is stable if and only if*

1. *all coefficients a_i in (9.3) are strictly positive, that is,*

$$\boxed{a_i > 0 \ (i = 1, \ldots, n)}$$

(this means that the Stodola's rule holds);

2.

$$\boxed{H_i^A > 0 \ (i = n-1, \quad n-3, \ldots,)} \tag{9.30}$$

where H_i^A are defined in (9.15).

Proof. As it has been shown before, a matrix A is stable if and only if the matrix V (9.20) is strictly positive definite. But, in view of the relation (9.21) to guarantee that all $H_i^A > 0$ $(i = 1, \ldots, n)$ are positive, it is sufficient to check only the positivity of $H_i^A > 0$ $(i = n-1, n-3, \ldots,)$. Theorem is proven. □

Example 9.3. *Consider the same example (9.2) with the polynomial (9.27). The Stodola rule implies that there should be*

$$1 - \alpha\beta > 0 \tag{9.31}$$

and the Liénard–Chipart criterion demands that

$$\mathrm{H}_3^A = 4\,(1 + \alpha\beta) > 0, \quad \mathrm{H}_1^A = 2 > 0$$

or, equivalently,

$$1 + \alpha\beta > 0 \tag{9.32}$$

Conditions (9.31) and (9.32) considered together lead to (9.29).

9.6 Geometric criteria

All criteria and rules concerning the matrix stability property presented above are given in the so-called *analytical form*. But there exists another form of the stability analysis called the *geometric* one. This form of the stability representation requires some preliminaries related to the *principle of argument variation* discussed below.

9.6.1 The principle of argument variation

Consider the characteristic polynomial $p_A(\lambda)$ (9.3) of a matrix A in the form

$$p_A(\lambda) = \prod_{j=1}^{n} [\lambda - \lambda_j(A)] \tag{9.33}$$

Suppose that all roots $\lambda_j(A)$ of this polynomial satisfy the condition

$$\mathrm{Re}\,\lambda_j(A) \neq 0, \quad j = 1, \ldots, n$$

This permits to represent (9.33) as

$$\boxed{p_A(\lambda) = \prod_{j=1}^{l} [\lambda - \lambda_j(A)] \prod_{k=1}^{r} [\lambda - \lambda_k(A)]} \tag{9.34}$$

where

l = the number of roots with negative (left) real parts
r = the number of roots with positive (right) real parts

Any complex number $z \in \mathbb{C}$ may be represented as

$$\boxed{z = |z|\, e^{i \arg z}} \tag{9.35}$$

where $\arg z$ is the angle formed by the vector z in the complex plane with the real axis measured in the clockwise (positive) direction. Moreover, if $|z| < \infty$ and $\operatorname{Re} z \neq 0$, then the simple complex function

$$f(i\omega) := i\omega - z = |i\omega - z|\, e^{i \arg(i\omega - z)}$$

has the following argument variation $\overset{\infty}{\underset{\omega=-\infty}{\Delta}} \arg f(i\omega)$ (when ω varies from $-\infty$ to ∞):

$$\overset{\infty}{\underset{\omega=-\infty}{\Delta}} \arg f(i\omega) = \begin{cases} \pi & \text{if} \quad \operatorname{Re} z < 0 \\ -\pi & \text{if} \quad \operatorname{Re} z > 0 \end{cases} \tag{9.36}$$

Lemma 9.5. (The principle of argument variation) *The polynomial (9.34) verifies the following*

$$\boxed{\overset{\infty}{\underset{\omega=-\infty}{\Delta}} \arg p_A(i\omega) = (l - r)\,\pi} \tag{9.37}$$

Proof. By the evaluation of (9.34) using (9.35) we have

$$\begin{aligned} p(i\omega) &= \prod_{j=1}^{l} [\lambda - \lambda_j(A)] \prod_{k=1}^{r} [\lambda - \lambda_k(A)] \\ &= \prod_{j=1}^{n} \left(|i\omega - \lambda_j(A)| \right) \exp i \left(\sum_{j=1}^{l} \arg(i\omega - \lambda_j(A)) + \sum_{k=1}^{r} \arg(i\omega - \lambda_k(A)) \right) \end{aligned}$$

So,

$$\arg p_A(i\omega) = \sum_{j=1}^{l} \arg\left(i\omega - \lambda_j(A)\right) + \sum_{k=1}^{r} \arg\left(i\omega - \lambda_k(A)\right)$$

and by (9.36) we derive (9.37). Lemma is proven. □

Corollary 9.3. *For any stable polynomial its* $\arg p_A(j\omega)$ *is a monotonically increasing function of* ω.

Corollary 9.4. *For the polynomials without neutral roots*

$$\boxed{\overset{\infty}{\underset{\omega=0}{\Delta}} \arg p_A(i\omega) = (l - r)\,\frac{\pi}{2}} \tag{9.38}$$

9.6.2 Mikhailov's criterion

Based on the previous lemma (9.5) we may present the following important result.

Theorem 9.4. (Mikhailov's criterion) *The polynomial $p_A(\lambda)$ (9.3) of order n is* ***Hurwitz*** *(or, equivalently the corresponding matrix A is* ***stable****) if and only if the godograph of $p_A(\lambda)$ (the corresponding curve in the coordinates)*

$$U(\omega) := \operatorname{Re} p_A(i\omega)$$

as ***the abscise*** *and*

$$V(\omega) := \operatorname{Im} p_A(i\omega)$$

as ***the ordinate*** *such that*

$$p_A(i\omega) = U(\omega) + iV(\omega)$$

has rotation in the clockwise (positive) direction and passes exactly n quadrants in the complex plane without crossing the origin when ω varies from 0 up to ∞.

Proof. By definition $p_A(\lambda)$ is Hurwitz if and only if it has the representation (9.34) with $l = n$. Hence, by lemma (9.5) in view of (9.37) it follows that

$$\overset{\infty}{\underset{\omega=\infty}{\Delta}} \arg p_A(i\omega) = n\pi$$

or, by corollary (9.38) to this lemma,

$$\overset{\infty}{\underset{\omega=0}{\Delta}} \arg p(i\omega) = n\frac{\pi}{2}$$

The criterion is proven. □

Consider now several examples illustrating the application of Mikhailov's criterion.

Example 9.4. *Let us consider the characteristic polynomial*

$$p_A(\lambda) = \lambda^5 + 5\lambda^4 + 10\lambda^3 + 11\lambda^2 + 7\lambda + 2 \tag{9.39}$$

We need to determine whether it is Hurwitz or not applying the geometric criterion. Taking $\lambda = i\omega$ for (9.39) one has

$$\begin{aligned} p_A(i\omega) &= i\omega^5 + 5\omega^4 - i10\omega^3 - 11\omega^2 + i7\omega + 2 \\ &= \left[5\omega^4 - 11\omega^2 + 2\right] + i\left[\omega\left(\omega^4 - 10\omega^2 + 7\right)\right] \end{aligned}$$

So,

$$\begin{aligned} U(\omega) &= 5\omega^4 - 11\omega^2 + 2 \\ V(\omega) &= \omega\left(\omega^4 - 10\omega^2 + 7\right) \end{aligned}$$

The corresponding godograph and its zoom-form are depicted at Figs. 9.1 and 9.2 correspondingly.

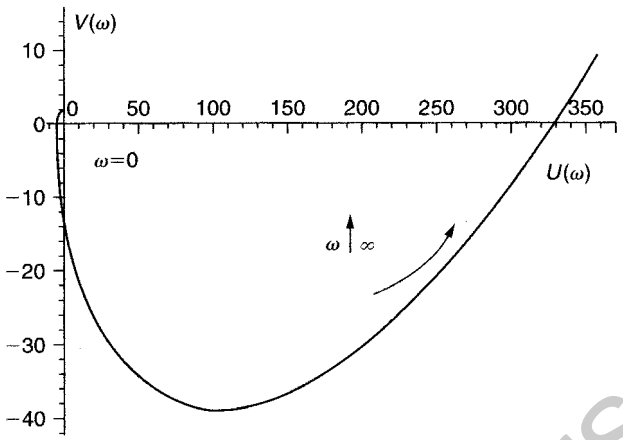

Fig. 9.1. The godograph of $p_A(i\omega)$.

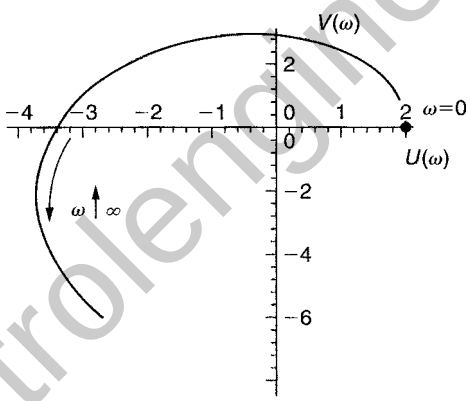

Fig. 9.2. The zoom-form of the godograph of $p_A(i\omega)$.

In view of the geometric form of this godograph we may conclude that this polynomial is Hurwitz.

Using this geometric approach one may determine not only if a polynomial is stable or not, but also determine the exact number (l) of stable, (r) unstable and (m) neutral roots.

Example 9.5. *Suppose that the polynomial* $p_A(\lambda)$ *has the order* $n = 5$ *and its godograph has the form as at Fig. 9.3. Notice that this godograph does not cross the origin* $(0, 0)$ *and therefore* $p_A(\lambda)$ *does not have neutral (with a real part equal to zero) roots, that is,* $m = 0$. *Hence by (9.37) we conclude that*

$$l + r + m = n = 5$$
$$m = 0, \quad l - r = 3$$

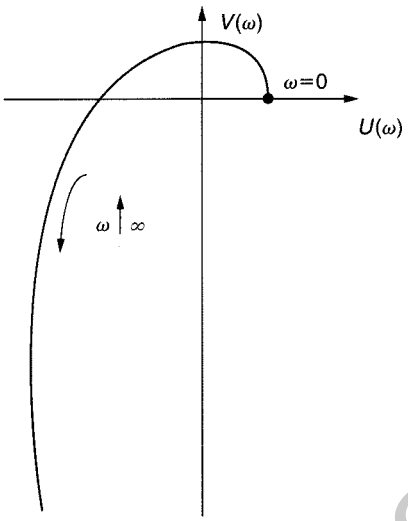

Fig. 9.3. The godograph of $p_A(i\omega)$.

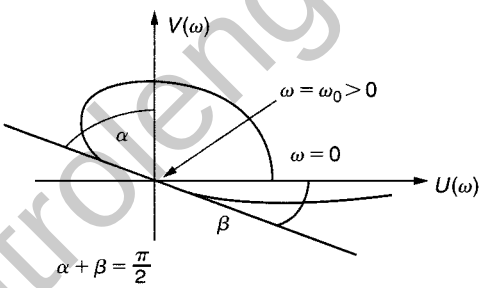

Fig. 9.4. The godograph of $p_A(i\omega)$.

which gives

$$l = 4, \quad r = 1, \quad m = 0$$

Example 9.6. *Suppose that the godograph of* $p_A(i\omega)$ *is as at Fig. 9.4 and corresponds to a polynomial* $p_A(\lambda)$ *of the order* $n = 6$. *One can see that this godograph crosses the origin* $(0, 0)$. *This means that* $p_A(\lambda)$ *has a root a with a real part equal to zero. But, as we have only one cross which corresponds to a frequency* $\omega = \omega_0 \neq 0$, *it means that there exist two complex conjugated roots such that*

$$\lambda_i(A) = i\omega_0, \quad \bar{\lambda}_i(A) = -i\omega_0$$

This means that $m = 2$. *So, by (9.37) we conclude that*

$$l + r + m = n = 6$$
$$\overset{\infty}{\underset{\omega=0}{\Delta}} \arg p(i\omega) = \pi, \quad l - r = 2, \quad m = 2$$

This finally gives:

$$l = 3, \quad r = 1, \quad m = 2$$

9.7 Polynomial robust stability

9.7.1 Parametric uncertainty and robust stability

As shown before, the stability property of a matrix $A \in \mathbb{R}^{n\times n}$ is characterized by the root's location of the corresponding characteristic polynomial $p_A(\lambda)$ (9.3). Evidently, any variations ΔA of the matrix A, namely, $A = A_0 + \Delta A$, are transformed into the variations of the coefficients a_j $(j = 1, \ldots, n)$ of the corresponding characteristic polynomial $p_A(\lambda)$. Denote the collection of its coefficients by

$$\boxed{a := (a_1, \ldots, a_n)^\top \in \mathbb{R}^n} \tag{9.40}$$

and suppose that this vector of coefficients belongs to a *connected set* $\mathcal{A} \in \mathbb{R}^n$ that corresponds to possible variations ΔA of the matrix A or maybe includes them, that is,

$$\boxed{a \in \mathcal{A} \in \mathbb{R}^n} \tag{9.41}$$

Definition 9.2. *A characteristic polynomial* $p_A(\lambda)$ *(9.3) is said to be* ***robust stable****, if for any* $a \in \mathcal{A}$ *the roots of the corresponding polynomial belongs to the left-hand side of the complex plane* $\mathbb{C}$, *i.e.,*

$$\boxed{\operatorname{Re}\lambda_j(A) < 0 \,(j = 1, \ldots, n)} \tag{9.42}$$

for all $a \in \mathcal{A}$.

Definition 9.3. *Denote by* $\mathcal{Q}_A(\omega)$ *the set of all values of the vector*

$$p_A(i\omega) = U(\omega) + iV(\omega)$$

given in $\mathbb{C}$ *under a fixed* $\omega \in [0, \infty)$ *when the parameters a take all possible values in* $\mathcal{A}$, *that is,*

$$\boxed{\mathcal{Q}_A(\omega) := \{z : z = p_A(i\omega) \mid a \in \mathcal{A}\}} \tag{9.43}$$

The next result represents the criterion of the polynomial robust stability and is a keystone in robust control theory.

Theorem 9.5. (The criterion of polynomial robust stability) *The characteristic polynomial* $p_A(\lambda)$ *(9.3) is* ***robust stable*** *if and only if*

1. *The class* $\mathcal{A}$ *of polynomials* $p_A(\lambda)$ *(9.3) contains at least one Hurwitz polynomial* $p_A^*(\lambda)$*, named a basic one.*
2. *The following* ***principle of "zero-excluding"*** *holds: the set* $\mathcal{Q}_A(\omega)$ *does not contain the origin ("zero-point"), i.e.,*

$$\boxed{0 \notin \mathcal{Q}_A(\omega)} \tag{9.44}$$

Proof. Since the vector $z = p_A(j\omega) \in \mathbb{C}$ is continually dependent on the vector parameter a, then a "transition" from stable polynomial to unstable one (when we are varying the coefficients a) may occur (this is always possible since the set $\mathcal{A}$ of parameters is a connected set) only when one of its roots crosses the imaginary axis, or, in other words, when there exists $\omega_0 \in [0, \infty)$ such that $p_A(i\omega_0) = U(\omega_0) + iV(\omega_0) = 0$. But this is equivalent to the following identity

$$U(\omega_0) = V(\omega_0) = 0$$

which means exactly that $0 \in \mathcal{Q}_A(\omega)$. Evidently, to avoid this effect it is necessary and sufficient to fulfill conditions 1 and 2 of this theorem. Theorem is proven. □

9.7.2 Kharitonov's theorem

Theorem 9.6. Kharitonov (1978) *Let the set* $\mathcal{A}$*, characterizing a parametric uncertainty, be defined as*

$$\boxed{\mathcal{A} := \left\{a \in \mathbb{R}^n : \; a_i^- \le a_i \le a_i^+ \;\; (i = 1, \ldots n)\right\}} \tag{9.45}$$

Then the polynomial $p_A(\lambda)$ *(9.3) is robust stable if and only if* ***four polynomials given below are stable*** *(Hurwitz):*

$$\boxed{\begin{array}{l} p_A^{(1)}(\lambda) := 1 + a_1^-\lambda + a_2^+\lambda^2 + a_3^+\lambda^3 + a_4^-\lambda^4 + a_5^-\lambda^5 + \cdots \\ p_A^{(2)}(\lambda) := 1 + a_1^+\lambda + a_2^+\lambda^2 + a_3^-\lambda^3 + a_4^-\lambda^4 + a_5^+\lambda^5 + \cdots \\ p_A^{(3)}(\lambda) := 1 + a_1^+\lambda + a_2^-\lambda^2 + a_3^-\lambda^3 + a_4^+\lambda^4 + a_5^+\lambda^5 + \cdots \\ p_A^{(4)}(\lambda) := 1 + a_1^-\lambda + a_2^-\lambda^2 + a_3^+\lambda^3 + a_4^+\lambda^4 + a_5^-\lambda^5 + \cdots \end{array}} \tag{9.46}$$

Proof. For any $a \in \mathcal{A}$

$$U(\omega) = 1 - a_2\omega^2 + a_4\omega^4 - \cdots$$
$$V(\omega) = a_1\omega - a_3\omega^3 + a_5\omega^5 \cdots$$

and hence for any $\omega \in [0, \infty)$

$$U^-(\omega) \le U(\omega) \le U^+(\omega) \quad \text{and} \quad V^-(\omega) \le V(\omega) \le V^+(\omega)$$

where

$$U^-(\omega) = 1 - a_2^+\omega^2 + a_4^-\omega^4 - \cdots$$
$$U^+(\omega) = 1 - a_2^-\omega^2 + a_4^+\omega^4 - \cdots$$

and

$$V^-(\omega) = a_1^-\omega - a_3^+\omega^3 + a_5^-\omega^5 \cdots$$
$$V^+(\omega) = a_1^+\omega - a_3^-\omega^3 + a_5^+\omega^5 \cdots$$

That's why for any $\omega \in [0, \infty)$ the set $\mathcal{Q}_A(\omega)$ (9.43) is rectangular (see Fig. 9.5) with width $[U^+(\omega) - U^-(\omega)]$ and height $[V^+(\omega) - V^-(\omega)]$ and with the center in the point $\mathring{p}_A(j\omega)$ corresponding to the stable polynomial with parameters $\mathring{a}_i = \frac{1}{2}(a_i^- + a_i^+)$.

Notice that the vertices of the set $\mathcal{Q}_A(\omega)$ correspond exactly to the polynomials (9.46). Suppose now that this rectangle touches the origin by one of its sides. Since the argument monotonically increases the vertices of this touching side will rotate in the clock-wise direction, and, hence, will become non-vertical which contradicts our previous concept. So, the direct application of the previous Theorem 9.5 leads to the formulated result. Theorem is proven. □

Example 9.7. *Let us find the parameter β for which the polynomial*

$$p_A(\lambda) = 1 + a_1\lambda + a_2\lambda^2 + a_3\lambda^3$$
$$1 - \beta \le a_1 \le 1 + \beta$$
$$1.5 \le a_2 \le 2, \quad a_3 = 1$$

is robust stable. To do this construct four polynomials (9.46):

$$p_A^{(1)}(\lambda) := 1 + (1 - \beta)\lambda + 2\lambda^2 + \lambda^3$$
$$p_A^{(2)}(\lambda) := 1 + (1 + \beta)\lambda + 2\lambda^2 + \lambda^3$$
$$p_A^{(3)}(\lambda) := 1 + (1 + \beta)\lambda + 1.5\lambda^2 + \lambda^3$$
$$p_A^{(4)}(\lambda) := 1 + (1 - \beta)\lambda + 1.5\lambda^2 + \lambda^3$$

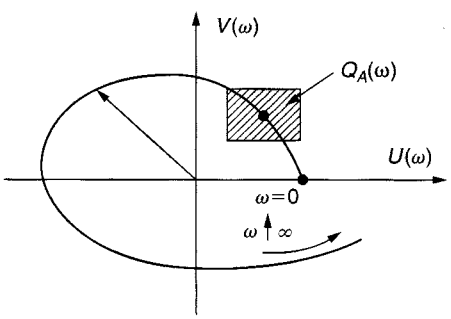

Fig. 9.5. Illustration of the Kharitonov's criterion.

The corresponding Hurwitz matrices H^A *are as follows:*

$$\begin{bmatrix} 1-\beta & 1 & 0 \\ 1 & 2 & 0 \\ 0 & (1-\beta) & 1 \end{bmatrix}, \quad \begin{bmatrix} 1+\beta & 1 & 0 \\ 1 & 2 & 0 \\ 0 & (1+\beta) & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1+\beta & 1 & 0 \\ 1 & 1.5 & 0 \\ 0 & 1+\beta & 1 \end{bmatrix}, \quad \begin{bmatrix} 1-\beta & 1 & 0 \\ 1 & 1.5 & 0 \\ 0 & 1-\beta & 1 \end{bmatrix}$$

By the Liénard–Chipart criterion we find that the conditions of the robust stability are

$$1-\beta > 0,\ 1+\beta > 0 \quad \textit{or, equivalently,} \quad |\beta| < 1$$

and

$$2(1-\beta)-1 > 0, \quad 2(1+\beta)-1 > 0$$

$$1.5(1+\beta)-1 > 0, \quad 1.5(1-\beta)-1 > 0$$

which leads to the following:

$$\beta < 0.5,\ \beta > -0.5, \quad \beta > \frac{2}{3}-1 = -\frac{1}{3}, \quad \beta < 1-\frac{2}{3} = \frac{1}{3}$$

or, equivalently,

$$|\beta| < 0.5, \quad |\beta| < \frac{1}{3}$$

Finally, all constraints taken together give

$$|\beta| < \frac{1}{3}$$

9.7.3 The Polyak–Tsypkin geometric criterion

Let the set $\mathcal{A}$ of all possible parameters a be defined as follows:

$$\boxed{\mathcal{A} := \left\{ a \in \mathbb{R}^n : |a_i - a_i^*| \le \gamma \alpha_i \ \ (i = 1, \ldots, n),\ \gamma > 0 \right\}} \tag{9.47}$$

Construct the following polynomials:

$$\boxed{\begin{aligned} U_a(\omega) &:= 1 - a_2\omega^2 + a_4\omega^4 - \cdots \\ V_a(\omega) &:= a_1\omega - a_3\omega^3 + a_5\omega^5 \cdots \\ T_a(\omega) &:= \frac{1}{\omega} V_a(\omega), \quad \omega > 0 \\ S(\omega) &:= 1 + \alpha_2\omega^2 + \alpha_4\omega^4 + \cdots \\ V(\omega) &:= \alpha_1\omega + \alpha_3\omega^3 + \alpha_5\omega^5 \cdots \\ T(\omega) &:= \frac{1}{\omega} V(\omega), \quad \omega > 0 \end{aligned}} \tag{9.48}$$

and define

$$\boxed{X(\omega) := \frac{U_{a^*}(\omega)}{S(\omega)}, \quad Y(\omega) := \frac{T_{a^*}(\omega)}{T(\omega)}} \tag{9.49}$$

Theorem 9.7. Polyak & Tsypkin (1990) *The characteristic polynomial* $p_A(\lambda)$ *(9.3) is* ***robust stable*** *if and only if the godograph*

$$\boxed{Z(i\omega) := X(\omega) + iY(\omega)} \tag{9.50}$$

passes exactly n *quadrants, when* ω *varies from* 0 *up to* ∞*, and does not cross the quadrant* Γ_γ *with the center in the origin and with the board equal to* 2γ *(see Fig. 9.6)*

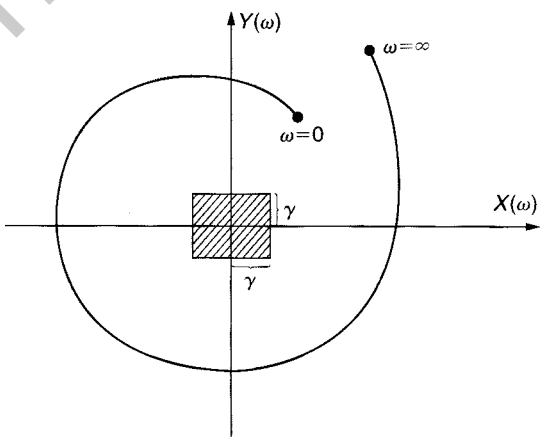

Fig. 9.6. Illustration of the Polyak–Tsypkin criterion.

such that

$$\boxed{|X(0)|, \quad |Y(0)|, \quad |X(\infty)|, \quad |Y(\infty)| > \gamma} \tag{9.51}$$

Proof. For any $a \in \mathcal{A}$ we have

$$|U_a(\omega) - U_{a^*}(\omega)| \leq \gamma S(\omega)$$
$$|V_a(\omega) - V_{a^*}(\omega)| \leq \gamma V(\omega)$$

or

$$|T_a(\omega) - T_{a^*}(\omega)| \leq \gamma T(\omega)$$

The condition that

$$0 \in \mathcal{Q}_A(\omega)$$

for some $\omega > 0$ means that $U_{\tilde{a}}(\omega) = V_{\tilde{a}}(\omega) = 0$ for some $\tilde{a} \in \mathcal{A}$ which implies

$$|U_{\tilde{a}}(\omega) - U_{a^*}(\omega)| = |U_{a^*}(\omega)| \leq \gamma S(\omega)$$
$$|V_{\tilde{a}}(\omega) - V_{a^*}(\omega)| = |V_{a^*}(\omega)| \leq \gamma V(\omega)$$

and

$$|T_{\tilde{a}}(\omega) - T_{a^*}(\omega)| = |T_{a^*}(\omega)| \leq \gamma T(\omega)$$

Since $S(\omega) > 0$ and $T(\omega) > 0$, we obtain

$$|X(\omega)| = \left|\frac{U_{a^*}(\omega)}{S(\omega)}\right| \leq \gamma, \quad |Y(\omega)| = \left|\frac{T_{a^*}(\omega)}{T(\omega)}\right| \leq \gamma$$

which means that

$$Z(i\omega) \in \Gamma_\gamma$$

Contrarily, the conditions that $0 \notin \mathcal{Q}_A(\omega_0)$ are

$$Z(i\omega) \notin \Gamma_\gamma$$
$$|X(0)|, \quad |Y(0)|, \quad |X(\infty)|, \quad |Y(\infty)| > \gamma$$

Then the result follows from Theorem 9.5. Theorem is proven. □

Remark 9.2. *The "maximal stability radius" $\gamma = \gamma_{\max}$ corresponds to the maximal quadrant $\Gamma_{\gamma_{\max}}$ which touches the godograph $Z(j\omega)$ from inside.*

9.8 Controllable, stabilizable, observable and detectable pairs

In this subsection we shall turn to some important concepts that will be used frequently in the following.

9.8.1 Controllability and a controllable pair of matrices

Definition 9.4. *The linear stationary system*

$$\boxed{\begin{array}{c}\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \\ \\ A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times r}\end{array}} \tag{9.52}$$

or the pair (A, B) *is said to be* ***controllable*** *on a time-interval* $[0, T]$ *if, for any initial state* x_0 *and any terminal state* x_T*, there exists a feasible (piecewise continuous) control* $u(t)$ *such that the solution of (9.52) satisfies*

$$x(T) = x_T \tag{9.53}$$

Otherwise, the system or pair (A, B) *is said to be* ***uncontrollable****.*

The next theorem represents some algebraic criteria (the necessary and sufficient conditions) of the controllability.

Theorem 9.8. (The criteria of the controllability) *The pair* (A, B) *is* ***controllable*** *if and only if one of the following properties holds:*

Criterion 1. *The* ***controllability grammian***

$$\boxed{G_c(t) := \int_{\tau=0}^{t} e^{A\tau} B B^{\mathsf{T}} e^{A^{\mathsf{T}}\tau} d\tau} \tag{9.54}$$

is positive definite for any $t \in [0, \infty)$*.*

Criterion 2. *The* ***controllability matrix***

$$\boxed{C := \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix}} \tag{9.55}$$

has full rank or, in other words,

$$\langle A \quad \operatorname{Im} B \rangle := \sum_{i=1}^{n} \operatorname{Im} \langle A^{i-1} B \rangle = \mathbb{R}^n \tag{9.56}$$

where $\operatorname{Im} B$ *is the image (range) of* $B : \mathbb{R}^r \longmapsto \mathbb{R}^n$ *defined by*

$$\operatorname{Im} B := \{ y \in \mathbb{R}^n : y = Bu, \ u \in \mathbb{R}^r \}\ B \tag{9.57}$$

Criterion 3. *The* ***Hautus matrix***

$$\boxed{\begin{bmatrix} A - \lambda I & \vdots & B \end{bmatrix}}$$

has full row rank for all $\lambda \in \mathbb{C}$*.*

Criterion 4. *For any left eigenvalues λ and the corresponding eigenvectors x of the matrix A, i.e., $x^*A = \lambda x^*$, the following property holds: $x^*B \neq 0$. In other words, all modes of A are B-controllable.*

Criterion 5. *The eigenvalues of the matrix $(A + BK)$ can be freely assigned by a suitable selection of K.*

Proof. **Criterion 1.**

(a) *Necessity.* Suppose that the pair (A, B) is controllable, but for some $t_1 \in [0, T]$ the grammian of controllability $G_c(T)$ is singular, that is, there exists a vector $x \neq 0$ such that

$$0 = x^\top \left[\int_{\tau=0}^{t_1} e^{A\tau} B B^\top e^{A^\top \tau} d\tau \right] x$$

$$= \left[\int_{\tau=0}^{t_1} x^\top e^{A\tau} B B^\top e^{A^\top \tau} x d\tau \right] = \int_{\tau=0}^{t_1} \left\| B^\top e^{A^\top \tau} x \right\|^2 d\tau$$

So,

$$x^\top e^{A\tau} B = 0 \tag{9.58}$$

for all $\tau \in [0, t_1]$. Select t_1 as a terminal instant, that is, $t_1 = T$ and $x(T) = x_T = 0$. Then by (9.59)

$$0 = x(t_1) = e^{At_1} x_0 + \int_{\tau=0}^{t_1} e^{A(t_1 - \tau)} B u(\tau) d\tau$$

and pre-multiplying the last equation by $x^\top$ we obtain

$$0 = x^\top x(t_1) = x^\top e^{At_1} x_0 + \int_{\tau=0}^{t_1} x^\top e^{A(t_1 - \tau)} B u(\tau) d\tau = x^\top e^{At_1} x_0$$

Selecting the initial conditions $x_0 = e^{-At_1} x$, we obtain $\|x\|^2 = 0$, or $x = 0$, which contradicts the assumption that $x \neq 0$.

(b) *Sufficiency.* Suppose conversely: $G_c(t) > 0$ for all $t \in [0, T]$. Hence, $G_c(T) > 0$. Define

$$u(t) := -B^\top e^{A^\top (T-t)} G_c^{-1}(T) \left[e^{AT} x_0 - x_T \right]$$

Then, by (9.52),

$$x(t) = e^{At} x_0 + \int_{\tau=0}^{t} e^{A(t-\tau)} B u(\tau) d\tau \tag{9.59}$$

which gives

$$x(T) = e^{AT}x_0 - \left[\int_{\tau=0}^{T} e^{A(T-\tau)}BB^{\mathrm{T}}e^{A^{\mathrm{T}}(T-t)}G_c^{-1}(T)\left[e^{AT}x_0 - x_T\right]d\tau\right]$$

$$= e^{AT}x_0 - \left[\int_{\tau=0}^{T} e^{A(T-\tau)}BB^{\mathrm{T}}e^{A^{\mathrm{T}}(T-t)}\,d\tau\right]G_c^{-1}(T)\left[e^{AT}x_0 - x_T\right]$$

$$\overset{T-\tau=s}{=} e^{AT}x_0$$

$$+\left[\int_{s=T}^{0} e^{As}BB^{\mathrm{T}}e^{A^{\mathrm{T}}s}\,ds\right]G_c^{-1}(T)\left[e^{AT}x_0 - x_T\right] = e^{AT}x_0$$

$$- G_c(T)\,G_c^{-1}(T)\left[e^{AT}x_0 - x_T\right] = x_T$$

So, the pair (A, B) is controllable. The first criterion is proven.

Criterion 2.

(a) *Necessity.* Suppose that $G_c(t) > 0$ for any $t \in [0, T]$, but the controllability matrix C has no full rank, that is, there exists a nonzero-vector $v \in \mathbb{R}^n$ such that

$$v^*A^iB = 0 \quad \text{for all} \quad i = 0, 1 \ldots, n-1$$

But by the Cayley–Hamilton theorem any matrix satisfies its own characteristic equation, namely, if

$$\det(A - \lambda I) = a_0\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n = 0, \quad a_0 \neq 0$$

then

$$a_0A^n + a_1A^{n-1} + \cdots + a_{n-1}A + a_nI = 0, \quad a_0 \neq 0$$

or, equivalently,

$$A^n = -\left(\frac{a_1}{a_0}A^{n-1} + \cdots + \frac{a_{n-1}}{a_0}A + \frac{a_n}{a_0}I\right)$$

and hence

$$v^*A^nB = -\left(\frac{a_1}{a_0}v^*A^{n-1}B + \cdots + \frac{a_{n-1}}{a_0}v^*AB + \frac{a_n}{a_0}v^*B\right) = 0$$

By the same reason

$$v^*A^{n+1}B = -\left(\frac{a_1}{a_0}v^*A^nB + \cdots + \frac{a_{n-1}}{a_0}v^*A^2B + \frac{a_n}{a_0}v^*AB\right) = 0$$

and so on. So,

$$v^*A^iB = 0 \quad \text{for any} \quad i \geq 0 \tag{9.60}$$

But since $e^{At} = \sum_{i=0}^{\infty} \frac{1}{i!} (At)^i$, in view of (9.60), for all $t \geq 0$ we have

$$v^* e^{At} B = \sum_{i=0}^{\infty} \frac{1}{i!} v^* A^i B t^i = 0$$

which implies

$$0 = v^* \int_{t=0}^{t_1 \leq T} e^{At} B B^\intercal e^{A^\intercal t} \, dt = v^* G_c(t_1)$$

for all $t_1 \leq T$ which is in contradiction with the assumption that $G_c(t_1)$ is nonsingular. So, C should have full rank.

(b) *Sufficiency*. Conversely, suppose now that C has full rank, but $G_c(t)$ is singular for some $t = t_1 \leq T$. Then, by (9.58), there exists a vector $x^\intercal \neq 0$ such that $x^\intercal e^{A\tau} B = 0$ for all $\tau \in [0, t_1]$. Taking $t = 0$, we get $x^\intercal B = 0$. Evaluating the ith derivatives at the point $t = 0$, we have

$$0 = x^\intercal \left(\frac{d}{d\tau} e^{A\tau} \right)_{t=0} B = x^\intercal A^i B, \quad i = 0, 1, \ldots, n-1$$

which implies

$$\begin{bmatrix} B & AB & A^2 B & \cdots & A^{n-1} B \end{bmatrix} = x^\intercal C = 0$$

It means that C has no full rank. This is in contradiction with the initial assumption that C has full rank. So, $G_c(t)$ should be nonsingular for all $t \in [0, T]$. The second criterion is proven too.

Criterion 3.

(a) *Necessity*. On the contrary, suppose that $[A - \lambda I \vdots B]$ has no full row rank for some $\lambda \in \mathbb{C}$, that is, there exists a vector $x^* \neq 0$ such that $x^*[A - \lambda I \vdots B] = 0$ but the system is controllable (C has full rank). This is equivalent to the following:

$$x^* A = \lambda x^*, \quad x^* B = 0$$

which results in

$$\begin{aligned} x^* C &= x^* \begin{bmatrix} B & AB & A^2 B & \cdots & A^{n-1} B \end{bmatrix} \\ &= \begin{bmatrix} \underbrace{x^* B}_{0} & \underbrace{\lambda x^* B}_{0} & \underbrace{\lambda^2 x^* B}_{0} & \cdots & \lambda^{n-1} \underbrace{x^* B}_{0} \end{bmatrix} = 0 \end{aligned}$$

But this is in contradiction with the assumption that C has full rank.

(b) *Sufficiency.* Suppose that $[A - \lambda I \vdots B]$ has full row rank for all $\lambda \in \mathbb{C}$, but $\mathcal{C}$ has no full rank, i.e., $x^*\mathcal{C} = 0$ for some $x \neq 0$. Representing this x as a linear combination of the eigenvectors x^{i*} of the matrix A as

$$x^* = \sum_{i=1}^{n} \alpha_i x^{i*}, \qquad \left(\sum_{i=1}^{n} \alpha_i^2 > 0 \right)$$

we get

$$\begin{aligned} 0 = x^*\mathcal{C} &= \sum_{i=1}^{n} \alpha_i x^{i*} \mathcal{C} \\ &= \sum_{i=1}^{n} \alpha_i x^{i*} \begin{bmatrix} B & AB & A^2B & \cdots & A^{n-1}B \end{bmatrix} \\ &= \sum_{i=1}^{n} \alpha_i x^{i*} \begin{bmatrix} B & \lambda_i B & \lambda_i^2 B & \cdots & \lambda_i^{n-1} B \end{bmatrix} \\ &= \sum_{i=1}^{n} \alpha_i x^{i*} \begin{bmatrix} I & \lambda_i I & \lambda_i^2 I & \cdots & \lambda_i^{n-1} I \end{bmatrix} B = \tilde{x}^* B \end{aligned}$$

where

$$\tilde{x}^* := \sum_{i=1}^{n} \alpha_i x^{i*} \begin{bmatrix} I & \lambda_i I & \lambda_i^2 I & \cdots & \lambda_i^{n-1} I \end{bmatrix}$$

So, there exists a vector $\tilde{x} \neq 0$ such that $\tilde{x}^* B = 0$ and

$$\begin{aligned} \tilde{x}^* A &= \sum_{i=1}^{n} \alpha_i x^{i*} \begin{bmatrix} I & \lambda_i I & \lambda_i^2 I & \cdots & \lambda_i^{n-1} I \end{bmatrix} A \\ &= \tilde{x}^* A = \sum_{i=1}^{n} \alpha_i x^{i*} A \begin{bmatrix} I & \lambda_i I & \lambda_i^2 I & \cdots & \lambda_i^{n-1} I \end{bmatrix} \\ &= \sum_{i=1}^{n} \alpha_i \lambda_i x^{i*} \begin{bmatrix} I & \lambda_i I & \lambda_i^2 I & \cdots & \lambda_i^{n-1} I \end{bmatrix} = \tilde{\lambda} \tilde{x}^* \end{aligned}$$

where

$$\tilde{\lambda} := \frac{\tilde{x}^* A \tilde{x}}{\tilde{x}^* \tilde{x}}$$

which is in contradiction with the assumption that the Hautus matrix $[A - \tilde{\lambda} I \vdots B]$ has full row rank.

Criterion 4. It directly follows from Criterion 3.

Criterion 5. The proof can be found in Zhou *et al.*, 1996.

Theorem is proven. □

9.8.2 Stabilizability and a stabilizable pair of matrices

Definition 9.5. *The linear stationary system (9.52) or the pair (A, B) is said to be* ***stabilizable*** *if there exists a state feedback $u(t) = Kx(t)$ such that the closed-loop system is stable, i.e., the matrix $A + BK$ is stable (Hurwitz). Otherwise, the system or pair (A, B) is said to be* ***unstabilizable****.*

Theorem 9.9. (Two criteria of stabilizability) *The pair (A, B) is stabilizable if and only if*

Criterion 1. *The Hautus matrix $[A - \lambda I \vdots B]$ has the full rank for all $\operatorname{Re}\lambda \geq 0$.*

Criterion 2. *For all λ and x such that $x^*A = \lambda x^*$ and $\operatorname{Re}\lambda \geq 0$, it follows that $x^*B \neq 0$.*

Proof. This theorem is a consequence of the previous one. □

9.8.3 Observability and an observable pair of matrices

Let us consider the following stationary linear system supplied by an output model:

$$\left.\begin{array}{c} \dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad t \in [0, \infty] \\ y(t) = Cx(t) \\ A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times r} \end{array}\right\} \tag{9.61}$$

where $y(t) \in \mathbb{R}^m$ is treated as an output vector and $C \in \mathbb{R}^{m\times n}$ is an output matrix.

Definition 9.6. *The stationary linear system (9.61) or the pair (C, A) is said to be* ***observable*** *if, for any time t_1, the initial state $x(0) = x_0$ can be determined from the history of the input $u(t)$ and the output $y(t)$ within the interval $[0, t_1]$. Otherwise, the system or pair (C, A) is said to be* ***unobservable****.*

Theorem 9.10. (The criteria of observability) *The pair (C, A) is observable if and only if one of the following criteria hold:*

Criterion 1. *The* ***observability grammian***

$$G_o(t) := \int_{\tau=0}^{t} e^{A^\top \tau} C^\top C e^{A\tau} d\tau \tag{9.62}$$

is positive definite for any $t \in [0, \infty)$.

Criterion 2. *The* ***observability matrix***

$$\boxed{\mathcal{O} := \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix}} \tag{9.63}$$

has the full column rank or, in other words,

$$\bigcap_{i=1}^{n} \operatorname{Ker}\left(CA^{i-1}\right) = 0 \tag{9.64}$$

where $\operatorname{Ker}(A)$ *is the* ***kernel*** *or* ***null space*** *of* $A : \mathbb{R}^n \longmapsto \mathbb{R}^m$ *defined by*

$$\operatorname{Ker}(A) = \mathcal{N}(A) := \{x \in \mathbb{R}^n : Ax = 0\} \tag{9.65}$$

Criterion 3. *The Hautus matrix*

$$\boxed{\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}}$$

has full column rank for all $\lambda \in \mathbb{C}$.

Criterion 4. *Let* λ *and* y *be any eigenvalue and any corresponding right eigenvector of* A, *that is,* $Ay = \lambda y$, *then* $Cy \neq 0$.

Criterion 5. *The eigenvalues of the matrix* $A + LC$ *can be freely assigned (complex eigenvalues are in conjugated pairs) by a suitable choice of* L.

Criterion 6. *The pair* $(A^\top, C^\top)$ *is controllable.*

Proof. **Criterion 1.**

(a) *Necessity.* Suppose that the pair (C, A) is observable, but for some t_1 the grammian of observability $G_0(t_1)$ is singular, that is, there exists a vector $x \neq 0$ such that

$$0 = x^\top \left[\int_{\tau=0}^{t_1} e^{A^\top \tau} C^\top C e^{A\tau}\, d\tau\right] x$$

$$= \left[\int_{\tau=0}^{t_1} x^\top e^{A^\top \tau} C^\top C e^{A\tau} x\, d\tau\right] = \int_{\tau=0}^{t_1} \left\|Ce^{A\tau}x\right\|^2 d\tau$$

So,

$$Ce^{A\tau}x = 0 \tag{9.66}$$

for all $\tau \in [0, t_1]$. Then by (9.59)

$$x(t_1) = e^{At_1}x_0 + \int_{\tau=0}^{t_1} e^{A(t_1-\tau)} Bu(\tau)\, d\tau$$

and, hence,

$$y(t_1) = Cx(t_1) = Ce^{At_1}x_0 + \int_{\tau=0}^{t_1} Ce^{A(t_1-\tau)}Bu(\tau)\,d\tau$$

or

$$v(t_1) := y(t_1) - \int_{\tau=0}^{t_1} Ce^{A(t_1-\tau)}Bu(\tau)\,d\tau = Ce^{At_1}x_0$$

Selecting the initial conditions $x_0 = 0$, we obtain $v(t_1) = 0$. But we have the same results for any $x_0 = x \neq 0$ satisfying (9.66) which means that x_0 cannot be determined from the history of the process. This contradicts that (C, A) is observable.

(b) *Sufficiency*. Suppose conversely: $G_o(t) > 0$ for all $t \in [0, \infty]$. Hence,

$$0 < x^T \left[\int_{\tau=0}^{t} e^{A^T\tau}C^TCe^{A\tau}\,d\tau\right] x$$

$$= \left[\int_{\tau=0}^{t_1} x^Te^{A^T\tau}C^TCe^{A\tau}x\,d\tau\right] = \int_{\tau=0}^{t_1} \left\|Ce^{A\tau}x\right\|^2 d\tau$$

which implies that there exists a time $\tau_0 \in [0, t]$ such that $\left\|Ce^{A\tau_0}x\right\|^2 > 0$ for any $x \neq 0$. This means that $Ce^{A\tau_0}$ is a full rank matrix $\left(e^{A^T\tau_0}C^TCe^{A\tau_0} > 0\right)$. Then

$$v(\tau_0) := y(\tau_0) - \int_{\tau=0}^{\tau_0} Ce^{A(\tau_0-\tau)}Bu(\tau)\,d\tau = Ce^{A\tau_0}x_0$$

and, hence,

$$e^{A^T\tau_0}C^Tv(\tau_0) = e^{A^T\tau_0}C^TCe^{A\tau_0}x_0$$

and

$$x_0 = \left[e^{A^T\tau_0}C^TCe^{A\tau_0}\right]^{-1} e^{A^T\tau_0}C^Tv(\tau_0)$$

So, the pair (C, A) is observable. The first criterion is proven.

Criterion 2.

(a) *Necessity*. Suppose that the pair (C, A) is observable, but that the observability matrix $\mathcal{O}$ does not have full column rank, i.e., there exists a vector $\tilde{x} \neq 0$ such that $\mathcal{O}\tilde{x} = 0$ or, equivalently,

$$CA^i\tilde{x} = 0 \;\forall i = 0, 1, \ldots, n-1$$

Suppose now that $x_0 = \tilde{x}$. Then, by the Cayley–Hamilton theorem

$$v(t) := y(t) - \int_{\tau=0}^{t} Ce^{A(t-\tau)}Bu(\tau)\,d\tau = Ce^{At}x_0$$
$$= C\sum_{i=0}^{\infty}\frac{1}{i!}(At)^i x_0 = \sum_{i=0}^{n-1}\frac{t^i}{i!}\underbrace{CA^i x_0}_{0} + \sum_{i=n}^{2n-1}\frac{t^i}{i!}\underbrace{CA^i x_0}_{0} + \underbrace{\ldots}_{0} = 0 \tag{9.67}$$

which implies

$$v(t) = Ce^{At}x_0 = 0$$

and, hence, x_0 cannot be determined from $v(t) \equiv 0$. We obtain the contradiction.

(b) *Sufficiency.* From (9.67) it follows that

$$\tilde{v} := \begin{bmatrix} v(0) \\ \dot{v}(0) \\ \ddot{v}(0) \\ \vdots \\ v^{(n-1)}(0) \end{bmatrix} = \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix} x_0 = \mathcal{O}x_0$$

and since $\mathcal{O}$ has a full rank, then $\mathcal{O}^\top\mathcal{O} > 0$ and hence

$$x_0 = [\mathcal{O}^\top\mathcal{O}]^{-1}\mathcal{O}^\top\tilde{v}$$

which means that x_0 may be uniquely defined. This completes the proof.

Criteria 3–6. They follows from by the duality of Criterion 6 to the corresponding criteria of controllability, since the controllability of the pair $(A^\top, C^\top)$ is equivalent to the existence of a matrix $L^\top$ such that $A^\top + C^\top L^\top$ is stable. But then it follows that

$$(A^\top + C^\top L^\top)^\top = A + LC$$

is also stable which coincides with Criteria 6 of observability.
Theorem is proven. □

9.8.4 Detectability and a detectable pair of matrices

Definition 9.7. *The stationary linear system (9.61) or the pair (C, A) is said to be **detectable** if the matrix $A + LC$ is stable (Hurwitz) for some L. Otherwise, the system or pair (C, A) is said to be **undetectable**.*

Theorem 9.11. (The criteria of detectability) *The pair (C, A) is detectable if and only if one of the following criteria holds:*

Criterion 1. *The Hautus matrix* $\begin{bmatrix} A - \lambda I \\ C \end{bmatrix}$ *has full column rank for all Re $\lambda \geq 0$.*

Criterion 2. *Let λ and y be any eigenvalue and any corresponding right eigenvector of A, such that $Ay = \lambda y$, $Re\,\lambda \geq 0$, then $Cy \neq 0$.*

Criterion 3. *There exists a matrix L such that the matrix $A + LC$ is stable.*

Criterion 4. *The pair $(A^\top, C^\top)$ is stabilizable.*

Proof. It follows from the duality of Criterion 4 of this theorem to the corresponding criterion of stabilizability. Theorem is proven. □

9.8.5 Popov–Belevitch–Hautus (PBH) test

Definition 9.8. *Let λ be an eigenvalue of the matrix A, or equivalently, a **mode** of the system (9.61). Then the mode λ is said to be*

*1. **controllable** if*

$$\boxed{x^* B \neq 0}$$

for all left eigenvectors x^* of the matrix A associated with this λ, i.e.,

$$\boxed{x^* A = \lambda x^*, \quad x^* \neq 0}$$

*2. **observable** if*

$$\boxed{Cx \neq 0}$$

for all right eigenvectors x of the matrix A associated with this λ, i.e.,

$$\boxed{Ax = \lambda x, \quad x \neq 0}$$

Otherwise, the mode is called uncontrollable (unobservable).

Using this definition we may formulate the following test-rule (Popov–Belevitch–Hautus PBH-test (Hautus & Silverman (1983)) for the verification of the properties discussed above.

Claim 9.1. (PBH test) *The system (9.61) is*

*1. **controllable** if and only if every mode is controllable;*
*2. **stabilizable** if and only if every unstable mode is controllable;*
*3. **observable** if and only if every mode is observable;*
*4. **detectable** if and only if every unstable mode is observable.*

10 Algebraic Riccati Equation

Contents

10.1 Hamiltonian matrix

Let us consider the *matrix Riccati equation*[1]

$$\boxed{PA + A^{\mathsf{T}}P + Q - PBR^{-1}B^{\mathsf{T}}P = 0} \tag{10.1}$$

and the associated $2n \times 2n$ *Hamiltonian matrix*:

$$\boxed{H := \begin{bmatrix} A & -BR^{-1}B^{\mathsf{T}} \\ -Q & -A^{\mathsf{T}} \end{bmatrix}} \tag{10.2}$$

Lemma 10.1. *The spectrum $\sigma(H)$ of the set of eigenvalues of H (10.2) is symmetric about the imaginary axis.*

Proof. To see this, introduce the $2n \times 2n$ matrix:

$$J := \begin{bmatrix} 0 & -I_{n\times n} \\ I_{n\times n} & 0 \end{bmatrix} \tag{10.3}$$

having the evident properties

$$J^2 = -I_{2n\times 2n}$$
$$J^{-1} = -J$$

So, we have

$$J^{-1}HJ = -JHJ = -H^{\mathsf{T}} \tag{10.4}$$

which implies that λ is an eigenvalue of H (10.2) if and only if $(-\bar{\lambda})$ is too (see Fig. 10.1). □

[1] In the Russian technical literature this equation is known as the matrix Lurie equation.

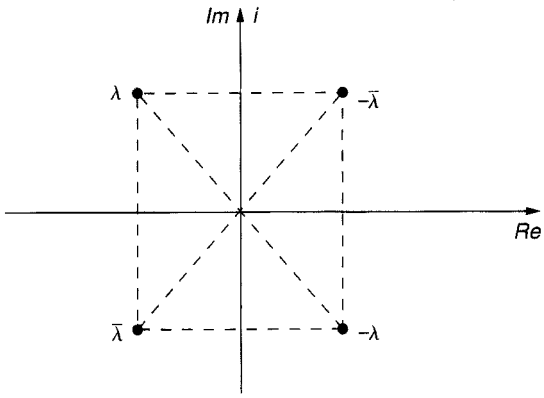

Fig. 10.1. Hamiltonian eigenvalues.

10.2 All solutions of the algebraic Riccati equation

10.2.1 Invariant subspaces

Definition 10.1. *Let $A: \mathbb{C}^n \longmapsto \mathbb{C}^n$ be a linear transformation (matrix), λ be an eigenvalue of A and x be a corresponding eigenvector of A, that is, $Ax = \lambda x$. So, $A(\alpha x) = \lambda(\alpha x)$ for any $\alpha \in \mathbb{R}$.*

1. *We say that the eigenvector x defines a* ***one-dimensional subspace*** *which is* ***invariant*** *with respect to pre-multiplication by A since*

$$A^k(\alpha x) = \lambda^k(\alpha x) \quad k = 1, \ldots$$

2. *Generalizing the definition before, we say that a* ***subspace*** *$S \subset \mathbb{C}^n$ is* ***invariant with respect to the transformation*** *A, or A-invariant, if*

$$Ax \in S \text{ for any } x \in S$$

or, in other words,

$$AS \subset S$$

3. *If one of the eigenvalues has a multiplicity l, i.e. $\lambda_1 = \lambda_2 = \ldots = \lambda_l$, then the* ***generalized eigenvectors*** *x_i $(i = 1, \ldots, l)$ are obtained by the following rule*

$$(A - \lambda_1 I)\, x_i = x_{i-1}, \quad i = 1, \ldots, l, \quad x_0 := 0 \tag{10.5}$$

10.2.2 Main theorems on the solution presentation

Theorem 10.1. *Let $\Theta \subset \mathbb{C}^{2n}$ be an n-dimensional invariant subspace of H, that is, if $z \in \Theta$ then $Hz \in \Theta$, and let $P_1, P_2 \in \mathbb{C}^{n \times n}$ be two complex matrices such that*

$$\Theta = \operatorname{Im} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

which means that the columns of $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ *may be considered as a basis in* Θ. *If* P_1 *is invertible, then*

$$\boxed{P = P_2 P_1^{-1}} \tag{10.6}$$

is a solution to the matrix Riccati equation (10.1) which is independent of a specific choice of bases of Θ.

Proof. Since Θ is an invariant subspace of H, there exists a matrix $\Lambda \in \mathbb{C}^{n\times n}$ such that

$$H \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda \tag{10.7}$$

Indeed, let the matrix $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ be formed by the eigenvectors of H such that

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = [v_1 \cdots v_n]$$

where each vector v_i satisfies the equation

$$H v_i = \lambda_i v_i$$

Here λ_i are the corresponding eigenvalues. Combining these equations for all $i = 1, \ldots, n$, we obtain

$$H \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = H [v_1 \cdots v_n] = [\lambda_1 v_1 \cdots \lambda_n v_n]$$

$$= [v_1 \cdots v_n] \begin{bmatrix} \lambda_1 & 0 \cdots 0 & 0 \\ 0 & \cdots & 0 \\ 0 & 0 \cdots 0 & \lambda_n \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda$$

$$\Lambda = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$$

In the extended form, the relation (10.7) is

$$\begin{bmatrix} A & -BR^{-1}B^{\mathsf{T}} \\ -Q & -A^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda \tag{10.8}$$

Post-multiplying this equation by P_1^{-1} we get

$$\begin{bmatrix} A & -BR^{-1}B^{\mathsf{T}} \\ -Q & -A^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I_{n\times n} \\ (P_2 P_1^{-1}) \end{bmatrix} = \begin{bmatrix} I_{n\times n} \\ (P_2 P_1^{-1}) \end{bmatrix} P_1 \Lambda P_1^{-1}$$

Then, the pre-multiplication of this equality by $\left[-\left(P_2 P_1^{-1}\right) \vdots I_{n\times n} \right]$ implies

$$\begin{aligned}
&\left[-\left(P_2 P_1^{-1}\right) \vdots I_{n\times n} \right] \begin{bmatrix} A & -BR^{-1}B^{\mathsf{T}} \\ -Q & -A^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I_{n\times n} \\ \left(P_2 P_1^{-1}\right) \end{bmatrix} \\
&= \left[-\left(P_2 P_1^{-1}\right) \vdots I_{n\times n} \right] \begin{bmatrix} A - BR^{-1}B^{\mathsf{T}}\left(P_2 P_1^{-1}\right) \\ -Q - A^{\mathsf{T}}\left(P_2 P_1^{-1}\right) \end{bmatrix} \\
&= -\left(P_2 P_1^{-1}\right) A + \left(P_2 P_1^{-1}\right) BR^{-1}B^{\mathsf{T}}\left(P_2 P_1^{-1}\right) - Q - A^{\mathsf{T}}\left(P_2 P_1^{-1}\right) \\
&= \left[-\left(P_2 P_1^{-1}\right) \vdots I_{n\times n} \right] \begin{bmatrix} I_{n\times n} \\ \left(P_2 P_1^{-1}\right) \end{bmatrix} P_1 \Lambda P_1^{-1} = 0
\end{aligned}$$

which means that $P := P_2 P_1^{-1}$ satisfies (10.1). Let T be a nonsingular matrix. Then any other basis from $\begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix}$ spanning Θ can be represented as

$$\begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} T = \begin{bmatrix} P_1 T \\ P_2 T \end{bmatrix}$$

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} \tilde{P}_1 \\ \tilde{P}_2 \end{bmatrix} T^{-1} = \begin{bmatrix} \tilde{P}_1 T^{-1} \\ \tilde{P}_2 T^{-1} \end{bmatrix}$$

and, hence,

$$\begin{aligned}
P = P_2 P_1^{-1} &= \left(\tilde{P}_2 T^{-1}\right)\left(\tilde{P}_1 T^{-1}\right)^{-1} \\
&= \tilde{P}_2 T^{-1} T \left(\tilde{P}_1\right)^{-1} = \tilde{P}_2 \left(\tilde{P}_1\right)^{-1}
\end{aligned}$$

which proves the theorem. □

Corollary 10.1. *The relation (10.8) implies*

$$\begin{aligned}
&AP_1 - BR^{-1}B^{\mathsf{T}} P_2 = P_1 \Lambda \\
&A - \left(BR^{-1}B^{\mathsf{T}}\right) P = P_1 \Lambda P_1^{-1}
\end{aligned} \tag{10.9}$$

Theorem 10.2. *If $P \in \mathbb{C}^{n\times n}$ is a solution to the matrix Riccati equation (10.1), then there exist matrices P_1, $P_2 \in \mathbb{C}^{n\times n}$ with P_1 invertible such that (10.6) holds, that is,*

$$\boxed{P = P_2 P_1^{-1}}$$

and the columns of $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ form a basis of Θ.

Proof. Define

$$\tilde{\Lambda} := A - \left(BR^{-1}B^{\mathsf{T}}\right) P$$

Pre-multiplying it by P gives

$$P\tilde{\Lambda} := PA - P\left(BR^{-1}B^{\mathsf{T}}\right) P = -Q - A^{\mathsf{T}}P$$

These two relations may be rewritten as

$$\begin{bmatrix} A & -BR^{-1}B^{\mathsf{T}} \\ -Q & -A^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} I_{n\times n} \\ P \end{bmatrix} = \begin{bmatrix} I_{n\times n} \\ P \end{bmatrix} \tilde{\Lambda}$$

Hence, the columns of $\begin{bmatrix} I_{n\times n} \\ P \end{bmatrix}$ span the invariant subspace Θ and defining $P_1 := I_{n\times n}$ and $P_2 = P$ completes the proof. □

10.2.3 Numerical example

Example 10.1. *Let*

$$A = \begin{bmatrix} -3 & 2 \\ -2 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad R = I_{2\times 2}, \quad Q = 0$$

$$H = \begin{bmatrix} -3 & 2 & 0 & 0 \\ -2 & 1 & 0 & -1 \\ 0 & 0 & 3 & 2 \\ 0 & 0 & -2 & -1 \end{bmatrix}$$

Then the eigenvalues of H are $1, 1, (-1), (-1)$ and the eigenvector and the generalized eigenvector (10.5) corresponding to $\lambda_1 = \lambda_2 = 1$ are

$$v_1 = (1, 2, 2, -2)^{\mathsf{T}}, \quad v_2 = (-1, -3/2, 1, 0)^{\mathsf{T}}$$

and the eigenvector and the generalized eigenvector corresponding to $\lambda_3 = \lambda_4 = -1$ are

$$v_3 = (1, 1, 0, 0)^{\mathsf{T}}, \quad v_4 = (1, 3/2, 0, 0)^{\mathsf{T}}$$

Several possible solutions of (10.1) are given below:

1. *span*$\{v_1, v_2\} := \{z \in \mathbb{C}^{2n\times 2n} : z = \alpha v_1 + \beta v_2,\ \alpha, \beta \in \mathbb{R}\}$ *is H-invariant: let*

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$$

then

$$P = P_2 P_1^{-1} = \begin{bmatrix} 2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 2 & -3/2 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 2 & 1 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} -3.0 & 2.0 \\ -4.0 & 2.0 \end{bmatrix} = \begin{bmatrix} -10.0 & 6.0 \\ 6.0 & -4.0 \end{bmatrix}$$

2. *span* $\{v_1, v_3\}$ *is H-invariant: let* $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} v_1 & v_3 \end{bmatrix}$, *then*

$$P = P_2 P_1^{-1} = \begin{bmatrix} 2 & 0 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}^{-1}$$

$$= \begin{bmatrix} 2 & 0 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} -1.0 & 1.0 \\ 2.0 & -1.0 \end{bmatrix} = \begin{bmatrix} -2.0 & 2.0 \\ 2.0 & -2.0 \end{bmatrix}$$

3. *span* $\{v_3, v_4\}$ *is H-invariant: let* $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} v_3 & v_4 \end{bmatrix}$, *then*

$$P = P_2 P_1^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 3/2 \end{bmatrix}^{-1} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

4. *Notice that here span* $\{v_1, v_4\}$, *span* $\{v_2, v_3\}$ *and span* $\{v_2, v_4\}$ *are not H-invariant.*

Remark 10.1. *If a collection of eigenvectors of H forms a basis in* $\mathbb{C}^n$ *which defines a solution of the Riccati matrix equation given by* $P = P_2 P_1^{-1}$, *then the number* N_{Ric} *of all possible solutions of this equation is*

$$N_{Ric} = \frac{(2n)!}{n!\,(2n-n)!} = \frac{(2n)!}{(n!)^2}$$

10.3 Hermitian and symmetric solutions

10.3.1 No pure imaginary eigenvalues

Theorem 10.3. *Let* $\Theta \subset \mathbb{C}^{2n}$ *be an n-dimensional invariant subspace of H and let* $P_1, P_2 \in \mathbb{C}^{n\times n}$ *be two complex matrices such that* $\Theta = \mathrm{Im} \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$. *Then the assumption*

$$\lambda_i + \bar{\lambda}_j \neq 0 \quad \textit{for all} \quad i, j = 1, \ldots, 2n \tag{10.10}$$

where $\lambda_i, \bar{\lambda}_j$ *are the eigenvalues of H, implies*

1. $P_1^* P_2$ *is* ***Hermitian****, that is,*

$$P_1^* P_2 = P_2^* P_1$$

2. *if, in addition,* P_1 *is nonsingular, the matrix* $P = P_2 P_1^{-1}$ *is Hermitian too, that is,*

$$P^* = \left(P_2 P_1^{-1}\right)^* = P$$

Remark 10.2. *The condition (10.10) is equivalent to the restriction*

$$\operatorname{Re} \lambda_i \neq 0 \quad \textit{for all} \quad i = 1, \ldots, 2n \tag{10.11}$$

which means that H ***has no eigenvalues on the imaginary axis****.*

Proof. Since Θ is an invariant subspace of H, then there exists a matrix Λ such that spectrums of the eigenvalues of Λ and H coincide, that is,

$$\sigma(\Lambda) = \sigma(H)$$

and (10.7) holds:

$$H \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda \tag{10.12}$$

Pre-multiplying this equation by $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* J$, we get

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* JH \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* J \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda$$

By (10.4), it follows that JH is symmetric and, hence, is Hermitian (since H is real). So, we obtain that the left-hand side is Hermitian, and, as a result, the right-hand side is Hermitian too:

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* J \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \Lambda = \Lambda^* \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* J^* \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

$$= -\Lambda^* \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}^* J \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

which implies

$$X\Lambda + \Lambda^* X = 0$$
$$X := \left(-P_1^* P_2 + P_2^* P_1\right)$$

But this is a Lyapunov equation which has a unique solution $X = 0$ if $\lambda_i(\Lambda) + \bar{\lambda}_j(\Lambda) \neq 0$. But, since the spectrum of eigenvalues of Λ and H coincides, we obtain the proof of the claim. Moreover, if P_1 is nonsingular, then for $P = P_2 P_1^{-1}$ it follows that

$$P^* = \left(P_2 P_1^{-1}\right)^* = \left(P_1^{-1}\right)^* P_2^* = \left(P_1^{-1}\right)^* \left(P_1^* P_2 P_1^{-1}\right)$$

$$= \left(P_1^*\right)^{-1} P_1^* P_2 P_1^{-1} = P_2 P_1^{-1} = P$$

Theorem is proven. □

Theorem 10.4. *Suppose a Hamiltonian matrix H (10.2) has no pure imaginary eigenvalues and $\mathcal{X}_-(H)$ and $\mathcal{X}_+(H)$ are n-dimensional invariant subspaces corresponding to eigenvalues $\lambda_i(H)$ $(i = 1, \ldots n)$ in $\operatorname{Re} s < 0$ and to $\lambda_i(H)$ $(i = n+1, \ldots, 2n)$ in $\operatorname{Re} s > 0$, respectively, that is, $\mathcal{X}_-(H)$ has the basis*

$$\begin{bmatrix} v_1 \cdots v_n \end{bmatrix} = \begin{bmatrix} v_{1,1} & \cdot & v_{n,1} \\ \cdot & \cdot & \cdot \\ v_{1,n} & \cdot & v_{n,n} \\ v_{1,n+1} & \cdot & v_{n,n+1} \\ \cdot & \cdot & \cdot \\ v_{1,2n} & \cdot & v_{n,2n} \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$$

$$P_1 = \begin{bmatrix} v_{1,1} & \cdot & v_{n,1} \\ \cdot & \cdot & \cdot \\ v_{1,n} & \cdot & v_{n,n} \end{bmatrix}, \quad P_2 = \begin{bmatrix} v_{1,n+1} & \cdot & v_{n,n+1} \\ \cdot & \cdot & \cdot \\ v_{1,2n} & \cdot & v_{n,2n} \end{bmatrix}$$

Then P_1 is invertible, i.e. P_1^{-1} exists if and only if the pair (A, B) is stabilizable.

Proof. Sufficiency. Let the pair (A, B) be stabilizable. We want to show that P_1 is nonsingular. Contrariwise, suppose that there exists a vector $x_0 \neq 0$ such that $P_1 x_0 = 0$. Then we have the following. First, notice that

$$x_0^* P_2^* \left(B R^{-1} B^\top\right) P_2 x_0 = \left\| R^{-1/2} B^\top P_2 x_0 \right\|^2 = 0 \tag{10.13}$$

or, equivalently,

$$R^{-1/2} B^\top P_2 x_0 = 0 \tag{10.14}$$

Indeed, the pre-multiplication of (10.12) by $\begin{bmatrix} I & 0 \end{bmatrix}$ implies

$$A P_1 - \left(B R^{-1} B^\top\right) P_2 = P_1 \Lambda \tag{10.15}$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix with elements from $\operatorname{Re} s < 0$. Then, pre-multiplying the last equality by $x_0^* P_2^*$, post-multiplying by x_0 and using the symmetricity of $P_2^* P_1 = P_1^* P_2$ we get

$$x_0^* P_2^* \left[A P_1 - \left(B R^{-1} B^\mathsf{T}\right) P_2\right] x_0$$
$$= -x_0^* P_2^* \left(B R^{-1} B^\mathsf{T}\right) P_2 x_0 = x_0^* P_2^* P_1 \Lambda x_0$$
$$= x_0^* P_1^* P_2 \Lambda x_0 = (P_1 x_0)^* P_2 \Lambda x_0 = 0$$

which implies (10.13). Pre-multiplying (10.12) by $[I\ 0]$, we get

$$-Q P_1 - A^\mathsf{T} P_2 = P_2 \Lambda \tag{10.16}$$

Post-multiplying (10.16) by x_0 we obtain

$$(-Q P_1 - A^\mathsf{T} P_2)\, x_0 = -A^\mathsf{T} P_2 x_0 = P_2 \Lambda x_0 = \lambda_0 P_2 x_0$$

where

$$\lambda_0 = \frac{x_0^* \Lambda x_0}{\|x_0\|^2}$$

which implies

$$0 = A^\mathsf{T} P_2 x_0 + P_2 \lambda_0 x_0 = (A^\mathsf{T} + \lambda_0 I)\, P_2 x_0$$

Taking into account that, by (10.13),

$$\left(B R^{-1} B^\mathsf{T}\right) P_2 x_0 = 0$$

it follows that

$$[\,(A^\mathsf{T} + \lambda_0 I) \vdots \left(B R^{-1} B^\mathsf{T}\right)\,] P_2 x_0 = 0$$

Then, the stabilizability of (A, B) (see Criterion 1 of stabilizability) implies that $P_2 x_0 = 0$. So,

$$\begin{bmatrix} P_1 \\ P_2 \end{bmatrix} x_0 = 0$$

and, since $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ forms the basis and, hence, has a full rank, we get $x_0 = 0$, which is a contradiction.

Necessity. Let P_1 be invertible. Hence, by (10.15)

$$A - \left(B R^{-1} B^\mathsf{T}\right) P_2 P_1^{-1} = P_1 \Lambda P_1^{-1}$$

Since the spectrum of the eigenvalues of $P_1 \Lambda P_1^{-1}$ coincides with one of Λ, we may conclude that the matrix

$$A_{closed} := A - \left(B R^{-1} B^\mathsf{T}\right) P_2 P_1^{-1}$$

is stable and, hence, the pair $\left(A, BR^{-1}B^{\mathsf{T}}\right)$ is stabilizable (in the corresponding definition $K = P_2P_1^{-1}$). It means that for all λ and x such that $Ax = \lambda x$ and $\mathrm{Re}\,\lambda \geq 0$, in other words, for all unstable modes of A

$$x^*BR^{-1}B^{\mathsf{T}} \neq 0 \tag{10.17}$$

which implies

$$x^*B \neq 0$$

Indeed, by the contradiction, assuming that $x^*B = 0$, we obtain $x^*BR^{-1}B^{\mathsf{T}} = 0$ which violates (10.17). □

Corollary 10.2. *The stabilizability of the pair* (A, B) *implies that the matrix*

$$A_{closed} := A - \left(BR^{-1}B^{\mathsf{T}}\right)P_2P_1^{-1} \tag{10.18}$$

is stable (Hurwitz).

Proof. Post-multiplying (10.12) by P_1^{-1} we get

$$H\begin{bmatrix} I \\ P \end{bmatrix} = \begin{bmatrix} I \\ P_2 \end{bmatrix} P_1\Lambda P_1^{-1}, \quad P = P_2P_1^{-1}$$

which after pre-multiplication by $\begin{bmatrix} I & 0 \end{bmatrix}$ gives

$$\begin{bmatrix} I & 0 \end{bmatrix} H \begin{bmatrix} I \\ P_2P_1^{-1} \end{bmatrix}$$

$$= \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} A - \left(BR^{-1}B^{\mathsf{T}}\right)P \\ -Q - A^{\mathsf{T}}P \end{bmatrix} = A - \left(BR^{-1}B^{\mathsf{T}}\right)P$$

$$= A_{closed} = \begin{bmatrix} I & 0 \end{bmatrix} \begin{bmatrix} I \\ P_2 \end{bmatrix} P_1\Lambda P_1^{-1} = P_1\Lambda P_1^{-1}$$

But $P_1\Lambda P_1^{-1}$ is stable, and hence A_{closed} is stable too. □

10.3.2 Unobservable modes

Theorem 10.5. *Assuming that the pair* (A, B) *is stabilizable, the Hamiltonian matrix* H *(10.2) has no pure imaginary eigenvalues if and only if the pair* (C, A)*, where* $Q = C^{\mathsf{T}}C$*, has no unobservable mode on the imaginary axis, that is, for all* λ *and* $x_1 \neq 0$ *such that* $Ax_1 = \lambda x_1$*,* $\lambda = i\omega$*, it follows that* $Cx_1 \neq 0$*.*

Proof. Suppose that $\lambda = i\omega$ is an eigenvalue and the corresponding eigenvector $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \neq 0$. Then

$$H \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} Ax_1 - BR^{-1}B^{\mathsf{T}}x_2 \\ -C^{\mathsf{T}}Cx_1 - A^{\mathsf{T}}x_2 \end{bmatrix} = i\omega \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} i\omega x_1 \\ i\omega x_2 \end{bmatrix}$$

After rearranging, we have

$$\begin{aligned} (A - i\omega I)\, x_1 &= BR^{-1}B^{\mathsf{T}}x_2 \\ -(A^{\mathsf{T}} - i\omega I)\, x_2 &= C^{\mathsf{T}}Cx_1 \end{aligned} \tag{10.19}$$

which implies

$$(x_2, (A - i\omega I)\, x_1) = \left(x_2, BR^{-1}B^{\mathsf{T}}x_2\right) = \left\|R^{-1/2}B^{\mathsf{T}}x_2\right\|$$

$$-(x_1, (A^{\mathsf{T}} - i\omega I)\, x_2) = -((A - i\omega I)\, x_1, x_2) = (x_1, C^{\mathsf{T}}Cx_1) = \|Cx_1\|^2$$

As a result, we get

$$\left\|R^{-1/2}B^{\mathsf{T}}x_2\right\| + \|Cx_1\|^2 = 0$$

and, hence,

$$B^{\mathsf{T}}x_2 = 0, \quad Cx_1 = 0$$

In view of this, from (10.19) it follows that

$$\begin{aligned} (A - i\omega I)\, x_1 &= BR^{-1}B^{\mathsf{T}}x_2 = 0 \\ -(A^{\mathsf{T}} - i\omega I)\, x_2 &= C^{\mathsf{T}}Cx_1 = 0 \end{aligned}$$

Combining the four last equations we obtain

$$x_2^* \left[(A - i\omega I)\; B\right] = 0$$

$$\begin{bmatrix} (A - i\omega I) \\ C \end{bmatrix} x_1 = 0$$

The stabilizability of (A, B) provides the full range for the matrix $\left[(A - i\omega I)\; B\right]$ and implies that $x_2 = 0$. So, it is clear that $i\omega$ is an eigenvalue of H if and only if it is an unobservable mode of (C, A), that is, the corresponding $x_1 = 0$ too. □

10.3.3 All real solutions

Theorem 10.6. *Let $\Theta \subset \mathbb{C}^{2n}$ be an n-dimensional invariant subspace of H and let $P_1, P_2 \in \mathbb{C}^{n\times n}$ be two complex matrices such that the columns of $\begin{bmatrix} P_1 \\ P_2 \end{bmatrix}$ form a basis of Θ and P_1 is nonsingular. Then $P = P_2 P_1^{-1}$ is* ***real*** *if and only if Θ is* ***conjugated symmetric****, i.e. $z \in \Theta$ implies that $\bar{z} \in \Theta$.*

Proof. Sufficiency. Since Θ is conjugated symmetric, then there exists a nonsingular matrix $\mathcal{N}$ such that

$$\begin{bmatrix} \bar{P}_1 \\ \bar{P}_2 \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \end{bmatrix} \mathcal{N}$$

Therefore,

$$\bar{P} = \bar{P}_2 \bar{P}_1^{-1} = (P_2\mathcal{N})(P_1\mathcal{N})^{-1} = P_2\mathcal{N}\mathcal{N}^{-1}P_1^{-1} = P_2P_1^{-1} = P$$

So, P is real.

Necessity. We have $\bar{P} = P$. By assumption $P \in \mathbb{R}^{n\times n}$ and, hence,

$$\operatorname{Im}\begin{bmatrix} I \\ P \end{bmatrix} = \Theta = \operatorname{Im}\begin{bmatrix} I \\ \bar{P} \end{bmatrix}$$

Therefore, Θ is a conjugated symmetric subspace. □

Remark 10.3. *Based on this theorem, we may conclude that* ***to form a basis in an invariant conjugated symmetric subspace we need to use the corresponding pairs of the complex conjugated symmetric eigenvectors or its linear nonsingular transformation (if n is odd then there exists a real eigenvalue to which an eigenvector should be added to complete a basis) which guarantees that P is real.***

10.3.4 Numerical example

Example 10.2. *Let*

$$A = \begin{bmatrix} -1 & 2 \\ 0 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad R = I_{2\times 2}, \quad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$H = \begin{bmatrix} -1 & 2 & 0 & 0 \\ 0 & 1 & 0 & -1 \\ -1 & 0 & 1 & 0 \\ 0 & -1 & -2 & -1 \end{bmatrix}$$

The eigenvalues λ_i and the corresponding eigenvectors v_i are as follows:

$$\lambda_1 = -1.4053 + 0.68902i \quad v_1 = \begin{bmatrix} -0.4172 - 0.50702i \\ 0.25921 - 4.0993 \times 10^{-2}i \\ -0.10449 - 0.24073i \\ 0.59522 - 0.27720i \end{bmatrix}$$

$$\lambda_2 = -1.4053 - 0.68902i \quad v_2 = \begin{bmatrix} -0.50702 - 0.4172i \\ -4.0993 \times 10^{-2} + 0.25921i \\ -0.24073 - 0.10449i \\ -0.27720 + 0.59522i \end{bmatrix}$$

$$\lambda_3 = 1.4053 + 0.68902i \quad v_3 = \begin{bmatrix} 2.9196 \times 10^{-2} + 0.44054i \\ -0.11666 + 0.53987i \\ -0.49356 - 0.24792i \\ 0.41926 - 0.1384i \end{bmatrix}$$

$$\lambda_4 = 1.4053 - 0.68902i \quad v_4 = \begin{bmatrix} -0.44054 - 2.9196 \times 10^{-2}i \\ -0.53987 + 0.11666i \\ 0.24792 + 0.49356i \\ 0.1384 - 0.41926i \end{bmatrix}$$

Notice that $(-i\,v_2) = \bar{v}_1$ and $(i\,v_4) = \bar{v}_3$ which corresponds to the fact that the eigenvectors stay the same being multiplied by a complex number. Then forming the basis in two-dimensional subspace as

$$\begin{bmatrix} v_1\ v_2 \end{bmatrix} = \begin{bmatrix} -0.4172 - 0.50702i & -0.50702 - 0.4172i \\ 0.25921 - 4.0993 \times 10^{-2}i & -4.0993 \times 10^{-2} + 0.25921i \\ -0.10449 - 0.24073i & -0.24073 - 0.10449i \\ 0.59522 - 0.27720i & -0.27720 + 0.59522i \end{bmatrix}$$

we may define

$$P_1 := \begin{bmatrix} -0.4172 - 0.50702i & -0.50702 - 0.4172i \\ 0.25921 - 4.0993 \times 10^{-2}i & -4.0993 \times 10^{-2} + 0.25921i \end{bmatrix}$$

$$P_1^{-1} = \begin{bmatrix} -0.13800 + 0.8726i & 1.7068 + 1.4045i \\ -0.8726 + 0.13800i & -1.4045 - 1.7068i \end{bmatrix}$$

and

$$P_2 := \begin{bmatrix} -0.10449 - 0.24073i & -0.24073 - 0.10449i \\ 0.59522 - 0.27720i & -0.27720 + 0.59522i \end{bmatrix}$$

Hence,

$$P = P_2 P_1^{-1} = \begin{bmatrix} 0.44896 & 0.31952 \\ 0.31949 & 2.8105 \end{bmatrix}$$

and we may see that P is a real matrix. Also we have

$$A_{closed} := A - \left(BR^{-1}B^{\intercal}\right) P_2 P_1^{-1}$$

$$= \begin{bmatrix} -1.0 & 2.0 \\ -0.31949 & -1.8105 \end{bmatrix}$$

with the eigenvalues:

$$\lambda_1 (A_{closed}) = -1.4053 + 0.68902i$$

$$\lambda_2 (A_{closed}) = -1.4053 - 0.68902i$$

10.4 Nonnegative solutions

10.4.1 Main theorems on the algebraic Riccati equation solution

Theorem 10.7. *The matrix Riccati equation (10.1)*

$$\boxed{PA + A^{\intercal}P + Q - PBR^{-1}B^{\intercal}P = 0} \tag{10.20}$$

has a ***unique nonnegative definite solution*** $P = P^{\intercal} \geq 0$ *which provides stability to the matrix*

$$\boxed{A_{closed} := A - BR^{-1}B^{\intercal}P} \tag{10.21}$$

corresponding to the original dynamic system

$$\boxed{\dot{x} = Ax + Bu} \tag{10.22}$$

closed by the linear feedback control given by

$$\boxed{u = -Kx = -R^{-1}B^{\intercal}Px \quad with \quad K = R^{-1}B^{\intercal}P} \tag{10.23}$$

if and only if the pair (A, B) is ***stabilizable*** *and the pair (C, A) where*

$$\boxed{Q = C^{\intercal}C} \tag{10.24}$$

has no unobservable mode on the imaginary axis.

Proof. The existence of $P = P_2 P_1^{-1}$ and its symmetricity and reality are already proven. We need to prove only that $P \geq 0$. Let us represent (10.1) in the following form

$$\begin{aligned} &PA + A^{\intercal}P + Q - K^{\intercal}RK = 0 \\ &RK = B^{\intercal}P \end{aligned} \tag{10.25}$$

By (10.25) it follows that

$$\begin{aligned} &PA_{closed} + A_{closed}^{\mathsf{T}} P = -(Q + K^{\mathsf{T}} RK) \\ &A_{closed} := A - BK, \quad K = R^{-1} B^{\mathsf{T}} P \end{aligned} \tag{10.26}$$

Since $(Q + K^{\mathsf{T}} RK) \geq 0$, by the Lyapunov Lemma 9.1 it follows that $P \geq 0$. □

Example 10.3. *Let us consider the following simple scalar dynamic system given by*

$$\dot{x} = ax + bu, \quad y = cx$$

with

$$a \neq 0, \quad b = 1, \quad c = 0$$

Notice that this system is completely unobservable! The corresponding Riccati equation (with $R = r = 1$) is $2ap - p^2 = p(2a - p) = 0$ and its solutions are $p_1 = 0$, $p_2 = 2a$. The case $a = 0$ corresponds to the case when the Hamiltonian (22.70) has the eigenvalues $(0, i0)$ on the imaginary axis. That's why this case is disregarded.

1. ***The case*** *$a < 0$. There exists the unique nonnegative solution $p = p_1 = 0$ of the Riccati equation which makes the closed-loop system stable. Indeed,*

$$a_{closed} := a - p = a < 0$$

2. ***The case*** *$a > 0$. Here the unique nonnegative solution of the Riccati equation making the closed-loop system stable is $p = p_2 = 2a$, since*

$$a_{closed} := a - p = -a < 0$$

So, ***the observability of a linear system is not necessary for making the closed-loop system stable with a stationary feedback designed as in (10.26)!***

Theorem 10.8. (On a positive definite solution) *If under the assumptions of the previous theorem additionally* ***the pair*** *(C, A)* ***is observable****, i.e. the matrix $\mathcal{O}$ (9.63) has the full column rank, then the solution P of the matrix Riccati equation (10.20) is strictly positive, that is, $P > 0$.*

Proof. Let us rewrite (10.20) as

$$PA + A^{\mathsf{T}} P - PBR^{-1} B^{\mathsf{T}} P = -Q \tag{10.27}$$

Suppose that for some vector $\tilde{x} \neq 0$ the condition $P\tilde{x} = 0$ holds. Then the post- and pre-multiplication of (10.27) by $\tilde{x}$ and $\tilde{x}^*$ leads to the following identity

$$0 = -\tilde{x}^* Q \tilde{x} = -\|C\tilde{x}\|^2 = 0$$

This means that $C\tilde{x} = 0$, or, equivalently, $\tilde{x}$ belongs to the unobservable subspace. Post-multiplying (10.27) by $\tilde{x}$ implies also that

$$PA\tilde{x} = 0$$

Using this fact, the post- and pre-multiplication of (10.27) by $A\tilde{x}$ and $\tilde{x}^* A^\intercal$ leads to the identity

$$0 = -\tilde{x}^* A^\intercal Q A\tilde{x} = -\|CA\tilde{x}\|^2 = 0$$

This means that $CA\tilde{x} = 0$, or, equivalently, $A\tilde{x}$ belongs to the unobservable subspace. Also it follows that

$$PA^2\tilde{x} = 0$$

Iterating this procedure we get that, for any $k = 0, 1, \ldots, n-1$, the following identities hold:

$$CA^k\tilde{x} = 0, \quad PA^k\tilde{x} = 0$$

This means exactly that $\mathcal{O}\tilde{x} := 0$ where $\tilde{x} \neq 0$. So, $\mathcal{O}$ is not a full column rank that contradicts the assumption of theorem, and, hence, $P > 0$. Theorem is proven. □

Corollary 10.3. *If there exists a vector $\tilde{x} \neq 0$ such that $P\tilde{x} = 0$, then* ***the pair*** (C, A) ***is unobservable.***

Summary 10.1. ***The matrix Riccati equation (10.1) has a unique positive definite solution:***

1. *if and only if* ***the pair*** (A, B) ***is stabilizable and the pair*** (C, A) ***has no neutral (on the imaginary axis) unobservable modes, that is,***

$$\text{if} \quad Ax = \lambda x, \quad \lambda = i\omega, \quad \text{then} \quad Cx \neq 0$$

2. *and if,* ***in addition, the pair*** (C, A) ***is observable, that is,***

$$\operatorname{rank} \mathcal{O} = n$$

The next simple example shows that the observability of the pair (C, A) **is not necessary** for the existence of a positive definite solution.

Example 10.4. (Zhou, Doyle & Glover (1996)) *Indeed, for*

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & 0 \end{bmatrix}$$

such that (A, B) is stabilizable, but (C, A) is not observable (even not detectable) the solution of the Riccati equation (10.1) is

$$P = \begin{bmatrix} 18 & -24 \\ -24 & 36 \end{bmatrix} > 0$$

11 Linear Matrix Inequalities

Contents

This chapter follows the fundamental book of Boyd *et al.* (1994) where there is shown that a wide variety of problems arising in system and control theory can be reduced to a few standard convex (or quasiconvex) optimization problems involving *linear matrix inequalities* (LMIs). Here we will also touch briefly the, so-called, *interior point method* Nesterov & Nemirovsky (1994) providing a powerful and efficient instrument to solve numerical LMIs arising in control theory.

11.1 Matrices as variables and LMI problem

11.1.1 Matrix inequalities

Definition 11.1. *A **linear matrix inequality** (LMI) has the block form*

$$0 < F(X) := \begin{bmatrix} F_{11}(X) & F_{12}(X) \\ F_{21}(X) & F_{22}(X) \end{bmatrix} = \begin{bmatrix} S_{11} + G_{11} X H_{11} + H_{11}^{\mathsf{T}} X G_{11}^{\mathsf{T}} & S_{12} + G_{12} X H_{12} + H_{12}^{\mathsf{T}} X G_{12}^{\mathsf{T}} \\ S_{21} + G_{21} X H_{21} + H_{21}^{\mathsf{T}} X G_{21}^{\mathsf{T}} & S_{22} + G_{22} X H_{22} + H_{22}^{\mathsf{T}} X G_{22}^{\mathsf{T}} \end{bmatrix} \tag{11.1}$$

*where the matrices $X \in \mathbb{R}^{n \times n}$, $S_{ii} \in \mathbb{R}^{n \times n}$ $(i = 1, 2)$ are **symmetric** and*

$$S_{ij} = S_{ji} \in \mathbb{R}^{n \times n}, \quad G_{ij} = G_{ji} \in \mathbb{R}^{n \times n}, \quad H_{ij} = H_{ji} \in \mathbb{R}^{n \times n} \; (i, j = 1, 2)$$

*such that each block $F_{ij}(X)$ is an **affine transformation** (mapping) from $\mathbb{R}^{n \times n}$ to $\mathbb{R}^{n \times n}$. This inequality means that $F(X)$ is positive definite, i.e. $u^{\mathsf{T}} F(X) u > 0$ for all nonzero $u \in \mathbb{R}^{2n}$. A **nonstrict LMI** has the form*

$$F(X) \geq 0 \tag{11.2}$$

Both inequalities (11.1) and (11.2) are closely related since the last one is equivalent to the following inequality

$$\tilde{F}(X) := F(X) + Q \geq Q > 0$$

where $Q \in \mathbb{R}^{2n\times 2n}$ is any positive definite matrix. So, without loss of generality we will consider below only strict LMIs (11.1).

Multiple LMIs

$$F^{(1)}(X) > 0, \ldots, F^{(p)}(X) > 0 \tag{11.3}$$

can be expressed as a single LMI

$$\operatorname{diag}\left(F^{(1)}(X), \ldots, F^{(p)}(X)\right) > 0 \tag{11.4}$$

Therefore we will make no distinction between a set of LMIs (11.3) and a single LMI (11.1).

Remark 11.1. *Nonlinear (convex) inequalities may be converted to LMI form using Schur complements. The basic idea of this relation is as follows: the LMI (11.1) is equivalent (see (7.14)) to the following systems of matrix inequalities:*

$$\boxed{\begin{array}{c} F_{11}(X) > 0 \\ F_{22}(X) > 0 \\ F_{11}(X) - F_{12}(X) F_{22}^{-1}(X) F_{12}^{\mathsf{T}}(X) > 0 \\ F_{22}(X) - F_{12}^{\mathsf{T}}(X) F_{11}^{-1}(X) F_{12}(X) > 0 \end{array}} \tag{11.5}$$

11.1.2 LMI as a convex constraint

Lemma 11.1. *The LMI (11.1) is a* ***convex*** *constraint on X, i.e., the set*

$$\{X \mid F(X) > 0\} \tag{11.6}$$

is convex.

Proof. Let for some $X, Y \in \mathbb{R}^{n\times n}$

$$F(X) > 0, \quad F(Y) > 0$$

Define $Z := \lambda X + (1-\lambda)Y$ for some $\lambda \in [0, 1]$. Then

$$
\begin{aligned}
F(Z) &= \begin{bmatrix} S_{11} + G_{11}ZH_{11} + H_{11}^{\mathsf{T}}ZG_{11}^{\mathsf{T}} & S_{12} + G_{12}ZH_{12} + H_{12}^{\mathsf{T}}ZG_{12}^{\mathsf{T}} \\ S_{21} + G_{21}ZH_{21} + H_{21}^{\mathsf{T}}ZG_{21}^{\mathsf{T}} & S_{22} + G_{22}ZH_{22} + H_{22}^{\mathsf{T}}ZG_{22}^{\mathsf{T}} \end{bmatrix} \\
&= \lambda \begin{bmatrix} S_{11} + G_{11}XH_{11} + H_{11}^{\mathsf{T}}XG_{11}^{\mathsf{T}} & S_{12} + G_{12}XH_{12} + H_{12}^{\mathsf{T}}XG_{12}^{\mathsf{T}} \\ S_{21} + G_{21}XH_{21} + H_{21}^{\mathsf{T}}XG_{21}^{\mathsf{T}} & S_{22} + G_{22}XH_{22} + H_{22}^{\mathsf{T}}XG_{22}^{\mathsf{T}} \end{bmatrix} \\
&\quad + (1-\lambda) \begin{bmatrix} S_{11} + \lambda G_{11}YH_{11} + \lambda H_{11}^{\mathsf{T}}YG_{11}^{\mathsf{T}} & S_{12} + \lambda G_{12}YH_{12} + \lambda H_{12}^{\mathsf{T}}YG_{12}^{\mathsf{T}} \\ S_{21} + G_{21}YH_{21} + H_{21}^{\mathsf{T}}XG_{21}^{\mathsf{T}} & S_{22} + G_{22}YH_{22} + H_{22}^{\mathsf{T}}YG_{22}^{\mathsf{T}} \end{bmatrix} \\
&= \lambda F(X) + (1-\lambda) F(Y) > 0
\end{aligned}
$$

which proves the result. □

11.1.3 Feasible and infeasible LMI

Given an LMI $F(X) > 0$ (11.1), the corresponding *LMI problem* is to find a feasible X^{feas} such that $F(X^{feas}) > 0$ or determine that this LMI is *infeasible*, or, in other words, LMI has no solution. Represent LMI $F(X) > 0$ (11.1) in the form

$$
\boxed{0 < F(X) = S + G(X)} \tag{11.7}
$$

where

$$
S := \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}
$$

$$
G(X) := \begin{bmatrix} G_{11}XH_{11} + H_{11}^{\mathsf{T}}XG_{11}^{\mathsf{T}} & G_{12}XH_{12} + H_{12}^{\mathsf{T}}XG_{12}^{\mathsf{T}} \\ G_{21}XH_{21} + H_{21}^{\mathsf{T}}XG_{21}^{\mathsf{T}} & G_{22}XH_{22} + H_{22}^{\mathsf{T}}XG_{22}^{\mathsf{T}} \end{bmatrix}
$$

LMI $F(X) > 0$ (11.1) is infeasible means that the affine set

$$
\left\{ F(X) > 0 \mid X \in \mathbb{R}^{n \times n} \right\}
$$

does not intersect the positive-definite cone. From convex analysis this is equivalent to the existence of a linear functional l that is positive on the positive-definite cone and nonpositive on the affine set of the matrix. By the Riss theorem 18.14 the linear functionals that are positive on the positive-definite cone are of the form $l = \mathrm{Tr}(LF)$, $L \geq 0$, $L \neq 0$. Here there is used the fact that the scalar product in the matrix space is $\mathrm{Tr}(\cdot)$. Since l is non-positive on the affine set $\left\{ F(X) > 0 \mid X \in \mathbb{R}^{n \times n} \right\}$ we may conclude that

$$
\mathrm{Tr}\,(LS) \leq 0, \quad \mathrm{Tr}\,(LG) = 0 \tag{11.8}
$$

So, *to prove that LMI* (11.1) *is infeasible means find a nonzero matrix* $L \geq 0$, $L \neq 0$ *which verifies* (11.8) *where S and G is the representation* (11.7) *of* $F(X)$.

11.2 Nonlinear matrix inequalities equivalent to LMI

11.2.1 Matrix norm constraint

The **matrix norm constraint**

$$\boxed{\|Z(X)\| < 1} \tag{11.9}$$

(where $Z(X) \in \mathbb{R}^{n\times q}$ depends affinely on X), or, equivalently,

$$I_{n\times n} - Z(X)Z^{\mathsf{T}}(X) > 0$$

is represented as

$$\boxed{\begin{bmatrix} I_{n\times n} & Z(X) \\ Z^{\mathsf{T}}(X) & I_{n\times n} \end{bmatrix} > 0} \tag{11.10}$$

11.2.2 Nonlinear weighted norm constraint

The **nonlinear weighted norm constraint**

$$\boxed{c^{\mathsf{T}}(X)P^{-1}(X)\,c(X) < 1} \tag{11.11}$$

(where $c(X) \in \mathbb{R}^n$, $0 < P(X) \in \mathbb{R}^{n\times n}$ depend affinely on X) is expressed as the following LMI

$$\boxed{\begin{bmatrix} P(X) & c(X) \\ c^{\mathsf{T}}(X) & 1 \end{bmatrix} > 0} \tag{11.12}$$

11.2.3 Nonlinear trace norm constraint

The **nonlinear trace norm constraint**

$$\boxed{\mathrm{Tr}\left(S^{\mathsf{T}}(X)P^{-1}(X)\,S(X)\right) < 1} \tag{11.13}$$

(where $S(X) \in \mathbb{R}^{n\times q}$, $0 < P(X) \in \mathbb{R}^{n\times n}$ depend affinely on X) is handled by introducing a new (slack) variable $Q = Q^{\mathsf{T}} \in \mathbb{R}^{p\times p}$ and the following LMI in X

$$\boxed{\mathrm{Tr}(Q) < 1, \quad \begin{bmatrix} Q & S^{\mathsf{T}}(X) \\ S(X) & P(X) \end{bmatrix} > 0} \tag{11.14}$$

11.2.4 Lyapunov inequality

The **Lyapunov inequality**

$$\boxed{XA + A^{\top}X < 0} \tag{11.15}$$

where $A \in \mathbb{R}^{n\times n}$ is a stable matrix, is equivalent to the following LMI

$$\boxed{\begin{bmatrix} -XA - A^{\top}X & 0_{n\times n} \\ 0_{n\times n} & I_{n\times n} \end{bmatrix} > 0} \tag{11.16}$$

11.2.5 Algebraic Riccati–Lurie's matrix inequality

The **algebraic Riccati–Lurie's matrix inequality**

$$\boxed{XA + A^{\top}X + XBR^{-1}B^{\top}X + Q < 0} \tag{11.17}$$

where $A, B, Q = Q^{\top}, R = R^{\top} > 0$ are given matrices of appropriate sizes and $X = X^{\top}$ is variable, is a quadratic matrix inequality in X. It may be represented as the following LMI:

$$\boxed{\begin{bmatrix} -XA - A^{\top}X - Q & XB \\ B^{\top}X & R \end{bmatrix} > 0} \tag{11.18}$$

11.2.6 Quadratic inequalities and S-procedure

Let us consider the quadratic functions given by

$$F_i(\zeta) = \zeta^{\top}A_i\zeta + 2b_i^{\top}\zeta + c_i \quad (i = 0, 1, \ldots, L)$$

with $A_i = A_i^{\top}$ $(i = 0, 1, \ldots, L)$. Consider also the following conditions on $F_0(\zeta)$, $F_1(\zeta), \ldots, F_L(\zeta)$:

$$\boxed{F_0(\zeta) \geq 0 \quad \text{for all} \quad \zeta \quad \text{such that} \quad F_i(\zeta) \geq 0 \ (i = 1, \ldots, L)} \tag{11.19}$$

Obviously if there exist such numbers $\tau_i \geq 0$ $(i = 1, \ldots, L)$ such that for all ζ

$$\boxed{F_0(\zeta) - \sum_{i=1}^{L} \tau_i F_i(\zeta) \geq 0} \tag{11.20}$$

then (11.19) holds.

Remark 11.2. *It is a nontrivial fact that for $L = 1$ the converse also holds, provided that there is some ζ^0 such that $F_0(\zeta^0) > 0$. The* ***Farkaš lemma*** *states the fact that in the general case, when $L \geq 1$ and when all functions $F_i(\zeta)$ are* ***affine****, i.e. $A_i = 0$ $(i = 0, 1, \ldots, L)$, (11.19) and (11.20) are equivalent.*

The last inequality (11.20) can be represented as

$$\boxed{\begin{array}{l} e^{\mathsf{T}} S e \geq 0, \quad e := \zeta / \|\zeta\|, \quad \zeta \neq 0 \\ S := \begin{bmatrix} A_0 & b_0 \\ b_0^{\mathsf{T}} & c_0 \end{bmatrix} - \sum_{i=1}^{L} \tau_i \begin{bmatrix} A_i & b_i \\ b_i^{\mathsf{T}} & c_i \end{bmatrix} \end{array}} \tag{11.21}$$

which is equivalent to the LMI inequality $S \geq 0$.

11.3 Some characteristics of linear stationary systems (LSS)

11.3.1 LSS and their transfer function

Let us consider a *linear stationary system* given by the following equations

$$\boxed{\left.\begin{array}{ll} \dot{x}(t) & = Ax(t) + B_w w(t) \\ x(0) & = x_0 \quad \text{is fixed} \\ z(t) & = C_{zx} x(t) + D_{zw} w(t) \end{array}\right\}} \tag{11.22}$$

where $x_t \in \mathbb{R}^n$ is the state of the system at time t, $z_t \in \mathbb{R}^m$ is its output and $w_t \in \mathbb{R}^k$ is an external input (or noise). The matrix A is assumed to be stable and the pair (A, B_w) is controllable, or, equivalently, the *controllability grammian* W_c defined as

$$\boxed{W_c := \int_{t=0}^{\infty} e^{At} B_w B_w^{\mathsf{T}} e^{A^{\mathsf{T}} t} \, dt} \tag{11.23}$$

is strictly positive definite, i.e. $W_c > 0$. Applying the Laplace transformation to (11.22) we found that the *transfer function* of this LSS is equal to the following matrix

$$\boxed{H(s) = C_{zx} (s I_{n \times n} - A)^{-1} B_w + D_{zw}} \tag{11.24}$$

where $s \in \mathbb{C}$.

11.3.2 H_2 norm

The H_2 *norm* of the LSS (11.22) is defined as

$$\boxed{\|H(s)\|_2 := \sqrt{\frac{1}{2\pi} \mathrm{Tr} \left(\int_{\omega=0}^{\infty} H(j\omega) H^*(j\omega) \, d\omega \right)}} \tag{11.25}$$

It is finite if and only if $D_{zw} = 0$. In this case it can be calculated as follows

$$\boxed{\|H(s)\|_2^2 = \mathrm{Tr}\left(C_{zx} W_c C_{zx}^{\mathsf{T}}\right)} \tag{11.26}$$

If C_{zx} is an affine function of some matrix K, i.e. $C_{zx} = C_{zx}(K)$, then the problem of finding some K fulfilling the inequality

$$\boxed{\mathrm{Tr}\left(C_{zx} W_c C_{zx}^{\mathsf{T}}\right) \le \gamma^2} \tag{11.27}$$

(here $\gamma > 0$ is a tolerance level of this LSS) is really LMI since by (11.14) the inequality (11.27) can be rewritten as

$$\boxed{\mathrm{Tr}(Q) \le 1, \quad \begin{bmatrix} \gamma^{-1} Q & C_{zx}(K) \\ C_{zx}^{\mathsf{T}}(K) & W_c^{-1} \end{bmatrix} > 0} \tag{11.28}$$

with a slack matrix variable Q.

11.3.3 Passivity and the positive-real lemma

The linear stationary system (11.22) with w_t and z_t of the same size is said to be *passive* if

$$\boxed{\int_{t=0}^{T} w_t^{\mathsf{T}} z_t \, dt \ge 0} \tag{11.29}$$

for all solutions of (11.22) (corresponding to all admissible $w_{(\cdot)}$) with $x_0 = 0$ and all $T \ge 0$. Passivity can be equivalently expressed in terms of the transfer function (11.24), namely, (11.22) is passive if and only if

$$\boxed{H(s) + H^*(s) = 2 \operatorname{Re} H(s) \ge 0 \quad \text{for all} \quad \operatorname{Re} s > 0} \tag{11.30}$$

that's why the passivity property is sometimes called *real-positiveness*. It is said that the system (11.22) has *dissipation* $\eta \ge 0$ if

$$\boxed{\int_{t=0}^{T} w_t^{\mathsf{T}} z_t \, dt \ge \eta \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt} \tag{11.31}$$

for all trajectories with $x_0 = 0$ and all $T \ge 0$.

Remark 11.3. *Evidently, if (11.22) has* ***dissipation*** $\eta = 0$, *then it is* ***passive*** *(but not inverse).*

Suppose that there exists a quadratic function $V(x) := x^{\mathsf{T}} P x$, $P > 0$, such that for all x_t and w_t, satisfying (11.22), the following inequality holds

$$\frac{d}{dt} V(x_t) - 2w_t^{\mathsf{T}} z_t + 2\eta w_t^{\mathsf{T}} w_t \leq 0 \tag{11.32}$$

Then, integrating this inequality within $[0, T]$-interval with $x_0 = 0$ yields

$$V(x_t) - \int_{t=0}^{T} w_t^{\mathsf{T}} z_t \, dt + \eta \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt \leq 0$$

and, since,

$$0 \leq V(x_T) \leq \int_{t=0}^{T} w_t^{\mathsf{T}} z_t \, dt - \eta \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt$$

we obtain (11.31). So, if (11.32) holds, then one may guarantee the η-dissipation for (11.22). Simple substitution

$$\begin{aligned} \frac{d}{dt} V(x_t) &= 2x_t^{\mathsf{T}} P \dot{x}_t = 2x_t^{\mathsf{T}} P [Ax_t + B_w w_t] \\ &= x_t^{\mathsf{T}} [PA + A^{\mathsf{T}} P] x_t + x_t^{\mathsf{T}} [PB_w] w_t + w_t^{\mathsf{T}} [B_w^{\mathsf{T}} P] x_t \end{aligned}$$

and

$$z_t = C_{zx} x_t + D_{zw} w_t$$

into (11.32) implies

$$\begin{pmatrix} x_t \\ w_t \end{pmatrix}^{\mathsf{T}} \begin{bmatrix} PA + A^{\mathsf{T}} P & PB_w - C_{zx} \\ B_w^{\mathsf{T}} P - C_{zx}^{\mathsf{T}} & 2\eta I_{n\times n} - (D_{zw}^{\mathsf{T}} + D_{zw}) \end{bmatrix} \begin{pmatrix} x_t \\ w_t \end{pmatrix} \leq 0$$

or, equivalently, as the following LMI

$$\boxed{\begin{bmatrix} -PA - A^{\mathsf{T}} P & -PB_w + C_{zx} \\ -B_w^{\mathsf{T}} P + C_{zx}^{\mathsf{T}} & -2\eta I_{n\times n} + (D_{zw}^{\mathsf{T}} + D_{zw}) \end{bmatrix} \geq 0} \tag{11.33}$$

So, if there exists a matrix $P = P^{\mathsf{T}} > 0$ satisfying (11.33) then the linear system (11.22) is η-dissipative.

Lemma 11.2. (The positive-real lemma) *Under the technical condition*

$$D_{zw}^{\mathsf{T}} + D_{zw} > 2\eta I_{n\times n} \tag{11.34}$$

the ***sufficient condition of*** *η-**dissipativity** (11.33) is equivalent to the existence of the positive definite solution P to the following Riccati inequality*

$$PA + A^{\mathsf{T}}P + [PB_w - C_{zx}]\left[\left(D_{zw}^{\mathsf{T}} + D_{zw}\right) - 2\eta I_{n\times n}\right]^{-1}\left[B_w^{\mathsf{T}}P - C_{zx}^{\mathsf{T}}\right] \leq 0 \tag{11.35}$$

Remark 11.4. *It is possible to show that LMI (11.35) is feasible if and only if (11.22) is passive.*

11.3.4 Nonexpansivity and the bounded-real lemma

The linear stationary system (11.22) is said to be *nonexpansive* if

$$\boxed{\int_{t=0}^{T} z_t^{\mathsf{T}} z_t \, dt \leq \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt} \tag{11.36}$$

for all solutions of (11.22) (corresponding to all admissible $w_{(\cdot)}$) with $x_0 = 0$ and all $T \geq 0$. *Nonexpansivity* can be equivalently expressed in terms of the transfer function (11.24), namely, (11.22) is nonexpansive if and only if the following *bounded-real condition* holds

$$\boxed{H^*(s)\,H(s) \leq I \quad \text{for all} \quad \operatorname{Re} s > 0} \tag{11.37}$$

that is why this condition is sometimes called nonexpansivity. This is sometimes expressed as

$$\|H\|_\infty \leq 1 \tag{11.38}$$

where

$$\begin{aligned}\|H\|_\infty &:= \sup\{\lambda_{\max}(H^*(s)\,H(s)) \mid \operatorname{Re} s > 0\} \\ &= \sup\{\lambda_{\max}(H^*(i\omega)\,H(i\omega)) \mid \omega \in (-\infty, \infty)\}\end{aligned} \tag{11.39}$$

Suppose that there exists a quadratic function $V(x) := x^{\mathsf{T}}Px$, $P > 0$, such that for all x_t and w_t, satisfying (11.22), the following inequality holds

$$\frac{d}{dt}V(x_t) - 2w_t^{\mathsf{T}}w_t + 2z_t^{\mathsf{T}}z_t \leq 0 \tag{11.40}$$

Then, integrating this inequality within the $[0, T]$-interval with $x_0 = 0$ yields

$$V(x_t) - \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt + \int_{t=0}^{T} z_t^{\mathsf{T}} z_t \, dt \leq 0$$

and, since,

$$0 \leq V(x_T) \leq \int_{t=0}^{T} w_t^{\mathsf{T}} w_t \, dt - \int_{t=0}^{T} z_t^{\mathsf{T}} z_t \, dt$$

we obtain (11.36). So, if (11.40) holds, then one may guarantee the nonexpansivity for (11.22). Simple substitution

$$\begin{aligned} \frac{d}{dt} V(x_t) &= 2x_t^{\mathsf{T}} P \dot{x}_t = 2x_t^{\mathsf{T}} P \left[A x_t + B_w w_t \right] \\ &= x_t^{\mathsf{T}} \left[PA + A^{\mathsf{T}} P \right] x_t + x_t^{\mathsf{T}} \left[PB_w \right] w_t + w_t^{\mathsf{T}} \left[B_w^{\mathsf{T}} P \right] x_t \end{aligned}$$

and

$$z_t = C_{zx} x_t + D_{zw} w_t$$

into (11.40) implies

$$\begin{pmatrix} x_t \\ w_t \end{pmatrix}^{\mathsf{T}} \begin{bmatrix} PA + A^{\mathsf{T}} P + C_{zx}^{\mathsf{T}} C_{zx} & PB_w + C_{zx}^{\mathsf{T}} D_{zw} \\ B_w^{\mathsf{T}} P + D_{zw}^{\mathsf{T}} C_{zx} & D_{zw}^{\mathsf{T}} D_{zw} - I \end{bmatrix} \begin{pmatrix} x_t \\ w_t \end{pmatrix} \leq 0$$

or, equivalently, as the LMI

$$\boxed{\begin{bmatrix} -PA - A^{\mathsf{T}} P - C_{zx}^{\mathsf{T}} C_{zx} & -PB_w - C_{zx}^{\mathsf{T}} D_{zw} \\ -B_w^{\mathsf{T}} P - D_{zw}^{\mathsf{T}} C_{zx} & I - D_{zw}^{\mathsf{T}} D_{zw} \end{bmatrix} \geq 0} \tag{11.41}$$

So, if there exists a matrix $P = P^{\mathsf{T}} > 0$ satisfying (11.41) then the linear system (11.22) is nonexpansive.

Lemma 11.3. (The bounded-real lemma) *Under the technical condition*

$$D_{zw}^{\mathsf{T}} D_{zw} \neq I \tag{11.42}$$

*the **sufficient condition of nonexpansivity** (11.36) is equivalent to the existence of the positive definite solution to the following Riccati inequality*

$$\begin{aligned} &PA + A^{\mathsf{T}} P + C_{zx}^{\mathsf{T}} C_{zx} \\ &\left[PB_w - C_{zx}^{\mathsf{T}} D_{zw} \right] \left[I - D_{zw}^{\mathsf{T}} D_{zw} \right]^{-1} \left[B_w^{\mathsf{T}} P - D_{zw}^{\mathsf{T}} C_{zx} \right] \leq 0 \end{aligned} \tag{11.43}$$

Remark 11.5. *It is possible to show that LMI (11.43) is feasible if and only if (11.22) is nonexpansive.*

11.3.5 H_∞ norm

The condition

$$\boxed{\|H\|_\infty \leq \gamma, \quad 0 < \gamma} \tag{11.44}$$

can be represented as

$$\left\|\tilde{H}\right\|_\infty \leq 1$$

with the transfer function $\tilde{H}(s)$ given by

$$\begin{aligned} &\tilde{H}(s) = \tilde{C}_{zx}\left(sI_{n\times n} - A\right)^{-1} B_w + \tilde{D}_{zw} \\ &\tilde{C}_{zx} := \gamma^{-1} C_{zx}, \ \tilde{D}_{zx} := \gamma^{-1} D_{zx} \end{aligned} \tag{11.45}$$

Therefore, based on the bounded-real lemma (see (11.41)), the constraint (11.44) would be valued if

$$\begin{bmatrix} PA + A^\intercal P + \tilde{C}_{zx}^\intercal \tilde{C}_{zx} & PB_w + \tilde{C}_{zx}^\intercal \tilde{D}_{zw} \\ B_w^\intercal P + \tilde{D}_{zw}^\intercal \tilde{C}_{zx} & \tilde{D}_{zw}^\intercal \tilde{D}_{zw} - I \end{bmatrix}$$

$$= \begin{bmatrix} PA + A^\intercal P + \gamma^{-2} C_{zx}^\intercal C_{zx} & PB_w + \gamma^{-2} C_{zx}^\intercal D_{zw} \\ B_w^\intercal P + \gamma^{-2} D_{zw}^\intercal C_{zx} & \gamma^{-2} D_{zw}^\intercal D_{zw} - I \end{bmatrix} \leq 0$$

which is equivalent to the feasibility of the following LMI

$$\boxed{\begin{bmatrix} \tilde{P}A + A^\intercal \tilde{P} + C_{zx}^\intercal C_{zx} & \tilde{P}B_w + C_{zx}^\intercal D_{zw} \\ B_w^\intercal \tilde{P} + D_{zw}^\intercal C_{zx} & D_{zw}^\intercal D_{zw} - \gamma^2 I \end{bmatrix} \leq 0 \\ 0 < \tilde{P} = \gamma^2 P} \tag{11.46}$$

11.3.6 γ-Entropy

The *γ-entropy* for the system (11.22) with the transfer function H (11.24) is defined in the following way:

$$\boxed{I_\gamma(H) := \\ \begin{cases} \dfrac{-\gamma^2}{2\pi} \displaystyle\int_{\omega=-\infty}^{\infty} \log\det\left(I - \gamma^2 H(j\omega) H^*(j\omega)\right) d\omega & \text{if } \|H\|_\infty < \gamma \\ \infty \text{ otherwise} \end{cases}} \tag{11.47}$$

When $\|H\|_\infty < \gamma$, $I_\gamma(H)$ can be calculated as

$$\boxed{I_\gamma(H) = \mathrm{Tr}\left(B_w^\intercal P B_w\right)} \tag{11.48}$$

where P is a symmetric matrix with smallest possible maximum singular value among all solutions of the following algebraic Riccati equation

$$\boxed{PA + A^\intercal P + C_{zx}^\intercal C_{zx} + \gamma^{-2} P B_w B_w^\intercal P = 0}$$

Therefore the γ-entropy constraint $I_\gamma(H) < \lambda$ is equivalent to LMI in P, γ and λ, namely,

$$\boxed{\begin{bmatrix} PA + A^\intercal P & PB_w & C_{zx}^\intercal \\ B_w^\intercal P & -\gamma^2 I & 0 \\ C_{zx} & 0 & -I \end{bmatrix} \le 0 \qquad \tilde{D}_{zw} = 0, \quad \mathrm{Tr}\left(B_w^\intercal P B_w\right) \le \lambda} \tag{11.49}$$

11.3.7 Stability of stationary time-delay systems

Consider a *stationary time-delay system* given by

$$\boxed{\dot{x}_t = Ax_t + \sum_{i=1}^{L} A_i x_{t-\tau_i}} \tag{11.50}$$

where $x_t \in \mathbb{R}^n$ and $\tau_i > 0$. If the Lyapunov–Krasovskii functional

$$\boxed{V(x,t) := x_t^\intercal P x_t + \sum_{i=1}^{L} \int_{s=t-\tau_i}^{t} x_s^\intercal P_i x_s \, ds \qquad P > 0, \quad P_i > 0 \ (i = 1, \ldots, L)} \tag{11.51}$$

satisfies

$$\frac{d}{dt} V(x,t) < 0$$

for every x_t satisfying (11.50), then this system is *asymptotically stable*, namely,

$$x_t \to 0 \text{ as } t \to \infty$$

This can be verified by the simple calculation

$$\boxed{\begin{gathered}\frac{d}{dt}V(x,t) = y_t^{\mathsf{T}} W y_t \\ W := \begin{bmatrix} \left[PA + A^{\mathsf{T}}P + \sum_{i=1}^{L} P_i\right] & PA_1 & \cdots & PA_L \\ A_1^{\mathsf{T}}P & -P_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ A_L^{\mathsf{T}}P & 0 & \cdots & -P_L \end{bmatrix} \\ y_t^{\mathsf{T}} = \left(x_t, x_{t-\tau_1}, \cdots, x_{t-\tau_L}\right)\end{gathered}} \tag{11.52}$$

providing that the matrices $P > 0$, $P_i > 0$ $(i = 1, \ldots, L)$ are satisfying LMI $W < 0$.

11.3.8 Hybrid time-delay linear stability

Let us consider the following *hybrid time-delay* linear system given by

$$\boxed{\begin{aligned}\dot{x}_1(t) &= A_0 x_1(t) + A_1 x_2(t-\tau) \\ x_2(t) &= A_2 x_1(t) + A_3 x_2(t-\tau) \\ x_1(0) = x_{10},\ & x_2(\theta) = \psi(\theta)\ \theta \in [-\tau, 0]\end{aligned}} \tag{11.53}$$

where $A_0 \in \mathbb{R}^{n\times n}$, $A_1 \in \mathbb{R}^{n\times m}$, $A_2 \in \mathbb{R}^{m\times n}$, $A_3 \in \mathbb{R}^{m\times m}$ are the given matrices of the corresponding dimensions and $\psi : \mathbb{R}^1 \to \mathbb{R}^m$ is a function from $C[-\tau, 0]$. Notice that the first equation in (11.53) is an ordinary differential equation and the second one is a difference equation in continuous time that justifies the name "hybrid time-delay system".

We are interested in finding the conditions of asymptotic stability for this system. Following Rasvan (1975), let us introduce the energetic (Lyapunov–Krasovskii-type) functional

$$\boxed{\begin{gathered}V(x_1(t), x_2) := x_1^{\mathsf{T}}(t) P x_1(t) + \int_{\theta=t-\tau}^{t} x_2^{\mathsf{T}}(\theta) S x_2(\theta)\, d\theta \\ 0 \le P = P^{\mathsf{T}} \in \mathbb{R}^{n\times n}, \quad 0 \le S = S^{\mathsf{T}} \in \mathbb{R}^{m\times m}\end{gathered}} \tag{11.54}$$

Its derivative on the trajectories of (11.53) is as follows:

$$\frac{d}{dt}V(x_1(t), x_2) = z^{\mathsf{T}}(t) W z(t) \tag{11.55}$$

where

$$z^{\mathsf{T}}(t) := (x_1(t), x_2(t-\tau)) \tag{11.56}$$

and

$$\boxed{W = \begin{bmatrix} A_0^{\mathsf{T}}P + PA_0 + A_2^{\mathsf{T}}SA_2 & PA_1 + A_2^{\mathsf{T}}SA_3 \\ A_1^{\mathsf{T}}P + A_3^{\mathsf{T}}SA_2 & A_3^{\mathsf{T}}SA_3 - S \end{bmatrix}} \tag{11.57}$$

As it is shown in Rasvan (1975), the existence of the matrices P and S such that the following LMI holds

$$W < 0$$

implies the asymptotic stability of (11.53).

11.4 Optimization problems with LMI constraints

11.4.1 Eigenvalue problem (EVP)

The *eigenvalue problem* (EVP) consists of the minimization of the maximum eigenvalue of an $n \times n$ matrix $A(P)$ that depends affinely on a variable, subject to LMI (symmetric) constraint $B(P) > 0$, i.e.,

$$\boxed{\begin{array}{c} \lambda_{\max}(A(P)) \to \min\limits_{P=P^{\mathsf{T}}} \\ B(P) > 0 \end{array}} \tag{11.58}$$

This problem can be equivalently represented as follows:

$$\boxed{\begin{array}{c} \lambda \to \min\limits_{\lambda, P=P^{\mathsf{T}}} \\ \begin{bmatrix} \lambda I_{n\times n} - A(P) > 0 & 0 \\ 0 & B(P) \end{bmatrix} > 0 \end{array}} \tag{11.59}$$

11.4.2 Tolerance level optimization

The *tolerance level optimization problem* can be represented in the following manner:

$$\boxed{\begin{array}{c} \gamma \to \min\limits_{0<\gamma, P=P^{\mathsf{T}}} \\ PA + A^{\mathsf{T}}P + C^{\mathsf{T}}C + \gamma^{-1}PBB^{\mathsf{T}}P < 0 \end{array}} \tag{11.60}$$

Equivalently, it can be rewritten as an optimization problem with LMI constraints:

$$\boxed{\begin{gathered}\gamma \to \min_{0<\gamma, 0<P=P^{\mathrm{T}}} \\ \begin{bmatrix} -PA - A^{\mathrm{T}}P - C^{\mathrm{T}}C & PB & 0 \\ B^{\mathrm{T}}P & \gamma I & 0 \\ 0 & 0 & P \end{bmatrix} > 0\end{gathered}} \tag{11.61}$$

11.4.3 Maximization of the quadratic stability degree

The *quadratic stability degree* of a stable $n \times n$ matrix A is defined as a positive value α satisfying the matrix inequality

$$A^{\mathrm{T}}P + PA < -\alpha P$$

for some positive definite matrix P. The problem of the maximization of the quadratic stability degree consists of the following optimization problem

$$\boxed{\begin{gathered}\alpha \to \max_{0<\alpha, 0<P=P^{\mathrm{T}}} \\ A^{\mathrm{T}}P + PA + \alpha P < 0\end{gathered}} \tag{11.62}$$

which can be expressed as an optimization with LMI constraint, namely,

$$\boxed{\begin{gathered}\alpha \to \max_{0<\alpha, P=P^{\mathrm{T}}} \\ \begin{bmatrix} -A^{\mathrm{T}}P - PA - \alpha P & 0 \\ 0 & P \end{bmatrix} > 0\end{gathered}} \tag{11.63}$$

11.4.4 Minimization of linear function $\mathrm{Tr}\,(CPC^{\mathrm{T}})$ *under the Lyapunov-type constraint*

Lemma 11.4. (Polyak & Sherbakov (2002)) *Let*

1. *the matrix* $A \in \mathbb{R}^{n\times n}$ *be Hurwitz;*
2. *the pair* (A, B) *be controllable, i.e. there exists a matrix* K *such that* $(A + KB)$ *is Hurwitz.*

Then for any matrix $C \in \mathbb{R}^{k\times n}$ *the solution of the problem*

$$\boxed{\mathrm{Tr}\,(CPC^{\mathrm{T}}) \to \min_{P\geq 0}} \tag{11.64}$$

under the constraint

$$\boxed{AP + PA^{\mathrm{T}} + BB^{\mathrm{T}} \leq 0} \tag{11.65}$$

is attained on the Lyapunov matrix equation

$$\boxed{AP^* + P^*A^\top + BB^\top = 0} \tag{11.66}$$

Proof. Suppose that the minimizing solution satisfies the equation

$$AP + PA^\top + BB^\top = -Q < 0$$

Then, by Lemma 9.1,

$$P = \int_{t=0}^{\infty} e^{At}\,(Q + BB^\top)\,e^{A^\top t}\,dt \geq \int_{t=0}^{\infty} e^{At}BB^\top e^{A^\top t}\,dt = P^*$$

and, hence,

$$\mathrm{Tr}\,(CPC^\top) = \mathrm{Tr}\,(CP^*C^\top) + \mathrm{Tr}\left(C\int_{t=0}^{\infty} e^{At}Qe^{A^\top t}dtC^\top\right) \geq \mathrm{Tr}\,(CP^*C^\top)$$

This means that P^* is a minimizer. Lemma is proven. □

11.4.5 The convex function log det $A^{-1}(X)$ *minimization*

First notice that log det $A^{-1}(X)$ is a convex function of A. We will encounter the following:

$$\boxed{\log\det A^{-1}(X) \to \min_{X = X^\top \in \mathbb{R}^{n\times n}}} \tag{11.67}$$

subjected to the following constraints

$$\boxed{A(X) > 0, \quad B(X) > 0} \tag{11.68}$$

where $A(X)$, $B(X)$ are symmetric matrices that depend affinely on X.

Example 11.1. *As an example of the problem (11.67)–(11.68) consider the following:* ***find a minimal ellipsoid***

$$\boxed{\varepsilon := \{z \mid z^\top Pz \leq 1\}, \quad P > 0} \tag{11.69}$$

containing the set of given points v_i $(i = 1, \ldots, L)$, *i.e.* $v_i \in \varepsilon$. *Since the volume of* ε *is proportional to* $(\det P)^{-1/2}$, *minimizing* $\log\det P^{-1}$ *is the same as minimizing the volume of* ε, *this problem is converted into the following:*

$$\boxed{\begin{array}{l}\log\det P^{-1} \to \min\limits_{P\in\mathbb{R}^{n\times n}} \\ P > 0, \quad v_i^\top P v_i \leq 1\end{array}} \tag{11.70}$$

11.5 Numerical methods for LMI resolution

11.5.1 What does it mean "to solve LMI"?

There exist several efficient methods for LMI resolution. By "solve an LMI" we mean here:

- determine whether or not the LMI (or the corresponding problem) is feasible;
- if it is, compute a feasible point with "an objective value" that exceeds the global minimum by less than some prespecified accuracy.

What does "an objective value" mean? It depends on each concrete problem to be solved. Here we will assume that the problem we are solving has at least one "optimal point", i.e., the constraints are feasible.

To realize the numerical methods described below, first, let us represent the matrix $X \in \mathbb{R}^{n\times n}$ as the corresponding extended vector $x \in \mathbb{R}^{n^2}$ obtained by the simple implementation of the operator col, that is,

$$\boxed{x := \operatorname{col} X} \tag{11.71}$$

11.5.2 Ellipsoid algorithm

In a feasible problem, we may consider any feasible point as being optimal. The **basic idea** of the *ellipsoid algorithm* is as follows:

1. One may start with an ellipsoid that is guaranteed to contain an optimal point.
2. Then the *cutting plane* for our problem is computed which passes through the center point $x^{(0)}$ of the initial ellipsoid $\varepsilon^{(0)}$. This means that we need to find a nonzero vector $g^{(0)}$ (namely, a vector orthogonal to the plane to be computed) such that an optimal point lies in the half-space

$$\left\{z \in \mathbb{R}^{n^2} \mid g^{(0)\mathsf{T}}\left(z - x^{(0)}\right) < 0\right\} \tag{11.72}$$

   (Below, we shall present some examples of how to calculate $g^{(0)}$ in some concrete problems.)
3. After this we may conclude that the sliced half-ellipsoid

$$\varepsilon^{(0)} \cap \left\{z \in \mathbb{R}^{n^2} \mid g^{(0)\mathsf{T}}\left(z - x^{(0)}\right) < 0\right\}$$

   contains an optimal point.
4. Then we compute the ellipsoid $\varepsilon^{(1)}$ of a minimum volume that contains this sliced half-ellipsoid. This ellipsoid $\varepsilon^{(1)}$ is guaranteed to contain an optimal point, but its volume is expected to be less than the volume of the initial ellipsoid $\varepsilon^{(0)}$.
5. The process is then iterated.

More explicitly, this algorithm may be described as follows. Any *ellipsoid* ε may be associated with some positive definite matrix, that is,

$$\boxed{\varepsilon := \left\{z \in \mathbb{R}^{n^2} \mid (z-a)^{\mathsf{T}} A^{-1} (z-a) \le 1\right\}} \tag{11.73}$$

where $A = A^{\mathsf{T}} > 0$. The *minimum volume ellipsoid* $\tilde{\varepsilon}$ containing the sliced half-ellipsoid

$$\left\{ z \in \mathbb{R}^{n^2} \mid (z-a)^{\mathsf{T}} A^{-1} (z-a) \leq 1,\ g^{\mathsf{T}} (z-a) < 0 \right\}$$

is given by the matrix $\tilde{A}$ and the vector $\tilde{a}$, namely,

$$\boxed{\begin{aligned} \tilde{\varepsilon} &:= \left\{ z \in \mathbb{R}^{n^2} \mid (z-\tilde{a})^{\mathsf{T}} \tilde{A}^{-1} (z-\tilde{a}) \leq 1 \right\} \\ \tilde{a} &= a - \frac{A\tilde{g}}{m+1}, \quad m := n^2 > 1 \\ \tilde{A} &= \frac{m^2}{m^2-1} \left(A - \frac{2}{m+1} A\tilde{g}\tilde{g}^{\mathsf{T}} A \right) \\ \tilde{g} &= \frac{g}{\sqrt{g^{\mathsf{T}} A g}} \end{aligned}} \tag{11.74}$$

(In the case of one variable ($m = 1$) the minimal length interval containing a half-interval is the half-interval itself.) So, the ellipsoid algorithm starts with the initial points $x^{(0)}$ and the initial matrix $A^{(0)}$. Then for each intermediate pair $x^{(k)}$ and $A^{(k)}$ ($k = 0, 1, 2, \ldots$) one may compute a vector $g^{(k)}$ and then calculate

$$x^{(k+1)} = x^{(k)} - \frac{A^{(k)}\tilde{g}}{m+1}, \quad m := n^2 > 1$$

$$A^{(k+1)} = \frac{m^2}{m^2-1} \left(A^{(k)} - \frac{2}{m+1} A^{(k)} \tilde{g}\tilde{g}^{\mathsf{T}} A^{(k)} \right)$$

$$\tilde{g} = \frac{g^{(k)}}{\sqrt{g^{(k)\mathsf{T}} A^{(k)} g^{(k)\mathsf{T}}}}$$

It turns out that the volume $\operatorname{vol} \varepsilon^{(k)} = \det A^{(k)}$ of these ellipsoids decreases geometrically, that is,

$$\operatorname{vol} \varepsilon^{(k+1)} \leq e^{-k/2m} \operatorname{vol} \varepsilon^{(k)}$$

This means that the recursion above generates a sequence of ellipsoids that is guaranteed to contain an optimal point and converges to it geometrically. It may be proven that this algorithm converges more quickly, namely, in "polynomial time" (see Nesterov & Nemirovsky (1994) and references within).

The next examples illustrate the rule of selection of the nonzero orthogonal vector g orthogonal to the cutting plane which is specified for each concrete problem.

Example 11.2. *If LMI is represented in the form*

$$\boxed{F(x) := F_0 + \sum_{i=1}^{m} x_i F_i > 0} \tag{11.75}$$

where F_i $(i = 0, 1, 2, \ldots, m)$ *are symmetric matrices. If* x *is infeasible, this means that there exists a nonzero vector* u *such that*

$$u^{\mathsf{T}} F(x)\, u \leq 0$$

Define $g = (g_1, \ldots, g_m)^{\mathsf{T}}$ *by*

$$\boxed{g_i = -u^{\mathsf{T}} F_i u} \tag{11.76}$$

Then for any z *satisfying* $g^{\mathsf{T}}(z - x) \geq 0$ *it follows*

$$u^{\mathsf{T}} F(z)\, u = u^{\mathsf{T}} \left[F_0 + \sum_{i=1}^{m} z_i F_i \right] u = u^{\mathsf{T}} F_0 u + \sum_{i=1}^{m} z_i u^{\mathsf{T}} F_i u$$

$$= u^{\mathsf{T}} F_0 u - \sum_{i=1}^{m} z_i g_i = u^{\mathsf{T}} F_0 u - g^{\mathsf{T}} z = u^{\mathsf{T}} F_0 u + g^{\mathsf{T}} x - g^{\mathsf{T}}(z - x)$$

$$= u^{\mathsf{T}} F(x)\, u - g^{\mathsf{T}}(z - x) \leq 0$$

So, any feasible point belongs to the half-plane

$$\{z \in \mathbb{R}^m \mid g^{\mathsf{T}}(z - x) < 0\}$$

or, in other words, this g*, given by (11.76), is a cutting plane for this LMI problem at the point* x*.*

Example 11.3. *If we deal with the minimization problem of linear function* $c^{\mathsf{T}} x$ *subjected LMI (11.75), that is,*

$$c^{\mathsf{T}} x \to \min_{x \in \mathbb{R}^{n^2}}$$

$$F(x) := F_0 + \sum_{i=1}^{m} x_i F_i > 0$$

we encounter two possible situations:

1. x *is infeasible, i.e.,* $F(x) \leq 0$*; in this case* g *can be taken as in the previous example (11.76) since we are discarding the half-plane*

$$\left\{ z \in \mathbb{R}^{n^2} \mid g^{\mathsf{T}}(z - x) > 0 \right\}$$

because all such points are infeasible;

2. *x is feasible, i.e., $F(x) > 0$; in this case g can be taken as*

$$g = c$$

since we are discarding the half-plane

$$\left\{ z \in \mathbb{R}^{n^2} \mid g^{\mathsf{T}} (z - x) > 0 \right\}$$

because all such points have an objective value larger than x and hence cannot be optimal.

11.5.3 Interior-point method

For the LMI problem

$$F(x) := F_0 + \sum_{i=1}^{m} x_i F_i > 0$$

let us define the, so-called, *barrier function* $\phi(x)$ for the feasible set:

$$\boxed{\phi(x) := \begin{cases} \log\det F^{-1}(x) & \text{if} \quad F(x) > 0 \\ \infty & \text{if} \quad F(x) \leq 0 \end{cases}} \tag{11.77}$$

Suppose then that the feasible set is nonempty and bounded. This implies that the matrices $F_1, \ldots, F_m$ are linearly independent (otherwise the feasible set will contain a line, i.e. be unbounded). It can be shown that $\phi(x)$ is strictly convex on the feasible set and, hence, it has a unique minimizer which we denote by x^*, that is,

$$x^* := \arg\min_x \phi(x)$$

This point is referred to as the *analytical center of the LMI* $F(x) > 0$. It is evident that

$$x^* := \arg\max_{F(x)>0} \det F(x)$$

Remark 11.6. *Two LMIs $F(x) > 0$ and $T^{\mathsf{T}} F(x)\, T > 0$ have the same analytical center provided T is nonsingular.*

Let us apply Newton's method for the search of the analytical center x^* of LMI, starting from a feasible initial point:

$$\boxed{x^{(k+1)} = x^{(k)} - \alpha^{(k)} H^{-1}\left(x^{(k)}\right) g\left(x^{(k)}\right)} \tag{11.78}$$

where $0 < \alpha^{(k)}$ is a damping factor at the kth iteration, $H\left(x^{(k)}\right)$ is the Hessian and $g\left(x^{(k)}\right)$ is the gradient, respectively, of $\phi(x)$ at the point $x^{(k)}$. In Nesterov & Nemirovsky (1994) it is shown that if the *damping factor* is

$$\boxed{\begin{aligned} \alpha^{(k)} &:= \begin{cases} 1 & \text{if } \delta\left(x^{(k)}\right) \leq 1/4 \\ 1/\left(1+\delta\left(x^{(k)}\right)\right) & \text{otherwise} \end{cases} \\ \delta\left(x^{(k)}\right) &:= \sqrt{g^{\mathrm{T}}\left(x^{(k)}\right) H^{-1}\left(x^{(k)}\right) g\left(x^{(k)}\right)} \end{aligned}} \tag{11.79}$$

then this step length always results in $x^{(k+1)}$, that is,

$$F\left(x^{(k+1)}\right) > 0$$

and convergence of $x^{(k)}$ to x^* when $k \to \infty$.

There exist other interior-point methods (for details, see Boyd *et al.* (1994)).

12 Miscellaneous

Contents

12.1 Λ-matrix inequalities

Lemma 12.1. *For any matrices* $X, Y \in \mathbb{R}^{n\times m}$ *and any symmetric positive definite matrix* $\Lambda \in \mathbb{R}^{n\times n}$ *the following inequalities hold*

$$\boxed{X^TY + Y^TX \le X^T\Lambda X + Y^T\Lambda^{-1}Y} \tag{12.1}$$

and

$$\boxed{(X+Y)^\intercal (X+Y) \le X^T (I_{n\times n} + \Lambda) X + Y^T \left(I_{n\times n} + \Lambda^{-1}\right) Y} \tag{12.2}$$

Proof. Define

$$H := X^T\Lambda X + Y^T\Lambda^{-1}Y - X^TY - Y^TX$$

Then for any vector v we may introduce the vectors

$$v_1 := \Lambda^{1/2}Xv \quad \text{and} \quad v_2 := \Lambda^{-1/2}Yv$$

which implies

$$v^THv = v_1^Tv_1 + v_2^Tv_2 - v_1^Tv_2 - v_2^Tv_1 = \|v_1 - v_2\|^2 \ge 0 \tag{12.3}$$

or, in matrix form:

$$H \ge 0$$

which is equivalent to (12.3). The inequality (12.2) is a direct consequence of (12.1). □

12.2 Matrix Abel identities

12.2.1 Matrix summation by parts

Lemma 12.2. (Matrix summation by parts) *For any matrices*

$$A_t \in \mathbb{R}^{m\times k}, \quad B_t \in \mathbb{R}^{k\times l}$$

and any integer numbers n_0 *and* $n \geq n_0$ *the following identity holds:*

$$\boxed{\sum_{t=n_0}^{n} A_t B_t = A_n \sum_{t=n_0}^{n} B_t - \sum_{t=n_0}^{n} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s} \tag{12.4}$$

(here $\sum_{s=n_0}^{t} B_s := 0$ *if* $t < n_0$*).*

Proof. Let us use induction. For $n = n_0$ the identity (12.4) is true since

$$\sum_{t=n_0}^{n_0} A_t B_t = A_{n_0} B_{n_0} = A_{n_0} \sum_{t=n_0}^{n_0} B_t - \sum_{t=n_0}^{n_0} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s$$

$$= A_{n_0} B_{n_0} - (A_{n_0} - A_{n_0-1}) \sum_{s=n_0}^{n_0-1} B_s = A_{n_0} B_{n_0}$$

Suppose now that it is valid for some $n > n_0$ and then prove that it is also true for $n+1$, we have

$$\sum_{t=n_0}^{n+1} A_t B_t = A_{n+1} B_{n+1} + \sum_{t=n_0}^{n} A_t B_t = A_{n+1} B_{n+1}$$

$$+ A_n \sum_{t=n_0}^{n} B_t - \sum_{t=n_0}^{n} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s = A_{n+1} B_{n+1} + A_{n+1} \sum_{t=n_0}^{n+1} B_t$$

$$- A_{n+1} \left(B_{n+1} + \sum_{t=n_0}^{n} B_t \right) + A_n \sum_{t=n_0}^{n} B_t - \sum_{t=n_0}^{n} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s$$

$$= A_{n+1} \sum_{t=n_0}^{n+1} B_t - (A_{n+1} - A_n) \sum_{t=n_0}^{n} B_t - \sum_{t=n_0}^{n} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s$$

$$= A_{n+1} \sum_{t=n_0}^{n+1} B_t - \sum_{t=n_0}^{n+1} (A_t - A_{t-1}) \sum_{s=n_0}^{t-1} B_s$$

Lemma is proven. □

12.2.2 Matrix product identity

Lemma 12.3. (Matrix product identity) *For any $n \times n$ matrices A_t $(t = t_0, \ldots, t_f)$ the following identity holds*

$$\boxed{\prod_{t=t_0}^{t_f} A_t + \sum_{t=t_0}^{t_f} \left[\left(\prod_{s=t+1}^{t_f} A_s \right) (I_{n\times n} - A_t) \right] = I_{n\times n}} \tag{12.5}$$

(here $\prod_s^t A_s := I_{n\times n}$ if $t < s$ and $\prod_{t=t_0}^{t_f} A_t := A_{t_f} \cdots A_{t_0}$).

Proof. Again let us use the induction method. For $t_f = t_0$ the identity (12.5) is valid since

$$\prod_{t=t_0}^{t_0} A_t + \sum_{t=t_0}^{t_0} \left[\left(\prod_{s=t+1}^{t_0} A_s \right) (I_{n\times n} - A_t) \right] = A_{t_0}$$

$$+ \left(\prod_{s=t_0+1}^{t_0} A_s \right) (I_{n\times n} - A_{t_0}) = A_{t_0} + I_{n\times n} (I_{n\times n} - A_{t_0}) = I_{n\times n}$$

Assuming that (12.5) is valid for some $t_f > t_0$ one can demonstrate that it is valid for $t_f + 1$. Indeed,

$$\prod_{t=t_0}^{t_f+1} A_t + \sum_{t=t_0}^{t_f+1} \left[\left(\prod_{s=t+1}^{t_f+1} A_s \right) (I_{n\times n} - A_t) \right] = A_{t_f+1} \prod_{t=t_0}^{t_f} A_t$$

$$+ \left(\prod_{s=t_f+2}^{t_f+1} A_s \right) (I_{n\times n} - A_{t_f+1}) + \sum_{t=t_0}^{t_f} \left[\left(\prod_{s=t+1}^{t_f+1} A_s \right) (I_{n\times n} - A_t) \right]$$

$$= A_{t_f+1} \prod_{t=t_0}^{t_f} A_t + (I_{n\times n} - A_{t_f+1})$$

$$+ A_{t_f+1} \sum_{t=t_0}^{t_f} \left[\left(\prod_{s=t+1}^{t_f} A_s \right) (I_{n\times n} - A_t) \right] = (I_{n\times n} - A_{t_f+1})$$

$$+ A_{t_f+1} \left(\prod_{t=t_0}^{t_f} A_t + \sum_{t=t_0}^{t_f} \left[\left(\prod_{s=t+1}^{t_f} A_s \right) (I_{n\times n} - A_t) \right] \right)$$

$$= (I_{n\times n} - A_{t_f+1}) + A_{t_f+1} I_{n\times n} = I_{n\times n}$$

Lemma is proven. □

12.3 S-procedure and Finsler lemma

12.3.1 Daneš' theorem

Let $\mathcal{F}(x) = x^{\mathsf{T}} F x$ and $\mathcal{G}(x) = x^{\mathsf{T}} G x$ be real quadratic forms with $F, G \in \mathbb{R}^{n\times n}$. Consider the mapping $\mathbb{R}^n \to \mathbb{R}^2$ defined by

$$\eta_1 = \mathcal{F}(x), \quad \eta_2 = \mathcal{G}(x) \tag{12.6}$$

which transforms any point from $\mathbb{R}^n$ into the real plane $\mathbb{R}^2$. The following theorem represents the important geometric result (Daneš 1972) arising in the theory of quadratic forms.

Theorem 12.1. (Daneš' theorem) *The range*

$$\mathcal{P} := \left\{(\eta_1, \eta_2) \in \mathbb{R}^2 \mid \eta_1 = \mathcal{F}(x),\ \eta_2 = \mathcal{G}(x),\ x \in \mathbb{R}^n\right\} \tag{12.7}$$

of the transformation (12.6) is a ***convex cone****, i.e., the set $\mathcal{P}$ together with $y \in \mathbb{R}^2$ contains also λy for any $\lambda \geq 0$ ($\mathcal{P}$ is a cone) and together with vectors $y^{(1)}, y^{(2)} \in \mathbb{R}^2$ contains also $y = \alpha y^{(1)} + (1-\alpha)\, y^{(2)}$ for any $\alpha \in [0, 1]$ ($\mathcal{P}$ is a convex set).*

Proof. Let $y = (\eta_1, \eta_2)^{\mathsf{T}}$ be a point in $\mathbb{R}^2$.

(a) Obviously, $\mathcal{P}$ is a cone since there exists a point $x \in \mathbb{R}^n$ such that

$$y = (\mathcal{F}(x),\ \mathcal{G}(x))$$

Then

$$\lambda y = \left(\mathcal{F}(\sqrt{\lambda}x),\ \mathcal{G}(\sqrt{\lambda}x)\right)$$

This means that $\lambda y \in \mathcal{P}$.

(b) Show now that $\mathcal{P}$ is a convex set. Let

$$y^{(1)} = \left(\mathcal{F}(x^{(1)}),\ \mathcal{G}(x^{(1)})\right), \quad y^{(2)} = \left(\mathcal{F}(x^{(2)}),\ \mathcal{G}(x^{(2)})\right)$$

In other words, we need to show that for any point $y^{(0)} = \left(\eta_1^{(0)}, \eta_2^{(0)}\right)^{\mathsf{T}} \in \left[y^{(1)}, y^{(2)}\right]$ there exists a vector $x^{(0)} \in \mathbb{R}^n$ such that

$$y^{(0)} = \left(\mathcal{F}(x^{(0)}),\ \mathcal{G}(x^{(0)})\right)$$

Let $A\eta_1 + B\eta_2 = C$ be the line $\mathcal{L}$ crossing the points $y^{(1)}$ and $y^{(2)}$. Consider the function

$$\varphi(\xi_1, \xi_2) := A\mathcal{F}(\xi_1 x^{(1)} + \xi_2 x^{(2)}) + B\mathcal{G}(\xi_1 x^{(1)} + \xi_2 x^{(2)})$$

and the mapping $\mathbb{R}^2 \to \mathbb{R}^2$:

$$y = \left(\mathcal{F}(\xi_1 x^{(1)} + \xi_2 x^{(2)}),\ \mathcal{G}(\xi_1 x^{(1)} + \xi_2 x^{(2)})\right)$$

Since the points $(1, 0)$ and $(0, 1)$ are transformed to the points $y^{(1)}$ and $y^{(2)}$, respectively, then in view of the properties

$$\mathcal{F}(x) = \mathcal{F}(-x), \quad \mathcal{G}(x) = \mathcal{G}(-x)$$

the points $(-1, 0)$ and $(0, -1)$ are transformed also to the same points $y^{(1)}$ and $y^{(2)}$. Consider then the set $\mathcal{M}$ defined as

$$\mathcal{M} := \{(\xi_1, \xi_2) : \varphi(\xi_1, \xi_2) = C\}$$

This set is nonempty since, evidently, $(\pm 1, 0), (0, \pm 1) \in \mathcal{M}$. The function $\varphi(\xi_1, \xi_2)$ may be represented as

$$\varphi(\xi_1, \xi_2) = \alpha\xi_1^2 + 2\beta\xi_1\xi_2 + \gamma\xi_2^2$$

where α, β, γ are some real numbers. If $\alpha = \beta = \gamma = 0$ (and, hence, $C = 0$), then $\mathcal{M}$ is the complete plane $\mathbb{R}^2$. Let $|\alpha| + |\beta| + |\gamma| > 0$. Denote $\delta := \alpha\gamma - \beta^2$. The curve $\varphi(\xi_1, \xi_2) = C$ is

- an ellipse, if $\delta > 0$;
- a hyperbola, if $\delta < 0$;
- a pair of parallel direct lines (maybe coincided), if $\delta = 0$.

Notice also that this curve is symmetric with respect to an origin $(0, 0)$. So, either $\mathcal{M}$ is a connected set or it is represented by two sets symmetric with respect to the origin.

Case 1: $\mathcal{M}$ is a connected set. Then the points $(1, 0)$ and $(0, 1)$ may be connected by a continuous curve without leaving the set $\mathcal{M}$. Let $\xi_1 = \xi_1(t)$, $\xi_2 = \xi_2(t)$ $(0 \le t \le 1)$ be this connecting curve and $y = y(t)$ be the corresponding continuous line. Obviously, this line lies in $\mathcal{L}$ and connects the points $y^{(1)}$ and $y^{(2)}$.

Case 2: $\mathcal{M}$ is two sets symmetric with respect the origin. In this case the points $(1, 0)$ and $(0, 1)$ lie in different connected parts of the curve $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$, say, $(1, 0) \in \mathcal{M}_1$ and $(0, 1) \in \mathcal{M}_2$. Then by symmetricity of the curve $\mathcal{M}$ with respect the origin, the points $(1, 0)$ and $(-1, 0)$ belong to the different connected curves, i.e., $(-1, 0) \in \mathcal{M}_2$. So, we may connect the points $(-1, 0) \in \mathcal{M}_2$ and $(0, 1) \in \mathcal{M}_2$ (whose images are $y^{(1)}$ and $y^{(2)}$, respectively) by a continuous curve $\xi_1 = \xi_1(t)$, $\xi_2 = \xi_2(t)$ $(0 \le t \le 1)$ staying in $\mathcal{M}_2$ such that $y = y(t)$ is again a continuous line lying in $\mathcal{L}$.

In both cases $y(t) \in \mathcal{L}$ $(0 \le t \le 1)$. This curve "covers" the interval $\left[y^{(1)}, y^{(2)}\right]$, that is, for any point $y^{(0)} \in \left[y^{(1)}, y^{(2)}\right]$ there exists $t_0 \in [0, 1]$, such that $y^{(0)} = y(t_0)$. The point $y^{(0)} = (\xi_1(t_0), \xi_2(t_0))$ will correspond to the vector $x^{(0)} = \xi_1(t_0)\, x^{(1)} + \xi_2(t_0)\, x^{(2)}$ verifying

$$y^{(0)} = \left(\mathcal{F}(x^{(0)}),\ \mathcal{G}(x^{(0)})\right)$$

Theorem is proven. □

12.3.2 S-procedure

In this subsection we follow Gelig *et al.* (1978).

Theorem 12.2. (S-procedure for two quadratic forms) *Let*

$$\boxed{\mathcal{F}(x) \geq 0 \quad \textit{for all} \quad x \quad \textit{where} \quad \mathcal{G}(x) \geq 0} \tag{12.8}$$

Then there exist real numbers

$$\boxed{\tau_1 \geq 0, \quad \tau_2 \geq 0, \quad \tau_1 + \tau_2 > 0} \tag{12.9}$$

such that for all $x \in \mathbb{R}^n$

$$\boxed{\tau_1 \mathcal{F}(x) - \tau_2 \mathcal{G}(x) \geq 0} \tag{12.10}$$

Proof. By Daneš' theorem the set $\mathcal{P}$ is a convex cone. Define $\mathcal{Q}$ as

$$\mathcal{Q} := \{(\eta_1, \eta_2) : \eta_1 < 0, \eta_2 \geq 0\}$$

By the assumptions of this theorem $\mathcal{P} \cap \mathcal{Q} = \varnothing$. Since $\mathcal{P}$ and $\mathcal{Q}$ are both convex cones, there exists (loosely) a plane separating them, that is, there exists τ_1, τ_1 (nonobligatory nonnegative) $|\tau_1| + |\tau_2| > 0$ such that

$$\begin{aligned} &\tau_1 \eta_1 - \tau_2 \eta_2 \leq 0 \quad \text{if} \quad (\eta_1, \eta_2) \in \mathcal{Q} \\ &{}_1\eta_1 - \tau_2 \eta_2 \geq 0 \quad \text{if} \quad (\eta_1, \eta_2) \in \mathcal{P} \end{aligned} \tag{12.11}$$

Taking into account that $(-1, 0) \in \mathcal{Q}$ and $(-\varepsilon, 1) \in \mathcal{Q}$ $(\varepsilon > 0)$, then $\tau_1 (-1) \leq 0$ implies

$$\tau_1 \geq 0 \quad \text{and} \quad -\tau_1 \varepsilon - \tau_2 \leq 0$$

Taking $\varepsilon \to 0$ gives $\tau_2 \geq 0$. Since $(\mathcal{F}(x), \mathcal{G}(x)) \in \mathcal{P}$ for any $x \in \mathbb{R}^n$, then the second inequality in (12.11) leads to (12.10). Theorem is proven. □

Corollary 12.1. *Let there exist a vector* $x^{(0)}$ *such that*

$$\boxed{\mathcal{G}(x^{(0)}) > 0} \tag{12.12}$$

Then the following two claims are equivalent:

1.

$$\boxed{\mathcal{F}(x) \geq 0 \quad \textit{for all} \quad x \quad \textit{where} \quad \mathcal{G}(x) \geq 0}$$

2. there exists $\tau \geq 0$ *such that*

$$\boxed{\mathcal{F}(x) - \tau \mathcal{G}(x \geq 0) \quad \textit{for all} \quad x \in \mathbb{R}^n} \tag{12.13}$$

Proof. Since (12.10) holds it is sufficient to show that $\tau_1 > 0$. Suppose the converse, namely, that $\tau_1 = 0$. Then by (12.10) it follows that $\tau_2 \mathcal{G}(x) \leq 0$ for all x, and hence, particularly, for $x = x^{(0)}$, that implies $\tau_2 = 0$. So, $\tau_1 = \tau_2 = 0$, but this contradicts the condition (12.9). Hence, $\tau_1 > 0$. Defining $\tau := \tau_2/\tau_1$ we obtain the main result. Corollary is proven. □

Corollary 12.2. (The case of the strict basic inequality) *Again let there exist a vector* $x^{(0)}$ *such that*

$$\boxed{\mathcal{G}(x^{(0)}) > 0}$$

Then the following two claims are equivalent:

1.

$$\boxed{\mathcal{F}(x) > 0 \quad \textit{for all} \quad x \quad \textit{where} \quad \mathcal{G}(x) \geq 0,\ x \neq 0} \tag{12.14}$$

2. there exists $\tau \geq 0$ *such that*

$$\boxed{\mathcal{F}(x) - \tau \mathcal{G}(x) > 0 \quad \textit{for all} \quad x \in \mathbb{R}^n,\ x \neq 0} \tag{12.15}$$

Proof. Evidently, (12.15) implies (12.14). Indeed, if $\tau = 0$, then $\mathcal{F}(x) > 0$ for any $x \neq 0$. If $\tau > 0$, then $\mathcal{F}(x) > \tau \mathcal{G}(x)$ and $\mathcal{F}(x) > 0$ for any $x \neq 0$ such that $\mathcal{G}(x) \geq 0$. Now, let (12.14) hold. Define the set $\mathcal{J} := \{x : \|x\| = 1,\ \mathcal{G}(x) \geq 0\}$. It is bounded and closed. So, by (12.14) $\mathcal{F}(x) > 0$ for any $x \in \mathcal{J}$. Hence, $\inf\limits_{x \in \mathcal{J}} \mathcal{F}(x) = \varepsilon > 0$, and, as a result, $\mathcal{F}(x) - \varepsilon \geq 0$ for any $x \in \mathcal{J}$. If $\mathcal{G}(x) \geq 0$, $x \neq 0$, then $\dfrac{x}{\|x\|} \in \mathcal{J}$. That's why $\mathcal{F}\left(\dfrac{x}{\|x\|}\right) - \varepsilon \geq 0$, or equivalently, $\mathcal{F}(x) - \varepsilon \|x\|^2 \geq 0$. So, $\tilde{\mathcal{F}}(x) := \mathcal{F}(x) - \varepsilon \|x\|^2 \geq 0$ under $\mathcal{G}(x) \geq 0$ and the previous Corollary (12.1) there exists $\tau \geq 0$ such that $\mathcal{F}(x) - \varepsilon \|x\|^2 - \tau \mathcal{G}(x) \geq 0$. Hence, $\mathcal{F}(x) - \tau \mathcal{G}(x) \geq \varepsilon \|x\|^2 > 0$. Corollary is proven. □

Remark 12.1. *The claim, analogous to (12.2) where nonstrict constraint* $\mathcal{G}(x) \geq 0$ *is changed to the strict one, i.e.,* $\mathcal{G}(x) > 0$, *is not correct which can be shown with a simple counterexample.*

The following matrix interpretation of Theorem 12.2 takes place.

Theorem 12.3. *Let inequalities*

$$\boxed{\mathcal{G}_i(x) := x^{\mathsf{T}} G_i x \leq \alpha_i \ (i = 1, \ldots, m)} \tag{12.16}$$

imply

$$\boxed{\mathcal{F}(x) := x^{\mathsf{T}} F x \leq \alpha_0} \tag{12.17}$$

where α_i $(i = 0, 1, \ldots, m)$ are some real numbers. If there exists $\tau_i \geq 0$ $(i = 1, \ldots, m)$ such that

$$\boxed{F \leq \sum_{i=1}^{m} \tau_i G_i, \quad \alpha_0 \geq \sum_{i=1}^{m} \tau_i \alpha_i} \tag{12.18}$$

then (12.16) implies (12.17). Inversely, if (12.16) implies (12.17) and, additionally, one of the following conditions is fulfilled:

1.

$$\boxed{m = 1}$$

2.

$$\boxed{m = 2, \quad n \geq 3}$$

and there exists a vector $x^{(0)}$, μ_1, μ_2 such that

$$\boxed{\begin{array}{c} \mathcal{G}_i(x^{(0)}) < \alpha_i \quad (i = 1, 2) \\ \mu_1 G_1 + \mu_2 G_2 > 0 \end{array}}$$

then there exists $\tau_i \geq 0$ $(i = 1, \ldots, m)$ such that (12.18) holds.

For $m > 2$ the analogue result is not true.

Proof. Sufficiency is trivial. Necessity follows from the previous Theorem 12.2 and Corollaries 12.1 and 12.2. The simple counterexample may show that this theorem is not valid for $m > 2$. □

12.3.3 Finsler lemma

The following statement is a partial case of Theorems 12.2 and 12.3.

Lemma 12.4. (Finsler 1937) *Let $\mathcal{F}(x) := x^{\mathsf{T}} F x \geq 0$ (or strictly > 0) for all $x \in \mathbb{R}^n$, $x \neq 0$ and such that*

$$\boxed{\mathcal{G}(x) := x^{\mathsf{T}} G x = 0} \tag{12.19}$$

Then there exists a real τ such that

$$\boxed{F + \tau G \geq 0 \ \ (or \ \ strictly > 0)} \tag{12.20}$$

Proof. This lemma is a partial case of Corollary 12.2 if we can show that in here the assumption (12.12) is not essential. Indeed, if (12.12) holds then Corollary 12.2 implies (12.20). If $x^{(0)}$ does not exist, where $\mathcal{G}(x^{(0)}) > 0$, we deal only with two situations:

(a) there exist $x^{(0)}$ where $\mathcal{G}(x^{(0)}) < 0$,

(b) $\mathcal{G}(x) = 0$ for all x.

In case (a) changing G to $(-G)$ we obtain the previous situation when (12.12) holds. In case (b) (12.20) holds automatically. □

Below we will illustrate the role of the Finsler lemma in the quadratic stabilization analysis (Polyak & Sherbakov 2002). Consider the linear plant

$$\dot{x} = Ax + Bu \tag{12.21}$$

with the linear feedback given by

$$u = Kx \tag{12.22}$$

The corresponding closed-loop system is

$$\begin{aligned} &\dot{x} = A_{cl}x \\ &A_{cl} := A + BK \end{aligned} \tag{12.23}$$

The quadratic form $V(x) = x^{\mathsf{T}}Px$ will be the Lyapunov function for (12.23) then and only then when

$$A_{cl}^{\mathsf{T}}P + PA_{cl} < 0, \quad P > 0$$

or, equivalently, when there exist matrices K and $P > 0$ such that

$$\boxed{(A + BK)^{\mathsf{T}} P + P(A + BK) < 0} \tag{12.24}$$

This relation represents a nonlinear matrix inequality with respect to two matrices K and P. Fortunately, the variable changing

$$\boxed{Y := KQ, \quad Q := P^{-1}} \tag{12.25}$$

transforms (12.24) into a linear one by pre- and post-multiplying (12.24) by Q:

$$QA^{\mathsf{T}} + AQ + Y^{\mathsf{T}}B^{\mathsf{T}} + BY < 0, \quad Q > 0 \tag{12.26}$$

Using the Finsler lemma 12.4 the variable Y may be excluded from (12.26). Indeed, the quadratic function

$$\mathcal{F}(x) := x^{\mathsf{T}}(Y^{\mathsf{T}}B^{\mathsf{T}} + BY)x = 2(B^{\mathsf{T}}x, Yx)$$

is equal to zero at the subspace $B^{\mathsf{T}}x = 0$, or when

$$\mathcal{G}(x) := (B^{\mathsf{T}}x, B^{\mathsf{T}}x) = x^{\mathsf{T}}(BB^{\mathsf{T}})x = 0$$

Supposing additionally that $\mathcal{G}(x^{(0)}) > 0$ for some $x^{(0)}$ (but this is not a real constraint) Lemma 12.4 implies that there exists a real τ such that

$$\mathcal{F}(x) + \tau\mathcal{G}(x) \geq 0$$

that is,

$$Y^\top B^\top + BY \geq -\tau BB^\top$$

We may take $\tau > 0$ since $BB^\top \geq 0$. So, (12.26) implies

$$QA^\top + AQ - \tau BB^\top \leq QA^\top + AQ + Y^\top B^\top + BY < 0, \quad Q > 0$$

But the left inequality is reachable if we take

$$Y := -\frac{\tau}{2}B^\top$$

Since $K = YQ^{-1}$ does not depend on τ, we may take $\tau = 2$ and obtain the following result.

Claim 12.1. *If Q is the solution of the Lyapunov inequality*

$$\boxed{QA^\top + AQ - 2BB^\top < 0}$$

then the regulator (12.22) with

$$\boxed{K = -B^\top Q^{-1}}$$

stabilizes the system (12.21) and the quadratic function

$$\boxed{V(x) = x^\top Q^{-1} x}$$

is the Lyapunov function for the closed system (12.23).

12.4 Farkaš lemma

12.4.1 Formulation of the lemma

The *Farkaš lemma* (Farkaš 1902) is a classical result belonging to a class of the, so-called, "theorem of the alternative" which characterizes the optimality conditions of different optimization problems.

Lemma 12.5. (Farkaš 1902) *Let A be a real $m \times n$ matrix and c be a real nonzero vector. Then*

1. either the primal system

$$\boxed{Ax \geq 0, \quad c^{\mathsf{T}}x < 0} \tag{12.27}$$

has a solution $x \in \mathbb{R}^n$

2. or the dual system

$$\boxed{A^{\mathsf{T}}y = c, \quad y \geq 0} \tag{12.28}$$

has a solution $y \in \mathbb{R}^m$,

but never both.

The question of which of the two systems is solvable is answered by considering the bounded least squares problem discussed below.

12.4.2 Axillary bounded least squares (LS) problem

Here we follow Dax (1997). Consider the following *bounded LS problem*:

$$\boxed{\text{minimize} \quad \|A^{\mathsf{T}}y - c\|^2 \quad \text{by} \quad y \in \mathbb{R}^m} \tag{12.29}$$

$$\boxed{\text{subject to} \quad y \geq 0} \tag{12.30}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Lemma 12.6. *The vector $y^* \in \mathbb{R}^m$ solves the problem (12.29)–(12.30) if and only if y^* and the residual vector*

$$\boxed{r^* := A^{\mathsf{T}}y^* - c} \tag{12.31}$$

satisfy the conditions

$$\boxed{\begin{array}{c} y^* \geq 0, \quad Ar^* \geq 0 \\ (y^*)^{\mathsf{T}}\, Ar^* = 0 \end{array}} \tag{12.32}$$

Proof.

(a) *Necessity.* Assume that y^* solves (12.29)–(12.30) and consider the one-parameter quadratic function

$$f_i(\theta) := \left\|A^{\mathsf{T}}\left(y^* + \theta e^{(i)}\right) - c\right\|^2 = \left\|\theta a^{(i)} + r^*\right\|^2, \quad i = 1, \ldots, m \tag{12.33}$$

where $\left(a^{(i)}\right)^{\mathsf{T}}$ is the ith row of A, θ is a real variable and $e^{(i)}$ denotes the ith column of the $m \times m$ unit matrix $I_{m\times m}$. Then, clearly, $\theta = 0$ solves the problem

$$\text{minimize} \quad f_i(\theta) = \theta^2 \left\|a^{(i)}\right\|^2 + 2\theta \left(a^{(i)}\right)^{\mathsf{T}} r^* + \|r^*\|^2$$

$$\text{subject to} \quad y_i^* + \theta \geq 0$$

since for any $y(\theta) := y^* + \theta e^{(i)}$

$$\|A^{\mathsf{T}} y(\theta) - c\|^2 \geq \|A^{\mathsf{T}} y^* - c\|^2$$

and $y(0) := y^*$. Therefore, taking into account that

$$f_i'(0) = \left(a^{(i)}\right)^{\mathsf{T}} r^*$$

we have that $y_i^* > 0$ implies $\left(a^{(i)}\right)^{\mathsf{T}} r^* = 0$, while $y_i^* = 0$ implies $\left(a^{(i)}\right)^{\mathsf{T}} r^* \geq 0$, which constitute (12.32).

(b) *Sufficiency*. Conversely, assume that (12.32) holds and let z be an arbitrary point in $\mathbb{R}^m$ such that $z \geq 0$. Define also $u := z - y^*$. Then $y_i^* = 0$ implies $u_i \geq 0$, while (12.32) leads to

$$u^{\mathsf{T}} A r^* \geq 0$$

Hence, the identity

$$\|A^{\mathsf{T}} z - c\|^2 = \|A^{\mathsf{T}} y^* - c\|^2 + 2u^{\mathsf{T}} A r^* + \|A^{\mathsf{T}} u\|^2$$

shows that

$$\|A^{\mathsf{T}} z - c\|^2 \geq \|A^{\mathsf{T}} y^* - c\|^2$$

Lemma is proven. □

12.4.3 Proof of Farkaš lemma

Notice that $c^{\mathsf{T}} x < 0$ implies $x \neq 0$, while $A^{\mathsf{T}} y = c$ means $y \neq 0$.

1. First, show that it is not possible that both systems are solvable. This can be seen from the following consideration: if both (12.27) and (12.28) hold then

$$c^{\mathsf{T}} x = (A^{\mathsf{T}} y)^{\mathsf{T}} x = y^{\mathsf{T}} A x \geq 0$$

which contradicts $c^{\mathsf{T}} x < 0$.

2. Assuming that y^* exists and combining (12.31) and (12.32) gives

$$c^{\mathsf{T}} r^* = (A^{\mathsf{T}} y^* - r^*)^{\mathsf{T}} r^* = (A^{\mathsf{T}} y^*)^{\mathsf{T}} r^* - (r^*)^{\mathsf{T}} r^* = -\|r^*\|^2$$

which leads to the following conclusion:

Conclusion 12.1. *Let y^* solve (12.29)–(12.30). If $r^* = 0$ then y^* solves (12.28). Otherwise, r^* solves (12.27) and $c^\top r^* = -\|r^*\|^2$.*

3. It remains to establish the existence of a point y^* solving (12.29)–(12.30). It follows from the observation that

$$Z := \{A^\top y \mid y \geq 0\}$$

is a closed set in $\mathbb{R}^n$. Using the closure of Z, we obtain that

$$\mathbb{B} := \{z \in Z \mid \|z - c\| \leq \|c\|\}$$

is a nonempty closed bounded set of $\mathbb{R}^n$. Note also that $\varphi(x) := \|x - c\|^2$ is a continuous function on x. Therefore, by the well-known Weierstrass' theorem, $\varphi(x)$ achieves its minimum over $\mathbb{B}$. Denote it by z^*. Since $\mathbb{B} \subseteq \mathbb{Z}$, there exists $y^* \in \mathbb{R}^m$ such that $y^* \geq 0$ and $z^* = A^\top y^*$. Therefore, y^* solves (12.29)–(12.30). However, y^* is not necessarily unique. By (6.29) from Chapter 6 dealing with the pseudoinverse, any vector

$$y^* = u + (A^\top)^+ c \quad \text{with any} \quad u \geq 0$$

is a solution of (12.29)–(12.30).

12.4.4 The steepest descent problem

Corollary 12.3. *Let y^* and $r^* \neq 0$ solve (12.29)–(12.30). Then the normalized vector $r^*/\|r^*\|$ solves the* ***steepest descent problem***

$$\boxed{\textit{minimize} \quad c^\top x}$$

$$\boxed{\textit{subject to} \quad Ax \geq 0 \quad \textit{and} \quad \|x\| = 1}$$

Proof. Let x satisfy the constraints above. Then

$$(y^*)^\top Ax \geq 0$$

while the Cauchy–Bounyakovski–Schwartz inequality gives

$$\left|(r^*)^\top x\right| \leq \|r^*\| \cdot \|x\| = \|r^*\|$$

Combining these two relations shows that

$$\begin{aligned} c^\top x &= (A^\top y^* - r^*)^\top x = (A^\top y^*)^\top x - (r^*)^\top x \\ &\geq -(r^*)^\top x \geq -\left|(r^*)^\top x\right| \geq -\|r^*\| \end{aligned}$$

Therefore, since $(A^\top y^*)^\top r^* = 0$ and

$$c^\top r^* / \|r^*\| = [(A^\top y^*)^\top r^* - (r^*)^\top r^*] / \|r^*\| = -\|r^*\|$$

the claim is proven. □

12.5 Kantorovich matrix inequality

Theorem 12.4. *If A is an $n \times n$ positive definite Hermitian matrix and $e \in \mathbb{C}^n$ is a unite vector (i.e. $e^*e = 1$), then*

$$\boxed{1 \le (e^*Ae)\left(e^*A^{-1}e\right) \le \frac{1}{4}\left(\sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}} + \sqrt{\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}}\right)^2} \tag{12.34}$$

Proof. The left-hand side of (12.34) follows from the Cauchy–Bounyakovski–Schwartz inequality

$$1 = \sqrt{e^*e} = \sqrt{e^*A^{1/2}A^{-1/2}e} = \sqrt{\left(A^{1/2}e, A^{-1/2}e\right)}$$

$$\le \left\|A^{1/2}e\right\| \cdot \left\|A^{-1/2}e\right\| = \sqrt{(e^*Ae)\left(e^*A^{-1}e\right)} \le (e^*Ae)\left(e^*A^{-1}e\right)$$

The right-hand side of (12.34) can be stated in the following manner using the matrix Λ inequality (valid in a scalar case for any $\varepsilon > 0$):

$$\sqrt{(e^*Ae)\left(e^*A^{-1}e\right)} \le \frac{1}{2}\left[(\varepsilon e^*Ae) + \varepsilon^{-1}\left(e^*A^{-1}e\right)\right]$$

$$= \frac{1}{2}e^*\left[\varepsilon A + \varepsilon^{-1}A^{-1}\right]e \le \frac{1}{2}\lambda_{\max}\left(\varepsilon A + \varepsilon^{-1}A^{-1}\right)$$

$$= \frac{1}{2}\lambda_{\max}\left(T\left[\varepsilon A + \varepsilon^{-1}A^{-1}\right]T^{-1}\right)$$

$$= \frac{1}{2}\lambda_{\max}\left(\operatorname{diag}\left[\varepsilon\lambda_i + \varepsilon^{-1}\lambda_i^{-1}\right]_{i=1,\ldots,n}\right) = \frac{1}{2}\max_i\left(\varepsilon\lambda_i + \varepsilon^{-1}\lambda_i^{-1}\right)$$

Here we have used $TAT^{-1} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$. The function

$$f(\lambda) := \varepsilon\lambda + \varepsilon^{-1}\lambda^{-1}$$

is convex for all $\lambda \in [\lambda_{\min}(A), \lambda_{\max}(A)]$ and therefore it takes its maximum in one of the boundary points, that is,

$$\max_i\left(\varepsilon\lambda_i + \varepsilon^{-1}\lambda_i^{-1}\right)$$
$$= \max\left\{\left(\varepsilon\lambda_{\max}(A) + \varepsilon^{-1}\lambda_{\max}^{-1}(A)\right),\ \left(\varepsilon\lambda_{\min}(A) + \varepsilon^{-1}\lambda_{\min}^{-1}(A)\right)\right\}$$

Taking $\varepsilon := \dfrac{1}{\sqrt{\lambda_{\max}(A)\,\lambda_{\min}(A)}}$ we have

$$\max\left\{\left(\varepsilon\lambda_{\max}(A)+\varepsilon^{-1}\lambda_{\max}^{-1}(A)\right),\ \left(\varepsilon\lambda_{\min}(A)+\varepsilon^{-1}\lambda_{\min}^{-1}(A)\right)\right\}$$
$$=\sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}}+\sqrt{\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}}$$

which implies

$$\sqrt{(e^*Ae)\left(e^*A^{-1}e\right)}\le\frac{1}{2}\left(\sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}}+\sqrt{\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}}\right)$$

and, as the result, we obtain (12.34). □

Remark 12.2. *The right-hand side of (12.34) is achievable for* $A=\lambda I$ *since in this case* $\lambda_{\max}(A)=\lambda_{\min}(A)=\lambda$ *and*

$$\frac{1}{4}\left(\sqrt{\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}}+\sqrt{\frac{\lambda_{\min}(A)}{\lambda_{\max}(A)}}\right)^2=1$$

PART II

Analysis

13 The Real and Complex Number Systems

Contents

In this part of the book we will follow the following classical publications: Rudin (1976), Apostol (1974), Fuchs & Shabat (1964) and Lavrentiev & Shabat (1987).

13.1 Ordered sets

13.1.1 Order

Definition 13.1. *Let S be a set of elements. An* ***order*** *on S is a relation, denoted by $<$, with the following properties:*

1. If $x \in S$ and $y \in S$ then one and only one of the statements is true:

$$x < y, \quad x = y, \quad y < x$$

2. If $x, y, z \in S$ and, in addition, $x < y$ and $y < z$, then $x < z$.

The statement "$x < y$" is referred to as "x is less (or smaller) than y". The notation "$x \leq y$" is the negation of "$y > x$".

Definition 13.2. *An* ***ordered set*** *is a set S in which an order is defined.*

13.1.2 Infimum and supremum

Definition 13.3. *Suppose S is an ordered set and $E \subset S$. If there exists an element $\beta \in S$ such that $x \leq \beta$ for any $x \in E$, we say that the subset E is* ***bounded above*** *and call β an upper bound of E. A* ***low bound*** *is defined in the same way with "$\geq$" in place of "$\leq$".*

Definition 13.4. *Suppose S is an ordered set and $E \subset S$ is bounded above. Suppose also there exists an $\alpha \in S$ such that*

(a) α is an upper bound for E.
(b) If $\gamma < \alpha$ then γ is not an upper bound of E.

*Then α is called the **least upper bound** of E or the **supremum** of E which will be written as*

$$\boxed{\alpha = \sup E = \sup_{x \in E} x} \tag{13.1}$$

*The **greatest lower bound**, or **infimum** of E, which is bound below, is defined in the same manner, namely,*

$$\boxed{\alpha = \inf E = \inf_{x \in E} x} \tag{13.2}$$

This means that α is a low bound of E and there is no $\beta > \alpha$ which is low bound too.

13.2 Fields

13.2.1 Basic definition and main axioms

Definition 13.5. *A **field** is a set $\mathcal{F}$ of elements with two operations:*

1. *Addition;*
2. *Multiplication.*

Both mentioned operations should satisfy the following "field axioms":

(A): **Axioms for addition**
- (A1) If $x, y \in \mathcal{F}$ then $(x + y) \in \mathcal{F}$.
- (A2) $x + y = y + x$ for all $x, y \in \mathcal{F}$ which means that addition is **commutative**.
- (A3) $(x + y) + z = y + (x + z)$ for all $x, y, z \in \mathcal{F}$ which means that addition is **associative**.
- (A4) $\mathcal{F}$ contains an element called 0 such that $x + 0 = x$ for all $x \in \mathcal{F}$.
- (A5) For any $x \in \mathcal{F}$ there exists an element $(-x) \in \mathcal{F}$ such that $x + (-x) = 0$.

(M): **Axioms for multiplication**
- (M1) If $x, y \in \mathcal{F}$ then $xy \in \mathcal{F}$.
- (M2) $xy = yx$ for all $x, y \in \mathcal{F}$ which means that multiplication is **commutative**.
- (M3) $(xy)z = y(xz)$ for all $x, y, z \in \mathcal{F}$ which means that multiplication is **associative**.
- (M4) $\mathcal{F}$ contains an element called 1 such that $1x = x$ for all $x \in \mathcal{F}$.
- (M5) For any $x \in \mathcal{F}$ and $x \neq 0$ there exists an element $1/x \in \mathcal{F}$ such that $x(1/x) = 1$.

(D): **The distributive law**

For all $x, y, z \in \mathcal{F}$ the following identity holds:

$$x_{\prime}(y + z) = xy + xz$$

13.2.2 Some important properties

Proposition 13.1. (Resulting from axioms (A))
(a) If $x + y = x + z$ *then* $y = z$ *(cancellation law).*
(b) If $x + y = x$ *then* $y = 0$.
(c) If $x + y = 0$ *then* $y = -x$.
(d) $-(-x) = x$.

Proposition 13.2. (Resulting from axioms (M))
(a) If $xy = xz$ *and* $x \neq 0$ *then* $y = z$.
(b) If $xy = x$ *and* $x \neq 0$ *then* $y = 1$.
(c) If $xy = 1$ *and* $x \neq 0$ *then* $y = 1/x$.
(d) If $x \neq 0$ *then* $1/(1/x) = x$.

Proposition 13.3. (Resulting from (A), (M) and (D))
(a) $0x = 0$.
(b) If $x \neq 0$ *and* $y \neq 0$ *then* $xy \neq 0$.
(c) $(-x)\, y = (-x)\, y = -(xy)$.
(d) $(-x)(-y) = xy$.

Definition 13.6. *An* ***ordered field*** *is a field* $\mathcal{F}$ *which is also an ordered set such that*
(a) $x + y < x + z$ *if* $x, y, z \in \mathcal{F}$ *and* $y < z$.
(b) $xy > 0$ *if* $x, y \in \mathcal{F}$ *and* $x > 0$, $y > 0$ *or* $x < 0$, $y < 0$.
If $x > 0$ *we call* x *positive and if* $x < 0$ *we call* x *negative.*

In every ordered field the following statements are true.

Proposition 13.4.
(a) If $x > 0$ *then* $-x < 0$.
(b) If $x > 0$ *and* $y < z$ *then* $xy < xz$.
(c) If $x < 0$ *and* $y < z$ *then* $xy > xz$.
(d) If $x \neq 0$ *then* $x^2 := xx > 0$.
(e) If $0 < x < y$ *then* $0 < 1/y < 1/x$.

Below we will deal with two commonly used fields: **real** and **complex**. It will be shown that the real field is an ordered field and the complex field is nonordered.

13.3 The real field

13.3.1 Basic properties

The following existence theorem holds.

Theorem 13.1. *There exists an ordered field* $\mathbb{R}$ *which possesses the following properties:*

- *If* $E \subset \mathbb{R}$ *and* E *is not empty and bounded above, then* $\sup E$ *exists in* $\mathbb{R}$;
- $\mathbb{R}$ *contains the set* Q *of all* ***rational numbers*** r *(*$r = m/n$ *where* m, n *are* ***integers*** $(\ldots -1, 0, 1, \ldots)$ *and* $n \neq 0$*) is a subfield.*

Proof of this theorem is rather long and tedious and therefore is omitted. It can be found in the appendix to Chapter 1 of Rudin (1976).

Definition 13.7. *The members of* $\mathbb{R}$ *are called* ***real numbers*** *and* $\mathbb{R}$ *itself is called the* ***real field****.*

The members of $\mathbb{R}$ have several simple properties given below.

Claim 13.1.

(a) **Archimedean property:** *If* $x, y \in \mathbb{R}$ *and* $x > 0$*, then there exists a positive integer number* n *such that* $nx > y$.

(b) Q**-density property in** $\mathbb{R}$**:** *If* $x, y \in \mathbb{R}$ *and* $x < y$*, then there exists a rational number* $p \in Q$ *such that* $x < p < y$.

(c) **The root existence:** *For any nonnegative real* $x \in \mathbb{R}$ $(x \geq 0)$ *and any integer* $n > 0$ *there is one and only one real* $y \in \mathbb{R}$ *such that* $y^n = x$*. This number* y *is written as*

$$y = \sqrt[n]{x} = x^{1/n}$$

13.3.2 Intervals

Definition 13.8.

- *The* ***open interval*** (a, b) *is the set of real numbers* x *such that* $a < x < b$*, i.e.,*

$$\boxed{(a, b) := \{x : a < x < b\}}$$

- *The* ***closed interval*** $[a, b]$ *is the set of real numbers* x *such that* $a \leq x \leq b$*, i.e.,*

$$\boxed{[a, b] := \{x : a \leq x \leq b\}}$$

- *The semi-open intervals* $[a, b)$ *and* $(a, b]$ *are the sets of real numbers such that* $a \leq x < b$ *and* $a < x \leq b$*, i.e.,*

$$\boxed{\begin{array}{l}[a, b) := \{x : a \leq x < b\} \\ (a, b] := \{x : a < x \leq b\}\end{array}}$$

13.3.3 Maximum and minimum elements

Definition 13.9. *Let* S *be a set of real numbers.*

(a) *If a smallest upper bound* $\alpha = \sup S$ *is also a member of* S *then* α *is called the* ***largest number*** *or the* ***maximum element*** *of* S *and denoted by* $\max S$*, that is, in this case*

$$\boxed{\alpha = \max S = \sup S} \tag{13.3}$$

(b) *If the greatest low bound* $\beta = \inf S$ *is also a member of* S *then* β *is called the* ***smallest number*** *or the* ***minimum element*** *of* S *and denoted by* $\max S$*, that is, in this case*

$$\boxed{\beta := \min S = \inf S} \tag{13.4}$$

Example 13.1.

1. For $S = [a, b]$ *it follows that*

$$\max S = \sup S = b$$
$$\min S = \inf S = a$$

2. For $S = (a, b)$ *it follows that*

$$\max S \quad \textit{does not exist}, \quad \sup S = b$$
$$\min S \quad \textit{does not exist}, \quad \inf S = a$$

3. For $S = [0, 1 - 1/2, 1 - 1/3, \ldots, 1 - 1/k, \ldots)$ *it follows that*

$$\max S \quad \textit{does not exist}, \quad \sup S = 1$$
$$\min S = \inf S = 0$$

13.3.4 Some properties of the supremum

Lemma 13.1. (Approximation property) *Let* S *be a nonempty set of real numbers with supremum, say* $b = \sup S$. *Then* S *contains numbers arbitrarily close to its supremum, that is, for every* $a < b$ *there is some* $x \in S$ *such that*

$$a < x \leq b$$

Proof. One has that $x \leq b$ for all $x \in S$. Supposing that $x \leq a$ for all $x \in S$ we obtain that a is an upper bound for S which is strictly less than b which contradicts the assumption that b is the lowest upper bound. So, $x > a$. □

Lemma 13.2. (Additive property) *Given nonempty sets* $\mathcal{A}$ *and* $\mathcal{B}$ *of* $\mathbb{R}$, *let* $\mathcal{C}$ *denote the set*

$$\mathcal{C} := \{z \in \mathbb{R} : z = x + y, x \in \mathcal{A}, y \in \mathcal{B}\}$$

If each of $\mathcal{A}$ *and* $\mathcal{B}$ *has a supremum, then* $\mathcal{C}$ *has a supremum too and*

$$\boxed{\sup \mathcal{C} = \sup \mathcal{A} + \sup \mathcal{B}} \tag{13.5}$$

Proof. Denoting $a := \sup \mathcal{A}$ and $b := \sup \mathcal{B}$ we have that $z = x + y \leq a + b$. Hence, $(a + b)$ is a supremum for $\mathcal{C}$. So, $\mathcal{C}$ has a supremum, say $c := \sup \mathcal{C}$ and $c \leq a + b$. Show next that $a + b \leq c$. By Lemma 13.1 it follows that there exist $x \in \mathcal{A}$ and $y \in \mathcal{B}$ such that

$$a - \varepsilon < x \leq a, \quad b - \varepsilon < y \leq b$$

for any chosen $\varepsilon > 0$. Adding these inequalities we find $a + b - 2\varepsilon < x + y$ or, equivalently, $a + b < x + y + 2\varepsilon$. So, $a + b < c + 2\varepsilon$. Taking $\varepsilon \to 0$, we obtain that $a + b \leq c$ and together with $c \leq a + b$ states that $c = a + b$. □

Lemma 13.3. (Comparison property) *Given nonempty sets S and T of $\mathbb{R}$ such that $s \leq t$ for any $s \in S$ and $t \in T$. If T has an infimum $\beta = \inf T$ then S has a supremum and*

$$\boxed{\sup S \leq \inf T} \tag{13.6}$$

Proof. $\sup S$ exists by the property $s \leq t$. Denote $\alpha := \sup S$. By Lemma 13.1 for any $\varepsilon > 0$ there exists $s \in S$ such that $\alpha - \varepsilon < s$ and there is t such that $\beta + \varepsilon > t$. So,

$$\alpha - \varepsilon < s \leq t < \beta + \varepsilon$$

or, $\alpha < \beta + 2\varepsilon$. Tending ε to zero leads to (13.6). □

Lemma 13.4. *If $\mathcal{A} \subseteq \mathcal{B} \subset \mathbb{R}$, then*

$$\boxed{\begin{array}{l} \sup \mathcal{A} \leq \sup \mathcal{B} \\ \inf \mathcal{A} \geq \inf \mathcal{B} \end{array}} \tag{13.7}$$

Proof. It evidently follows from (13.1) and (13.2). □

13.3.5 Absolute value and the triangle inequality

Definition 13.10. *For any real number x the absolute value of x, denoted by $|x|$, is defined as follows:*

$$\boxed{|x| = \begin{cases} x & if \quad x \geq 0 \\ -x & if \quad x < 0 \end{cases}} \tag{13.8}$$

Evidently, $|x| \geq 0$ always.

Lemma 13.5. (The fundamental inequality) *If $|x| \leq a$ then*

$$\boxed{-a \leq x \leq a} \tag{13.9}$$

Proof. This is a simple consequence of (13.8). □

Theorem 13.2. (The triangle inequality) *For any real $x, y \in \mathbb{R}$ we have*

$$\boxed{|x + y| \leq |x| + |y|} \tag{13.10}$$

Proof. Adding two inequalities

$$-|x| \leq x \leq |x|$$
$$-|y| \leq y \leq |y|$$

gives

$$-|x| - |y| \leq x + y \leq |x| + |y|$$

which implies

$$|x + y| \leq |x| + |y|$$

Theorem is proven. □

Corollary 13.1. *For any real* $x, y, z \in \mathbb{R}$ *we have*
1.

$$\boxed{|x - z| \leq |x - y| + |y - z|} \tag{13.11}$$

2.

$$\boxed{|x \pm y| \geq |x| - |y|} \tag{13.12}$$

3. For any real numbers $x_i \in \mathbb{R}$ $\left(i = \overline{1, n}\right)$

$$\boxed{\left|\sum_{i=1}^{n} x_i\right| \geq |x_1| - |x_2| - \ldots - |x_n|}$$

Proof. The inequality (13.11) follows from (13.10) written as

$$|\tilde{x} + \tilde{y}| \leq |\tilde{x}| + |\tilde{y}| \tag{13.13}$$

if we take $\tilde{x} := x - z$ and $\tilde{y} := -(y - z)$. The inequality (13.12) follows from (13.13) if we take $\tilde{x} := x \pm y$ and $\tilde{y} := \mp y$. The third inequality may be easily proven by induction. □

13.3.6 The Cauchy–Schwarz inequality

Theorem 13.3. (The Cauchy–Schwarz inequality) *For any real numbers* $x_i, y_i \in \mathbb{R}$ $\left(i = \overline{1, n}\right)$ *the following inequality holds*

$$\boxed{\left(\sum_{i=1}^{n} x_i y_i\right)^2 \leq \left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i^2\right)} \tag{13.14}$$

Proof. For any $z \in \mathbb{R}$ we have

$$0 \leq \sum_{i=1}^{n} (x_i z + y_i)^2 = \sum_{i=1}^{n} \left[z^2 (x_i)^2 + 2z x_i y_i + y_i^2\right]$$
$$= Az^2 + 2Bz + C$$

where

$$A := \sum_{i=1}^{n} x_i^2, \quad B := \sum_{i=1}^{n} x_i y_i, \quad C := \sum_{i=1}^{n} y_i^2$$

This quadratic polynomial may be nonnegative for any $z \in \mathbb{R}$ if and only if

$$B^2 - AC \leq 0$$

which is equivalent to (13.14). □

13.3.7 The extended real number system

Definition 13.11. *The extended real number system consists of the real field $\mathbb{R}$ and two symbols: $+\infty$ (or simply ∞) and $-\infty$ which possess the following properties:*

(a) for any real $x \in \mathbb{R}$

$$-\infty < x < \infty$$

(b) for any real $x \in \mathbb{R}$

$$x + \infty = \infty$$
$$x - \infty = -\infty$$
$$\frac{x}{\infty} = \frac{x}{-\infty} = 0$$

(c) if $x > 0$ then

$$x \cdot \infty = \infty \quad \textit{and} \quad x \cdot (-\infty) = -\infty$$

if $x < 0$ then

$$x \cdot \infty = -\infty \quad \textit{and} \quad x \cdot (-\infty) = \infty$$

13.4 Euclidean spaces

Let us consider an integer positive k and let $\mathbb{R}^k$ be the set of all ordered k-tuples

$$\mathbf{x} := (x_1, x_2, \ldots, x_k)$$

where x_i $(i = 1, \ldots, k)$ are real numbers, called the *coordinates* of $\mathbf{x}$. The elements of $\mathbb{R}^k$ are called points (or vectors). Defining two operations

$$\boxed{\begin{aligned}\mathbf{x} + \mathbf{y} &:= (x_1 + y_1, x_2 + y_2, \ldots, x_k + y_k)\\ \alpha\mathbf{x} &:= (\alpha x_1, \alpha x_2, \ldots, \alpha x_k),\ \alpha \in \mathbb{R}\end{aligned}} \tag{13.15}$$

it is easy to see that they satisfy the commutative, associative and distributive law that make $\mathbb{R}^k$ into a vector space over the vector field with $\mathbf{0}$ elements all of whose coordinates are 0.

We also may define the, so-called, *inner (scalar) product* of two vectors $\mathbf{x}$ and $\mathbf{y}$ by

$$\boxed{\mathbf{x} \cdot \mathbf{y} = (\mathbf{x}, \mathbf{y}) := \sum_{i=1}^{k} x_i y_i} \tag{13.16}$$

and the corresponding norm of $\mathbf{x}$ by

$$\boxed{\|\mathbf{x}\| = \sqrt{(\mathbf{x}, \mathbf{x})} := \left(\sum_{i=1}^{k} x_i^2\right)^{1/2}} \tag{13.17}$$

Definition 13.12. *The vector space $\mathbb{R}^k$ with the above inner product (13.16) and norm (13.17) is called* ***Euclidean k-space.***

The following properties of the norm (13.17) hold.

Remark 13.1. *The Cauchy–Schwarz inequality (13.14) for the Euclidean k-space $\mathbb{R}^k$ may be rewritten as*

$$\boxed{(\mathbf{x}, \mathbf{y})^2 \leq \|\mathbf{x}\|^2 \, \|\mathbf{y}\|^2} \tag{13.18}$$

or, equivalently,

$$\boxed{|(\mathbf{x}, \mathbf{y})| \leq \|\mathbf{x}\| \, \|\mathbf{y}\|} \tag{13.19}$$

13.5 The complex field

13.5.1 Basic definition and properties

Definition 13.13.

(a) A ***complex number*** *is an "ordered" pair $x = (x_1, x_2)$ of real numbers where the first member x_1 is called the* ***real part*** *of the complex number and the second member x_2 is called the* ***imaginary part.*** *"Ordered" means that (x_1, x_2) and (x_2, x_1) are regarded as distinct if $x_1 \neq x_2$.*

(b) *Two complex numbers* $x = (x_1, x_2)$ *and* $y = (y_1, y_2)$ *are said to be* ***equal*** *(we write* $x = y$*) if and only if*

$$\boxed{x_1 = y_1 \quad \text{and} \quad x_2 = y_2}$$

(c) *The sum* $(x + y)$ *and the product* (xy) *are defined by the equations*

$$\boxed{\begin{aligned} x + y &= (x_1 + y_1, x_2 + y_2) \\ xy &= (x_1 y_1 - x_2 y_2, x_1 y_2 + x_2 y_1) \end{aligned}} \tag{13.20}$$

The set of all complex numbers is denoted by $\mathbb{C}$.

It is easy to check that the main field operations, namely, addition and multiplication (13.20) satisfy the commutative, associative and distributive laws.

The following properties hold in $\mathbb{C}$.

Proposition 13.5.

1.

$$\boxed{\begin{aligned} (x_1, x_2) + (0, 0) &= (x_1, x_2) \\ (x_1, x_2)\,(0, 0) &= (0, 0) \\ (x_1, x_2)\,(1, 0) &= (x_1, x_2) \\ (x_1, x_2) + (-x_1, -x_2) &= (0, 0) \end{aligned}}$$

2. *Given two complex numbers* $x = (x_1, x_2)$ *and* $y = (y_1, y_2)$ *there exists a complex number* $z = (z_1, z_2)$ *such that* $x + z = y$. *In fact,*

$$\boxed{z := y - x = (y_1 - x_1, y_2 - x_2)}$$

The complex number $(-x_1, -x_2)$ *is denoted by* $(-x)$.

3. *For any two complex numbers* x *and* y *we have*

$$\boxed{(-x)\,y = x\,(-y) = -(xy) = (-1, 0)\,(xy)}$$

4. *Given two complex numbers* $x = (x_1, x_2) \neq (0, 0)$ *and* $y = (y_1, y_2)$ *there exists a complex number* $z = (z_1, z_2)$ *such that*

$$\boxed{\begin{aligned} &xz = y, \text{ namely, } z := y/x = yx^{-1} \\ &x^{-1} := \left(\frac{x_1}{x_1^2 + x_2^2}, -\frac{x_2}{x_1^2 + x_2^2} \right) \end{aligned}}$$

Remark 13.2. *The complex number* $(x_1, 0) = x_1$ *is the real number* x_1. *This identification gives us the real field as a subfield of the complex field.*

13.5.2 The imaginary unite

Definition 13.14. *The **imaginary unite** i is the complex number* $(0, 1)$, *that is,*

$$\boxed{i := (0, 1)} \tag{13.21}$$

Lemma 13.6.

$$\boxed{i^2 = -1} \tag{13.22}$$

Proof. Indeed, by (13.20) and (13.2)

$$i^2 = (0, 1)\,(0, 1) = (-1, 0) = -1$$

Lemma is proven. □

Lemma 13.7. *Any complex number $x = (x_1, x_2)$ can be represented as*

$$\boxed{x = x_1 + ix_2} \tag{13.23}$$

Proof. Since by (13.2)

$$x_1 = (x_1, 0)$$
$$ix_2 = (0, 1)\,(x_2, 0) = (0, x_2)$$

it follows that

$$x_1 + ix_2 = (x_1, 0) + (0, x_2) = (x_1, x_2) = x$$

Lemma is proven. □

13.5.3 The conjugate and absolute value of a complex number

Definition 13.15. *If a and b are real and $z = a + ib$, then*
(a) the complex number

$$\boxed{\bar{z} = a - ib} \tag{13.24}$$

*is called the **conjugate** of z and*

$$\boxed{a = \operatorname{Re} z \quad \text{and} \quad b = \operatorname{Im} z} \tag{13.25}$$

*are referred to as the **real** and **imaginary** parts of z;*

(b) the nonnegative real number $|z|$ *given by*

$$\boxed{|z| = \sqrt{a^2 + b^2}} \tag{13.26}$$

is called the ***absolute value*** *(or* ***module****) of the complex number* z.

Proposition 13.6. *If* x *and* z *are complex then*

1.

$$\boxed{\overline{x + z} = \bar{x} + \bar{z}}$$

2.

$$\boxed{\overline{xz} = \bar{x}\bar{z}}$$

3.

$$\boxed{\begin{aligned} z + \bar{z} &= 2\,\mathrm{Re}\, z \\ z - \bar{z} &= 2\,\mathrm{Im}\, z \end{aligned}}$$

4.

$$\boxed{z\bar{z} = |z|^2}$$

5. The identity $|z| = 0$ *implies that* $z = 0 = (0, 0)$.

6.

$$\boxed{\begin{aligned} |\bar{z}| &= |z| \\ |xz| &= |x|\,|z| \end{aligned}}$$

7.

$$\boxed{|\mathrm{Re}\, z| \leq |z|}$$

8.

$$\boxed{|x/z| = |x| \,/\, |z| \quad \textit{for} \quad z \neq 0}$$

9. The ***triangle inequality*** *holds:*

$$\boxed{|x + z| \leq |x| + |z|}$$

Proof. Propositions 1–7 can be checked directly using the definition only. The proof of 8 follows from the identity

$$|x/z| = \sqrt{\left(\frac{x}{z}\right)\overline{\left(\frac{x}{z}\right)}} = \sqrt{\left(\frac{x}{z}\right)\overline{\left(\frac{x}{z}\right)}} = \sqrt{\frac{x\bar{x}}{z\bar{z}}} = \sqrt{\frac{|x|^2}{|z|^2}}$$

To prove 9 notice that $x\bar{z}$ is the conjugate of $\bar{x}z$, which is why by property 3

$$x\bar{z} + \bar{x}z = 2\,\mathrm{Re}\,(x\bar{z})$$

and, hence,

$$\begin{aligned} |x+z|^2 &= (x+z)(\bar{x}+\bar{z}) \\ &= x\bar{x} + z\bar{z} + x\bar{z} + \bar{x}z = |x|^2 + |z|^2 + 2\,\mathrm{Re}\,(x\bar{z}) \\ &\le |x|^2 + |z|^2 + 2|x\bar{z}| = |x|^2 + |z|^2 + 2|x|\,|\bar{z}| \\ &= (|x| + |z|)^2 \end{aligned}$$

which proves 9. □

Theorem 13.4. (Schwarz inequality for complex numbers)
If $a_i, b_i \in \mathbb{C}$ $(i = \overline{1,n})$ *then*

$$\boxed{\left|\sum_{i=1}^{n} a_i \bar{b}_i\right|^2 \le \left(\sum_{i=1}^{n} |a_i|^2\right)\left(\sum_{i=1}^{n} |b_i|^2\right)} \tag{13.27}$$

Proof. Denote

$$A := \sum_{i=1}^{n} |a_i|^2, \quad B := \sum_{i=1}^{n} |b_i|^2, \quad C := \sum_{i=1}^{n} a_i \bar{b}_i$$

Notice that A, B are real and C is complex. If $B = 0$ then all $b_i = 0$ and the inequality is trivial. Assume now that $B > 0$. Then by (13.6)

$$\begin{aligned} 0 \le \sum_{i=1}^{n} |\, Ba_i - Cb_i\,|^2 &= \sum_{i=1}^{n} (Ba_i - Cb_i)\left(B\bar{a}_i - C\bar{b}_i\right) \\ &= B^2 \sum_{i=1}^{n} |a_i|^2 - B\bar{C} \sum_{i=1}^{n} a_i \bar{b}_i - BC \sum_{i=1}^{n} \bar{a}_i b_i + |C|^2 \sum_{i=1}^{n} |b_i|^2 \\ &= B^2 A - B\,|C|^2 - B\,|C|^2 + |C|^2 B = B^2 A - B\,|C|^2 \\ &= B\left(BA - |C|^2\right) \end{aligned}$$

So, $BA - |C|^2 \ge 0$ coincides with (13.27). Theorem is proven. □

13.5.4 The geometric representation of complex numbers

Let us consider the plane with Cartesian (Decartes) coordinates x (as the abscise) and y (as the ordinate). So the complex number $z = (x, y)$ may be considered as the point on this plane or, equivalently, as the vector with the coordinates x and y (see Fig. 13.1). In the polar coordinates (r, φ) the same vector is expressed as

$$\boxed{z = x + iy = r(\cos\varphi + i\sin\varphi)} \tag{13.28}$$

where $r = |z|$ is the *module* of the complex number z and φ is its *argument* (or phase) denoted by Arg z, that is,

$$\boxed{\varphi = \operatorname{Arg} z := \begin{cases} \arctan\left(\dfrac{y}{x}\right) + 2\pi k & \text{for I and IV quadrants} \\ \arctan\left(\dfrac{y}{x}\right) + (2k+1)\pi & \text{for II and III quadrants} \end{cases}} \tag{13.29}$$

where $\arctan\left(\dfrac{y}{x}\right)$ means the principal (main) value of $\operatorname{Arctan}\left(\dfrac{y}{x}\right)$, i.e., the value which is more than $(-\pi/2)$ and does not exceed $(\pi/2)$, and $k = 0, 1, 2, \ldots$ is any integer number. As it follows from the definitions above, the module is uniquely defined while the *argument is not uniquely defined.*

Proposition 13.7.

1. For any $z_1, z_2 \in \mathbb{C}$

$$\boxed{z_1 z_2 = |z_1|\,|z_2|\,(\cos(\varphi_1 + \varphi_2) + i\sin(\varphi_1 + \varphi_2))} \tag{13.30}$$

$$\boxed{\frac{z_1}{z_2} = \frac{|z_1|}{|z_2|}(\cos(\varphi_1 - \varphi_2) + i\sin(\varphi_1 - \varphi_2)),\ z_2 \neq 0} \tag{13.31}$$

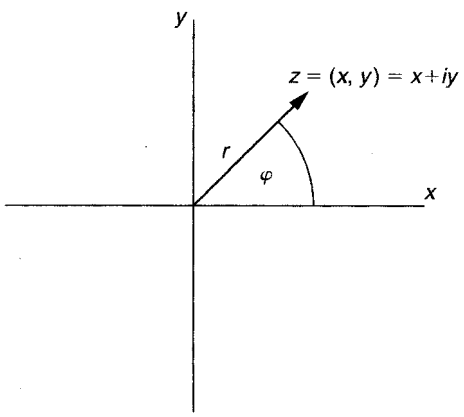

Fig. 13.1. The complex number z in the polar coordinates.

2. *For any* $z_1, z_2, \ldots, z_n \in \mathbb{C}$

$$\boxed{\prod_{i=1}^{n} z_i = \prod_{i=1}^{n} |z_i| \left(\cos\left(\sum_{i=1}^{n} \varphi_i \right) + i \sin\left(\sum_{i=1}^{n} \varphi_i \right) \right)} \tag{13.32}$$

Proof. The property (13.30) follows from (13.28) and the identities

$$\cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2) = \cos(\varphi_1 + \varphi_2)$$
$$\cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2) = \sin(\varphi_1 + \varphi_2)$$

Indeed,

$$\begin{aligned} z_1 z_2 &= r_1 r_2 (\cos\varphi_1 + i \sin\varphi_1)(\cos\varphi_2 + i \sin\varphi_2) \\ &= r_1 r_2 [\cos(\varphi_1)\cos(\varphi_2) - \sin(\varphi_1)\sin(\varphi_2) \\ &\quad i(\cos(\varphi_1)\sin(\varphi_2) + \sin(\varphi_1)\cos(\varphi_2))] \\ &= r_1 r_2 (\cos(\varphi_1 + \varphi_2) + i \sin(\varphi_1 + \varphi_2)) \end{aligned}$$

The property (13.31) related to the quotient $\frac{z_1}{z_2}$ results from the relations

$$\begin{aligned} \frac{z_1}{z_2} &= \frac{z_1 \bar{z}_2}{z_2 \bar{z}_2} = \frac{1}{|z_2|^2} z_1 \bar{z}_2 \\ z_1 \bar{z}_2 &= r_1 r_2 (\cos\varphi_1 - i \sin\varphi_1)(\cos\varphi_2 - i \sin\varphi_2) \\ &= r_1 r_2 (\cos\varphi_1 + i \sin(-\varphi_1))(\cos\varphi_2 + i \sin(-\varphi_2)) \end{aligned}$$

with the following application of (13.30). The identity (13.32) results from (13.30) by induction. □

Example 13.2.

$$\boxed{z^{-1} = r^{-1}(\cos\varphi + i \sin(-\varphi)) = r^{-1}(\cos\varphi - i \sin\varphi)} \tag{13.33}$$

13.6 Some simple complex functions

13.6.1 Power

Definition 13.16. *The nth* ***power*** *of the complex number z is the product*

$$\boxed{\begin{aligned} z^n &:= z^{n-1} z, \quad z^0 = 1, \quad n = 0, 1, 2, \ldots \\ z^{-n} &:= (z^{-1})^n, \quad z \neq 0, \quad n = 1, 2, \ldots \end{aligned}} \tag{13.34}$$

By (13.32) and (13.33) it follows that

$$\begin{aligned} z^n &= r^n \left[\cos(n\varphi) + i \sin(n\varphi)\right] \\ z^{-n} &= r^{-n} \left[\cos(n\varphi) - i \sin(n\varphi)\right] \end{aligned} \tag{13.35}$$

Example 13.3.

(a)

$$i^{4k+3} = i^{4k} i^3 = \left(i^4\right)^k (-i) = -i$$

(b)

$$\begin{aligned} (1-i)^{-2} &= \left(\frac{1}{1-i}\right)^2 = \left(\frac{1+i}{(1-i)(1+i)}\right)^2 \\ &= \left(\frac{1+i}{2}\right)^2 = \frac{1}{4}\left(1+2i+i^2\right) = \frac{i}{2} \end{aligned}$$

13.6.2 Roots

Definition 13.17. *If two complex numbers w and z are related by the equation*

$$\boxed{w^n = z, \quad n = 1, 2, \ldots} \tag{13.36}$$

then w is called a root of degree n of the number z and denoted as

$$\boxed{w := \sqrt[n]{z}} \tag{13.37}$$

Lemma 13.8. (The Moivre–Laplace formula) *There exist exactly n roots of $\sqrt[n]{z}$ which may be expressed as*

$$\boxed{\begin{aligned} &w_k := \sqrt[n]{r}\left[\cos\left(\frac{\varphi + 2\pi k}{n}\right) + i \sin\left(\frac{\varphi + 2\pi k}{n}\right)\right] \\ &\text{for} \quad z = r(\cos\varphi + i \sin\varphi), \quad k = 0, 1, 2, \ldots, n-1 \end{aligned}} \tag{13.38}$$

Proof. Denoting $w = \rho(\cos\theta + i \sin\theta)$, by (13.36) and (13.35) we derive

$$w^n = \rho^n \left[\cos(n\theta) + i \sin(n\theta)\right] = r(\cos\varphi + i \sin\varphi) = z$$

This leads to the following relations:

$$\rho = \sqrt[n]{r}, \quad n\theta = \varphi + 2\pi k$$

which completes the proof. □

Example 13.4.

$$\sqrt[4]{-1} = \cos\left(\frac{\pi + 2\pi k}{n}\right) + i \sin\left(\frac{\pi + 2\pi k}{n}\right)$$

$$= \begin{cases} \cos\left(\frac{\pi}{4}\right) + i \sin\left(\frac{\pi}{4}\right) & k = 0 \\ \cos\left(\frac{\pi}{4} + \frac{\pi}{2}\right) + i \sin\left(\frac{\pi}{4} + \frac{\pi}{2}\right) & k = 1 \\ \cos\left(\frac{\pi}{4} + \pi\right) + i \sin\left(\frac{\pi}{4} + \pi\right) & k = 2 \\ \cos\left(\frac{\pi}{4} + \frac{3}{2}\pi\right) + i \sin\left(\frac{\pi}{4} + \frac{3}{2}\pi\right) & k = 3 \end{cases}$$

$$= \begin{cases} w_1 = (1+i)/\sqrt{2}, & k = 0 \\ w_2 = (-1+i)/\sqrt{2}, & k = 1 \\ w_3 = (-1-i)/\sqrt{2}, & k = 2 \\ w_4 = (1-i)/\sqrt{2}, & k = 3 \end{cases}$$

The roots in the complex plane are depicted at Fig. 13.2.

13.6.3 Complex exponential

Definition 13.18. (Euler's formula) *If $z = x + iy$ is a complex number, we define the* ***complex exponent*** $e^z = e^{x+iy}$ *to be the complex number*

$$\boxed{e^z = e^x (\cos y + i \sin y)} \tag{13.39}$$

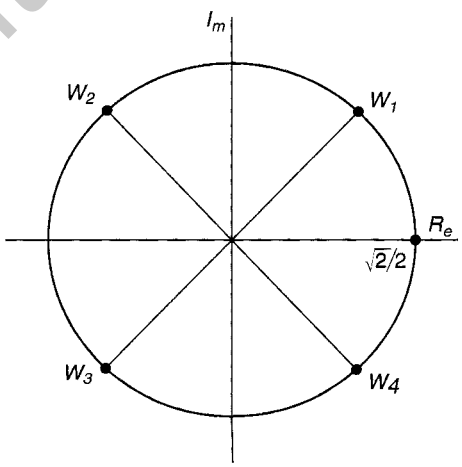

Fig. 13.2. The roots of $\sqrt[4]{-1}$.

Evidently, the complex exponent e^z possesses the following properties which can be easily proven using only the definition (13.39).

Proposition 13.8. *For any complex numbers z, z_1 and z_2*

1.

$$\boxed{e^{z_1+z_2} = e^{z_1} e^{z_2}}$$

2.

$$\boxed{\begin{gathered} e^z e^{-z} = e^0 = 1 \\ e^z \neq 0 \end{gathered}}$$

3.

$$\boxed{\left|e^{iy}\right| = 1}$$

4. $e^z = 1$ *if and only if* $z = 2\pi n \cdot i$ *(n is an integer).*
5. $e^{z_1} = e^{z_2}$ *if and only if* $z_1 - z_2 = 2\pi n \cdot i$ *(n is an integer).*
6.

$$\boxed{\begin{gathered} z = |z|\, e^{i \operatorname{Arg} z} = |z|\, e^i \\ \arg z := \operatorname{arctg}(y/x) \end{gathered}} \tag{13.40}$$

13.6.4 Complex logarithms

Definition 13.19. *The number w is called the (natural) logarithm of the complex number $z \neq 0$ (the notation is $w = \operatorname{Ln} z$) if $e^w = z$.*

Putting $w = u + iv$ from the definition above it follows that $z = e^u e^{iv}$. Comparing this with (13.39) implies

$$\begin{aligned} |z| &= e^u \\ v &= \operatorname{Arg} z = \arg z + 2\pi k \end{aligned}$$

So, $u = \ln |z|$ and, thus,

$$\boxed{w = \operatorname{Ln} z = \ln |z| + i \operatorname{Arg} z = \ln |z| + i\,(\arg z + 2\pi k)} \tag{13.41}$$

Formula (13.41) defines an infinite number of complex numbers which are logarithms of the nonzero $z \in \mathbb{C}$. Of these, the particular value corresponding to $k = 0$ is called the *principal value of the complex logarithm* and is denoted by

$$\boxed{\ln z := \ln |z| + i \arg z} \tag{13.42}$$

Example 13.5.

1.

$$\operatorname{Ln} i = \left(\frac{\pi}{2} + 2\pi k\right) i$$
$$\ln i = \frac{\pi}{2} i$$

2.

$$\operatorname{Ln}(-1) = (2k+1)\pi i$$
$$\ln(-1) = \pi i$$

Lemma 13.9. *If* $z_1 z_2 \neq 0$ *then*

$$\boxed{\begin{aligned}\operatorname{Ln}(z_1 z_2) &= \operatorname{Ln} z_1 + \operatorname{Ln} z_1 \\ &= \ln|z_1| + \ln|z_2| + i\left[\arg z_1 + \arg z_1 + 2\pi\,(k_1 + k_2)\right]\end{aligned}}$$

where k_1, k_2 *are integers.*

Proof.

$$\begin{aligned}\operatorname{Ln}(z_1 z_2) &= \ln|z_1 z_2| + i\operatorname{Arg}(z_1 z_2) \\ &= \ln|z_1| + \ln|z_2| + i\left[\operatorname{Arg}(z_1) + \operatorname{Arg}(z_1)\right]\end{aligned}$$

□

13.6.5 Complex sines and cosines

Taking in Euler's formula (13.39) $x = 0$ we have

$$e^{iy} = \cos y + i \sin y$$
$$e^{-iy} = \cos y - i \sin y$$

which implies

$$\cos y = \frac{e^{iy} + e^{-iy}}{2}, \quad \sin y = \frac{e^{iy} - e^{-iy}}{2i}$$

valid for any real $y \in \mathbb{R}$. Extending these formulas to the complex plane $\mathbb{C}$ one may suggest the following definition.

Definition 13.20. *Given a complex number* z*, we define*

$$\boxed{\begin{aligned}\cos z &= \frac{e^{iz} + e^{-iz}}{2} \\ \sin z &= \frac{e^{iz} - e^{-iz}}{2i}\end{aligned}} \tag{13.43}$$

Lemma 13.10. *For* $z = x + iy$

$$\boxed{\begin{aligned}\cos z &= \cos x \cosh y - i \sin x \sinh y \\ \sin z &= \sin x \cosh y + i \cos x \sinh y\end{aligned}} \tag{13.44}$$

where

$$\cosh y := \frac{e^y + e^{-y}}{2}, \quad \sinh y := \frac{e^y - e^{-y}}{2}$$

Proof. The result follows from the identities

$$\begin{aligned}2\cos z = e^{iz} + e^{-iz} &= e^{-y+ix} + e^{y-ix} \\ &= e^{-y}[\cos x + i \sin x] + e^{y}[\cos x - i \sin x] \\ &= \cos x \left(e^y + e^{-y}\right) - i \sin x \left(e^y - e^{-y}\right)\end{aligned}$$

which gives the first representation in (13.44). The proof for $\sin z$ is similar. □

Exercise 13.1.

1. *Defining*

$$\boxed{\tan z := \frac{\sin z}{\cos z}} \tag{13.45}$$

it is easy to show by direct calculation that

$$\boxed{\tan z = \frac{\sin 2x + i \sinh 2y}{\cos 2x + \cosh 2y}} \tag{13.46}$$

2. *For any complex* z *and* $n = 1, 2, \ldots$

$$\boxed{z^n - 1 = \prod_{k=1}^{n} \left(z - e^{2\pi k i/n}\right)} \tag{13.47}$$

14 Sets, Functions and Metric Spaces

Contents

14.1 Functions and sets

14.1.1 The function concept

Definition 14.1. *Let us consider two sets $\mathcal{A}$ and $\mathcal{B}$ whose elements may be any objects whatsoever. Suppose that with each element $x \in \mathcal{A}$ there is associated, in some manner, an element $y \in \mathcal{B}$ which we denote by $y = f(x)$.*

1. *Then f is said to be a **function** from $\mathcal{A}$ to $\mathcal{B}$ or a **mapping** of $\mathcal{A}$ into $\mathcal{B}$.*
2. *If $\mathcal{E} \subset \mathcal{A}$ then $f(\mathcal{E})$ is defined to be the set of all elements $f(x)$, $x \in \mathcal{E}$ and it is called the image of $\mathcal{E}$ under f. The notation $f(\mathcal{A})$ is called the **range** of f (evidently, $f(\mathcal{A}) \subseteq \mathcal{B}$). If $f(\mathcal{A}) = \mathcal{B}$ we say that f maps $\mathcal{A}$ onto $\mathcal{B}$.*
3. *For $\mathcal{D} \subset \mathcal{B}$ the notation $f^{-1}(\mathcal{D})$ denotes the set of all $x \in \mathcal{A}$ such that $f(x) \in \mathcal{B}$. We call $f^{-1}(\mathcal{D})$ the **inverse image** of $\mathcal{D}$ under f. So, if $y \in \mathcal{D}$ then $f^{-1}(y)$ is the set of all $x \in \mathcal{A}$ such that $f(x) = y$. If for each $y \in \mathcal{B}$ the set $f^{-1}(y)$ consists of at most one element of $\mathcal{A}$ then f is said to be **one-to-one mapping** of $\mathcal{A}$ to $\mathcal{B}$.*

The one-to-one mapping f means that $f(x_1) \neq f(x_2)$ if $x_1 \neq x_2$ for any $x_1, x_2 \in \mathcal{A}$. We will often use the following notation for the mapping f:

$$\boxed{f : \mathcal{A} \to \mathcal{B}} \tag{14.1}$$

If, in particular, $\mathcal{A} = \mathbb{R}^n$ and $\mathcal{B} = \mathbb{R}^m$ we will write

$$\boxed{f : \mathbb{R}^n \to \mathbb{R}^m} \tag{14.2}$$

Definition 14.2. *If for two sets $\mathcal{A}$ and $\mathcal{B}$ there exists a one-to-one mapping then we say that these sets are **equivalent** and we write*

$$\boxed{\mathcal{A} \sim \mathcal{B}} \tag{14.3}$$

Claim 14.1. *The relation of equivalency* ($\sim$) *clearly has the following properties:*

(a) it is ***reflexive****, i.e.,* $\mathcal{A} \sim \mathcal{A}$;
(b) it is ***symmetric****, i.e., if* $\mathcal{A} \sim \mathcal{B}$ *then* $\mathcal{B} \sim \mathcal{A}$;
(c) it is ***transitive****, i.e., if* $\mathcal{A} \sim \mathcal{B}$ *and* $\mathcal{B} \sim \mathcal{C}$ *then* $\mathcal{A} \sim \mathcal{C}$.

14.1.2 Finite, countable and uncountable sets

Denote by $\mathcal{J}_n$ the set of positive numbers $1, 2, \ldots, n$, that is,

$$\boxed{\mathcal{J}_n = \{1, 2, \ldots, n\}}$$

and by $\mathcal{J}$ we will denote the set of all positive numbers, namely,

$$\boxed{\mathcal{J} = \{1, 2, \ldots\}}$$

Definition 14.3. *For any* $\mathcal{A}$ *we say:*

1. $\mathcal{A}$ *is* ***finite*** *if*

$$\mathcal{A} \sim \mathcal{J}_n$$

for some finite n *(the* ***empty set*** $\varnothing$*, which does not contain any element, is also considered as finite);*

2. $\mathcal{A}$ *is* ***countable*** *(enumerable or denumerable) if*

$$\mathcal{A} \sim \mathcal{J}$$

3. $\mathcal{A}$ *is* ***uncountable*** *if it is neither finite nor countable;*
4. $\mathcal{A}$ *is* ***at most countable*** *if it is both finite or countable.*

Evidently, if $\mathcal{A}$ is infinite then it is equivalent to one of its subsets. Also it is clear that any infinite subset of a countable set is countable.

Definition 14.4. *By* ***a sequence*** *we mean a function* f *defined on the set* $\mathcal{J}$ *of all positive integers. If* $x_n = f(n)$ *it is customary to denote the corresponding sequence by*

$$\boxed{\{x_n\} := \{x_1, x_2, \ldots\}}$$

(sometimes this sequence starts with x_0 *but not with* x_1*).*

Claim 14.2.

1. *The set* $\mathcal{N}$ *of all integers is countable;*
2. *The set* $\mathcal{Q}$ *of all rational numbers is countable;*
3. *The set* $\mathbb{R}$ *of all real numbers is uncountable.*

14.1.3 Algebra of sets

Definition 14.5. *Let $\mathcal{A}$ and Ω be sets. Suppose that with each element $\alpha \in \mathcal{A}$ there is associated a subset $\mathcal{E}_\alpha \subset \Omega$. Then*

(a) The ***union*** *of the sets $\mathcal{E}_\alpha$ is defined to be the set $\mathcal{S}$ such that $x \in S$ if and only if $x \in \mathcal{E}_\alpha$ at least for one $\alpha \in \mathcal{A}$. It will be denoted by*

$$\boxed{\mathcal{S} := \bigcup_{\alpha \in A} \mathcal{E}_\alpha} \tag{14.4}$$

If $\mathcal{A}$ consists of all integers $(1, 2, \ldots, n)$, which means $\mathcal{A} = \mathcal{J}_n$, we will use the notation

$$\boxed{\mathcal{S} := \bigcup_{\alpha=1}^{n} \mathcal{E}_\alpha} \tag{14.5}$$

and if $\mathcal{A}$ consists of all integers $(1, 2, \ldots)$, which means $\mathcal{A} = \mathcal{J}$, we will use the notation

$$\boxed{\mathcal{S} := \bigcup_{\alpha=1}^{\infty} \mathcal{E}_\alpha} \tag{14.6}$$

(b) The ***intersection*** *of the sets $\mathcal{E}_\alpha$ is defined as the set $\mathcal{P}$ such that $x \in \mathcal{P}$ if and only if $x \in \mathcal{E}_\alpha$ for every $\alpha \in \mathcal{A}$. It will be denoted by*

$$\boxed{\mathcal{S} := \bigcap_{\alpha \in \mathcal{A}} \mathcal{E}_\alpha} \tag{14.7}$$

If $\mathcal{A}$ consists of all integers $(1, 2, \ldots, n)$, which means $\mathcal{A} = \mathcal{J}_n$, we will use the notation

$$\boxed{\mathcal{S} := \bigcap_{\alpha=1}^{n} \mathcal{E}_\alpha} \tag{14.8}$$

and if $\mathcal{A}$ consists of all integers $(1, 2, \ldots)$, which means $\mathcal{A} = \mathcal{J}$, we will use the notation

$$\boxed{\mathcal{S} := \bigcap_{\alpha=1}^{\infty} \mathcal{E}_\alpha} \tag{14.9}$$

If for two sets $\mathcal{A}$ and $\mathcal{B}$ we have $\mathcal{A} \cap \mathcal{B} = \varnothing$, we say that these two sets are ***disjoint****.*

(c) The ***complement*** *of $\mathcal{A}$ relative to $\mathcal{B}$, denoted by $\mathcal{B} - \mathcal{A}$, is defined to be the set*

$$\boxed{\mathcal{B} - \mathcal{A} := \{x : x \in \mathcal{B}, \text{ but } x \notin \mathcal{A}\}} \tag{14.10}$$

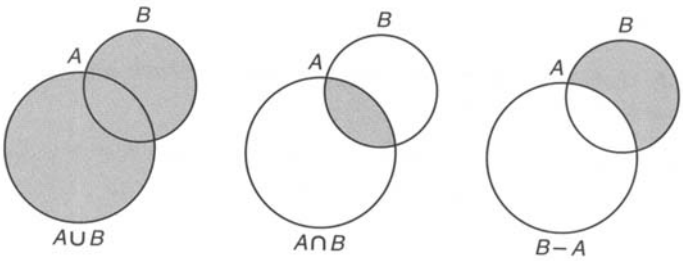

Fig. 14.1. Two sets relations.

The sets $\mathcal{A} \cup \mathcal{B}$, $\mathcal{A} \cap \mathcal{B}$ and $\mathcal{B} - \mathcal{A}$ are illustrated at Fig. 14.1. Using these graphic illustrations it is possible to prove easily the following set-theoretical identities for union and intersection.

Proposition 14.1.

1.

$$\boxed{\mathcal{A} \cup (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cup \mathcal{B}) \cup \mathcal{C}, \quad \mathcal{A} \cap (\mathcal{B} \cap \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cap \mathcal{C}}$$

2.

$$\boxed{\mathcal{A} \cap (\mathcal{B} \cup \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C})}$$

3.

$$\boxed{(\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{A} \cup \mathcal{C}) = \mathcal{A} \cup (\mathcal{B} \cap \mathcal{C})}$$

4.

$$\boxed{(\mathcal{A} \cup \mathcal{B}) \cap (\mathcal{B} \cup \mathcal{C}) \cap (\mathcal{C} \cup \mathcal{A}) = (\mathcal{A} \cap \mathcal{B}) \cup (\mathcal{A} \cap \mathcal{C}) \cup (\mathcal{B} \cap \mathcal{C})}$$

5.

$$\boxed{\mathcal{A} \cap (\mathcal{B} - \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) - (\mathcal{A} \cap \mathcal{C})}$$

6.

$$\boxed{(\mathcal{A} - \mathcal{C}) \cap (\mathcal{B} - \mathcal{C}) = (\mathcal{A} \cap \mathcal{B}) - \mathcal{C}}$$

7.

$$(\mathcal{A} - \mathcal{B}) \cup \mathcal{B} = \mathcal{A}$$

if and only if $\mathcal{B} \subseteq \mathcal{A}$.

8.

$$\boxed{\mathcal{A} \subset \mathcal{A} \cup \mathcal{B}, \quad \mathcal{A} \cap \mathcal{B} \subset \mathcal{A}}$$

9.

$$\boxed{\mathcal{A} \cup \varnothing = \mathcal{A}, \quad \mathcal{A} \cap \varnothing = \varnothing}$$

10.

$$\boxed{\mathcal{A} \cup \mathcal{B} = \mathcal{B}, \quad \mathcal{A} \cap \mathcal{B} = \mathcal{A}}$$

if $\mathcal{A} \subset \mathcal{B}$.

The next relations generalize the previous unions and intersections to arbitrary ones.

Proposition 14.2.

1. Let $f : \mathcal{S} \to \mathcal{T}$ *be a function and* $\mathcal{A}, \mathcal{B}$ *any subsets of* S. *Then*

$$\boxed{f(\mathcal{A} \cup \mathcal{B}) = f(\mathcal{A}) \cup f(\mathcal{B})}$$

2. For any $\mathcal{Y} \subseteq \mathcal{T}$ *define* $f^{-1}(\mathcal{Y})$ *as the largest subset of* S *which* f *maps into* $\mathcal{Y}$. *Then*
(a)

$$\boxed{\mathcal{X} \subseteq f^{-1}(f(\mathcal{X}))}$$

(b)

$$\boxed{f\left(f^{-1}(\mathcal{Y})\right) \subseteq \mathcal{Y}}$$

and

$$\boxed{f\left(f^{-1}(\mathcal{Y})\right) = \mathcal{Y}}$$

if and only if $\mathcal{T} = f(\mathcal{S})$.
(c)

$$\boxed{f^{-1}(\mathcal{Y}_1 \cup \mathcal{Y}_2) = f^{-1}(\mathcal{Y}_1) \cup f^{-1}(\mathcal{Y}_2)}$$

(d)

$$\boxed{f^{-1}(\mathcal{Y}_1 \cap \mathcal{Y}_2) = f^{-1}(\mathcal{Y}_1) \cap f^{-1}(\mathcal{Y}_2)}$$

(e)

$$\boxed{f^{-1}(\mathcal{T} - \mathcal{Y}) = \mathcal{S} - f^{-1}(\mathcal{Y})}$$

and for subsets $\mathcal{B} \subseteq \mathcal{A} \subseteq \mathcal{S}$ *it follows that*

$$\boxed{f(\mathcal{A} - \mathcal{B}) = f(\mathcal{A}) - f(\mathcal{B})}$$

14.2 Metric spaces

14.2.1 Metric definition and examples of metrics

Definition 14.6. *A set $\mathcal{X}$, whose elements we shall call points, is said to be a* ***metric space*** *if with any two points p and q of $\mathcal{X}$ there is associated a real number $d(p,q)$, called a* ***distance*** *between p and q, such that*
(a)

$$\boxed{\begin{array}{c} d(p,q) > 0 \quad if \quad p \neq q \\ d(p,p) = 0 \end{array}} \tag{14.11}$$

(b)

$$\boxed{d(p,q) = d(q,p)} \tag{14.12}$$

(c) for any $r \in \mathcal{X}$ the following "triangle inequality" holds:

$$\boxed{d(p,q) \leq d(p,r) + d(r,q)} \tag{14.13}$$

Any function with these properties is called a ***distance function*** *or a* ***metric****.*

Example 14.1. *The following functions are metrics:*

1. *For any p, q from the* ***Euclidean space*** $\mathbb{R}^n$
   (a) the ***Euclidean metric****:*

$$\boxed{d(p,q) = \|p - q\|} \tag{14.14}$$

   (b) the ***discrete metric****:*

$$\boxed{d(p,q) = \begin{cases} 0 & if \quad p = q \\ 1 & if \quad p \neq q \end{cases}} \tag{14.15}$$

   (c) the ***weighted metric****:*

$$\boxed{\begin{array}{c} d(p,q) = \|p - q\|_Q := \sqrt{(p-q)^\top Q (p-q)} \\ Q = Q^\top > 0 \end{array}} \tag{14.16}$$

   (d) the ***module metric****:*

$$\boxed{d(p,q) = \sum_{i=1}^{n} |p_i - q_i|} \tag{14.17}$$

   (e) the ***Chebyshev's metric****:*

$$\boxed{d(p,q) = \max\{|p_1 - q_1|, \ldots, |p_n - q_n|\}} \tag{14.18}$$

(f) the ***Prokhorov's metric****:*

$$\boxed{d(p,q) = \frac{\|p-q\|}{1+\|p-q\|} \in [0,1)} \tag{14.19}$$

2. For any z_1 and z_2 of the ***complex plane*** $\mathbb{C}$

$$\boxed{\begin{aligned} d(z_1,z_2) &= |z_1 - z_2| \\ &= \sqrt{(\mathrm{Re}\,(z_1-z_2))^2 + (\mathrm{Im}\,(z_1-z_2))^2} \end{aligned}} \tag{14.20}$$

14.2.2 Set structures

Let $\mathcal{X}$ be a metric space. All points and sets mentioned below will be understood to be elements and subsets of $\mathcal{X}$.

Definition 14.7.

(a) A ***neighborhood*** *of a point x is a set $\mathcal{N}_r(x)$ consisting of all points y such that $d(x,y) < r$ where the number r is called the radius of $\mathcal{N}_r(x)$, that is,*

$$\boxed{\mathcal{N}_r(x) := \{y \in \mathcal{X} : d(x,y) < r\}} \tag{14.21}$$

(b) A point $x \in \mathcal{X}$ is a ***limit point*** *of the set $\mathcal{E} \subset \mathcal{X}$ if every neighborhood of x contains a point $y \neq x$ such that $y \in \mathcal{E}$.*
(c) If $x \in \mathcal{E}$ and x is not a limit point of $\mathcal{E}$ then x is called an ***isolated point*** *of $\mathcal{E}$.*
(d) $\mathcal{E} \subset \mathcal{X}$ is ***closed*** *if every limit of elements from $\mathcal{E}$ is a point of $\mathcal{E}$.*
(e) A point $x \in \mathcal{E}$ is an ***interior point*** *of $\mathcal{E}$ if there is a neighborhood of $\mathcal{N}_r(x)$ of x such that $\mathcal{N}_r(x) \subset \mathcal{E}$.*
(f) $\mathcal{E}$ is ***open*** *if every point of $\mathcal{E}$ is an interior point of $\mathcal{E}$.*
(g) The ***complement*** *$\mathcal{E}^c$ of $\mathcal{E}$ is the set of all points $x \in \mathcal{X}$ such that $x \notin \mathcal{E}$.*
(h) $\mathcal{E}$ is ***bounded*** *if there exist a real number M and a point $x \in \mathcal{E}$ such that $d(x,y) < M$ for all $y \in \mathcal{E}$.*
(i) $\mathcal{E}$ is ***dense*** *in $\mathcal{X}$ if every point $x \in \mathcal{X}$ is a limit point of $\mathcal{E}$, or a point of $\mathcal{E}$, or both.*
(j) $\mathcal{E}$ is ***connected*** *in $\mathcal{X}$ if it is not a union of two nonempty* ***separated sets****, that is, $\mathcal{E}$ cannot be represented as $\mathcal{E} = \mathcal{A} \cup \mathcal{B}$ where $\mathcal{A} \neq \varnothing$, $\mathcal{B} \neq \varnothing$ and $\mathcal{A} \cap \mathcal{B} = \varnothing$.*

Example 14.2. *The set $J_{open}(p)$ defined as*

$$\boxed{J_{open} := \{x \in \mathcal{X},\ d(x,p) < r\}}$$

is an open set but the set $J_{closed}(p)$ defined as

$$\boxed{J_{closed}(p) := \{x \in \mathcal{X},\ d(x,p) \leq r\}}$$

is closed.

The following claim seems to be evident and that is why they are given without proofs.

Claim 14.3.

1. *Every neighborhood* $\mathcal{N}_r(x) \subset \mathcal{E}$ *is an open set.*
2. *If* x *is a limit point of* $\mathcal{E}$ *then every neighborhood* $\mathcal{N}_r(x) \subset \mathcal{E}$ *contains infinitely many points of* $\mathcal{E}$.
3. *A finite point set has no limit points.*

Let us prove the following lemma concerning complement sets.

Lemma 14.1. *Let* $\{\mathcal{E}_\alpha\}$ *be a collection (finite or infinite) of sets* $\mathcal{E}_\alpha \subseteq \mathcal{X}$. *Then*

$$\boxed{\left(\bigcup_\alpha \mathcal{E}_\alpha\right)^c = \bigcap_\alpha \mathcal{E}_\alpha^c} \tag{14.22}$$

Proof. If $x \in \left(\bigcup_\alpha \mathcal{E}_\alpha\right)^c$ then, evidently, $x \notin \bigcup_\alpha \mathcal{E}_\alpha$ and, hence, $x \notin \mathcal{E}_\alpha$ for any α. This means that $x \in \bigcap_\alpha \mathcal{E}_\alpha^c$. Thus,

$$\left(\bigcup_\alpha \mathcal{E}_\alpha\right)^c \subseteq \bigcap_\alpha \mathcal{E}_\alpha^c \tag{14.23}$$

Conversely, if $x \in \bigcap_\alpha \mathcal{E}_\alpha^c$ then $x \in \mathcal{E}_\alpha^c$ for every α and, hence, $x \notin \bigcup_\alpha \mathcal{E}_\alpha$. So, $x \in \left(\bigcup_\alpha \mathcal{E}_\alpha\right)^c$ which implies

$$\bigcap_\alpha \mathcal{E}_\alpha^c \subseteq \left(\bigcup_\alpha \mathcal{E}_\alpha\right)^c \tag{14.24}$$

Combining (14.23) and (14.24) gives (14.22). Lemma is proven. □

This lemma provides the following corollaries.

Corollary 14.1.

(a) A set $\mathcal{E}$ *is open if and only if its complement* $\mathcal{E}^c$ *is closed.*
(b) A set $\mathcal{E}$ *is closed if and only if its complement* $\mathcal{E}^c$ *is open.*
(c) For any collection $\{\mathcal{E}_\alpha\}$ *of open sets* $\mathcal{E}_\alpha$ *the set* $\bigcup_\alpha \mathcal{E}_\alpha$ *is open.*
(d) For any collection $\{\mathcal{E}_\alpha\}$ *of closed sets* $\mathcal{E}_\alpha$ *the set* $\bigcap_\alpha \mathcal{E}_\alpha^c$ *is closed.*
(e) For any finite collection $\{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$ *of open sets* $\mathcal{E}_\alpha$ *the set* $\bigcap_\alpha \mathcal{E}_\alpha^c$ *is open too.*
(f) For any finite collection $\{\mathcal{E}_1, \ldots, \mathcal{E}_n\}$ *of closed sets* $\mathcal{E}_\alpha$ *the set* $\bigcup_\alpha \mathcal{E}_\alpha$ *is closed too.*

Definition 14.8. *Let* $\mathcal{X}$ *be a metric space and* $\mathcal{E} \subset \mathcal{X}$. *Denote by* $\mathcal{E}'$ *the set of all limit points of* $\mathcal{E}$. *Then the set* $\mathrm{cl}\,\mathcal{E}$ *defined as*

$$\boxed{\mathrm{cl}\,\mathcal{E} := \mathcal{E} \cup \mathcal{E}'} \tag{14.25}$$

is called the ***closure*** *of* $\mathcal{E}$.

The next properties seem to be logical consequences of this definition.

Proposition 14.3. *If $\mathcal{X}$ is a metric space and $\mathcal{E} \subset \mathcal{X}$, then*

(a) $\operatorname{cl}\mathcal{E}$ *is closed;*
(b) $\mathcal{E} = \operatorname{cl}\mathcal{E}$ *if and only if* $\mathcal{E}$ *is closed;*
(c) $\operatorname{cl}\mathcal{E} \subset \mathcal{P}$ *for every closed set* $\mathcal{P} \subset \mathcal{X}$ *such that* $\mathcal{E} \subset \mathcal{P}$*;*
(d) If $\mathcal{E}$ *is a nonempty set of real numbers which is bounded above, i.e.,* $\varnothing \neq \mathcal{E} \subset \mathbb{R}$ *and* $y := \sup\mathcal{E} < \infty$*. Then* $y \in \operatorname{cl}\mathcal{E}$ *and, hence,* $y \in \mathcal{E}$ *if* $\mathcal{E}$ *is closed.*

Proof.

(a) If $x \in \mathcal{X}$ and $x \notin \operatorname{cl}\mathcal{E}$ then x is neither a point of $\mathcal{E}$ nor a limit point of $\mathcal{E}$. Hence x has a neighborhood which does not intersect $\mathcal{E}$. Therefore the complement $\mathcal{E}^c$ of $\mathcal{E}$ is an open set. So, $\operatorname{cl}\mathcal{E}$ is closed.
(b) If $\mathcal{E} = \operatorname{cl}\mathcal{E}$ then by (a) it follows that $\mathcal{E}$ is closed. If $\mathcal{E}$ is closed then for $\mathcal{E}'$, defined in (14.8), we have that $\mathcal{E}' \subset \mathcal{E}$. Hence, $\mathcal{E} = \operatorname{cl}\mathcal{E}$.
(c) $\mathcal{P}$ is closed and $\mathcal{P} \supset \mathcal{E}$ (defined in (14.8)) then $\mathcal{P} \supset \mathcal{P}'$ and, hence, $\mathcal{P} \supset \mathcal{E}'$. Thus $\mathcal{P} \supset \operatorname{cl}\mathcal{E}$.
(d) If $y \in \mathcal{E}$ then $y \in \operatorname{cl}\mathcal{E}$. Assume $y \notin \mathcal{E}$. Then for any $\varepsilon > 0$ there exists a point $x \in \mathcal{E}$ such that $y - \varepsilon < x < y$, otherwise $(y - \varepsilon)$ would be an upper bound of $\mathcal{E}$ that contradicts the supposition $\sup\mathcal{E} = y$. Thus y is a limit point of $\mathcal{E}$. Hence, $y \in \operatorname{cl}\mathcal{E}$.

The proposition is proven. □

Definition 14.9. *Let $\mathcal{E}$ be a set of a metric space $\mathcal{X}$. A point $x \in \mathcal{E}$ is called a* ***boundary point*** *of $\mathcal{E}$ if any neighborhood $\mathcal{N}_r(x)$ of this point contains at least one point of $\mathcal{E}$ and at least one point of $\mathcal{X} - \mathcal{E}$. The set of all boundary points of $\mathcal{E}$ is called the* ***boundary of the set*** *$\mathcal{E}$ and is denoted by $\partial\mathcal{E}$.*

It is not difficult to verify that

$$\boxed{\partial\mathcal{E} = \operatorname{cl}\mathcal{E} \cap \operatorname{cl}(\mathcal{X} - \mathcal{E})} \tag{14.26}$$

Denoting by

$$\boxed{\operatorname{int}\mathcal{E} := \mathcal{E} - \partial\mathcal{E}} \tag{14.27}$$

the set of all internal points of the set $\mathcal{E}$, it is easily verified that

$$\boxed{\begin{array}{c} \operatorname{int}\mathcal{E} = \mathcal{X} - \operatorname{cl}(\mathcal{X} - \mathcal{E}) \\ \operatorname{int}(\mathcal{X} - \mathcal{E}) = \mathcal{X} - \operatorname{cl}\mathcal{E} \\ \operatorname{int}(\operatorname{int}\mathcal{E}) = \operatorname{int}\mathcal{E} \\ \text{If} \quad \operatorname{cl}\mathcal{E} \cap \operatorname{cl}\mathcal{D} = \varnothing \quad \text{then} \quad \partial(\mathcal{E} \cup \mathcal{D}) = \partial\mathcal{E} \cup \partial\mathcal{D} \end{array}} \tag{14.28}$$

14.2.3 Compact sets

Definition 14.10.

1. *By an* ***open cover of a set*** $\mathcal{E}$ *in a metric space* $\mathcal{X}$ *we mean a collection* $\{\mathcal{G}_\alpha\}$ *of open subsets of* $\mathcal{X}$ *such that*

$$\boxed{\mathcal{E} \subset \bigcup_{\alpha} \mathcal{G}_\alpha} \tag{14.29}$$

2. *A subset* $\mathcal{K}$ *of a metric space* $\mathcal{X}$ *is said to be* ***compact*** *if every open cover of* $\mathcal{K}$ *contains a finite subcover; more exactly, there are a finite number of indices* $\alpha_1, \ldots, \alpha_n$ *such that*

$$\boxed{\mathcal{K} \subset \mathcal{G}_{\alpha_1} \cup \cdots \cup \mathcal{G}_{\alpha_n}} \tag{14.30}$$

Remark 14.1. *Evidently, every finite set is compact.*

Theorem 14.1. *A set* $\mathcal{K} \subset \mathcal{Y} \subset \mathcal{X}$ *is a compact relative to* $\mathcal{X}$ *if and only if* $\mathcal{K}$ *is a compact relative to* $\mathcal{Y}$.

Proof. Necessity. Suppose $\mathcal{K}$ is a compact relative to $\mathcal{X}$. Hence, by the definition (14.30) there exists its finite subcover such that

$$\mathcal{K} \subset \mathcal{G}_{\alpha_1} \cup \cdots \cup \mathcal{G}_{\alpha_n} \tag{14.31}$$

where $\mathcal{G}_{\alpha_i}$ is an open set with respect to $\mathcal{X}$. On the other hand $\mathcal{K} \subset \bigcup_{\alpha} \mathcal{V}_\alpha$ where $\{\mathcal{V}_\alpha\}$ is a collection of sets open with respect to $\mathcal{Y}$. But any open set $\mathcal{V}_\alpha$ can be represented as $\mathcal{V}_\alpha = \mathcal{Y} \cap \mathcal{G}_\alpha$. So, (14.31) implies

$$\mathcal{K} \subset \mathcal{V}_{\alpha_1} \cup \cdots \cup \mathcal{V}_{\alpha_n} \tag{14.32}$$

Sufficiency. Conversely, if $\mathcal{K}$ is a compact relative to $\mathcal{Y}$ then there exists a finite collection $\{\mathcal{V}_\alpha\}$ of open sets in $\mathcal{Y}$ such that (14.32) holds. Putting $\mathcal{V}_\alpha = \mathcal{Y} \cap \mathcal{G}_\alpha$ for a special choice of indices $\alpha_1, \ldots, \alpha_n$ it follows that $\mathcal{V}_\alpha \subset \mathcal{G}_\alpha$ which implies (14.31). Theorem is proven. □

Theorem 14.2. *Compact sets of metric spaces are closed.*

Proof. Suppose $\mathcal{K}$ is a compact subset of a metric space $\mathcal{X}$. Let $x \in \mathcal{X}$ but $x \notin \mathcal{K}$ and $y \in \mathcal{K}$. Consider the neighborhoods $\mathcal{N}_r(x)\, \mathcal{N}_r(y)$ of these points with $r < \frac{1}{2} d(x, y)$. Since $\mathcal{K}$ is a compact there are finitely many points $y_1, \ldots, y_n$ such that

$$\mathcal{K} \subset \mathcal{N}_r(y_1) \cup \cdots \cup \mathcal{N}_r(y_n) = \mathcal{N}$$

If $\mathcal{V} = \mathcal{N}_{r_1}(x) \cap \cdots \cap \mathcal{N}_{r_n}(x)$, then evidently $\mathcal{V}$ is a neighborhood of x which does not intersect $\mathcal{N}$ and, hence, $\mathcal{V} \subset \mathcal{K}^c$. So, x is an interior point of $\mathcal{K}^c$. Theorem is proven. □

The following two propositions seem to be evident.

Proposition 14.4.

1. *Closed subsets of compact sets are compacts too.*
2. *If $\mathcal{F}$ is closed and $\mathcal{K}$ is compact then $\mathcal{F} \cap \mathcal{K}$ is compact.*

Theorem 14.3. *If $\mathcal{E}$ is an infinite subset of a compact set $\mathcal{K}$ then $\mathcal{E}$ has a limit point in $\mathcal{K}$.*

Proof. If no point of $\mathcal{K}$ were a limit point of $\mathcal{E}$ then $y \in \mathcal{K}$ would have a neighborhood $\mathcal{N}_r(y)$ which contains at most one point of $\mathcal{E}$ (namely, y if $y \in \mathcal{E}$). It is clear that no finite subcollection $\{\mathcal{N}_{r_k}(y)\}$ can cover $\mathcal{E}$. The same is true of $\mathcal{K}$ since $\mathcal{E} \subset \mathcal{K}$. But this contradicts the compactness of $\mathcal{K}$. Theorem is proven. □

The next theorem explains the compactness property especially in $\mathbb{R}^n$ and is often applied in a control theory analysis.

Theorem 14.4. *If a set $\mathcal{E} \subset \mathbb{R}^n$ then the following three properties are equivalent:*

(a) $\mathcal{E}$ is closed and bounded.
(b) $\mathcal{E}$ is compact.
(c) Every infinite subset of $\mathcal{E}$ has a limit point in $\mathcal{E}$.

Proof. It is the consequence of all previous theorems and propositions and left for readers' consideration. The details of the proof can be found in Chapter 2 of Rudin (1976). □

Remark 14.2. *Notice that properties (b) and (c) are equivalent in any metric space, but (a) is not.*

14.2.4 Convergent sequences in metric spaces

14.2.4.1 Convergence

Definition 14.11. *A sequence $\{x_n\}$ in a metric space $\mathcal{X}$ is said to* ***converge*** *if there is a point $x \in \mathcal{X}$ which for any $\varepsilon > 0$ there exists an integer n_ε such that $n \geq n_\varepsilon$ implies that $d(x_n, x) < \varepsilon$. Here $d(x_n, x)$ is the metric (distance) in $\mathcal{X}$. In this case we say that $\{x_n\}$ converges to x, or that x is a limit of $\{x_n\}$, and we write*

$$\boxed{\lim_{n\to\infty} x_n = x \quad or \quad x_n \underset{n\to\infty}{\to} x} \tag{14.33}$$

If $\{x_n\}$ does not converge, it is usually said to ***diverge****.*

Example 14.3. *The sequence $\{1/n\}$ converges to 0 in $\mathbb{R}$, but fails to converge in $\mathbb{R}_+ := \{x \in \mathbb{R} \mid x > 0\}$.*

Theorem 14.5. *Let $\{x_n\}$ be a sequence in a metric space $\mathcal{X}$.*

1. *$\{x_n\}$ converges to $x \in \mathcal{X}$ if and only if every neighborhood $\mathcal{N}_\varepsilon(x)$ of x contains all but (excluding) finitely many of the terms of $\{x_n\}$.*
2. *If $x', x'' \in \mathcal{X}$ and*

$$\boxed{x_n \underset{n\to\infty}{\to} x' \quad and \quad x_n \underset{n\to\infty}{\to} x''}$$

then

$$\boxed{x' = x''}$$

3. *If* $\{x_n\}$ *converges then* $\{x_n\}$ *is bounded.*
4. *If* $\mathcal{E} \subset \mathcal{X}$ *and* x *is a limit point of* $\mathcal{E}$ *then there is a sequence* $\{x_n\}$ *in* $\mathcal{E}$ *such that* $x = \lim_{n\to\infty} x_n$.

Proof.

1. (a) *Necessity.* Suppose $x_n \underset{n\to\infty}{\to} x$ and let $\mathcal{N}_\varepsilon(x)$ (for some $\varepsilon > 0$) be a neighborhood of x. The conditions $d(y, x) < \varepsilon$, $y \in \mathcal{X}$ imply $y \in \mathcal{N}_\varepsilon(x)$. Corresponding to this ε there exists a number n_ε such that for any $n \geq n_\varepsilon$ it follows that $d(x_n, x) < \varepsilon$. Thus, $x_n \in \mathcal{N}_\varepsilon(x)$. So, all x_n are bounded.
   (b) *Sufficiency.* Conversely, suppose every neighborhood of x contains all but finitely many of the terms of $\{x_n\}$. Fixing $\varepsilon > 0$ denoting by $\mathcal{N}_\varepsilon(x)$ the set of all $y \in \mathcal{X}$ such that $d(y, x) < \varepsilon$. By assumption there exists n_ε such that for any $n \geq n_\varepsilon$ it follows that $x_n \in \mathcal{N}_\varepsilon(x)$. Thus $d(x_n, x) < \varepsilon$ if $n \geq n_\varepsilon$ and, hence, $x_n \underset{n\to\infty}{\to} x$.
2. For the given $\varepsilon > 0$ there exist integers n' and n'' such that $n \geq n'$ implies $d(x_n, x') < \varepsilon/2$ and $n \geq n''$ implies $d(x_n, x'') < \varepsilon/2$. So, for $n \geq \max\{n', n''\}$ it follows $d(x', x'') \leq d(x', x_n) + d(x_n, x'') < \varepsilon$. Taking ε small enough we conclude that $d(x', x'') = 0$.
3. Suppose $x_n \underset{n\to\infty}{\to} x$. Then, evidently there exists an integer n_0 such that for all $n \geq n_0$ we have $d(x_n, x) < 1$. Define $r := \max\{1, d(x_1, x), \ldots, d(x_{n_0}, x)\}$. Then $d(x_n, x) < r$ for all $n = 1, 2, \ldots$.
4. For any integer $n = 1, 2, \ldots$ there exists a point $x_n \in \mathcal{E}$ such that $d(x_n, x) < 1/n$. For any given $\varepsilon > 0$ define n_ε such that $\varepsilon n_\varepsilon > 1$. Then for $n \geq n_\varepsilon$ one has $d(x_n, x) < 1/n < \varepsilon$ which means that $x_n \underset{n\to\infty}{\to} x$.

This completes the proof. □

14.2.4.2 Subsequences

Definition 14.12. *Given a sequence* $\{x_n\}$ *let us consider a sequence* $\{n_k\}$ *of positive integers satisfying* $n_1 < n_2 < \cdots$. *Then the sequence* $\{x_{n_k}\}$ *is called a* ***subsequence*** *of* $\{x_n\}$.

Claim 14.4. *If a sequence* $\{x_n\}$ *converges to* x *then any subsequence* $\{x_{n_k}\}$ *of* $\{x_n\}$ *converges to the same limit point* x.

Proof. This result can be easily proven by contradiction. Indeed, assuming that two different subsequences $\{x_{n_k}\}$ and $\{x_{n_j}\}$ have different limit points x' and x'', it follows that there exist $0 < \varepsilon < d(x', x'')$ and a number k_ε such that for all $k \geq k_\varepsilon$ we shall have: $d(x_{n_k}, x_{n_j}) > \varepsilon$ which is in contradiction with the assumption that $\{x_n\}$ converges. □

Theorem 14.6.

(a) If $\{x_n\}$ *is a sequence in a compact metric space* $\mathcal{X}$ *then it contains some subsequence* $\{x_{n_k}\}$ *convergent to a point of* $\mathcal{X}$.
(b) Any bounded sequence in $\mathbb{R}^n$ *contains a convergent subsequence.*

Proof.

(a) Let $\mathcal{E}$ be the range of $\{x_n\}$. If $\{x_n\}$ converges then the desired subsequence is this sequence itself. Suppose that $\{x_n\}$ diverges. If $\mathcal{E}$ is finite then there is a point $x \in \mathcal{E}$ and numbers $n_1 < n_2 < \cdots$ such that $x_{n_2} = x_{n_2} = \cdots = x$. The subsequence $\{x_{n_k}\}$ so obtained converges evidently to x. If $\mathcal{E}$ is infinite then by Theorem (14.3) $\mathcal{E}$ has a limit point $x \in \mathcal{X}$. Choose n_1 so that $d(x_{n_1}, x) < 1$, and, hence, there are integers $n_i > n_{i-1}$ such that $d(x_{n_i}, x) < 1/i$. This means that x_{n_i} converges to x.

(b) This follows from (a) since Theorem (14.4) implies that every bounded subset of $\mathbb{R}^n$ lies in a compact subset of $\mathbb{R}^n$.

Theorem is proven. □

14.2.4.3 Cauchy sequences

Definition 14.13. *A sequence $\{x_n\}$ in a metric space $\mathcal{X}$ is said to be a* ***Cauchy (fundamental) sequence*** *if for every $\varepsilon > 0$ there is an integer n_ε such that $d(x_n, x_m) < \varepsilon$ if both $n \geq n_\varepsilon$ and $m \geq n_\varepsilon$.*

Defining the *diameter* of $\mathcal{E}$ as

$$\boxed{\operatorname{diam} \mathcal{E} := \sup_{x,y \in \mathcal{E}} d(x, y)} \tag{14.34}$$

one may conclude that if $\mathcal{E}_{n_\varepsilon}$ consists of the points $\{x_{n_\varepsilon}, x_{n_\varepsilon+1}, \ldots\}$ then $\{x_n\}$ is a Cauchy sequence if and only if

$$\boxed{\lim_{n_\varepsilon \to \infty} \operatorname{diam} \mathcal{E} = 0} \tag{14.35}$$

Theorem 14.7.

(a) If $\operatorname{cl} \mathcal{E}$ is the closure of a set $\mathcal{E}$ in a metric space $\mathcal{X}$ then

$$\boxed{\operatorname{diam} \mathcal{E} = \operatorname{diam} \operatorname{cl} \mathcal{E}} \tag{14.36}$$

(b) If $\{\mathcal{K}_n\}$ is a sequence of compact sets in $\mathcal{X}$ such that $\mathcal{K}_n \supset \mathcal{K}_{n-1}$ $(n = 2, 3, \ldots)$ then the set $\mathcal{K} := \bigcap_{n=1}^{\infty} \mathcal{K}_n$ consists exactly of one point.

Proof.

(a) Since $\mathcal{E} \subseteq \operatorname{cl} \mathcal{E}$ it follows that

$$\operatorname{diam} \mathcal{E} \leq \operatorname{diam} \operatorname{cl} \mathcal{E} \tag{14.37}$$

Fix $\varepsilon > 0$ and select $x, y \in \operatorname{cl} \mathcal{E}$. By definition (14.25) there are two points $x', y' \in \mathcal{E}$ such that both $d(x, x') < \varepsilon$ and $d(y, y') < \varepsilon$ which implies

$$\begin{aligned} d(x, y) &\leq d(x, x') + d(x', y') + d(y', y) \\ &< 2\varepsilon + d(x', y') \leq 2\varepsilon + \operatorname{diam} \mathcal{E} \end{aligned}$$

As a result, we have

$$\operatorname{diam} \operatorname{cl} \mathcal{E} \leq 2\varepsilon + \operatorname{diam} \mathcal{E}$$

and since ε is arbitrary it follows that

$$\operatorname{diam} \operatorname{cl} \mathcal{E} \leq \operatorname{diam} \mathcal{E} \tag{14.38}$$

The inequalities (14.37) and (14.38) give (14.36).

(b) If $\mathcal{K}$ contains more than one point then diam $\mathcal{K} > 0$. But for each n we have that $\mathcal{K}_n \supset \mathcal{K}$, so that diam $\mathcal{K}_n \geq$ diam $\mathcal{K}$. This contradicts that diam $\mathcal{K}_n \underset{n\to\infty}{\to} 0$.

Theorem is proven. □

The next theorem explains the importance of fundamental sequence in the analysis of metric spaces.

Theorem 14.8.

(a) Every convergent sequence $\{x_n\}$ given in a metric space $\mathcal{X}$ is a Cauchy sequence.
(b) If $\mathcal{X}$ is a compact metric space and if $\{x_n\}$ is a Cauchy sequence in $\mathcal{X}$ then $\{x_n\}$ converges to some point in $\mathcal{X}$.
(c) In $\mathbb{R}^n$ a sequence converges if and only if it is a Cauchy sequence.

Usually, claim *(c)* is referred to as the **Cauchy criterion**.

Proof.

(a) If $x_n \to x$ then for any $\varepsilon > 0$ there exists an integer n_ε such that $d(x_n, x) < \varepsilon$ for all $n \geq n_\varepsilon$. So, $d(x_n, x_m) \leq d(x_n, x) + d(x, x_m) < 2\varepsilon$ if $n, m \geq n_\varepsilon$. Thus $\{x_n\}$ is a Cauchy sequence.

(b) Let $\{x_n\}$ be a Cauchy sequence and the set $\mathcal{E}_{n_\varepsilon}$ contains the points $x_{n_\varepsilon}, x_{n_\varepsilon+1}, x_{n_\varepsilon+2}, \ldots$ Then by Theorem (14.7) and in view of (14.35) and (14.36)

$$\lim_{n_\varepsilon\to\infty} \operatorname{diam} \operatorname{cl} \mathcal{E}_{n_\varepsilon} = \lim_{n_\varepsilon\to\infty} \operatorname{diam} \mathcal{E}_{n_\varepsilon} = 0 \tag{14.39}$$

Being a closed subset of the compact space $\mathcal{X}$ each cl $\mathcal{E}_{n_\varepsilon}$ is compact (see Proposition 14.4). And since $\mathcal{E}_n \supset \mathcal{E}_{n+1}$ then cl $\mathcal{E}_n \supset$ cl $\mathcal{E}_{n+1}$. By Theorem (14.7b), there is a unique point $x \in \mathcal{X}$ which lies in cl $\mathcal{E}_n$. The expression (14.39) means that for any $\varepsilon > 0$ there exists an integer n_ε such that diam cl $\mathcal{E}_n < \varepsilon$ if $n \geq n_\varepsilon$. Since $x \in$ cl $\mathcal{E}_n$ then $d(x, y) < \varepsilon$ for any $y \in$ cl $\mathcal{E}_n$ which is equivalent to the following: $d(x, x_n) < \varepsilon$ if $n \geq n_\varepsilon$. But this means that $x_n \to x$.

(c) Let $\{x_n\}$ be a Cauchy sequence in $\mathbb{R}^n$ and define $\mathcal{E}_{n_\varepsilon}$ as in statement (b) but with $\mathbf{x}_n \in \mathbb{R}^n$ instead of x_n. For some n_ε we have that diam $\mathcal{E}_{n_\varepsilon} < 1$. The range of $\{\mathbf{x}_n\}$ is the union of $\mathcal{E}_n$ and the finite set $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_\varepsilon-1}\}$. Hence, $\{\mathbf{x}_n\}$ is bounded and since every bounded subset in $\mathbb{R}^n$ has a compact closure in $\mathbb{R}^n$, the statement follows from statement (b).

Theorem is proven. □

Definition 14.14. *A metric space where each Cauchy sequence converges is said to be* ***complete****.*

Example 14.4.

1. *By Theorem 14.8 it follows that all Euclidean spaces are complete.*
2. *The space of all rational numbers with the metric $d(x, y) = |x - y|$ is not complete.*
3. *In $\mathbb{R}^n$ any convergent sequence is bounded but not any bounded sequence obligatory converges.*

There is a special case when bounded sequence obligatory converges. The next theorem specifies such sequences.

Theorem 14.9. (Weierstrass theorem) *Any **monotonic sequence** $\{s_n\}$ of real numbers, namely, when*

*(a) $\{s_n\}$ is **monotonically nondecreasing**: $s_n \leq s_{n+1}$;*
*(b) $\{s_n\}$ is **monotonically nonincreasing**: $s_n \geq s_{n+1}$;*

converges if and only if it is bounded.

Proof. If $\{s_n\}$ converges it is bounded by Theorem 14.5, claim 3. Suppose that $\{s_n\}$ is bounded, namely, $\sup s_n = s < \infty$. Then $s_n \leq s$ and for every $\varepsilon > 0$ there exists an integer n_ε such that $s - \varepsilon \leq s_n \leq s$ for otherwise $s - \varepsilon$ would be an upper bound for $\{s_n\}$. Since $\{s_n\}$ increases and ε is arbitrarily small this means $s_n \to s$. The case $s_n \geq s_{n+1}$ is considered analogously. Theorem is proven. □

14.2.4.4 Upper and lower limits in $\mathbb{R}$

Definition 14.15. *Let $\{s_n\}$ be a sequence of real numbers in $\mathbb{R}$.*

(a) If for every real M there exists an integer n_M such that $s_n \geq M$ for all $n \geq n_M$ we then write

$$\boxed{s_n \to \infty} \tag{14.40}$$

(b) If for every real M there exists an integer n_M such that $s_n \leq M$ for all $n \geq n_M$ we then write

$$\boxed{s_n \to -\infty} \tag{14.41}$$

*(c) Define the **upper limit** of a sequence $\{s_n\}$ as*

$$\boxed{\limsup_{n\to\infty} s_n := \lim_{t\to\infty} \sup_{n\geq t} s_n} \tag{14.42}$$

which may be treated as the biggest limit of all possible subsequences.

*(d) Define the **lower limit** of a sequence $\{s_n\}$ as*

$$\boxed{\liminf_{n\to\infty} s_n := \lim_{t\to\infty} \inf_{n\geq t} s_n} \tag{14.43}$$

which may be treated as a lowest limit of all possible subsequences.

The following theorem, whose proof is quite trivial, is often used in many practical problems.

Theorem 14.10. *Let $\{s_n\}$ and $\{t_n\}$ be two sequences of real numbers in $\mathbb{R}$. Then the following properties hold:*

1.

$$\boxed{\liminf_{n\to\infty} s_n \leq \limsup_{n\to\infty} s_n} \tag{14.44}$$

2.

$$\boxed{\begin{array}{l}\limsup_{n\to\infty} s_n = \infty \quad \text{if} \quad s_n \to \infty \\ \liminf_{n\to\infty} s_n = -\infty \quad \text{if} \quad s_n \to -\infty\end{array}} \tag{14.45}$$

3.

$$\boxed{\limsup_{n\to\infty} (s_n + t_n) \leq \limsup_{n\to\infty} s_n + \limsup_{n\to\infty} t_n} \tag{14.46}$$

4.

$$\boxed{\liminf_{n\to\infty} (s_n + t_n) \geq \liminf_{n\to\infty} s_n + \liminf_{n\to\infty} t_n} \tag{14.47}$$

5. If $\lim_{n\to\infty} s_n = s$ then

$$\boxed{\liminf_{n\to\infty} s_n = \limsup_{n\to\infty} s_n = s} \tag{14.48}$$

6. If $s_n \leq t_n$ for all $n \geq M$ which is fixed then

$$\boxed{\begin{array}{l}\limsup_{n\to\infty} s_n \leq \limsup_{n\to\infty} t_n \\ \liminf_{n\to\infty} s_n \leq \liminf_{n\to\infty} t_n\end{array}} \tag{14.49}$$

Example 14.5.

1.

$$\limsup_{n\to\infty} \sin\left(\frac{\pi}{2}n\right) = 1, \quad \liminf_{n\to\infty} \sin\left(\frac{\pi}{2}n\right) = -1$$

2.

$$\limsup_{n\to\infty} \tan\left(\frac{\pi}{2}n\right) = \infty, \quad \liminf_{n\to\infty} \tan\left(\frac{\pi}{2}n\right) = -\infty$$

3. *For* $s_n = \dfrac{(-1)^n}{1 + 1/n}$

$$\limsup_{n\to\infty} s_n = 1, \qquad \liminf_{n\to\infty} s_n = -1$$

14.2.5 Continuity and function limits in metric spaces

14.2.5.1 Continuity and limits of functions

Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces and $\mathcal{E} \subset \mathcal{X}$, f maps $\mathcal{E}$ into $\mathcal{Y}$ and $p \in \mathcal{X}$.

Definition 14.16.

(a) We write

$$\boxed{\lim_{x\to p} f(x) = q} \tag{14.50}$$

if there is a point $q \in \mathcal{Y}$ such that for every $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon, p) > 0$ for which $d_{\mathcal{Y}}(f(x), q) < \varepsilon$ for all $x \in \mathcal{E}$ for which $d_{\mathcal{X}}(x, p) < \delta$. The symbols $d_{\mathcal{Y}}$ and $d_{\mathcal{X}}$ are referred to as the distance in $\mathcal{X}$ and $\mathcal{Y}$, respectively. Notice that f may be not defined at p since p may not belong to $\mathcal{E}$.

(b) If, in addition, $p \in \mathcal{E}$ and $d_{\mathcal{Y}}(f(x), f(p)) < \varepsilon$ for every $\varepsilon > 0$ and for all $x \in \mathcal{E}$ for which $d_{\mathcal{X}}(x, p) < \delta = \delta(\varepsilon)$ then f is said to be ***continuous*** *at the point p.*

(c) If f is continuous at every point of $\mathcal{E}$ then f is said to be ***continuous on*** *$\mathcal{E}$.*

(d) If for any $x, y \in \mathcal{E} \subseteq \mathcal{X}$

$$\boxed{d_{\mathcal{Y}}(f(x), f(y)) \leq L_f d_{\mathcal{X}}(x, y), \quad L_f < \infty} \tag{14.51}$$

then f is said to be ***Lipschitz continuous on*** *$\mathcal{E}$.*

Remark 14.3. *If p is a limit point of $\mathcal{E}$ then f is continuous at the point p if and only if*

$$\boxed{\lim_{x\to p} f(x) = f(p)} \tag{14.52}$$

The proof of this result follows directly from the definition above.

The following properties related to continuity are evidently fulfilled.

Proposition 14.5.

1. If for metric spaces $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$ the following mappings are defined:

$$f : \mathcal{E} \subset \mathcal{X} \to \mathcal{Y}, \quad g : f(\mathcal{E}) \to \mathcal{Z}$$

and

$$h(x) := g(f(x)), \quad x \in \mathcal{E}$$

then h is continuous at a point $p \in \mathcal{E}$ if f is continuous at p and g is continuous at $f(p)$.

2. *If $f : \mathcal{X} \to \mathbb{R}^n$ and $f(x) := (f_1(x), \ldots, f_n(x))$ then f is continuous if and only if all $f_i(x)$ $\left(i = \overline{1,n}\right)$ are continuous.*
3. *If $f, g : \mathcal{X} \to \mathbb{R}^n$ are continuous mappings then $f + g$ and (f, g) are continuous too on $\mathcal{X}$.*
4. *A mapping $f : \mathcal{X} \to \mathcal{Y}$ is continuous on $\mathcal{X}$ if and only if $f^{-1}(\mathcal{V})$ is open (closed) in $\mathcal{X}$ for every open (closed) set $\mathcal{V} \subset \mathcal{Y}$.*

14.2.5.2 Continuity, compactness and connectedness

Theorem 14.11. *If $f : \mathcal{X} \to \mathcal{Y}$ is a continuous mapping of a compact metric space $\mathcal{X}$ into a metric space $\mathcal{Y}$ then $f(\mathcal{X})$ is compact.*

Proof. Let $\{\mathcal{V}_\alpha\}$ be an open cover of $f(\mathcal{X})$. By continuity of f and in view of Proposition 14.5 it follows that each of the sets $f^{-1}(\mathcal{V}_\alpha)$ is open. By the compactness of $\mathcal{X}$ there are finitely many indices $\alpha_1, \ldots, \alpha_n$ such that

$$\mathcal{X} \subset \bigcup_{i=1}^{n} f^{-1}(\mathcal{V}_{\alpha_i}) \tag{14.53}$$

Since $f\left(f^{-1}(\mathcal{E})\right) \subset \mathcal{E}$ for any $\mathcal{E} \subset \mathcal{Y}$ it follows that (14.53) implies that $f(\mathcal{X}) \subset \bigcup_{\alpha=1}^{n} \mathcal{V}_{\alpha_i}$. This completes the proof. □

Corollary 14.2. *If $f : \mathcal{X} \to \mathbb{R}^n$ is a continuous mapping of a compact metric space $\mathcal{X}$ into $\mathbb{R}^n$ then $f(\mathcal{X})$ is closed and bounded, that is, it contains all its limit points and $\|f(x)\| \leq M < \infty$ for any $x \in \mathcal{X}$.*

Proof. It follows directly from Theorems 14.11 and 14.4. □

The next theorem is particularly important when f is real.

Theorem 14.12. (Weierstrass theorem) *If $f : \mathcal{X} \to \mathbb{R}^n$ is a **continuous mapping** of a **compact metric space** $\mathcal{X}$ into $\mathbb{R}$ and*

$$M = \sup_{x \in \mathcal{X}} f(x), \quad m = \inf_{x \in \mathcal{X}} f(x)$$

then there exist points $x_M, x_m \in \mathcal{X}$ such that

$$M = f(x_M), \quad m = f(x_m)$$

This means that f attains its maximum (at x_M) and its minimum (at x_m), that is,

$$\boxed{M = \sup_{x \in \mathcal{X}} f(x) = \max_{x \in \mathcal{X}} f(x), \quad m = \inf_{x \in \mathcal{X}} f(x) = \min_{x \in \mathcal{X}} f(x)}$$

Proof. By Theorem 14.11 and its corollary it follows that $f(\mathcal{X})$ is a closed and bounded set (say, $\mathcal{E}$) of real numbers. So, if $M \in \mathcal{E}$ then $M \in \mathrm{cl}\,\mathcal{E}$. Suppose $M \notin \mathcal{E}$. Then for any $\varepsilon > 0$ there is a point $y \in \mathcal{E}$ such that $M - \varepsilon < y < M$, for otherwise $(M - \varepsilon)$ would be an upper bound. Thus y is a limit point of $\mathcal{E}$. Hence, $y \in \mathrm{cl}\,\mathcal{E}$ proves the theorem. □

The next theorem deals with the continuity property for inverse continuous one-to-one mappings.

Theorem 14.13. *If $f : \mathcal{X} \to \mathcal{Y}$ is a continuous one-to-one mapping of a compact metric space $\mathcal{X}$ into a metric space $\mathcal{Y}$ then the inverse mapping $f^{-1} : \mathcal{Y} \to \mathcal{X}$ defined by*

$$f^{-1}(f(x)) = x \in \mathcal{X}$$

is a continuous mapping too.

Proof. By Proposition 14.4, applied to f^{-1} instead of f, one can see that it is sufficient to prove that $f(\mathcal{V})$ is an open set of $\mathcal{Y}$ for any open set $\mathcal{V} \subset \mathcal{X}$. Fixing a set $\mathcal{V}$ we may conclude that the complement $\mathcal{V}^c$ of $\mathcal{V}$ is closed in $\mathcal{X}$ and, hence, by Proposition 14.5 it is a compact. As the result, $f(\mathcal{V}^c)$ is a compact subset of $\mathcal{Y}$ (14.11) and so, by Theorem 14.2, it is closed in $\mathcal{Y}$. Since f is one-to-one and onto, $f(\mathcal{V})$ is the complement of $f(\mathcal{V}^c)$ and, hence, it is open. This completes the proof. □

14.2.5.3 Uniform continuity

Definition 14.17. *Let $f : \mathcal{X} \to \mathcal{Y}$ be a mapping of a space $\mathcal{X}$ into a metric space $\mathcal{Y}$. A mapping f is said to be*

(a) ***uniformly continuous*** *on $\mathcal{X}$ if for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that $d_{\mathcal{Y}}(f(x), f(x')) < \varepsilon$ for all $x, x' \in \mathcal{X}$ for which $d_{\mathcal{X}}(x, x') < \delta$.*

(b) ***uniformly Lipschitz continuous*** *on a (x, z)-set $\mathcal{E}$ with respect to x, if there exists a positive constant $L_f < \infty$ such that*

$$d_{\mathcal{Y}}\left(f(x, z), f\left(x', z\right)\right) \leq L_f d_{\mathcal{X}}\left(x, x'\right)$$

for all $x, x', z \in \mathcal{E}$.

Remark 14.4. *The difference between the concepts of* ***continuity*** *and* ***uniform continuity*** *concerns two aspects:*

(a) uniform continuity is a property of a function on a set, whereas continuity is defined for a function in a single point;

(b) δ, participating in the definition (14.50) of continuity, is a function of ε and a point p, that is, $\delta = \delta(\varepsilon, p)$, whereas δ, participating in the definition (14.17) of the uniform continuity, is a function of ε only serving for all points of a set (space) $\mathcal{X}$, that is, $\delta = \delta(\varepsilon)$.

Evidently, any uniformly continued function is continuous but not inverse. The next theorem shows when both concepts coincide.

Theorem 14.14. *If $f : \mathcal{X} \to \mathcal{Y}$ is a continuous mapping of a compact metric space $\mathcal{X}$ into a metric space $\mathcal{Y}$ then f is uniformly continuous on $\mathcal{X}$.*

Proof. Continuity means that for any point $p \in \mathcal{X}$ and any $\varepsilon > 0$ we can associate a number $\delta(\varepsilon, p)$ such that

$$x \in \mathcal{X},\ d_{\mathcal{X}}(x, p) < \delta(\varepsilon, p) \quad \text{implies} \quad d_{\mathcal{Y}}(f(x), f(p)) < \varepsilon/2 \tag{14.54}$$

Define the set

$$\mathcal{J}(p) := \{x \in \mathcal{X} : d_{\mathcal{X}}(x, p) < \delta(\varepsilon, p)/2\}$$

Since $p \in \mathcal{J}(p)$ the collection of all sets $\mathcal{J}(p)$ is an open cover of $\mathcal{X}$ and by the compactness of $\mathcal{X}$ there are a finite set of points $p_1, \ldots, p_n$ such that

$$\mathcal{X} \subset \mathcal{J}(p_1) \cup \cdots \cup \mathcal{J}(p_n) \tag{14.55}$$

Put

$$\tilde{\delta}(\varepsilon) := \frac{1}{2} \min\{\delta(\varepsilon, p_1), \ldots, \delta(\varepsilon, p_n)\} > 0$$

Now let $x \in \mathcal{X}$ satisfy the inequality $d_{\mathcal{X}}(x, p) < \tilde{\delta}(\varepsilon)$. By the compactness (namely, by (14.55)) there is an integer m $(1 \le m \le n)$ such that $p \in \mathcal{J}(p_m)$ implies

$$d_{\mathcal{X}}(x, p_m) < \frac{1}{2}\delta(\varepsilon, p_m)$$

and, as the result,

$$d_{\mathcal{X}}(x, p_m) \le d_{\mathcal{X}}(x, p) + d_{\mathcal{X}}(p, p_m) \le \tilde{\delta}(\varepsilon) + \frac{1}{2}\delta(\varepsilon, p_m) \le \delta(\varepsilon, p_m)$$

Finally, by (14.54)

$$d_{\mathcal{Y}}(f(x), f(p)) \le d_{\mathcal{Y}}(f(x), f(p_m)) + d_{\mathcal{Y}}(f(p_m), f(p)) \le \varepsilon$$

which completes the proof. □

Remark 14.5. *The alternative proof of this theorem may be obtained in the following manner: assuming that f is not uniformly continuous we conclude that there exists $\varepsilon > 0$ and the sequences $\{x_n\}, \{p_n\}$ on $\mathcal{X}$ such that $d_{\mathcal{X}}(x_n, p_n) \underset{n\to\infty}{\to} 0$ but $d_{\mathcal{Y}}(f(x_n), f(p_n)) > \varepsilon$. The last is in contradiction with Theorem 14.3.*

Next examples show that compactness is essential in the hypotheses of the previous theorems.

Example 14.6. *If $\mathcal{E}$ is a noncompact in $\mathbb{R}$ then*

1. There is a continuous function on $\mathcal{E}$ which is not bounded, for example,

$$f(x) = \frac{1}{x - 1}, \quad \mathcal{E} := \{x \in \mathbb{R} : |x| < 1\}$$

Here, $\mathcal{E}$ is a noncompact, $f(x)$ is continuous on $\mathcal{E}$, but evidently unbounded. It is easy to check that it is not uniformly continuous.

2. *There exists a continuous and bounded function on $\mathcal{E}$ which has no maximum, for example,*

$$f(x) = \frac{1}{1 + (x-1)^2}, \quad \mathcal{E} := \{x \in \mathbb{R} : |x| < 1\}$$

Evidently,

$$\sup_{x \in \mathcal{E}} f(x) = 1$$

whereas $\frac{1}{2} \leq f(x) < 1$ *and, hence, has no maximum on* $\mathcal{E}$.

14.2.5.4 Continuity of a family of functions: equicontinuity

Definition 14.18. *A family $\boldsymbol{F}$ of functions $f(x)$ defined on some x set $\mathcal{E}$ is said to be* ***equicontinuous*** *if for any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon)$, the same for all class $\boldsymbol{F}$, such that $d_{\mathcal{X}}(x, y) < \delta$ implies $d_{\mathcal{Y}}(f(x), f(y)) < \varepsilon$ for all $x, y \in \mathcal{E}$ and any $f \in \boldsymbol{F}$.*

The most frequently encountered equicontinuous families $\boldsymbol{F}$ occur when $f \in \boldsymbol{F}$ are uniformly Lipschitz continuous on $\mathcal{X} \subseteq \mathbb{R}^n$ and there exists an $L_f > 0$ which is a Lipschitz constant for all $f \in \boldsymbol{F}$. In this case $\delta = \delta(\varepsilon)$ can be chosen as $\delta = \varepsilon / L_f$.

The following claim can be easily proven.

Claim 14.5. *If a sequence of continuous functions on a compact set $\mathcal{X} \subseteq \mathbb{R}^n$ is uniformly convergent on $\mathcal{X}$, then it is uniformly bounded and equicontinuous.*

The next two assertions are usually referred to as the *Ascoli–Arzelà's theorems* (see the reference in Hartman (2002)). They will be used below for the analysis of ordinary differential equations.

Theorem 14.15. (on the propagation, Ascoli–Arzelà, 1883–1895) *Let, on a compact x-set of $\mathcal{E}$, the sequence of functions $\{f_n(x)\}_{n=1,2,\ldots}$ be equicontinuous and convergent on a dense subset of $\mathcal{E}$. Then there exists a subsequence $\{f_{n_k}(x)\}_{k=1,2,\ldots}$ which is uniformly convergent on $\mathcal{E}$.*

Another version of the same fact is as follows.

Theorem 14.16. (on the selection, Ascoli–Arzelà, 1883–1895) *Let, on a compact x-set of $\mathcal{E} \subset \mathbb{R}^n$, the sequence of functions $\{f_n(x)\}_{n=1,2,\ldots}$ be uniformly bounded and equicontinuous. Then there exists a subsequence $\{f_{n_k}(x)\}_{k=1,2,\ldots}$ which is uniformly convergent on $\mathcal{E}$.*

Proof. Let us consider the set of all rational numbers $\boldsymbol{R} \subseteq \mathcal{E}$. Since $\boldsymbol{R}$ is countable, all of its elements can be designated by numbers, i.e., $\boldsymbol{R} = \{r_j\}$ $(j = 1, \ldots)$. The numerical vector-sequence $\{f_n(r_1)\}_{n=1,2,\ldots}$ is norm-bounded, say, $\|f_n(r_1)\| \leq M$. Hence, we can

choose a convergent sequence $\{f_{n_k}(r_2)\}_{k=1,2,\ldots}$ which is also bounded by the same M. Continuing this process we obtain a subsequence $\{f_p(r_q)\}_{p=1,2,\ldots}$ that converges in a point r_q, $q = 1, 2, \ldots$. Let $f_p := f_p(r_p)$. Show that the sequence $\{f_p\}$ is uniformly convergent on $\mathcal{E}$ to a continuous function $f \in C(\mathcal{E})$. In fact, $\{f_p\}$ converges in any point of $\boldsymbol{R}$ by the construction. To establish its convergence in any point of $\mathcal{E}$, it is sufficient to show that for any fixed $x \in \mathcal{E}$ the sequence $\{f_p(x)\}$ converges on itself. Since $\{f_p(x)\}$ is equicontinuous, for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that for $\|x - x'\| < \delta$ and $x, x' \in \mathcal{E}$ there is $\|f_p(x) - f_p(x')\| < \varepsilon$. Choose r_j such that $\|x - r_j\| < \delta$ implies $\|f_p(x) - f_p(r_j)\| < \varepsilon$. But the sequence $\{f_p(r_j)\}$ converges on itself. Hence, there is a number p_0 such that $\|f_p(x) - f_{p'}(x')\| < \varepsilon$ whenever $p, p' > p_0$. So,

$$\begin{aligned}\|f_p(x) - f_{p'}(x')\| &\le \|f_p(x) - f_p(r_j)\| \\ &+ \|f_p(r_j) - f_{p'}(r_j)\| + \|f_{p'}(r_j) - f_{p'}(x')\| \le 3\varepsilon\end{aligned}$$

Thus $\{f_p(x)\}$ converges at each $x \in \mathcal{E}$. It remains to prove that $\{f_p(x)\}$ converges uniformly on $\mathcal{E}$ and, therefore, its limit f is from $C(\mathcal{E})$. Again, by the assumption on equicontinuity, one can cover the set $\mathcal{E}$ with the finite δ-set containing, say, l-subsets. In each of them select rational numbers, say, $r_1, \ldots, r_l$. By the convergence of $\{f_p(x)\}$ there exists p_0 such that $\|f_p(r_j) - f_{p'}(r_j)\| < \varepsilon$ whenever $p, p' > p_0$, so that

$$\begin{aligned}\|f_p(x) - f_{p'}(x)\| &\le \|f_p(x) - f_p(r_j)\| \\ &+ \|f_p(r_j) - f_{p'}(r_j)\| + \|f_{p'}(r_j) - f_{p'}(x)\| \le 3\varepsilon\end{aligned}$$

where j is selected in such a way that r_j belongs to the same δ-subset as x. Taking $p' \to \infty$, this inequality implies $\|f_p(x) - f(x)\| \le 3\varepsilon$ for all x from the considered δ-subset, but this means the uniform converges on $\{f_p(x)\}$ exactly. Theorem is proven. □

14.2.5.5 Connectedness

The definition of the connectedness of a set $\mathcal{E}$ has been given in Definition 14.7. Here we will discuss its relation with the continuity property of a function f.

Lemma 14.2. *If $f : \mathcal{X} \to \mathcal{Y}$ is a continuous mapping of a metric space $\mathcal{X}$ into a metric space $\mathcal{Y}$, and if $\mathcal{E}$ is a connected subset of $\mathcal{X}$, then $f(\mathcal{E})$ is connected.*

Proof. On the contrary, assume that $f(\mathcal{E}) = \mathcal{A} \cup \mathcal{B}$ with nonempty sets $\mathcal{A}, \mathcal{B} \subset \mathcal{Y}$ such that $\mathcal{A} \cap \mathcal{B} = \varnothing$. Put $\mathcal{G} = \mathcal{E} \cap f^{-1}(\mathcal{A})$ and $\mathcal{H} = \mathcal{E} \cap f^{-1}(\mathcal{B})$. Then $\mathcal{E} = \mathcal{G} \cup \mathcal{H}$ and both $\mathcal{G}$ and $\mathcal{H}$ are nonempty. Since $\mathcal{A} \subset \operatorname{cl}\mathcal{A}$ it follows that $\mathcal{G} \subset f^{-1}(\operatorname{cl}\mathcal{A})$ and $f(\operatorname{cl}\mathcal{G}) \subset \operatorname{cl}\mathcal{A}$. Taking into account that $f(\mathcal{H}) = \mathcal{B}$ and $\operatorname{cl}\mathcal{A} \cap \mathcal{B} = \varnothing$ we may conclude that $\mathcal{G} \cap \mathcal{H} = \varnothing$. By the same argument we conclude that $\mathcal{G} \cap \operatorname{cl}\mathcal{H} = \varnothing$. Thus, $\mathcal{G}$ and $\mathcal{H}$ are separated which is impossible if $\mathcal{E}$ is connected. Lemma is proven. □

This theorem serves as an instrument to state the important result in $\mathbb{R}$ which is known as the Bolzano theorem which concerns a global property of real-valued functions continuous on a compact interval $[a, b] \in \mathbb{R}$: if $f(a) < 0$ and $f(b) > 0$ then the graph of the function $f(x)$ must cross the x-axis somewhere in between. But this theorem as well

as other results concerning the analysis of functions given on $\mathbb{R}^n$ will be considered in detail below in Chapter 16.

14.2.5.6 Homeomorphisms

Definition 14.19. *Let $f : \mathcal{S} \to \mathcal{T}$ be a function mapping points from one metric space $(\mathcal{S}, d_{\mathcal{S}})$ to another $(\mathcal{T}, d_{\mathcal{T}})$ such that it is one-to-one mapping or, in other words, $f^{-1} : \mathcal{T} \to \mathcal{S}$ exists. If additionally f is continuous on $\mathcal{S}$ and f^{-1} on $\mathcal{T}$ then such mapping f is called a* ***topological mapping*** *or* ***homeomorphism****, and the spaces $(\mathcal{S}, d_{\mathcal{S}})$ and $(f(\mathcal{S}), d_{\mathcal{T}})$ are said to be* ***homeomorphic****.*

It is clear from this definition that if f is homeomorphic then f^{-1} is homeomorphic too. The important particular case of a homeomorphism is the so-called *isometry*, i.e., it is a one-to-one continuous mapping which preserves the metric, namely, which for all $x, x' \in \mathcal{S}$ keeps the identity

$$\boxed{d_{\mathcal{T}}(f(x), f(x')) = d_{\mathcal{S}}(x, x')} \tag{14.56}$$

14.2.6 The contraction principle and a fixed point theorem

Definition 14.20. *Let $\mathcal{X}$ be a metric space with a metric d. If φ maps $\mathcal{X}$ into $\mathcal{X}$ and if there is a number $c \in [0, 1)$ such that*

$$\boxed{d(\varphi(x), \varphi(x')) \leq c d(x, x')} \tag{14.57}$$

for all $x, x' \in \mathcal{X}$, then φ is said to be a ***contraction*** *of $\mathcal{X}$ into $\mathcal{X}$.*

Theorem 14.17. (The fixed point theorem) *If $\mathcal{X}$ is a complete metric space and if φ is a contraction of $\mathcal{X}$ into $\mathcal{X}$, then there exists one and only one point $x \in \mathcal{X}$ such that*

$$\boxed{\varphi(x) = x} \tag{14.58}$$

Proof. Pick $x_0 \in \mathcal{X}$ arbitrarily and define the sequence $\{x_n\}$ recursively by setting $x_{n+1} = \varphi(x_n)$, $n = 0, 1, \ldots$. Then, since φ is a contraction, we have

$$\begin{aligned} d(x_{n+1}, x_n) &= d(\varphi(x_n), \varphi(x_{n-1})) \\ &\leq c d(x_n, x_{n-1}) \leq \cdots \leq c^n d(x_1, x_0) \end{aligned}$$

Taking $m > n$ and in view of the triangle inequality, it follows that

$$\begin{aligned} d(x_m, x_n) &\leq \sum_{i=n+1}^{m} d(x_i, x_{i-1}) \leq \left(c^{m-1} + \cdots + c^n\right) d(x_1, x_0) \\ &\leq c^n \left(c^{m-1-n} + \cdots + 1\right) d(x_1, x_0) \leq c^n (1 - c)^{-1} d(x_1, x_0) \end{aligned}$$

Thus $\{x_n\}$ is a Cauchy sequence, and since $\mathcal{X}$ is a complete metric space, it should converge, that is, there exists $\lim_{n\to\infty} x_n := x$. And, since φ is a contraction, it is continuous (in fact, uniformly continuous). Therefore $\varphi(x) = \lim_{n\to\infty} \varphi(x_n) = \lim_{n\to\infty} x_n = x$. The uniqueness follows from the following consideration. Assume that there exists another point $y \in \mathcal{X}$ such that $\varphi(y) = y$. Then by (14.57) it follows that $d(x, y) \leq cd(\varphi(x), \varphi(y)) = cd(x, y)$ which may only happen if $d(x, y) = 0$ which proves the theorem. □

14.3 Summary

The properties of sets which remain invariant under every topological mapping are usually called the *topological properties*. Thus properties of being open, closed, or compact are topological properties.

15 Integration

Contents

15.1 Naive interpretation

15.1.1 What is the Riemann integration?

It is well known from elementary calculus that to find the area of the region under the graph of a positive function f on the closed interval $[a, b]$, one needs to subdivide the interval into a finite number of subintervals, say n, with the kth subinterval Δx_k and to consider the sums I_n^R defined as

$$I_n^R := \sum_{i=1}^{n} f(t_k)\,\Delta x_k, \quad t_k \in [x_{k-1}, x_k) \qquad x_0 = a < x_1 < \cdots < x_n = b, \quad \Delta x_k := x_k - x_{k-1} \tag{15.1}$$

Such sum is suggested to be considered as an approximation of the area by means of rectangles (see Fig. 15.1).

Making the successive subdivisions finer and finer, or, in other words, taking $n \to \infty$ and if there exists some hope that these sums will tend to a limit $I^R(f)$ then such sums will converge to a real value of the square of the area under consideration. This, roughly

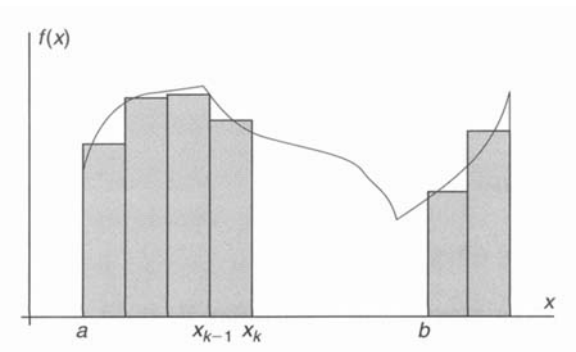

Fig. 15.1. Riemann's type of integration.

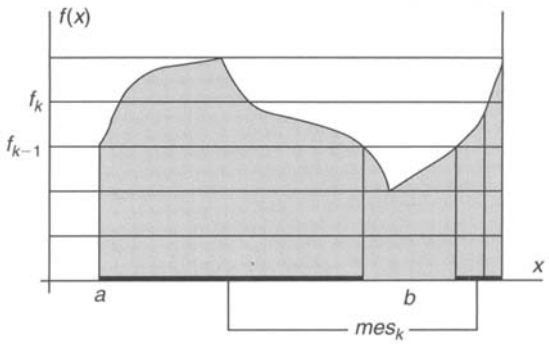

Fig. 15.2. Lebesgue's type of integration.

speaking, is what is involved in *Riemann's* definition of the definite *integral* $\int_{x=a}^{b} f(x)\,dx$ which is studied in detail within the elementary calculus course. This type of integration is well defined for the class of continuous or partially continuous functions.

15.1.2 What is the Lebesgue integration?

If a function has some more complex structure and admits any discontinuity of a more complex nature then another generalizing integration method is required. One of such method is the *Lebesgue integration*. It corresponds to the following approximation scheme (see Fig. 15.2):

$$\boxed{\begin{array}{c} I_n^L := \sum_{i=1}^{n} (f_k - f_{k-1})\,\mathrm{mes}_k \\ f_0 := \inf_{x\in[a,b]} f(x) < f_1 < \cdots < f_n := \sup_{x\in[a,b]} f(x) \\ \mathrm{mes}_k \text{ is the longitude of all intervals where} \\ f_{k-1} \le f(x) < f_k \end{array}} \tag{15.2}$$

If a limit I^L of I_n^L (when $n \to \infty$) exists it is called the Lebesgue integral of $f(x)$ on $[a, b]$ and is denoted by $I^L(f) := \int_{x=a}^{b} f$. It is closely related to *a measure of a set*. This chapter considers both integration schemes in detail and rigorously from a mathematical point of view.

15.2 The Riemann–Stieltjes integral

15.2.1 Riemann integral definition

Let $[a, b]$ be a given interval and a **partition** P_n of $[a, b]$ be defined as a finite collection of points

$$a = x_0 < x_1 < \cdots < x_n = b$$

We write

$$\Delta x_i := x_i - x_{i-1},\ i = 1, \ldots, n$$

Definition 15.1. *Suppose f is a bounded real function defined on $[a, b]$ and*

$$\boxed{M_i := \sup_{x \in [x_{i-1}, x_i)} f(x), \quad m_i := \inf_{x \in [x_{i-1}, x_i)} f(x)} \tag{15.3}$$

Then

$$\boxed{U(P_n, f) := \sum_{i=1}^{n} M_i \Delta x_i} \tag{15.4}$$

and

$$\boxed{L(P_n, f) := \sum_{i=1}^{n} m_i \Delta x_i} \tag{15.5}$$

are called the ***upper*** *and* ***lower Darboux sums****, respectively.*

Definition 15.2.

1. *The* ***upper Riemann integral*** $I^U(f)$ *is defined as follows:*

$$\boxed{I^u(f) := \limsup_{n \to \infty} \sup_{P_n} U(P_n, f)} \tag{15.6}$$

where $\sup_{P_n}$ *is taken over all partitions P_n of the interval $[a, b]$.*

2. *The* ***lower Riemann integral*** $I^L(f)$ *is defined as follows:*

$$\boxed{I^l(f) := \liminf_{n \to \infty} \inf_{P_n} L(P_n, f)} \tag{15.7}$$

where $\inf_{P_n}$ *is taken over all partitions P_n of the interval $[a, b]$.*

3. *If*

$$I^l(f) = I^u(f)$$

then the ***Riemann integral*** $I^R(f)$*, often written as*

$$I^R(f) = \int_{x=a}^{b} f(x)\, dx$$

is defined by

$$\boxed{I^R(f) = \int_{x=a}^{b} f(x)\, dx := I^l(f) = I^u(f)} \tag{15.8}$$

Notice that for any partition P_n we have

$$\min_{i=1,\ldots,n} M_i := m \leq L(P_n, f) \leq U(P_n, f) \leq M := \max_{i=1,\ldots,n} M_i$$

So, the numbers $L(P_n, f)$ and $U(P_n, f)$ form a bounded set and, hence, are correctly defined. The question of the integrability of f (when (15.8) holds) is a delicate question which will be discussed below.

15.2.2 Definition of Riemann–Stieltjes integral

We shall be working here with a compact set $[a, b] \in \mathbb{R}$ and all functions will be assumed to be real-valued functions defined on $[a, b]$. Complex-valued functions will be considered below in Chapter 18.

Let $P_n := \{a = x_0, x_1, \ldots, x_n = b\}$ be a partition of $[a, b]$, t_k be a point within the interval $[x_{k-1}, x_k)$ and $\Delta\alpha_k := \alpha(x_k) - \alpha(x_{k-1})$ where $\alpha : \mathbb{R} \to \mathbb{R}$ is a real function defined on $[a, b]$.

Definition 15.3.

1. *A partition P' of $[a, b]$ is said to be **finer** than P (or a **refinement** of P) if*

$$\boxed{P \subseteq P'} \tag{15.9}$$

2. *A sum of the form*

$$\boxed{S(P_n, f, \alpha) := \sum_{k=1}^{n} f(t_k)\,\Delta\alpha_k} \tag{15.10}$$

*is called a **Riemann–Stieltjes sum** of f with respect to α corresponding to a given partition P_n.*

3. *We say that f is **integrable in the Riemann sense** with respect to α on $[a, b]$ and we will write $f \in \mathcal{R}_{[a,b]}(\alpha)$ if there exists a number I^{R-S} having the following property: for any $\varepsilon > 0$ there exists a partition P_ε of $[a, b]$ such that for any partition P finer than P_ε and for any choice of the points $t_k \in [x_{k-1}, x_k)$ we have*

$$\boxed{\left|S(P, f, \alpha) - I^{R-S}\right| < \varepsilon} \tag{15.11}$$

4. *When such number I^{R-S} exists, it is uniquely determined and is denoted by*

$$\boxed{I^{R-S} := \int_{x=a}^{b} f(x)\,d\alpha(x)} \tag{15.12}$$

*This is the **Riemann–Stieltjes integral** (or simply the **Stieltjes integral**) of f with respect to α on $[a, b]$.*

5. *The functions f and α are referred to as the **integrand** and the **integrator**, respectively.*

Remark 15.1. *The letter x in (15.12) is a "dummy variable", so it may be replaced by any other convenient symbol, for example,*

$$\boxed{I^{R-S} := \int_{x=a}^{b} f(x)\, d\alpha(x) = \int_{\zeta=a}^{b} f(\zeta)\, d\alpha(\zeta)} \tag{15.13}$$

Remark 15.2. *By taking $\alpha(x) = x$, the Riemann integral (15.8) is seen to be a special (partial) case of the Riemann–Stieltjes integral (15.12).*

15.2.3 Main properties of the Riemann–Stieltjes integral

Theorem 15.1. (on linear properties)

1. If $f, g \in \mathcal{R}_{[a,b]}(\alpha)$ then for any $c_1, c_2 \in \mathbb{R}$

$$\boxed{c_1 f + c_2 g \in \mathcal{R}_{[a,b]}(\alpha)} \tag{15.14}$$

and

$$\boxed{\begin{aligned}&\int_{x=a}^{b} [c_1 f(x) + c_2 g(x)]\, d\alpha(x) \\ &\quad = c_1 \int_{x=a}^{b} f(x)\, d\alpha(x) + c_2 \int_{x=a}^{b} g(x)\, d\alpha(x)\end{aligned}} \tag{15.15}$$

2. If $f \in \mathcal{R}_{[a,b]}(\alpha)$ and at the same time $f \in \mathcal{R}_{[a,b]}(\beta)$ then for any $c_1, c_2 \in \mathbb{R}$

$$\boxed{f \in \mathcal{R}_{[a,b]}(c_1\alpha + c_2\beta)} \tag{15.16}$$

and

$$\boxed{\begin{aligned}&\int_{x=a}^{b} f(x)\, d[c_1\alpha(x) + c_2\beta(x)] \\ &\quad = c_1 \int_{x=a}^{b} f(x)\, d\alpha(x) + c_2 \int_{x=a}^{b} f(x)\, d\beta(x)\end{aligned}} \tag{15.17}$$

Proof. It follows directly from the linear property for the Riemann–Stieltjes sums (15.10) $S(P, c_1 f + c_2 g, \alpha)$ and $S(P, h, c_1\alpha + c_2\beta)$. □

Theorem 15.2. (on intervals summation) *Assume that* $c \in [a, b]$. *If two of the three integrals in the next identity exist then the third one also exists and*

$$\boxed{\int_{x=a}^{b} f(x)\, d\alpha(x) = \int_{x=a}^{c} f(x)\, d\alpha(x) + \int_{x=c}^{b} f(x)\, d\alpha(x)} \tag{15.18}$$

Proof. If P is a partition of $[a, b]$ and $c \in P$ then we may introduce the corresponding partitions of $[a, c]$ and $[c, b]$, respectively, as follows:

$$P' := P \cap [a, c], \quad P'' := P \cap [c, b]$$

Then by the linear property for the Riemann–Stieltjes sums (15.10) we have

$$S(P, f, \alpha) = S\left(P', f, \alpha\right) + S\left(P'', f, \alpha\right)$$

which implies the proof of the desired result. □

Corollary 15.1.

1. If $a < b$ *and* $f \in \mathcal{R}_{[a,b]}(\alpha)$ *then*

$$\boxed{\int_{x=b}^{a} f(x)\, d\alpha(x) = -\int_{x=a}^{b} f(x)\, d\alpha(x)} \tag{15.19}$$

whenever $\int_{x=a}^{b} f(x)\, d\alpha(x)$ *exists.*

2.

$$\boxed{\int_{x=a}^{a} f(x)\, d\alpha(x) = 0} \tag{15.20}$$

3. The identity (15.18) can be represented as

$$\boxed{\int_{x=a}^{b} f(x)\, d\alpha(x) + \int_{x=b}^{c} f(x)\, d\alpha(x) + \int_{x=c}^{a} f(x)\, d\alpha(x) = 0} \tag{15.21}$$

Theorem 15.3. (on integration by parts) *If* $f \in \mathcal{R}_{[a,b]}(\alpha)$ *then*

1.

$$\boxed{\alpha \in \mathcal{R}_{[a,b]}(f)} \tag{15.22}$$

2.

$$\boxed{\int_{x=a}^{b} f(x)\,d\alpha(x) + \int_{x=a}^{b} \alpha(x)\,df(x) = f(b)\,\alpha(b) - f(a)\,\alpha(a)} \tag{15.23}$$

Proof. Since $\int_{x=a}^{b} f(x)\,d\alpha(x)$ exists then for every $\varepsilon > 0$ there is a partition P_ε of $[a, b]$ such that for every $P' \supseteq P_\varepsilon$ we have

$$\left| S(P', f, \alpha) - \int_{x=a}^{b} f(x)\,d\alpha(x) \right| < \varepsilon$$

Then for an arbitrary $(P \supseteq P_\varepsilon)$ Riemann–Stieltjes sum it follows that

$$S(P_n, \alpha, f) = \sum_{k=1}^{n} \alpha(t_k)\,\Delta f_k = \sum_{k=1}^{n} \alpha(t_k)\,f(x_k) - \sum_{k=1}^{n} \alpha(t_k)\,f(x_{k-1})$$

Define

$$A := f(b)\,\alpha(b) - f(a)\,\alpha(a) = \sum_{k=1}^{n} f(x_k)\,\alpha(x_k) - \sum_{k=1}^{n} f(x_{k-1})\,\alpha(x_{k-1})$$

Subtracting the last two equations we derive

$$S(P_n, \alpha, f) - A = \sum_{k=1}^{n} f(x_k)\,[\alpha(t_k) - \alpha(x_k)] + \sum_{k=1}^{n} f(x_{k-1})\,[\alpha(x_{k-1}) - \alpha(t_k)]$$

Two sums in the right-hand side can be considered as a single one of the form $S(P', f, \alpha)$ where P' is a partition of $[a, b]$ obtained by taking the points x_k and t_k together. So, for such a partition it follows $P' \supseteq P_\varepsilon$ and, hence,

$$A - S(P, \alpha, f) - \int_{x=a}^{b} f(x)\,d\alpha(x) = \sum_{k=1}^{n} f(x_k)\,[\alpha(x_k) - \alpha(t_k)] + \sum_{k=1}^{n} f(x_{k-1})\,[\alpha(t_k) - \alpha(x_{k-1})] - \int_{x=a}^{b} f(x)\,d\alpha(x)$$

$$= S(P', \alpha, f) - \int_{x=a}^{b} f(x)\,d\alpha(x)$$

As the result, we obtain

$$\left| A - S(P_n, \alpha, f) - \int_{x=a}^{b} f(x)\, d\alpha(x) \right| = \left| S(P', \alpha, f) - \int_{x=a}^{b} f(x)\, d\alpha(x) \right| < \varepsilon$$

which is exactly the statement that $\int_{x=a}^{b} \alpha(x)\, df(x)$ exists and equals $A - \int_{x=a}^{b} f(x)\, d\alpha(x)$. The theorem is proven. □

Theorem 15.4. (on the change of variables) *Let $f \in \mathcal{R}_{[a,b]}(\alpha)$ and let $g : \mathbb{R} \to \mathbb{R}$ be a strictly monotonic increasing (or decreasing) function defined on the interval $[c, d]$, that is,*

$$g(x) < g(x') \quad \text{if} \quad x < x'$$

Assume that

$$\boxed{g(c) = a, \quad g(d) = b} \tag{15.24}$$

and

$$\boxed{h(x) := f(g(x)), \quad \beta(x) := \alpha(g(x))} \tag{15.25}$$

are the composite functions defined for any $x \in [c, d]$. Then

$$\boxed{h \in \mathcal{R}_{[c,d]}(\beta) \quad \text{and} \quad \int_{x=a}^{b} f(x)\, d\alpha(x) = \int_{x=c}^{d} h(x)\, d\beta(x)} \tag{15.26}$$

or, equivalently,

$$\boxed{\int_{g(c)}^{g(d)} f(t)\, d\alpha(t) = \int_{x=c}^{d} f(g(x))\, d\alpha(g(x))} \tag{15.27}$$

Proof. By strict monotonicity it follows that for every partition $P_n := \{y_0, y_1, \ldots, y_n\}$ of $[c, d]$ there corresponds one and only one partition $P_n' := \{x_0, x_1, \ldots, x_n\}$ of $[a, b]$. In fact,

$$P_n' = g(P_n), \quad P_n = g^{-1}(P_n')$$

and, moreover, any refinement of P_n produces a corresponding refinement of P_n' and conversely. If $P_n \supseteq P_\varepsilon$ $(\varepsilon > 0)$ then let us consider

$$S(P_n, h, \beta) := \sum_{k=1}^{n} h(u_k)\,\Delta\beta_k, \quad u_k \in [y_{k-1}, y_k), \quad \Delta\beta_k := \beta(y_k) - \beta(y_{k-1})$$

If putting $t_k := g(u_k)$ and $x_k := g(y_k)$ we obtain $P_n' \supseteq P_\varepsilon'$ for which

$$\left| S(P_n', f, \alpha) - \int_a^b f(t)\,d\alpha(t) \right| < \varepsilon$$

Then

$$\begin{aligned} S(P_n, h, \beta) &= \sum_{k=1}^{n} h(g(u_k))\,[\alpha(g(x_k)) - \alpha(g(x_{k-1}))] \\ &= \sum_{k=1}^{n} f(t_k)\,(\alpha(x_k) - \alpha(x_{k-1})) = S(P_n', f, \alpha) \end{aligned}$$

since $t_k \in [x_k, x_{k-1})$. Therefore,

$$\begin{aligned} &\left| S(P_n, h, \beta) - \int_a^b f(t)\,d\alpha(t) \right| \\ &\quad = \left| S(P_n', f, \alpha) - \int_a^b f(t)\,d\alpha(t) \right| < \varepsilon \end{aligned}$$

which completes the proof of this theorem. □

Exercise 15.1. If $f, f^2 \in \mathcal{R}_{[a,b]}(\alpha)$ and $g, g^2 \in \mathcal{R}_{[a,b]}(\alpha)$ then

$$\begin{aligned} &\frac{1}{2}\int_{x=a}^{b}\left[\int_{y=a}^{b}\left(\det\begin{bmatrix} f(x) & g(x) \\ f(y) & g(y) \end{bmatrix}\right)^2 d\alpha(y)\right] d\alpha(x) \\ &\quad = \left[\int_{x=a}^{b} f^2(x)\,d\alpha(x)\right]\left[\int_{x=a}^{b} g^2(x)\,d\alpha(x)\right] \\ &\qquad - \left[\int_{x=a}^{b} f(x)\,g(x)\,d\alpha(x)\right]^2 \end{aligned} \tag{15.28}$$

Exercise 15.2. If $f, g, f \cdot g \in \mathcal{R}_{[a,b]}(\alpha)$ then

$$\begin{aligned}\frac{1}{2}\int_{x=a}^{b}\left[\int_{y=a}^{b}(f(y)-f(x))(g(y)-f(x))\,d\alpha(y)\right]d\alpha(x)\\ =[\alpha(b)-\alpha(a)]\int_{x=a}^{b}f(x)\,g(x)\,d\alpha(x)\\ -\left[\int_{x=a}^{b}f(x)\,d\alpha(x)\right]\left[\int_{x=a}^{b}g(x)\,d\alpha(x)\right]\end{aligned} \tag{15.29}$$

15.2.4 Different types of integrators

15.2.4.1 Differentiable integrators

Theorem 15.5. (on a reduction to the Riemann integral) *Given $f \in \mathcal{R}_{[a,b]}(\alpha)$ assume that the integrator $\alpha(x)$ can be represented as*

$$\alpha(x)=\int_{t=a}^{x}\alpha'(t)\,dt \tag{15.30}$$

where $\alpha'(t)$, called ***the derivative*** *of $\alpha(x)$ at the point $x \in [a,b]$, is a continuous function on $[a,b]$. Then*

1. *There exists the Riemann integral*

$$\int_{a}^{b}f(x)\,\alpha'(x)\,dx$$

2. *The following identity holds*

$$\int_{a}^{b}f(x)\,d\alpha(x)=\int_{a}^{b}f(x)\,\alpha'(x)\,dx \tag{15.31}$$

Proof. For a partition P_n of $[a,b]$ define

$$S(P_n,g,x):=\sum_{k=1}^{n}g(t_k)\,\Delta x_k$$

$$g(t_k):=f(t_k)\,\alpha'(t_k),\qquad \Delta x_k:=x_k-x_{k-1}$$

$$S(P_n,f,\alpha):=\sum_{k=1}^{n}f(t_k)\,\Delta\alpha_k$$

By continuity of $\alpha'(x)$ it follows that

$$\Delta\alpha_k = \alpha'\left(v^k\right)\Delta x_k, \quad v^k \in [x_{k-1}, x_k]$$

and that $\alpha'(x)$ is uniformly continuous on $[a, b]$, that is, for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that $0 < |x - x'| < \delta$ implies $|\alpha'(x) - \alpha'(x')| < \varepsilon$. Hence,

$$|S(P_n, g, x) - S(P_n, f, \alpha)| = \left|\sum_{k=1}^{n} f(t_k)\left[\alpha'(t_k) - \alpha'\left(v^k\right)\right]\Delta x_k\right|$$
$$\leq \sum_{k=1}^{n} |f(t_k)|\left|\alpha'(t_k) - \alpha'\left(v^k\right)\right||\Delta x_k| \leq \varepsilon \sum_{k=1}^{n} |f(t_k)||\Delta x_k| \leq \varepsilon M(b-a)$$

where $M := \sup_{x\in[a,b]} f(x)$. On the other hand, since $f \in \mathcal{R}_{[a,b]}(\alpha)$, there exists a partition P_n finer than P_ε such that

$$\left|S(P, f, \alpha) - \int_a^b f(x)\, d\alpha(x)\right| < \varepsilon$$

which leads to the following:

$$\left|S(P_n, g, x) - \int_a^b f(x)\, d\alpha(x)\right| = \left|[S(P_n, g, x) - S(P_n, f, \alpha)]\right.$$
$$\left. + \left[S(P_n, f, \alpha) - \int_a^b f(x)\, d\alpha(x)\right]\right| \leq |S(P_n, g, x) - S(P_n, f, \alpha)|$$
$$+ \left|S(P_n, f, \alpha) - \int_a^b f(x)\, d\alpha(x)\right| \leq \varepsilon M(b-a) + \varepsilon = \varepsilon[1 + M(b-a)]$$

The arbitrary of ε implies (15.31) which completes the proof. □

15.2.4.2 Step functions

If the integrator $\alpha(x)$ is a constant over the interval $[a, b]$ then the integral $\int_a^b f(x)\, d\alpha(x)$ exists and is equal to zero for any partially continuous function $f(x)$. However, if $\alpha(x)$ is a constant except for a jump discontinuity at one point, then the integral $\int_a^b f(x)\, d\alpha(x)$ not obligatory exists, but if it does exist, its value need not be zero. The next theorem clarifies this situation.

Theorem 15.6. (on a single jump integrator) *Given $a < c < b$ let us assume that*

(a) the values $\alpha(a)$, $\alpha(c)$ and $\alpha(b)$ are arbitrary;
(b) $\alpha(x)$ defined on $[a, b]$ is a step function, i.e.,

$$\alpha(x) = \begin{cases} \alpha(a) & \text{if } a \leq x < c \\ \alpha(b) & \text{if } c < x \leq b \end{cases}$$

(c) $f : \mathbb{R} \to \mathbb{R}$ *is defined on* $[a, b]$ *in such a way that at least one of the functions* f *or* α *is continuous from left at* c *(this means that there exists* $\lim_{x \to c-0} f(x) = f(c)$ *or* $\lim_{x \to c-0} \alpha(x) = \alpha(c)$*) and at least one is continuous from right at* c*.*

Then the integral $\int_a^b f(x)\, d\alpha(x)$ *exists, that is,* $f \in \mathcal{R}_{[a,b]}(\alpha)$ *and*

$$\boxed{\begin{array}{l} \displaystyle\int_a^b f(x)\, d\alpha(x) = f(c)\, \Delta\alpha \\ \Delta\alpha := \alpha(c+0) - \alpha(c-0) \end{array}} \tag{15.32}$$

where $\Delta\alpha$ *is* ***the jump of the function*** $\alpha(x)$ *at the point* c*.*

Remark 15.3. *The result also holds if* $c = a$ *provided that*

$$\alpha(c-0) := \alpha(c)$$

and if $c = b$ *defining*

$$\alpha(c+0) := \alpha(b)$$

Proof. Supposing $c \in P_n$ implies

$$\begin{aligned} S(P_n, f, \alpha) &= f(t_{k-1})\,[\alpha(c) - \alpha(c-0)] \\ &\quad + f(t_k)\,[\alpha(c+0) - \alpha(c)] = [f(t_{k-1}) - f(c)]\,[\alpha(c) - \alpha(c-0)] \\ &\quad + [f(t_k) - f(c)]\,[\alpha(c+0) - \alpha(c)] + \gamma \\ \gamma &:= f(c)\ \ [\alpha(c+0) - \alpha(c-0)] \end{aligned}$$

where $t_{k-1} \le c \le t_k$. So, for $\Delta := S(P_n, f, \alpha) - \gamma$ one has

$$\begin{aligned} |\Delta| &\le \left|[f(t_{k-1}) - f(c)]\,[\alpha(c) - \alpha(c-0)]\right| \\ &\quad + [f(t_k) - f(c)]\,[\alpha(c+0) - \alpha(c)] \le \varepsilon\left(\left|[\alpha(c) - \alpha(c-0)]\right|\right) \\ &\quad + \varepsilon\left(\left|[\alpha(c+0) - \alpha(c)]\right|\right) \le \varepsilon \cdot \text{const} \end{aligned}$$

if $|f(t_{k-1}) - f(c)| < \varepsilon$ and $|f(t_k) - f(c)| < \varepsilon$ which may be done by the corresponding partitioning of $[a, b]$. This proves the theorem. □

Next, let us consider a *step function* $f(x)$ defined on $[a, b]$ by a partition $P_n := \{a = x_0, x_1, \ldots, x_n = b\}$ such that $\alpha(x)$ is a constant on each open subinterval (x_{k-1}, x_k) and has jumps

$$\Delta\alpha_k := \alpha(x_k + 0) - \alpha(x_k - 0), \quad k = 2, \ldots, n-1$$

$$\Delta\alpha_1 := \alpha(x_1 + 0) - \alpha(x_1)$$

$$\Delta\alpha_n := \alpha(x_n) - \alpha(x_n - 0)$$

Then the following theorem provides the connecting link between the Riemann–Stieltjes integral and finite sums depending on values of the integrator function jumps.

Theorem 15.7. (on multiple jumps) *If a function f defined on $[a, b]$ in such a way that neither f nor α are discontinuous from the right or from the left at each jump point x_k then $\int_a^b f(x)\, d\alpha(x)$ exists, that is, $f \in \mathcal{R}_{[a,b]}(\alpha)$ and*

$$\boxed{\int_a^b f(x)\, d\alpha(x) = \sum_{k=1}^{n} f(x_k)\, \Delta\alpha_k} \tag{15.33}$$

Proof. Evidently, by the additivity property (15.18) the integral $\int_a^b f(x)\, d\alpha(x)$ can be rewritten as a sum of integrals with a single jump that proves the theorem. □

Example 15.1. *Denote by $[x]$ the, so-called,* ***greatest-integer function*** *defined as the unique integer satisfying the inequality*

$$\boxed{[x] \le x < [x] + 1} \tag{15.34}$$

Then any finite sum $\sum_{k=1}^{n} a_k$ can be represented as a Riemann–Stieltjes integral as follows:

$$\boxed{\begin{gathered} \sum_{k=1}^{n} a_k = \int_{x=0}^{n} f(x)\, d[x] \\ f(x) = a_k \text{ if } x \in (k-1, k], \quad f(0) = 0 \end{gathered}} \tag{15.35}$$

Example 15.2. (Euler's summation formula) *If f has a continuous derivative f' (see (15.30)) on $[a, b]$ then*

$$\boxed{\begin{aligned} \sum_{k=[a]+1}^{[b]} f(n) = \int_{x=a}^{b} f(x)\, dx + \int_{x=a}^{b} f'(x)\, d\left(x - [x]\right) \\ + f(a)\left(a - [a]\right) - f(b)\left(b - [b]\right) \end{aligned}} \tag{15.36}$$

If a and b are integers then (15.36) becomes

$$\boxed{\begin{aligned} \sum_{k=a}^{b} f(n) = \int_{x=a}^{b} f(x)\, dx + \int_{x=a}^{b} f'(x)\left(x - [x] - 1/2\right) dx \\ + [f(a) + f(b)]/2 \end{aligned}} \tag{15.37}$$

Notice that (15.36) may be obtained using integration by part (15.23):

$$\begin{aligned}
&\int_{x=a}^{b} f(x)\, d\left(x-[x]\right) \\
&\quad = -\int_{x=a}^{b} \left(x-[x]\right) df(x) + f(b)\left(b-[b]\right) - f(a)\left(a-[a]\right) \\
&\quad = -\int_{x=a}^{b} \left(x-[x]\right) f'(x)\, dx + f(b)\left(b-[b]\right) - f(a)\left(a-[a]\right)
\end{aligned}$$

15.2.4.3 Monotonically nondecreasing integrators

When α is nondecreasing on $[a, b]$, i.e., when $\alpha(x) \leq \alpha(x')$ if $x \leq x'$, the differences $\Delta\alpha_k$ which appear in the Riemann–Stieltjes integral are all nonnegative which plays a vital role in the development of the integration theory. For brevity, we will use the abbreviation

$$\boxed{\text{“}\alpha \uparrow \text{ on } [a, b]\text{”}} \tag{15.38}$$

to mean that $\alpha(x)$ is nondecreasing on $[a, b]$. The following properties seem to be evident.

Proposition 15.1. *Assume $\alpha \uparrow$ on $[a, b]$ and $f, g \in \mathcal{R}_{[a,b]}(\alpha)$. Then*

1. *If $f(x) \leq g(x)$ for all $x \in [a, b]$ we have*

$$\boxed{\int_{x=a}^{b} f(x)\, d\alpha(x) \leq \int_{x=a}^{b} g(x)\, d\alpha(x)} \tag{15.39}$$

2. *If $g(x) \geq 0$ for all $x \in [a, b]$ it follows*

$$\boxed{0 \leq \int_{x=a}^{b} g(x)\, d\alpha(x)} \tag{15.40}$$

which can be obtained from (15.39) taking $f(x) = 0$;

3.

$$\boxed{|f| \in \mathcal{R}_{[a,b]}(\alpha)} \tag{15.41}$$

and

$$\boxed{\left|\int_{x=a}^{b} f(x)\, d\alpha(x)\right| \leq \int_{x=a}^{b} |f(x)|\, d\alpha(x)} \tag{15.42}$$

which follows from the inequality

$$||f(x)| - |f(y)|| \leq |f(x) - f(y)|$$

4.

$$\boxed{f^2 \in \mathcal{R}_{[a,b]}(\alpha)} \tag{15.43}$$

5.

$$\boxed{f \cdot g \in \mathcal{R}_{[a,b]}(\alpha)} \tag{15.44}$$

which follows from the identity

$$2f(x)\,g = (f(x) + g(x))^2 - f^2(x) - g^2(x)$$

Proposition 15.2. (The Cauchy–Schwarz inequality) If $f,\ f^2,\ g,\ g^2 \in \mathcal{R}_{[a,b]}(\alpha)$ and, in addition, $\alpha \uparrow$ on $[a, b]$, then

$$\boxed{\begin{aligned}\left[\int_{x=a}^{b} f(x)\,g(x)\,d\alpha(x)\right]^2 \\ \leq \left[\int_{x=a}^{b} f^2(x)\,d\alpha(x)\right]\left[\int_{x=a}^{b} g^2(x)\,d\alpha(x)\right]\end{aligned}} \tag{15.45}$$

Proof. It follows directly from (15.28) since the left-hand side of this identity is nonnegative. □

Proposition 15.3. *If $f, g, f \cdot g \in \mathcal{R}_{[a,b]}(\alpha)$, both f and g are either nondecreasing or nonincreasing and $\alpha \uparrow$ on $[a, b]$, then*

$$\boxed{\begin{aligned}\left[\int_{x=a}^{b} f(x)\,d\alpha(x)\right]\left[\int_{x=a}^{b} g(x)\,d\alpha(x)\right] \\ \leq [\alpha(b) - \alpha(a)]\int_{x=a}^{b} f(x)\,g(x)\,d\alpha(x)\end{aligned}} \tag{15.46}$$

Proof. It follows directly from (15.29) since the left-hand side of this identity is nonnegative. □

15.2.4.4 Integrators of bounded variation

Definition 15.4. *If P_n is a partition of a compact interval $[a, b]$, $\Delta\alpha_k := \alpha(x_k) - \alpha(x_{k-1})$ and there exists a positive number M such that*

$$\boxed{\sum_{k=1}^{n} |\Delta\alpha_k| \leq M} \tag{15.47}$$

for all partitions P_n of the interval $[a, b]$, then α is said to be of ***bounded variation*** *on $[a, b]$.*

Lemma 15.1. *If α is monotonic on $[a, b]$ then it is of bounded variation on $[a, b]$.*

Proof. Let α be nondecreasing. Then $\Delta\alpha_k \geq 0$ for all $k = 1, \ldots, n$ and, hence,

$$\sum_{k=1}^{n} |\Delta\alpha_k| = \sum_{k=1}^{n} \Delta\alpha_k = \alpha(x_n) - \alpha(x_0) = \alpha(b) - \alpha(a)$$

If f is nonincreasing then $\Delta\alpha_k \leq 0$ and $|\Delta\alpha_k| = -\Delta\alpha_k$ which gives

$$\sum_{k=1}^{n} |\Delta\alpha_k| = \alpha(a) - \alpha(b)$$

Lemma is proven. □

Lemma 15.2. *If α is continuous on $[a, b]$ and if α' exists and is bounded (say, $\sup_{x\in[a,b]} |\alpha'(x)| \leq M < \infty$) then α is of bounded variation on $[a, b]$.*

Proof. Since $\Delta\alpha_k = \alpha(x_k) - \alpha(x_{k-1}) = \alpha'(t_k)(x_k - x_{k-1})$ where $t_k \in (x_{k-1}, x_k)$ it follows that

$$\sum_{k=1}^{n} |\Delta\alpha_k| = \sum_{k=1}^{n} |\alpha'(t_k)(x_k - x_{k-1})| = \sum_{k=1}^{n} |\alpha'(t_k)|(x_k - x_{k-1})$$
$$\leq M \sum_{k=1}^{n} (x_k - x_{k-1}) = M(b - a) < \infty$$

which completes the proof. □

Lemma 15.3. *If α is of bounded variation on $[a, b]$, say $\sum_{k=1}^{n} |\Delta\alpha_k| \leq M$ for all partitions of $[a, b]$, then α is bounded on $[a, b]$, namely,*

$$\boxed{\alpha(x) \leq \alpha(a) + M} \tag{15.48}$$

Proof. For any $x \in (a, b)$, using the special partition $P := \{a, x, b\}$, we find

$$|\alpha(x) - \alpha(a)| + |\alpha(b) - \alpha(x)| \leq M$$

which implies $|\alpha(x) - \alpha(a)| \leq M$, or, equivalently, $\alpha(x) \leq \alpha(a) + M$. The same inequality is valid if $x = a$ or $x = b$. Lemma is proven. □

To work more exactly with functions of bounded variations we need the following definition.

Definition 15.5. *For a function α of bounded variation on $[a, b]$ the number*

$$\boxed{V_\alpha[a, b] := \sup_P \left\{ \sum_{k=1}^{n} |\Delta\alpha_k| \right\}} \tag{15.49}$$

(where sup *is taken over all possible partitions of $[a, b]$) is called the* ***total variation*** *of α on the interval $[a, b]$.*

The following properties of $V_\alpha[a, b]$ are evident:

1. Since α is of bounded variation the number $V_\alpha[a, b]$ is finite;
2.

$$\boxed{V_\alpha[a, b] \geq 0} \tag{15.50}$$

3.

$$V_\alpha[a, b] = 0$$

if and only if $\alpha(x) = \text{const}$ on $[a, b]$;

4.

$$\boxed{V_{\alpha+\beta}[a, b] \leq V_\alpha[a, b] + V_\beta[a, b]} \tag{15.51}$$

5.

$$\boxed{\begin{array}{l} V_{\alpha\cdot\beta}[a, b] \leq AV_\alpha[a, b] + BV_\beta[a, b] \\ A := \sup_{x\in[a,b]} |\beta(x)|, \quad B := \sup_{x\in[a,b]} |\alpha(x)| \end{array}} \tag{15.52}$$

6. If $c \in (a, b)$ then

$$\boxed{V_\alpha[a, b] = V_\alpha[a, c] + V_\alpha[c, b]} \tag{15.53}$$

7. If $x \in (a, b)$ then the function

$$\boxed{V(x) := V_\alpha[a, x]} \tag{15.54}$$

possesses the following properties:

(a)

$$V(a) = 0$$

(b) $V(x)$ is a nondecreasing function on $[a, b]$;
(c) $[V(x) - \alpha(x)]$ is a nondecreasing function on $[a, b]$;
(d) Any point of continuity of $\alpha(x)$ is a point of continuity of $V(x)$ and inversely.

The following theorem gives the simple and elegant characterization of functions of bounded variations.

Theorem 15.8. (on a difference of increasing functions) *Let α be defined on $[a, b]$. Then α is of bounded variation on $[a, b]$ if and only if α can be represented as the difference of two nondecreasing functions, namely, if and only if*

$$\boxed{\alpha(x) = \alpha^+(x) - \alpha^-(x)} \tag{15.55}$$

where $\alpha^+ \uparrow$ on $[a, b]$ and $\alpha^- \uparrow$ on $[a, b]$.

Proof. Define $\alpha^+(x) = V(x)$, where $V(x)$ is the function (15.54), and $\alpha^-(x) := V(x) - \alpha(x)$. By the statement 7(b–c) of the previous claim it follows that both $\alpha^+(x)$ and $\alpha^-(x)$ are nondecreasing which proves the theorem. □

Corollary 15.2. *If $\alpha(x)$ is continuous at the point x, then $\alpha^+(x)$ and $\alpha^-(x)$ are also continuous at x.*

Example 15.3. *Consider the function (see Fig. 15.3)*

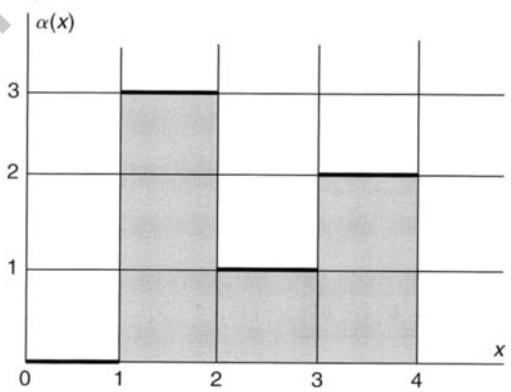

Fig. 15.3. The function of bounded variation.

$$\alpha(x) := \begin{cases} 0 & \text{if } 0 \le x < 1 \\ 3 & \text{if } 1 \le x < 2 \\ 1 & \text{if } 2 \le x < 3 \\ 2 & \text{if } 3 \le x \le 4 \end{cases}$$

Define (see Fig. 15.4)

$$\alpha^{+}(x) := \begin{cases} 0 & \text{if } 0 \le x < 1 \\ 3 & \text{if } 1 \le x < 2 \\ 4 & \text{if } 2 \le x < 3 \\ 6 & \text{if } 3 \le x \le 4 \end{cases}$$

and (see Fig. 15.5)

$$\alpha^{-}(x) := \begin{cases} 0 & \text{if } 0 \le x < 2 \\ 3 & \text{if } 3 \le x < 3 \\ 4 & \text{if } 3 \le x \le 4 \end{cases}$$

Then, it is clear that $\alpha(x) = \alpha^{+}(x) - \alpha^{-}(x)$.

Corollary 15.3. (Royden 1968) *For any function* $\alpha(x)$ *of bounded variation on* $[a, b]$ *and for each point* $c \in (a, b)$ *there exist* $\lim_{x \to c-0} \alpha(x)$ *and* $\lim_{x \to c+0} \alpha(x)$.

Corollary 15.4. (Royden 1968) *Any monotone function and, hence, any function of bounded variation on* $[a, b]$ *can have only a countable number of discontinuities.*

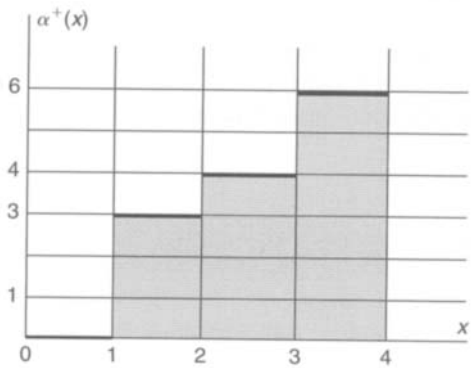

Fig. 15.4. The first nondecreasing function.

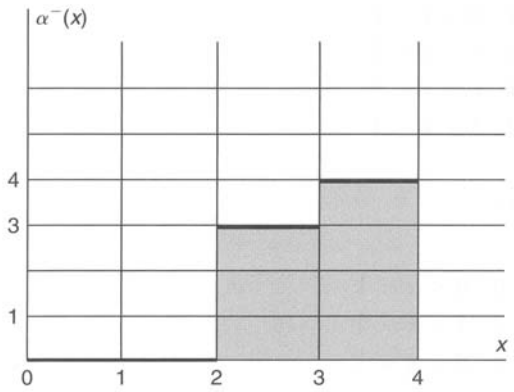

Fig. 15.5. The second nondecreasing function.

Proof. It follows from the fact that for any monotone function $\alpha(x)$ the number of points where

$$|\alpha(x+0) - \alpha(x-0)| > 1/s_n$$

$$s_n := \sum_{k=1}^{n} [\Delta\alpha_k]_-, \quad [x]_- := \min\{0; x\}$$

for any partition P_n is finite. □

Corollary 15.5. (Royden 1968) *If $\alpha(x)$ is a function of bounded variation on $[a, b]$, then $\alpha'(x)$ exists for almost all $x \in [a, b]$, that is, $\alpha(x)$ is differentiable almost everywhere on $[a, b]$.*

15.3 The Lebesgue–Stieltjes integral

The purpose of this section is to present the fundamental concepts of the Lebesgue theory of measure and integration and to prove some crucial theorems in a rather general setting without obscuring the main lines of the developments by a mass of comparatively trivial detail.

15.3.1 Algebras, σ-algebras and additive functions of sets

Definition 15.6. *A family $\mathfrak{F}$ of subsets of Ω is called an* ***algebra*** *(or a* ***ring****)* ***generated by*** *Ω, if for any finite $n < \infty$ and for any subsets $\mathcal{A}_i \in \Omega$ $(i = 1, \ldots, n)$*

1.

$$\boxed{\Omega \in \mathfrak{F}} \tag{15.56}$$

2.

$$\boxed{\bigcup_{k=1}^{n} \mathcal{A}_k \in \mathfrak{F}} \tag{15.57}$$

3.

$$\boxed{\bigcap_{k=1}^{n} \mathcal{A}_k \in \mathfrak{F}} \tag{15.58}$$

4. *for any* $\mathcal{A} \in \Omega$

$$\boxed{\overline{\mathcal{A}} \triangleq \Omega - \mathcal{A} \in \mathfrak{F}} \tag{15.59}$$

Definition 15.7. *A system* $\mathcal{F}$ *of subsets of* Ω *is called* ***an*** σ***-algebra*** *(or a* σ***-ring)*** ***generated by*** Ω*, if*

1. *it is algebra;*
2. *for any sequences of subsets* $\{\mathcal{A}_i\}$, $\mathcal{A}_i \in \mathcal{F}$

$$\boxed{\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}, \quad \bigcap_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}} \tag{15.60}$$

Definition 15.8.

1. *A set function* $\phi : \mathcal{F} \to \mathbb{R}$ *defined for every* $\mathcal{A} \in \mathcal{F}$ *is said to be* ***additive*** *if* $\mathcal{A} \cap \mathcal{B} = \varnothing$ *and* $\mathcal{B} \in \mathcal{F}$ *implies*

$$\boxed{\phi(\mathcal{A} \cup \mathcal{B}) = \phi(\mathcal{A}) + \phi(\mathcal{B})} \tag{15.61}$$

2. *A set function* $\phi : \mathcal{F} \to \mathbb{R}$ *defined for every* $\mathcal{A}$ *from a* σ*-algebra* $\mathcal{F}$ *is said to be* ***countably additive*** *if* $\mathcal{A}_i \bigcap\limits_{i \neq j} \mathcal{A}_j = \varnothing$ *and* $\mathcal{A}_i, \mathcal{A}_j \in \mathcal{F}$ *implies*

$$\boxed{\phi\left(\bigcup_{i=1}^{\infty} \mathcal{A}_i\right) = \sum_{i=1}^{\infty} \phi(\mathcal{A}_i)} \tag{15.62}$$

We shall also assume that

- the range of ϕ does not contain both $(+\infty)$ and $(-\infty)$, for if it did the right-hand side of (15.61) could become meaningless;
- we exclude functions whose only value is $(+\infty)$ or $(-\infty)$.

Assuming, in addition, for an *additive* ϕ that

(a)

$$\boxed{\phi(\varnothing) = 0} \tag{15.63}$$

(b) for all $\mathcal{A} \in \mathcal{F}$

$$\boxed{\phi(\mathcal{A}) \geq 0} \tag{15.64}$$

the following properties are easily verified.

Proposition 15.4.

1. If $\mathcal{A}_i \underset{i \neq j}{\cap} \mathcal{A}_j = \varnothing$ *and* $\mathcal{A}_i, \mathcal{A}_j \in \mathcal{F}$ *then*

$$\boxed{\phi\left(\bigcup_{i=1}^{n} \mathcal{A}_i\right) = \sum_{i=1}^{n} \phi(\mathcal{A}_i)} \tag{15.65}$$

2. For any $\mathcal{A}_1, \mathcal{A}_2 \in \mathcal{F}$ *we have*

$$\boxed{\phi(\mathcal{A}_1 \cup \mathcal{A}_2) + \phi(\mathcal{A}_1 \cap \mathcal{A}_2) = \phi(\mathcal{A}_1) + \phi(\mathcal{A}_2)} \tag{15.66}$$

3. If $\mathcal{A}_1 \subset \mathcal{A}_2 \in \mathcal{F}$ *then*

$$\boxed{\phi(\mathcal{A}_1) \leq \phi(\mathcal{A}_2)} \tag{15.67}$$

4. If $\mathcal{B} \subset \mathcal{A} \in \mathcal{F}$ *and* $\phi(\mathcal{B}) < \infty$ *then*

$$\boxed{\phi(\mathcal{A} - \mathcal{B}) = \phi(\mathcal{A}) - \phi(\mathcal{B})} \tag{15.68}$$

For *countably additive* ϕ the following result holds.

Theorem 15.9. *Suppose* ϕ *is* ***countably additive*** *on* σ*-algebra* $\mathcal{F}$ *and* $\mathcal{A}_i \in \mathcal{F}$ $(i = 1, 2, \ldots)$, $\mathcal{A}_1 \subset \mathcal{A}_2 \subset \cdots$, $\mathcal{A} := \bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathcal{F}$. *Then, as* $n \to \infty$

$$\boxed{\phi(\mathcal{A}_n) \to \phi(\mathcal{A})} \tag{15.69}$$

Proof. Define $\mathcal{B}_1 := \mathcal{A}_1$ and $\mathcal{B}_n := \mathcal{A}_n - \mathcal{A}_{n-1}$ $(n = 2, 3, \ldots)$. Then

$$\mathcal{B}_i \underset{i \neq j}{\cap} \mathcal{B}_j = \varnothing, \quad \mathcal{A}_n = \mathcal{B}_1 \cup \mathcal{B}_2 \cup \cdots \cup \mathcal{B}_n, \quad \mathcal{A} = \bigcup_{i=1}^{\infty} \mathcal{B}_i$$

Hence, $\phi(\mathcal{A}_n) = \sum_{i=1}^{n} \phi(\mathcal{B}_i)$ and $\phi(\mathcal{A}) = \sum_{i=1}^{\infty} \phi(\mathcal{B}_i)$. Theorem is proven. □

15.3.2 Measure theory

This subsection deals with construction of the, so-called, *Lebesgue measure* which plays the key role in the definition of the Lebesgue–Stieltjes integral.

15.3.2.1 Intervals

Definition 15.9. ***Intervals*** *in p-dimensional Euclidean space* $\mathbb{R}^p$ *are defined as follows:*

1. *the* ***closed interval***

$$[a,b] := \left\{x := \{x_1,\ldots,x_p\} : a_i \le x_i \le b_i \ \ \overline{(i=1,f)}\right\}$$

2. *the* ***semi-open interval***

$$[a,b) := \left\{x := \{x_1,\ldots,x_p\} : a_i \le x_i < b_i \ \ \overline{(i=1,f)}\right\}$$

or

$$(a,b] := \left\{x := \{x_1,\ldots,x_p\} : a_i < x_i \le b_i \ \ \overline{(i=1,f)}\right\}$$

3. *the* ***open interval***

$$(a,b) := \left\{x := \{x_1,\ldots,x_p\} : a_i < x_i < b_i \ \ \overline{(i=1,f)}\right\}$$

The possibility that $a_i = b_i$ for any value of i is not ruled out; in particular, the empty set is included among the intervals. If $\mathcal{A}$ is a union of a finite number of intervals it is called an *elementary set*.

15.3.2.2 Additive set functions

Definition 15.10. *If I is an interval, we define*

$$\boxed{m(I) := \prod_{i=1}^{p} (b_i - a_i)} \tag{15.70}$$

and if $\mathcal{A} = I_1 \cup \cdots \cup I_p$ *then we set*

$$\boxed{m(\mathcal{A}) := m(I_1) + \cdots + m(I_p)} \tag{15.71}$$

We let $\mathcal{E}$ *denote the family of elementary subsets of* $\mathbb{R}^p$.

At this point, the following properties should be easily verified.

Proposition 15.5.

1. $\mathcal{E}$ *is algebra (ring), but not a* σ*-algebra;*
2. *If* $\mathcal{A} \in \mathcal{E}$ *then* $\mathcal{A}$ *is a union of a finite number of* ***disjoint*** *intervals;*
3. *If* $\mathcal{A} \in \mathcal{E}$ *then* $m(\mathcal{A})$ *is well defined by (15.71) on* $\mathcal{E}$*;*
4. m *is additive on* $\mathcal{E}$.

Remark 15.4. *If* $p = 1, 2, 3$ *then* m *is* ***length***, ***area*** *and* ***volume***, *respectively.*

Definition 15.11. *A nonnegative additive set function $\phi : \mathcal{E} \to \mathbb{R}$ defined on $\mathcal{E}$ is said to be* ***regular*** *if for every $\mathcal{A} \in \mathcal{E}$ and every $\varepsilon > 0$ there exist sets $\mathcal{F}, \mathcal{G} \in \mathcal{E}$ such that $\mathcal{F}$ is closed, $\mathcal{G}$ is open, $\mathcal{F} \subset \mathcal{A} \subset \mathcal{G}$ and*

$$\boxed{\phi(\mathcal{G}) - \varepsilon \leq \phi(\mathcal{A}) \leq \phi(\mathcal{F}) + \varepsilon} \tag{15.72}$$

Example 15.4.

1. *If $\mathcal{A} = I$ is an interval then m (15.70) is regular.*
2. *For $p = 1$ let $\alpha : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function possibly having discontinuity points. Put*

$$\boxed{\begin{array}{l} \mu([a,b]) := \alpha(b+0) - \alpha(a-0) \\ \mu([a,b)) := \alpha(b-0) - \alpha(a-0) \\ \mu((a,b]) := \alpha(b+0) - \alpha(a+0) \\ \mu((a,b)) := \alpha(b-0) - \alpha(a+0) \end{array}} \tag{15.73}$$

μ is regular on $\mathcal{E}$.

15.3.2.3 Countably additive set functions

Our next objective is to show that every regular set function on $\mathcal{E}$ can be extended to a countably additive set function on σ-algebra containing $\mathcal{E}$.

Definition 15.12. *Define*

$$\boxed{\mu^*(\mathcal{E}) := \inf \sum_{n=1}^{\infty} \mu(\mathcal{A}_n)} \tag{15.74}$$

where $\bigcup_{i=1}^{\infty} \mathcal{A}_i$ is a countable covering of $\mathcal{E} \subset \mathbb{R}^p$ by open elementary sets $\mathcal{A}_n$, that is, $\mathcal{E} \subset \bigcup_{i=1}^{\infty} \mathcal{A}_i$, μ is additive, regular, nonnegative and finite on $\mathcal{E}$, and inf *being taken over all countable coverings of $\mathcal{E}$ by open elementary set. $\mu^*(\mathcal{E})$ is called the* ***outer measure*** *of $\mathcal{E}$ corresponding to μ.*

Theorem 15.10.

1. *For every $\mathcal{A} \in \mathcal{E}$*

$$\boxed{\mu^*(\mathcal{A}) = \mu(\mathcal{A})} \tag{15.75}$$

2. *The following* ***subadditivity*** *property holds: if $E = \bigcup_{i=1}^{\infty} E_i$ then*

$$\boxed{\mu^*(E) \leq \sum_{i=1}^{\infty} \mu^*(E_i)} \tag{15.76}$$

Proof.

1. Choose $\mathcal{A} \in \mathcal{E}$ and $\varepsilon > 0$. By the regularity of μ, $\mathcal{A}$ is contained in an open elementary set $\mathcal{G}$ such that $\mu(\mathcal{G}) \leq \mu(\mathcal{A}) + \varepsilon$. Since $\mu^*(\mathcal{A}) \leq \mu(\mathcal{G})$ and arbitrarity it follows that

$$\mu^*(\mathcal{A}) \leq \mu(\mathcal{A}) \tag{15.77}$$

By the definition (15.74) there is a sequence $\{\mathcal{A}_n\}$ of open elementary sets whose union contains $\mathcal{A}$ such that $\sum_{n=1}^{\infty} \mu(\mathcal{A}_n) \leq \mu^*(\mathcal{A}) + \varepsilon$. The regularity of μ shows also that $\mathcal{A}$ contains a closed elementary set such that $\mu(\mathcal{F}) \geq \mu(\mathcal{A}) - \varepsilon$. Since $\mathcal{F}$ is a compact we have $\mathcal{F} \subset \mathcal{A}_1 \cup \cdots \cup \mathcal{A}_N$ for some N. Hence,

$$\begin{aligned} \mu(\mathcal{A}) &\leq \mu(\mathcal{F}) + \varepsilon \leq \mu(\mathcal{A}_1 \cup \cdots \cup \mathcal{A}_N) + \varepsilon \\ &\leq \sum_{n=1}^{N} \mu(\mathcal{A}_n) + \varepsilon \leq \mu^*(\mathcal{A}) + 2\varepsilon \end{aligned} \tag{15.78}$$

which, in conjunction with (15.77), proves (15.75).

2. Suppose $E = \bigcup_{i=1}^{\infty} E_i$ and $\mu^*(E_n) < \infty$ for all n. Given $\varepsilon > 0$ there are covering sets $\{A_{nk}\}_{k=1,2,\ldots}$ of E_n by open elementary sets such that $\sum_{k=1}^{\infty} \mu(A_{nk}) \leq \mu^*(E_n) + 2^{-n}\varepsilon$ which leads to the inequality

$$\begin{aligned} \mu^*(E) &\leq \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \mu(A_{nk}) \leq \sum_{n=1}^{\infty} \mu^*(E_n) \\ &+ \sum_{n=1}^{\infty} 2^{-n}\varepsilon = \sum_{n=1}^{\infty} \mu^*(E_n) + \varepsilon \end{aligned}$$

and (15.76) follows. In the excluded case when $\mu^*(E_n) = \infty$ for some n, (15.76) trivially holds. Theorem is proven.

□

15.3.2.4 μ-measurable sets

Definition 15.13.

1. *For any $\mathcal{A}, \mathcal{B} \subset \mathbb{R}^p$ let us define the set $S(\mathcal{A}, \mathcal{B})$, called the **symmetric difference** of $\mathcal{A}$ and $\mathcal{B}$, as*

$$\boxed{S(\mathcal{A}, \mathcal{B}) := (\mathcal{A} - \mathcal{B}) \cup (\mathcal{B} - \mathcal{A})} \tag{15.79}$$

2. *The **distance function** (metric) is defined as follows*

$$\boxed{d(\mathcal{A}, \mathcal{B}) := \mu^*(S(\mathcal{A}, \mathcal{B}))} \tag{15.80}$$

3. *We will write* $\mathcal{A}_n \to \mathcal{A}$ *when* $n \to \infty$ *if*

$$\boxed{\lim_{n\to\infty} d(\mathcal{A}_n, \mathcal{A}) = 0} \tag{15.81}$$

4. *If there is a sequence* $\{\mathcal{A}_n\}$ *of elementary sets such that* $\mathcal{A}_n \to \mathcal{A}$*, we say that* $\mathcal{A}$ *is* ***finitely*** μ***-measurable*** *and write*

$$\boxed{\mathcal{A} \in \mathfrak{M}_{fin}(\mu)} \tag{15.82}$$

5. *If* $\mathcal{A}$ *is the union of a countable collection of finitely* μ*-measurable sets, we say that* $\mathcal{A}$ *is* μ***-measurable*** *and write*

$$\boxed{\mathcal{A} \in \mathfrak{M}(\mu)} \tag{15.83}$$

Some properties of $S(\mathcal{A}, \mathcal{B})$ are summarized in the following claim.

Claim 15.1.

1.

$$\boxed{S(\mathcal{A}, \mathcal{A}) = \varnothing} \tag{15.84}$$

2.

$$\boxed{S(\mathcal{A}, \mathcal{B}) = S(\mathcal{B}, \mathcal{A})} \tag{15.85}$$

3.

$$\boxed{S(\mathcal{A}, \mathcal{B}) \subset S(\mathcal{A}, \mathcal{C}) \cup S(\mathcal{C}, \mathcal{B})} \tag{15.86}$$

which follows from

$$(\mathcal{A} - \mathcal{B}) \subset (\mathcal{A} - \mathcal{C}) \cup (\mathcal{C} - \mathcal{B})$$
$$(\mathcal{B} - \mathcal{A}) \subset (\mathcal{C} - \mathcal{A}) \cup (\mathcal{B} - \mathcal{C})$$

4.

$$\boxed{\left.\begin{array}{l} S(\mathcal{A}_1 \cup \mathcal{A}_2, \mathcal{B}_1 \cup \mathcal{B}_2) \\ S(\mathcal{A}_1 \cap \mathcal{A}_2, \mathcal{B}_1 \cap \mathcal{B}_2) \\ S(\mathcal{A}_1 - \mathcal{A}_2, \mathcal{B}_1 - \mathcal{B}_2) \end{array}\right\} \subset S(\mathcal{A}_1, \mathcal{B}_1) \cup S(\mathcal{A}_2, \mathcal{B}_2)} \tag{15.87}$$

which follows from

$$(\mathcal{A}_1 \cup \mathcal{A}_2) - (\mathcal{B}_1 \cup \mathcal{B}_2) \subset (\mathcal{A}_1 - \mathcal{B}_1) \cup (\mathcal{A}_2 - \mathcal{B}_2)$$

$$S(\mathcal{A}_1 \cap \mathcal{A}_2, \mathcal{B}_1 \cap \mathcal{B}_2) = S\left(\mathcal{A}_1^c \cup \mathcal{A}_2^c, \mathcal{B}_1^c \cup \mathcal{B}_2^c\right) \subset$$

$$S\left(\mathcal{A}_1^c, \mathcal{B}_1^c\right) \cup S\left(\mathcal{A}_2^c, \mathcal{B}_2^c\right) = S(\mathcal{A}_1, \mathcal{B}_1) \cup S(\mathcal{A}_2, \mathcal{B}_2)$$

where $\mathcal{A}^c := \mathbb{R}^p - \mathcal{A}$ *is the complement of* $\mathcal{A}$

$$\mathcal{A}_1 - \mathcal{A}_2 = \mathcal{A}_1 \cap \mathcal{A}_2^c$$

The next properties of $d(\mathcal{A}, \mathcal{B})$ can be checked directly from the definition (15.80).

Claim 15.2.

1.

$$\boxed{d(\mathcal{A}, \mathcal{A}) = 0} \tag{15.88}$$

2.

$$\boxed{d(\mathcal{A}, \mathcal{B}) = d(\mathcal{B}, \mathcal{A})} \tag{15.89}$$

3.

$$\boxed{d(\mathcal{A}, \mathcal{B}) \leq d(\mathcal{A}, \mathcal{C}) + d(\mathcal{C}, \mathcal{B})} \tag{15.90}$$

4.

$$\boxed{\left.\begin{array}{l} d(\mathcal{A}_1 \cup \mathcal{A}_2, \mathcal{B}_1 \cup \mathcal{B}_2) \\ d(\mathcal{A}_1 \cap \mathcal{A}_2, \mathcal{B}_1 \cap \mathcal{B}_2) \\ d(\mathcal{A}_1 - \mathcal{A}_2, \mathcal{B}_1 - \mathcal{B}_2) \end{array}\right\} \leq d(\mathcal{A}_1, \mathcal{B}_1) + d(\mathcal{A}_2, \mathcal{B}_2)} \tag{15.91}$$

which follows from (15.87);

5.

$$\boxed{|\mu^*(\mathcal{A}) - \mu^*(\mathcal{B})| \leq d(\mathcal{A}, \mathcal{B})} \tag{15.92}$$

*6. If $d(\mathcal{A}, \mathcal{B}) = 0$ this **does not imply** $\mathcal{A} = \mathcal{B}$. By this property $d(\mathcal{A}, \mathcal{B})$ is "quasi-metric".*

The next theorem will enable us to obtain the desired extension of the measure μ (15.73).

Theorem 15.11. (The main theorem on a measure extension) *$\mathfrak{M}(\mu)$, defined by (15.83), is a σ-algebra (σ-ring) and μ^* (15.74) is countably additive on $\mathfrak{M}(\mu)$.*

Proof.

(a) Let $\mathcal{A}, \mathcal{B} \in \mathfrak{M}_{fin}(\mu)$. Choose $\{\mathcal{A}_n\}$, $\{\mathcal{B}_n\}$ such that $\mathcal{A}_n, \mathcal{B}_n \in \mathcal{E}$ and $\mathcal{A}_n \to \mathcal{A}$, $\mathcal{B}_n \to \mathcal{B}$. Then by (15.91) and (15.92) $\mathcal{A}_n \cup \mathcal{B}_n \to \mathcal{A} \cup \mathcal{B}$, $\mathcal{A}_n \cap \mathcal{B}_n \to \mathcal{A} \cap \mathcal{B}$, $\mathcal{A}_n - \mathcal{B}_n \to \mathcal{A} - \mathcal{B}$, $\mu^*(\mathcal{A}_n) \to \mu^*(\mathcal{A})$. Also, $\mu^*(\mathcal{A}) < \infty$ since $d(\mathcal{A}_n, \mathcal{A}) \to 0$. This implies that $\mathfrak{M}_{fin}(\mu)$ is an algebra (ring). By (15.66) we have $\mu(\mathcal{A}_n) + \mu(\mathcal{B}_n) = \mu(\mathcal{A}_n \cup \mathcal{B}_n) + \mu(\mathcal{A}_n \cap \mathcal{B}_n)$. Letting $n \to \infty$ we obtain $\mu^*(\mathcal{A}) + \mu^*(\mathcal{B}) = \mu^*(\mathcal{A} \cup \mathcal{B}) + \mu^*(\mathcal{A} \cap \mathcal{B})$. If $\mathcal{A} \cap \mathcal{B} = \varnothing$ then $\mu^*(\mathcal{A} \cap \mathcal{B}) = 0$. So, μ^* is additive on $\mathfrak{M}_{fin}(\mu)$.

(b) Let $\mathcal{A} \in \mathfrak{M}(\mu)$. Then $\mathcal{A}$ can be represented as the union of countable collection of disjoint sets of $\mathfrak{M}_{fin}(\mu)$ and for $\mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{A}'_n$ with $\mathcal{A}'_n \in \mathfrak{M}(\mu)$. Define $\mathcal{A}_1 = \mathcal{A}'_1$ and

$$\mathcal{A}_n := \left(\mathcal{A}'_1 \cup \mathcal{A}'_2 \cup \cdots \cup \mathcal{A}'_n\right) - \left(\mathcal{A}'_1 \cup \mathcal{A}'_2 \cup \cdots \cup \mathcal{A}'_{n-1}\right)$$

Then $\mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{A}_n$ is the required representation. By the subadditivity property (15.76) $\mu^*(\mathcal{A}) \leq \sum_{i=1}^{\infty} \mu^*(\mathcal{A}_i)$. On the other hand,

$$\mathcal{A} \supset (\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_n)$$

and, by the additivity of μ^*on $\mathfrak{M}_{fin}(\mu)$, we have

$$\mu^*(\mathcal{A}) \geq \mu^*(\mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_n) = \sum_{i=1}^{n} \mu^*(\mathcal{A}_i)$$

which implies

$$\mu^*(\mathcal{A}) = \sum_{i=1}^{\infty} \mu^*(\mathcal{A}_i) \tag{15.93}$$

Suppose $\mu^*(\mathcal{A})$ is finite. Put $\mathcal{B}_n := \mathcal{A}_1 \cup \mathcal{A}_2 \cup \cdots \cup \mathcal{A}_n$. Then

$$d(\mathcal{A}, \mathcal{B}_n) := \mu^*\left(\bigcup_{i=n+1}^{\infty} \mathcal{A}_i\right) = \sum_{i=n+1}^{\infty} \mu^*(\mathcal{A}_i) \to 0$$

as $n \to \infty$. This means that $\mathcal{B}_n \to \mathcal{A}$ and, since $\mathcal{B}_n \in \mathfrak{M}_{fin}(\mu)$, it is easily seen that $\mathcal{A} \in \mathfrak{M}_{fin}(\mu)$. So, we have thus shown that $\mathcal{A} \in \mathfrak{M}_{fin}(\mu)$ if $\mathcal{A} \in \mathfrak{M}(\mu)$ and $\mu^*(\mathcal{A}) < \infty$. Now it is evident that μ^* is countably additive on $\mathfrak{M}(\mu)$. For if $\mathcal{A} = \bigcup_{i=1}^{\infty} \mathcal{A}_i$ where $\{\mathcal{A}_i\}$ is a sequence of disjoint sets of $\mathfrak{M}(\mu)$ we have just shown that (15.93) holds if $\mu^*(\mathcal{A}_n) < \infty$ for any $n = 1, 2, \ldots$, and, in other cases, it looks trivial.

(c) Finally, we have to show that $\mathfrak{M}(\mu)$ is σ-algebra (σ-ring). If $\mathcal{A}_n \in \mathfrak{M}(\mu)$, it is clear that $\bigcup_{i=1}^{\infty} \mathcal{A}_i \in \mathfrak{M}(\mu)$. Suppose $\mathcal{A}, \mathcal{B} \in \mathfrak{M}(\mu)$ where $\mathcal{A} = \bigcup_{i=1}^{\infty} \mathcal{A}_i$, $\mathcal{B} = \bigcup_{i=1}^{\infty} \mathcal{B}_i$ and $\mathcal{A}_n, \mathcal{B}_n \in \mathfrak{M}_{fin}(\mu)$. Then the identity $\mathcal{A}_n \cap \mathcal{B} = \bigcup_{i=1}^{\infty} (\mathcal{A}_n \cap \mathcal{B}_i)$ implies that $(\mathcal{A}_n \cap \mathcal{B}) \in \mathfrak{M}(\mu)$. In view of $\mu^*(\mathcal{A}_n \cap \mathcal{B}) \leq \mu^*(\mathcal{A}_n) < \infty$ we have $(\mathcal{A}_n \cap \mathcal{B}) \in \mathfrak{M}_{fin}(\mu)$ and, hence, $(\mathcal{A}_n - \mathcal{B}) \in \mathfrak{M}_{fin}(\mu)$, and $(\mathcal{A} - \mathcal{B}) \in \mathfrak{M}(\mu)$ since $(\mathcal{A} - \mathcal{B}) = \bigcup_{n=1}^{\infty} (\mathcal{A}_n - \mathcal{B})$.

Theorem is proven. □

So, now we may replace $\mu^*(\mathcal{A})$ by $\mu(\mathcal{A})$ if $\mathcal{A} \in \mathfrak{M}(\mu)$ and thus μ, originally defined only on $\mathcal{E}$, is extended to a countable additive set function defined on the σ-algebra $\mathfrak{M}(\mu)$.

Corollary 15.6.

(a) If $\mathcal{A}$ is open, then $\mathcal{A} \in \mathfrak{M}(\mu)$ since for every open set in $\mathbb{R}^p$ there is the union of a countable collection of open intervals.
(b) Every closed set $\mathcal{A}$ is also in $\mathfrak{M}(\mu)$ which follows from previous comment by taking complements.
(c) If $\mathcal{A} \in \mathfrak{M}(\mu)$ and $\varepsilon > 0$ there exist a closed set $\mathcal{F}$ and an open set $\mathcal{G}$ such that

$$\boxed{\mathcal{F} \subset \mathcal{A} \subset \mathcal{G}} \tag{15.94}$$

and

$$\boxed{\mu(\mathcal{G} - \mathcal{A}) < \varepsilon,\ \mu(\mathcal{A} - \mathcal{F}) < \varepsilon} \tag{15.95}$$

Now we are ready to give the main definition of this section.

Definition 15.14.

*(a) Such extended set function μ^* (15.74) is called a **countably additive measure**.*
*(b) The special case $\mu = m$ (see (15.74)) is called the **Lebesgue measure** on $\mathbb{R}^p$.*

15.3.2.5 Borel sets

Definition 15.15. *$\mathcal{E}$ is said to be a **Borel set** if $\mathcal{E}$ can be obtained by a countable number of operations, starting from **open** sets, each operation consisting of taking unions, intersections, or complements.*

The difference between a Borel set and σ-algebra (ring) (15.7) is that Ω in the case of a Borel set must be an open set.

The following facts take place for Borel sets.

Claim 15.3.

1. *The collection $\mathfrak{B}$ of all Borel sets in $\mathbb{R}^p$ is a σ-algebra (ring). In fact, it is the smallest σ-algebra (ring) which contains all open sets, that is, if $\mathcal{E} \in \mathfrak{B}$ then $\mathcal{E} \in \mathfrak{M}(\mu)$.*
2. *If $\mathcal{A} \in \mathfrak{M}(\mu)$, there exist Borel sets $\mathcal{F}$ and $\mathcal{G}$ such that $\mathcal{F} \subset \mathcal{A} \subset \mathcal{G}$ and*

$$\boxed{\mu(\mathcal{G} - \mathcal{A}) = \mu(\mathcal{A} - \mathcal{F}) = 0} \tag{15.96}$$

   This follows from (15.95) if we take $\varepsilon = 1/n$ and let $n \to \infty$.
3. *If $\mathcal{A} = \mathcal{F} \cup (\mathcal{A} - \mathcal{F})$ one can see that $\mathcal{A} \in \mathfrak{M}(\mu)$ is the union of a Borel set and a set of **measure zero**.*
4. *Borel sets are μ-measurable for every μ (for details see below), but the sets of measure zero (that is, the sets $\mathcal{E}$ for which $\mu^*(\mathcal{E}) = 0$) may be different for different μ's.*
5. *For every μ the sets of measure zero from σ-algebra (ring).*
6. *In the case of the Lebesgue measure ($\mu = m$) every countable set has measure zero. But there are uncountable (in fact, perfect) sets of measure zero (see Rudin (1976) Chapter 11 with the Cantor set as an example).*

15.3.3 Measurable spaces and functions

15.3.3.1 Measurable spaces

Consider $\mathcal{X}$ which is a set, not necessarily a subset, of a Euclidean space, or indeed of any metric space.

Definition 15.16.

- *$\mathcal{X}$ is said to be a **measure space** if there exist a σ-algebra (ring) $\mathfrak{M}$ of subsets of $\mathcal{X}$ (which are called measurable sets) and a nonnegative countable additive function μ (which is called a **measure**) defined on $\mathfrak{M}$.*
- *If, in addition, $\mathcal{X} \in \mathfrak{M}$ then $\mathcal{X}$ is called a **measurable space**.*

Example 15.5.

1. *Take $\mathcal{X} = \mathbb{R}^p$, then $\mathfrak{M}$ is the collection of all Lebesgue measurable subsets of $\mathbb{R}^p$ and μ is the Lebesgue measure.*
2. *Let $\mathcal{X}$ be the set of all positive integers. Then $\mathfrak{M}$ is the collection of all subsets of $\mathcal{X}$ and $\mu(\mathcal{E})$ is the number of elements of $\mathcal{E}$.*
3. *Another example is provided by probability theory where events are considered as sets and the corresponding probability of the occurrence of events is a countably additive set function.*

15.3.3.2 Measurable functions

Definition 15.17. *Let $f : \mathcal{X} \to \mathbb{R}$ be a function, defined on the measurable space $\mathcal{X}$, with values in $\mathbb{R}$. The function f is said to be **measurable** if the set $\{x \mid f(x) > a\}$ is measurable for every real a, that is, when for any $a \in \mathbb{R}$*

$$\boxed{\{x \mid f(x) > a\} \subset \mathcal{X}} \tag{15.97}$$

Example 15.6. *If $\mathcal{X} = \mathbb{R}^p$ with $\mathfrak{M} = \mathfrak{M}(\mu)$ defined in (15.83) then any continuous function f is measurable, since (15.97) is an open set.*

Lemma 15.4. *Each of the following four conditions implies the other three:*

1. *for every real a*

$$\{x \mid f(x) > a\} \text{ is measurable} \tag{15.98}$$

2. *for every real a*

$$\{x \mid f(x) \geq a\} \text{ is measurable} \tag{15.99}$$

3. *for every real a*

$$\{x \mid f(x) < a\} \text{ is measurable} \tag{15.100}$$

4. for every real a

$$\{x \mid f(x) \leq a\} \text{ is measurable} \tag{15.101}$$

Proof. The relations

$$\{x \mid f(x) \geq a\} = \bigcap_{n=1}^{\infty} \{x \mid f(x) \geq a - 1/n\}$$
$$\{x \mid f(x) < a\} = \mathcal{X} - \{x \mid f(x) \geq a\}$$
$$\{x \mid f(x) \leq a\} = \bigcap_{n=1}^{\infty} \{x \mid f(x) < a - 1/n\}$$
$$\{x \mid f(x) > a\} = \mathcal{X} - \{x \mid f(x) \leq a\}$$

being applied successfully demonstrate that (15.15) implies (15.99), (15.99) implies (15.100), (15.100) implies (15.101) and (15.101) implies (15.15). Lemma is proven. □

Lemma 15.5. *If f is measurable then $|f|$ is measurable.*

Proof. It follows from the relation

$$\{x \mid |f(x)| < a\} = \{x \mid f(x) < a\} \cap \{x \mid f(x) > -a\}$$

and the previous lemma. □

Theorem 15.12. *Let $\{f_n\}$ be a sequence of measurable functions. Then*

$$g(x) := \sup_n f_n(x)$$

and

$$h(x) := \limsup_n f_n(x)$$

are measurable too.

Proof. Indeed,

$$\{x \mid g(x) > a\} = \bigcup_{n=1}^{\infty} \{x \mid f_n(x) > a\}$$

and

$$h(x) = \lim_{m \to \infty} \sup_{n \geq m} f_n(x) = \inf_n \sup_{n \geq m} f_n(x)$$

which implies the desired result. □

Corollary 15.7.

(a) If f and g are measurable then $\max\{f, g\}$ *and* $\min\{f, g\}$ *are measurable.*
(b) If

$$\boxed{f^+ := \max\{f, 0\} \quad \textit{and} \quad f^- := -\min\{f, 0\}} \tag{15.102}$$

then it follows that f^+ and f^- are measurable.
(c) The limit of a convergent sequence of measurable functions is measurable.

Theorem 15.13. *Let f and g be measurable real-valued functions defined on $\mathcal{X}$, and let F be real and continuous on $\mathbb{R}^2$. Put*

$$h(x) := F(f(x), g(x))$$

Then h is measurable and, in particular, $(f + g)$ and $(f \cdot g)$ are measurable.

Proof. Define $\mathcal{G}_n := \{(u, v) \mid F(u, v) > a\}$. Then $\mathcal{G}_n$ is an open subset of $\mathbb{R}^2$ which can be represented as $\mathcal{G}_n = \bigcup_{n=1}^{\infty} \{x \mid f_n(x) > a\}$ where

$$I_n := \{(u, v) \mid a_n < u < b_n,\ c_n < v < d_n\}$$

Since the set

$$\{x \mid a_n < f(x) < b_n\} = \{x \mid f(x) > a_n\} \cap \{x \mid f(x) < b_n\}$$

is measurable, it follows that the set

$$\{x \mid (f(x), g(x)) \in I_n\} = \{x \mid a_n < f(x) < b_n\} \cap \{x \mid c_n < g(x) < d_n\}$$

is measurable too. Hence, the same is true for the set

$$\begin{aligned}\{x \mid h(x) > a\} &= \{x \mid (f(x), g(x)) \in G_n\} \\ &= \bigcup_{n=1}^{\infty} \{x \mid (f(x), g(x)) \in I_n\}\end{aligned}$$

which completes the proof. □

Summary 15.1.

(a) *Summing up, we may say that all ordinary operations of analysis, including limit operations, being applied to measurable functions, lead to measurable functions as well. In other words, all functions that are ordinarily met with are measurable. But this is, however, only a rough statement since, for example, the function $h(x) = f(g(x))$, where f is measurable and g is continuous, is not necessarily measurable.*

(b) *The concrete measure has not been mentioned in the discussions above. In fact, the class of measurable functions on $\mathcal{X}$ depends only on the σ-algebra (ring) $\mathfrak{M}$. That's why we may speak of Borel-measurable functions on $\mathbb{R}^p$, that is, of functions for which the set $\{x \mid f(x) > a\}$ is always a Borel set, without reference to any particular measure.*

15.3.4 The Lebesgue–Stieltjes integration

15.3.4.1 Simple functions

Definition 15.18.

1. *If the range of a real-valued function $s : \mathcal{X} \to \mathbb{R}$, defined on $\mathcal{X}$, is finite, we say that s is a* ***simple function****.*
2. *Define the* ***characteristic function*** *χ_E of a set $\mathcal{E} \subset \mathcal{X}$ as follows:*

$$\boxed{\chi_{\mathcal{E}}(x) := \begin{cases} 1 & if \quad x \in \mathcal{E} \\ 0 & if \quad x \notin \mathcal{E} \end{cases}} \tag{15.103}$$

It is evident that if the range of a simple function s consists of the distinct numbers $c_1, c_2, \ldots, c_n$ then s can be represented as a finite linear combination of characteristic functions, namely,

$$\boxed{s(x) = \sum_{i=1}^{n} c_i \chi_{\mathcal{E}_i}(x)} \tag{15.104}$$

where

$$\boxed{\mathcal{E}_i := \{x \mid s(x) = c_i\}, \ i = 1, 2, \ldots n} \tag{15.105}$$

It is clear by the construction that s is measurable if and only if the sets $\mathcal{E}_i$ $(i = 1, 2, \ldots n)$ are measurable.

The next theorem shows that any function can be approximated by simple functions.

Theorem 15.14. *Let $f : \mathcal{X} \to \mathbb{R}$ be a real function on $\mathcal{X}$. There exists a sequence $\{s_n\}$ of simple functions such that $s_n(x) \to f(x)$ as $n \to \infty$ for every $x \in \mathcal{X}$. If f is measurable, $\{s_n\}$ can be chosen to be a sequence of measurable functions. If $f \geq 0$, $\{s_n\}$ can be chosen as a monotonically nondecreasing sequence.*

Proof. For $f \geq 0$ define

$$\mathcal{E}_{n,i}(x) := \{x \mid (i-1)/2^n \leq f(x) \leq i/2^n\}$$
$$\mathcal{F}_n(x) := \{x \mid f(x) \geq n\}$$

for $i = 1, 2, \ldots, n2^n$ and $n = 1, 2, \ldots$. Put

$$s_n(x) := \sum_{i=1}^{n2^n} \frac{(i-1)}{2^n} \chi_{\mathcal{E}_{n,i}}(x) + n\chi_{\mathcal{F}_n}(x)$$

It is not difficult to see that such constructed $s_n(x)$ converges to f. In the general case, let

$$\boxed{f = f^+ - f^-} \tag{15.106}$$

and apply the preceding construction for f^+ and f^-. Theorem is proven. □

Remark 15.5. *The sequence $s_n(x)$ converges monotonically to f if f is bounded.*

15.3.4.2 Integration

Here we shall define integration on a measurable space $\mathcal{X}$ in which $\mathfrak{M}$ is the σ-algebra (ring) of measurable sets and μ is the measure.

Definition 15.19. (Integral of a nonnegative function) *Suppose*

$$\boxed{s(x) = \sum_{i=1}^{n} c_i \chi_{E_i}(x), \quad x \in \mathcal{X}, c_i \geq 0} \tag{15.107}$$

is measurable, and suppose $\mathcal{E} \subset \mathfrak{M}$. Define

$$\boxed{I_{\mathcal{E}}(s) := \sum_{i=1}^{n} c_i \mu(\mathcal{E} \cap \mathcal{E}_i)} \tag{15.108}$$

If f is measurable and nonnegative, we define

$$\boxed{\int_{\mathcal{E}} f \, d\mu := \sup I_{\mathcal{E}}(s)} \tag{15.109}$$

where the sup *is taken over all measurable simple functions such that $0 \leq s(x) \leq f(x)$ for all $x \in \mathcal{X}$. The left-hand side member of (15.109) is called the* ***Lebesgue–Stieltjes*** *(or, simply,* ***Lebesgue****)* ***integral*** *of f with respect to measure μ over the set $\mathcal{E}$. It should be noted that integrals may have the value $(+\infty)$.*

Claim 15.4. *It is easy to verify that*

$$\boxed{\int_{\mathcal{E}} s \, d\mu = I_{\mathcal{E}}(s)} \tag{15.110}$$

Definition 15.20. (Integral of a measurable function) *Let f be measurable. Consider two Lebesgue integrals*

$$\int_{\mathcal{E}} f^+ \, d\mu \quad \text{and} \quad \int_{\mathcal{E}} f^- \, d\mu \tag{15.111}$$

where f^+ and f^- are defined by (15.102). If at least one of the integrals in (15.111) is finite, we may define

$$\int_{\mathcal{E}} f\, d\mu := \int_{\mathcal{E}} f^+\, d\mu - \int_{\mathcal{E}} f^-\, d\mu \tag{15.112}$$

If both integrals in (15.102) are finite then the left-hand side in (15.112) is finite too, and we say that f is ***integrable*** *(or summable) on $\mathcal{E}$ in* ***the Lebesgue sense*** *with respect to the measure μ. We write*

$$f \in \mathcal{L}(\mu) \text{ on } \mathcal{E} \tag{15.113}$$

Proposition 15.6. *The following properties of the Lebesgue integral are evident:*

1. *If f is measurable and bounded on $\mathcal{E}$ and if $\mu(\mathcal{E}) < \infty$, then $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$.*
2. *If $a \leq f(x) \leq b$ on $\mathcal{E}$ and if $\mu(\mathcal{E}) < \infty$, then*

$$a\mu(\mathcal{E}) \leq \int_{\mathcal{E}} f\, d\mu \leq b\mu(\mathcal{E}) \tag{15.114}$$

3. *If $f, g \in \mathcal{L}(\mu)$ on $\mathcal{E}$ and if $f(x) \leq g(x)$ for all $x \in \mathcal{E}$, then*

$$\int_{\mathcal{E}} f\, d\mu \leq \int_{\mathcal{E}} g\, d\mu \tag{15.115}$$

4. *If $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$, then $cf \in \mathcal{L}(\mu)$ for every finite constant c, and*

$$\int_{\mathcal{E}} cf\, d\mu = c\int_{\mathcal{E}} f\, d\mu \tag{15.116}$$

5. *If $\mu(\mathcal{E}) = 0$ and f is measurable, then*

$$\int_{\mathcal{E}} f\, d\mu = 0 \tag{15.117}$$

6. *If $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$, $\mathcal{A} \in \mathfrak{M}$ and $\mathcal{A} \subset \mathcal{E}$, then $f \in \mathcal{L}(\mu)$ on $\mathcal{A}$.*

Theorem 15.15.

(a) Suppose f is measurable and nonnegative on $\mathcal{X}$. For $\mathcal{A} \in \mathfrak{M}$ define

$$\phi(\mathcal{A}) = \int_{\mathcal{A}} f\, d\mu \tag{15.118}$$

Then ϕ is countably additive on $\mathfrak{M}$.

(b) The same conclusion is valid if $f \in \mathcal{L}(\mu)$ on $\mathcal{X}$.

Proof. Claim (b) follows from (a) if we write $f = f^+ - f^-$ and apply (a) to f^+ and f^-. To prove (a) we have to show that $\phi(\mathcal{A})$ can be represented as $\phi(\mathcal{A}) = \sum_{n=1}^{\infty} \phi(\mathcal{A}_n)$ if $\mathcal{A}_n \in \mathfrak{M}$, $\mathcal{A}_i \cap \mathcal{A}j = \varnothing$ if $i \neq j$ and $\mathcal{A} = \bigcup_{n=1}^{\infty} \mathcal{A}_n$. This can be done if the simple function approximation is applied that proves (a). □

Corollary 15.8. *If $\mathcal{A} \in \mathfrak{M}$, $\mathcal{B} \subset \mathcal{A}$ and $\mu(\mathcal{A} - \mathcal{B}) = 0$, then*

$$\boxed{\int_{\mathcal{A}} f \, d\mu = \int_{\mathcal{B}} f \, d\mu} \tag{15.119}$$

that is, ***the sets of measure zero are negligible in integration.***

Proof. It follows from Remark (15.117) and the representation $\mathcal{A} = \mathcal{B} \cup (\mathcal{A} - \mathcal{B})$. □

Lemma 15.6. *If $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$ then $|f| \in \mathcal{L}(\mu)$ on $\mathcal{E}$, and*

$$\boxed{\left| \int_{\mathcal{E}} f d\mu \right| \leq \int_{\mathcal{E}} |f| \, d\mu} \tag{15.120}$$

Proof. Let us represent $\mathcal{E}$ as $\mathcal{E} = \mathcal{A} \cup \mathcal{B}$ where $f(x) \geq 0$ on $\mathcal{A}$ and $f(x) < 0$ on $\mathcal{B}$. Then by Theorem 15.15 it follows that

$$\int_{\mathcal{E}} |f| \, d\mu = \int_{\mathcal{A}} |f| \, d\mu + \int_{\mathcal{B}} |f| \, d\mu = \int_{\mathcal{A}} f^+ \, d\mu + \int_{\mathcal{A}} f^- \, d\mu < \infty$$

so that $|f| \in \mathcal{L}(\mu)$ on $\mathcal{E}$. Since $f \leq |f|$ and $-f \leq |f|$ one can see that

$$\int_{\mathcal{E}} f \, d\mu \leq \int_{\mathcal{E}} |f| \, d\mu \quad \text{and} \quad -\int_{\mathcal{E}} f \, d\mu \leq \int_{\mathcal{E}} |f| \, d\mu$$

which proves (15.120). □

Lemma 15.7. *Suppose f is measurable on $\mathcal{E}$, $|f| \leq g$ and $g \in \mathcal{L}(\mu)$ on $\mathcal{E}$. Then $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$.*

Proof. It follows from the inequalities $f^+ \leq g$ and $f^- \leq g$. □

15.3.5 The "almost everywhere" concept

Definition 15.21. *Let us write*

$$\boxed{f \sim g \quad on \quad \mathcal{E}} \tag{15.121}$$

if the set $\{x \mid f(x) \neq g(x)\} \cap \mathcal{E}$ *has measure zero:*

$$\boxed{\mu(\{x \mid f(x) \neq g(x)\} \cap \mathcal{E}) = 0} \tag{15.122}$$

Proposition 15.7. *It is evident that on* $\mathcal{E}$

1. $f \sim f$;
2. $f \sim g$ *implies* $g \sim f$;
3. $f \sim g$ *and* $g \sim h$ *imply* $f \sim h$ *which means that the relation "*$\sim$*" is an equivalence relation.*
4. *If* $f \sim g$ *on* $\mathcal{E}$ *then*

$$\boxed{\int_{\mathcal{A}} f \, d\mu = \int_{\mathcal{A}} g \, d\mu} \tag{15.123}$$

provided the integrals exist for every $\mathcal{A} \subset \mathcal{E}$.

Definition 15.22. (The "almost everywhere" concept) *If some property* P *holds for every* $x \in \mathcal{E} - \mathcal{A}$ *and if* $\mu(\mathcal{A}) = 0$ *then it is customary to say that* P ***holds for almost all*** $x \in \mathcal{E}$, *or that* P *holds* ***almost everywhere*** *on* $\mathcal{E}$.

This concept depends, of course, on the particular measure to be in use. In the literature, unless something is said to the contrary, it usually refers to the Lebesgue measure.

Example 15.7. *If* $f \in \mathcal{L}(\mu)$ *on* $\mathcal{E}$ *it is clear that* $f(x)$ *must be finite almost everywhere on* $\mathcal{E}$.

15.3.5.1 Essential supremum and infimum

Definition 15.23. *Let us consider a measurable function* $f : \mathbb{R} \to \mathbb{R}$ *defined on* $\mathcal{X}$.

(a) *The* ***essential supremum*** *"ess* sup f*" of* f *(sometimes denoted also by "vrai* max f*") is defined as follows:*

$$\boxed{ess \sup f := \inf_{c \in \mathbb{R}} c} \tag{15.124}$$

such that

$$\boxed{\mu(\{x \mid f(x) > c\}) = 0} \tag{15.125}$$

(b) The ***essential infimum*** *"ess* inf *f"* *of* *f* *(sometimes denoted also by "vrai* min *f")* *is defined as follows:*

$$\boxed{ess\ \inf\ f := \sup_{c\in\mathbb{R}} c} \tag{15.126}$$

such that

$$\boxed{\mu\left(\{x \mid f(x) < c\}\right) = 0} \tag{15.127}$$

Example 15.8. *Let us consider the function* $f : \mathbb{R} \to \mathbb{R}$ *defined on* $[0, 2\pi]$ *as*

$$f(x) := \begin{cases} 5 & if \quad x = 0 \\ \sin x & if \quad x \in (0, \pi) \\ -2 & if \quad x = \pi \\ \sin x & if \quad x \in (\pi, 2\pi) \\ 3 & if \quad x = 2\pi \end{cases}$$

We have

$$\sup_{x\in[0,2\pi]} f(x) = \max_{x\in[0,2\pi]} f(x) = 5$$

$$ess\ \sup_{x\in[0,2\pi]} f(x) = 1$$

$$\inf_{x\in[0,2\pi]} f(x) = \min_{x\in[0,2\pi]} f(x) = -2$$

$$ess\ \min_{x\in[0,2\pi]} f(x) = -1$$

15.3.6 "Atomic" measures and δ-function

15.3.6.1 The "delta-function"

Definition 15.24. *The "Dirac delta-function"* $\delta(x - x_0)$ *(which is not in reality a function, but a distribution or a measure) is defined as follows:*

$$\boxed{\int_{\mathcal{X}\subset\mathbb{R}} f(x)\,\delta(x - x_0)\,dx := f(x_0)} \tag{15.128}$$

where the integral is intended in Riemann sense and $f : \mathbb{R} \to \mathbb{R}$ *is any continuous function.*

15.3.6.2 "Atomic" measures

Let us consider a continuous function $f : \mathbb{R} \to \mathbb{R}$ which takes some fixed values $\{c_1, \ldots, c_n\}$ in the points $\{x_1, \ldots, x_n\}$, that is,

$$\boxed{f(x_i) = c_i} \tag{15.129}$$

Consider also the sum

$$S := \sum_{i=1}^{n} c_i \mu_i, \quad \mu_i \geq 0 \ (i = 1, \ldots, n) \tag{15.130}$$

In fact, if some multipliers μ_i are negative, one can rewrite the product $c_i \mu_i$ as

$$c_i \mu_i = (-c_i)(-\mu_i) = (-c_i)\,|\mu_i|$$

obtaining the previous case with nonnegative weights.

Using (15.128) and in view of the additivity property of the Riemann integral, S can be represented as

$$S := \sum_{i=1}^{n} c_i \mu_i = \int_{-\infty}^{\infty} f(x) \sum_{i=1}^{n} \mu_i \delta(x - x_i)\, dx \tag{15.131}$$

Let us consider also the *step function* $\alpha(x)$ defined on $[a, b]$ by a partition $P_n := \{a = x_0, x_1, \ldots, x_n = b\}$ such that $\alpha(x)$ is a constant on each open subinterval (x_{k-1}, x_k) and has jumps

$$\begin{aligned} \mu_k &:= \alpha(x_k + 0) - \alpha(x_k - 0), \ k = 2, \ldots, n-1 \\ \mu_1 &:= \alpha(x_1 + 0) - \alpha(x_1) \\ \mu_n &:= \alpha(x_n) - \alpha(x_n - 0) \end{aligned} \tag{15.132}$$

Then, using the Riemann–Stieltjes integral representation (15.12) with the integrator $\alpha(x)$ of a *step-function type,* we can represent (15.131) as follows:

$$S := \sum_{i=1}^{n} c_i \mu_i = \int_{-\infty}^{\infty} f(x) \sum_{i=1}^{n} \mu_i \delta(x - x_i)\, dx = \int_{a}^{b} f(x)\, d\alpha(x) \tag{15.133}$$

So, symbolically, we can write

$$d\alpha(x) := \sum_{i=1}^{n} \mu_i \delta(x - x_i)\, dx \tag{15.134}$$

and

$$\alpha'(x) := \begin{cases} 0 & \text{if} \quad x \in (x_{k-1}, x_k) \\ \mu_i \delta(x - x_i) & \text{if} \quad x = x_k \end{cases} \tag{15.135}$$

associating $\alpha(x)$ with the "measure" of points $x_i \in [a, x)$ supplied by the weights μ_i. In fact, $\alpha(x)$ is the *atomic measure concentrated in the isolated points* $\{x_1, \ldots, x_n\}$.

15.4 Summary

Based on the presentations above, we may conclude that any sum S (15.130) with finite or infinite n (if it exists) can be represented by the Riemann–Stieltjes integral (15.133) with the integrator $\alpha(x)$ as the step-function (15.132). The same sum S (15.130) can be *symbolically* treated as the Lebesgue integral with the measure $\mu(x) = \alpha(x)$ referred to as the "atomic" measure concentrated in the points $\{x_1, \ldots, x_n\}$ with the corresponding nonnegative weights $\{\mu_1, \ldots, \mu_n\}$.

16 Selected Topics of Real Analysis

Contents

16.1 Derivatives

16.1.1 Basic definitions and properties

16.1.1.1 Definition of a derivative

Definition 16.1. *Let $f : \mathcal{S} \to \mathbb{R}$ be defined as a closed interval $\mathcal{S} \subset \mathbb{R}$ and assume that f is continuous at the point $c \in \mathcal{S}$. Then*

(a) f is said to have ***a right-hand derivative*** *at c if the right-hand limit*

$$\boxed{\lim_{x \to c+0} \frac{f(x) - f(c)}{x - c}} \tag{16.1}$$

exists as a finite value, or if the limit is $(+\infty)$ or $(-\infty)$. This limit will be denoted as $f'_+(c)$;

(b) f is said to have ***a left-hand derivative*** *at c if the right-hand limit*

$$\boxed{\lim_{x \to c-0} \frac{f(x) - f(c)}{x - c}} \tag{16.2}$$

exists as a finite value, or if the limit is $(+\infty)$ or $(-\infty)$. This limit will be denoted as $f'_-(c)$;

(c) f is said to have ***a derivative*** *$f'(c)$ (or be* ***differentiable****) at c if*

$$\boxed{f'_+(c) = f'_-(c) := f'(c)} \tag{16.3}$$

and $|f'(c)| < \infty$; we say that $f'(c) = +\infty$ (or $-\infty$) if both the right- and left-hand derivatives at c are $+\infty$ (or $-\infty$);

(d) f is said to be ***differentiable*** *on $\mathcal{S}$ if it is differentiable at each point $c \in \mathcal{S}$.*

16.1.1.2 Differentiability and continuity

Lemma 16.1. *Let f be defined on $[a, b]$. If f is differentiable at a point $c \in [a, b]$, then f is continuous at c.*

Proof. By the definition (16.3), for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that the inequality $|x - c| \leq \delta$ implies

$$\left| \frac{f(x) - f(c)}{x - c} - f'(c) \right| \leq \varepsilon$$

which is equivalent to the following inequalities

$$f(x) - f(c) \leq [f'(c) + \varepsilon](x - c) \leq [|f'(c)| + \varepsilon]|x - c| \leq \delta[|f'(c)| + \varepsilon]$$

Hence, $|f(x) - f(c)| \leq \varepsilon$, if take $\delta := \varepsilon / [|f'(c)| + \varepsilon]$ which proves the lemma. □

Remark 16.1. *The converse of Lemma 16.1 is not true. To see this it is sufficient to construct the continuous function which fails to be differential at an isolated point. For example, $f(x) = |x|$ which is not differentiable at the point $x = 0$ since*

$$-1 = f'_-(c) \neq f'_+(c) = 1$$

The next claim describes the usual formulas for differentiation of the sum, difference, product, quotient of two functions and function composition.

Claim 16.1. *Suppose f and g are defined on $[a, b]$ and are differentiable at a point $x \in [a, b]$. Then*

(a) for any α, β

$$\boxed{[\alpha f(x) \pm \beta g(x)]' = \alpha f'(x) \pm \beta g'(x)} \tag{16.4}$$

(b)

$$\boxed{[f(x)\, g(x)]' = f'(x)\, g(x) + f(x)\, g'(x)} \tag{16.5}$$

(c) if $g(x) \neq 0$

$$\boxed{[f(x)/g(x)]' = \frac{f'(x)\, g(x) - g'(x)\, f(x)}{g^2(x)}} \tag{16.6}$$

Proof.

(a) is evident by the property (16.3). For $h = fg$ we have

$$h(t) - h(x) = f(t)[g(t) - g(x)] + g(x)[f(t) - f(x)]$$

If we divide by $(t-x)$ and notice that $f(t) \to f(x)$ as $t \to x$ (b) follows. Taking $h = f/g$, (c) follows from the identity

$$\frac{h(t)-h(x)}{t-x} = \frac{1}{g(t)\,g(x)}\left[g(x)\frac{f(t)-f(x)}{t-x} - f(x)\frac{g(t)-g(x)}{t-x}\right]$$

letting $t \to x$. □

Claim 16.2. (The chain rule) *Suppose f is defined on $[a,b]$ and g is defined on an interval, containing the range of f, and g is differentiable at the point $f(x)$. Then, the function $h(x) = g(f(x))$ is differentiable at the point x and*

$$\boxed{h'(x) = g'(f(x))\,f'(x)} \tag{16.7}$$

Proof. Let $y = f(x)$. By the derivative definition, we have

$$\begin{aligned} f(t)-f(x) &= (t-x)[f'(x)+u(t)], \quad u(t) \underset{t\to x}{\to} 0 \\ g(s)-g(y) &= (s-y)[g'(y)+v(s)], \quad v(t) \underset{s\to y}{\to} 0 \end{aligned}$$

which leads to the following identity

$$\begin{aligned} h(t)-h(x) &= g(f(t)) - g(f(x)) \\ &= [f(t)-f(x)][g'(y)+v(s)] \\ &= (t-x)[f'(x)+u(t)][g'(y)+v(s)] \end{aligned}$$

or, if $t \neq x$

$$\frac{h(t)-h(x)}{(t-x)} = [f'(x)+u(t)][g'(y)+v(s)]$$

Letting $t \to x$ in view of the continuity of $f(x)$ we obtain (16.7). □

16.1.1.3 Higher order derivatives

Definition 16.2. *If f has a derivative on an interval, and if f' is itself differentiable, we denote the derivative of f' by f'' and call f'' the second derivative of f. Continuing in this manner, namely,*

$$\boxed{f^{(n)} = \left(f^{(n-1)}\right)', \quad n = 1, 2, \ldots} \tag{16.8}$$

we obtain the functions $f', f'', \ldots, f^{(n-1)}, f^{(n)}$ each of which is the derivative of the previous one.

In order for $f^{(n)}(x)$ to exist at a point x, $f^{(n-1)}(t)$ must exist in a neighborhood of x (or in a one-side neighborhood, if x is an endpoint of the interval on which f is defined). Sure, since $f^{(n-1)}(t)$ must exist in a neighborhood of x, $f^{(n-2)}(t)$ must be differentiable in that neighborhood.

16.1.1.4 Rolle's and generalized mean-value theorems

Theorem 16.1. (Rolle) *Assume that f has a derivative (finite or infinite) at each point of an open interval (a, b), and assume that f is continuous at both endpoints a and b. If $f(a) = f(b)$ then there exists at least one interior point c at which $f'(c) = 0$.*

Proof. Suppose that $f'(x) \neq 0$ in (a, b) and show that we obtain a contradiction. Indeed, since f is continuous on a compact set $[a, b]$, it attains its maximum M and its minimum m somewhere in $[a, b]$. But, by the assumption, neither extreme value attains an interior point (otherwise f' would vanish there). Since $f(a) = f(b)$ it follows that $M = m$, and hence f is constant on $[a, b]$ which contradicts the assumption that $f'(x) \neq 0$ on (a, b). Therefore, $f'(c) = 0$ at least one point in (a, b). □

This theorem serves as an instrument for proving the next important result.

Theorem 16.2. (The generalized mean-value theorem) *Let f and g be two functions each having a derivative (finite or infinite) at each point of an open interval (a, b) and each is continuous at the endpoints a and b. Assume also that there is no interior point x at which both $f'(x)$ and $g'(x)$ are infinite. Then for some interior point $c \in (a, b)$ the following identity holds*

$$\boxed{f'(c)\,[g(b) - g(a)] = g'(c)\,[f(b) - f(a)]} \tag{16.9}$$

Proof. Let

$$h(x) := f(x)\,[g(b) - g(b)] - g(x)\,[f(b) - f(a)]$$

Then $h'(x)$ is finite if both $f'(x)$ and $g'(x)$ are finite and $h'(x)$ is infinite if one of $f'(x)$ or $g'(x)$ is infinite. Also, $h(x)$ is continuous at the endpoints so that

$$h(a) = h(b) = f(a)g(b) - g(a)f(b)$$

By Rolle's theorem 16.1 we have that $h'(c)$ for some interior point which proves the assertion. □

Corollary 16.1. (The mean-value theorem) *Let f be a function having a derivative (finite or infinite) at each point of an open interval (a, b) and is continuous at the endpoints a and b. Then there exists a point $c \in (a, b)$ such that*

$$\boxed{f(b) - f(a) = f'(c)\,(b - a)} \tag{16.10}$$

Proof. It is sufficient to take $g(x) = x$ in (16.9). □

16.1.1.5 Taylor's formula with remainder

Theorem 16.3. (Taylor) *Suppose f is a real function on $[a, b]$, n is a positive integer, $f^{(n-1)}$ is continuous on $[a, b]$, $f^{(n)}(t)$ exists for every $t \in (a, b)$. Let x and c be distinct points of $[a, b]$, and define*

$$\boxed{P(t) := \sum_{k=0}^{n-1} \frac{f^{(k)}(c)}{k!}(t - c)^k} \tag{16.11}$$

Then there exists a point θ between x and c such that

$$\boxed{\begin{aligned} f(x) &= P(x) + \frac{f^{(n)}(\theta)}{k!}(x-c)^n \\ &= \sum_{k=0}^{n-1} \frac{f^{(k)}(c)}{k!}(x-c)^k + \frac{f^{(n)}(\theta)}{k!}(x-c)^n \end{aligned}} \tag{16.12}$$

Proof. Let M be the number defined by

$$f(x) = P(x) + M(x-c)^n$$

and put

$$g(t) := f(t) - P(t) - M(t-c)^n, \quad t \in [a,b] \tag{16.13}$$

We have to show that $n!M = f^{(n)}(\theta)$ for some $\theta \in (x,c)$. By (16.11) and (16.13) it follows that

$$g^{(n)}(t) = f^{(n)}(t) - n!M$$

Hence to complete the proof we have to show that $g^{(n)}(\theta) = 0$ for some $\theta \in (x,c)$. The choice of M, which we have done above, shows that $g(x) = 0$, so that $g'(x_1) = 0$ for some $x_1 \in (x,c)$ by the mean-value theorem 16.1. Since $g'(c) = 0$, we may conclude similarly that $g''(x_2) = 0$ for some $x_2 \in (x_1,c)$. After n steps we arrive at the conclusion that $g^{(n)}(x_n) = 0$ for some $x_n \in (x_{n-1},c)$, that is, between x and c. Theorem is proven. □

Remark 16.2. *For $n = 1$ Taylor's formula (16.12) is just the mean-value theorem 16.1.*

16.1.2 Derivative of multivariable functions

Definition 16.3. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a real function mapping an open set $\mathcal{E} \subset \mathbb{R}^n$ into $\mathbb{R}$.*

1. If for $e^i := \left(\underbrace{0,0,\ldots,0,1}_{i},0,\ldots 0\right) \in \mathbb{R}^n$ and some $x \in \mathbb{R}^n$ there exists the limit

$$\boxed{\frac{\partial}{\partial x_i} f(x) := \lim_{t\to\infty} \frac{f\left(x + te^i\right) - f(x)}{t}} \tag{16.14}$$

then $\dfrac{\partial}{\partial x_i} f(x)$ (sometimes denoted also as $D_i f(x)$) is called ***the partial derivative*** *of the function $f(x)$ at the point x.*

2. *If there exist the partial derivatives* $\dfrac{\partial}{\partial x_i} f(x)$ *of the function* $f(x)$ *at the point* x *for all* $i = 1, \ldots, n$, *then the vector*

$$\boxed{\nabla f(x) := \left(\frac{\partial}{\partial x_1} f(x), \ldots, \frac{\partial}{\partial x_n} f(x) \right)^{\mathsf{T}}} \tag{16.15}$$

(often denoted also as $\dfrac{\partial}{\partial x} f(x)$*) is called* ***the gradient*** *of the function* $f(x)$ *at the point* x.

3. *If for some* $u \in \mathbb{R}^n$ *and some* $x \in \mathbb{R}^n$ *there exists a vector* $a \in \mathbb{R}^n$ *such that*

$$\boxed{\lim_{t \to \infty} \left| \frac{f(x+tu) - f(x)}{t} - a^{\mathsf{T}} u \right| = 0} \tag{16.16}$$

then the number $a^{\mathsf{T}} u$ *(often denoted also as* $D_u f(x)$*) is called the* ***directional derivative*** *of the function* $f(x)$ *in the direction* u.

Remark 16.3. *In fact, the vector* a *in (16.16) is the gradient* $\nabla f(x)$, *that is,* $\mathbf{a} = \nabla f(x)$ *and, therefore,*

$$\boxed{D_u f(x) = \nabla^{\mathsf{T}} f(x)\, u = (\nabla f(x), u)} \tag{16.17}$$

Indeed, any $u \in \mathbb{R}^n$ *can be represented as* $u = \sum_{i=1}^{n} u_i e^i$. *Taking* $u_j = \delta_{i,j}$ *we obtain*

$$\frac{f(x+tu) - f(x)}{t} - a^{\mathsf{T}} u = \frac{f(x+tu) - f(x)}{t}$$

$$-\sum_{i=1}^{n} u_i \left(a, e^i\right) = \frac{f(x+tu) - f(x)}{t} - a_i$$

which, according to the definition (16.14), implies the identity $a_i = \dfrac{\partial}{\partial x_i} f(x)$ *that proves (16.17).*

16.1.2.1 Mixed partial derivatives

Definition 16.4. *We call the function*

$$\boxed{D_k \left(D_i f(x)\right) = \frac{\partial}{\partial x_k} \left(\frac{\partial}{\partial x_i} f(x) \right) = \frac{\partial^2}{\partial x_k \partial x_i} f(x)} \tag{16.18}$$

the ***second order*** ik***-partial derivative*** *of the function* $f(x)$ *at point* x. *Higher order partial derivatives are similarly defined.*

The following example shows that in general

$$\frac{\partial^2}{\partial x_k \partial x_i} f(x) \neq \frac{\partial^2}{\partial x_i \partial x_k} f(x)$$

Example 16.1. *Let us consider the function*

$$f(x, y) = \begin{cases} xy\dfrac{x^2 - y^2}{x^2 + y^2} & if \quad x^2 + y^2 > 0 \\ 0 & if \quad x = y = 0 \end{cases}$$

Then one has

$$\frac{\partial}{\partial x} f(x, y) = \begin{cases} y\dfrac{x^4 + 4x^2y^2 - y^4}{\left(x^2 + y^2\right)^2} & if \quad x^2 + y^2 > 0 \\ 0 & if \quad x = y = 0 \end{cases}$$

Hence, $\dfrac{\partial}{\partial x} f(0, y) = -y$ *and*

$$\frac{\partial^2}{\partial y \partial x} f(0, y) = -1$$

On the other hand,

$$\frac{\partial}{\partial y} f(x, y) = \begin{cases} x\dfrac{x^4 + 4x^2y^2 - y^4}{\left(x^2 + y^2\right)^2} & if \quad x^2 + y^2 > 0 \\ 0 & if \quad x = y = 0 \end{cases}$$

which implies $\dfrac{\partial}{\partial y} f(x, 0) = x$ *and* $\dfrac{\partial^2}{\partial x \partial y} f(x, 0) = 1$. *So, we see that*

$$-1 = \frac{\partial^2}{\partial y \partial x} f(0, 0) \neq \frac{\partial^2}{\partial x \partial y} f(0, 0) = 1$$

It is not so difficult to prove the following result (see Theorem 12.13 in Apostol (1974)).

Theorem 16.4. *If both partial derivatives* $\dfrac{\partial^2}{\partial y \partial x} f(x, y)$ *and* $\dfrac{\partial^2}{\partial x \partial y} f(x, y)$ *exist in a neighborhood of the point* (x, y) *and both are* ***continuous*** *at this point, then*

$$\frac{\partial^2}{\partial y \partial x} f(x, y) = \frac{\partial^2}{\partial x \partial y} f(x, y)$$

Corollary 16.2. *A differential* $[P(x, y)\, dx + Q(x, y)\, dy]$, *where* $\dfrac{\partial}{\partial y} P(x, y)$ *and* $\dfrac{\partial}{\partial x} Q(x, y)$ *exist and are continuous, can be represented as a* ***complete differential*** *of some function* $f(x, y)$, *namely,*

$$\boxed{P(x, y)\, dx + Q(x, y)\, dy = df(x, y)} \tag{16.19}$$

if and only if

$$\boxed{\frac{\partial}{\partial y} P(x, y) = \frac{\partial}{\partial x} Q(x, y)} \tag{16.20}$$

Proof.

(a) *Necessity.* If $df(x, y)$ is a complete differential satisfying (16.19) then

$$\begin{aligned} df(x, y) &= \frac{\partial}{\partial x} f(x, y)\, dx + \frac{\partial}{\partial y} f(x, y)\, dy \\ &= P(x, y)\, dx + Q(x, y)\, dy \end{aligned}$$

and, hence,

$$P(x, y) = \frac{\partial}{\partial x} f(x, y),\ Q(x, y) = \frac{\partial}{\partial y} f(x, y)$$

But, by the condition of this corollary, both derivatives $\frac{\partial}{\partial y} P(x, y)$ and $\frac{\partial}{\partial x} Q(x, y)$ exist and are continuous. So, by Theorem 16.4,

$$\frac{\partial}{\partial y} P(x, y) = \frac{\partial^2}{\partial y \partial x} f(x, y) = \frac{\partial^2}{\partial x \partial y} f(x, y) = \frac{\partial}{\partial x} Q(x, y) \tag{16.21}$$

(b) *Sufficiency.* Suppose (16.20) holds and there exists a function $f(x, y)$ such that $P(x, y) = \frac{\partial}{\partial x} f(x, y)$ and $Q(x, y) = \frac{\partial}{\partial y} f(x, y)$. If so, one has

$$\frac{\partial}{\partial y} P(x, y) = \frac{\partial^2}{\partial y \partial x} f(x, y), \quad \frac{\partial}{\partial x} Q(x, y) = \frac{\partial^2}{\partial x \partial y} f(x, y)$$

which gives

$$\frac{\partial^2}{\partial y \partial x} f(x, y) = \frac{\partial}{\partial y} P(x, y) = \frac{\partial}{\partial x} Q(x, y) = \frac{\partial^2}{\partial x \partial y} f(x, y)$$

This means that any function $f(x, y)$, for which (16.21) holds, exists which completes the proof. □

16.1.2.2 Multivariable mean-value theorem

Theorem 16.5. *Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable at each point of an open convex set $\mathcal{E} \subset \mathbb{R}^n$. Denote by $L(x, y) \subseteq \mathcal{E}$ the line segment joining two points $x, y \in \mathbb{R}^n$, namely,*

$$L(x, y) := \{tx + (1 - t)\, y \mid t \in [0, 1]\} \tag{16.22}$$

Then for any $x, y \in \mathbb{R}^n$ there exists a point $z \in \mathcal{L}(x, y)$ such that

$$\boxed{f(y) - f(x) = (\nabla f(z), y - x)} \tag{16.23}$$

Proof. Let $u := y - x$. Since $\mathcal{E}$ is open and $L(x, y) \subseteq \mathcal{E}$, then there is a $\delta > 0$ such that $x + tu \in \mathcal{E}$ for some $t \in (-\delta, 1 + \delta)$. Define on $(-\delta, 1 + \delta)$ the real function $F(t)$ by the relation $F(t) := f(x + tu)$. Then $F(t)$ is differentiable on $(-\delta, 1 + \delta)$ and by (16.16) for any $t \in (-\delta, 1 + \delta)$

$$F'(t) = (\nabla f(x + tu), u)$$

By the usual mean-value theorem 16.1, we have

$$F(1) - F(0) = F'(\theta), \theta \in (0, 1)$$

or, equivalently,

$$f(x + u) - f(x) = f(y) - f(x) = (\nabla f(x + \theta u), u)$$

$$= (\nabla f(x(1 - \theta) + \theta y), y - x) = (\nabla f(z), y - x)$$

which proves the theorem. □

16.1.2.3 Taylor's formula

Theorem 16.6. *Assume that f and all its partial (mixed) derivatives of order less than m are differentiable at each point of an open set $\mathcal{S} \subset \mathbb{R}^n$. If x and y are two points of $\mathcal{S}$ such that $L(x, y) \subset \mathcal{S}$ ($L(x, y)$ is defined by (16.22)), then there exists a point $z \in \mathcal{L}(x, y)$ such that*

$$\begin{aligned} f(y) - f(x) &= (\nabla f(x), y - x) \\ &+ \frac{1}{2}\left(y - x, \left\| \frac{\partial^2}{\partial x_i \partial x_k} f(x) \right\|_{i,k=1,\ldots,n} (y - x)\right) + \ldots \\ &+ \frac{1}{(m-1)!} \sum_{i_1=1}^{n} \cdots \sum_{i_{m-1}=1}^{n} \frac{\partial^{(m-1)}}{\partial x_{i_1} \ldots \partial x_{i_{m-1}}} f(x) \prod_{s=1}^{m-1} (y_{i_s} - x_{i_s}) \\ &+ \frac{1}{(m)!} \sum_{i_1=1}^{n} \cdots \sum_{i_{m-1}=1}^{n} \frac{\partial^{(m-1)}}{\partial x_{i_1} \ldots \partial x_{i_{m-1}}} f(z) \prod_{s=1}^{m-1} (y_{i_s} - x_{i_s}) \end{aligned} \tag{16.24}$$

Proof. Define $g(t) := f(x + t(y - x))$. Then $f(y) - f(x) = g(1) - g(0)$. By applying the one-dimensional Taylor formula (16.12) we obtain

$$g(1) - g(0) = \sum_{k=1}^{m-1} \frac{1}{k!} g^{(k)}(0) + \frac{1}{m!} g^{(m)}(\theta), \quad \theta \in (0, 1)$$

Applying the chain rule (see Claim 16.2) we obtain the result. □

16.1.2.4 Lemma on a finite increment

Lemma 16.2. (on a finite increment) *If f is differentiable in open set $\mathcal{S} \subset \mathbb{R}^n$ and its gradient $\nabla f(x)$ satisfies the* ***Lipschitz condition*** *on $\mathcal{S}$, that is, for all $x, y \in \mathcal{S}$ there exists a positive constant L_∇ such that*

$$\|\nabla f(x) - \nabla f(y)\| \leq L_\nabla \|x - y\|$$

then for all $x, y \in \mathcal{S}$ the following inequality holds

$$\boxed{|f(y) - f(x) - (\nabla f(x), y - x)| \leq \frac{L_\nabla}{2} \|x - y\|^2} \tag{16.25}$$

Proof. It follows from the identities

$$\int_{t=0}^{1} (\nabla f(x + t(y - x)), y - x)\, dt = \int_{t=0}^{1} d[f(x + t(y - x))]$$

$$= f(y) - f(x)\, f(y) - f(x) - (\nabla f(x), y - x)$$

$$= \int_{t=0}^{1} (\nabla f(x + t(y - x)) - \nabla f(x), y - x)\, dt$$

Taking the module of both parts and applying the Cauchy–Schwartz inequality, we get

$$|f(y) - f(x) - (\nabla f(x), y - x)|$$

$$= \left| \int_{t=0}^{1} (\nabla f(x + t(y - x)) - \nabla f(x), y - x)\, dt \right|$$

$$\leq \int_{t=0}^{1} \|\nabla f(x + t(y - x)) - \nabla f(x)\| \, \|y - x\|\, dt$$

$$\leq \int_{t=0}^{1} L_\nabla t \, \|y - x\|^2\, dt = \frac{L_\nabla}{2} \|x - y\|^2$$

Lemma is proven. □

16.1.3 Inverse function theorem

Theorem 16.7. (on the inverse function) *Suppose f is a continuously differentiable mapping from an open set $\mathcal{E} \subset \mathbb{R}^n$ into $\mathbb{R}^n$, the matrix $\frac{\partial}{\partial x} f(x) := \left\| \frac{\partial}{\partial x_k} f_i(x) \right\|_{i,k=1,\ldots,n}$ is invertible in a point $x = a \in \mathcal{E}$ and $f(a) = b$. Then*

(a) there exist open sets $\mathcal{U}$ and $\mathcal{V}$ in $\mathbb{R}^n$ such that $a \in \mathcal{U}$, $b \in \mathcal{V}$, f is one-to-one on $\mathcal{U}$ and

$$\boxed{f(\mathcal{U}) = \mathcal{V}} \tag{16.26}$$

(b) if f^{-1} is the inverse of f (which exists by (a)), defined on $\mathcal{V}$ by

$$\boxed{f^{-1}(f(x)) = x, \quad x \in \mathcal{U}} \tag{16.27}$$

then f^{-1} is continuously differentiable on $\mathcal{V}$.

Proof.

(a) Denote $A := \frac{\partial}{\partial x} f(a)$ and choose λ so that

$$2\lambda \left\| A^{-1} \right\| = 1 \tag{16.28}$$

where $\left\| A^{-1} \right\| := \sqrt{\lambda_{\max}\left(A^{-1}\left(A^{-1}\right)^{\mathrm{T}}\right)}$. Since $\frac{\partial}{\partial x} f(x)$ is continuous in a, there exists a ball $\mathcal{U} \subset \mathcal{E}$, with center in a, such that for all $x \in \mathcal{U}$

$$\left\| \frac{\partial}{\partial x} f(x) - A \right\| < \lambda \tag{16.29}$$

Let us associate to each point $y \in \mathbb{R}^n$ the function φ_y defined by

$$\varphi_y(x) := x + A^{-1}(y - f(x)) \tag{16.30}$$

Note that $y = f(x)$ if and only if x is a fixed point of φ_y. (16.28) and (16.29) imply

$$\left\| \frac{\partial}{\partial x} \varphi_y(x) \right\| = \left\| I - A^{-1} \frac{\partial}{\partial x} f(x) \right\| = \left\| A^{-1} \left(A - \frac{\partial}{\partial x} f(x) \right) \right\|$$

$$\leq \left\| A^{-1} \right\| \left\| A - \frac{\partial}{\partial x} f(x) \right\| < \left\| A^{-1} \right\| \lambda = \frac{1}{2}$$

Hence, by the mean-value theorem 16.5 it follows that

$$\left\| \varphi_y(x') - \varphi_y(x'') \right\| \leq \sup_{\theta \in \mathcal{U}} \left\| \frac{\partial}{\partial x} f(\theta) \right\| \left\| x' - x'' \right\| < \frac{1}{2} \left\| x' - x'' \right\| \tag{16.31}$$

which means, by Theorem 14.17, that $\varphi_y(x)$ has at most one fixed point in $\mathcal{U}$. So, $y = f(x)$ for at most one point $x \in \mathcal{U}$. And, since $\frac{\partial}{\partial x} f(x)$ is invertible in $\mathcal{U}$, we conclude that f is one-to-one in $\mathcal{U}$. Next, put $f(\mathcal{U}) = \mathcal{V}$ and pick $y_0 \in \mathcal{V}$. Then $y_0 = f(x_0)$ for some $x_0 \in \mathcal{U}$. Let $\mathcal{B}$ be an open ball with the center in x_0 and radius $r > 0$, so small that its closure $\bar{\mathcal{B}}$ lies in $\mathcal{U}$. Let us show that $y \in \mathcal{V}$ whenever $\|y - y_0\| < \lambda r$. Fix y such that $\|y - y_0\| < \lambda r$. By (16.30) we have

$$\|\varphi_y(x_0) - x_0\| = \|A^{-1}(y - y_0)\| \le \|A^{-1}\| \, \|y - y_0\| < \|A^{-1}\| \lambda r = r/2$$

If $x \in \bar{\mathcal{B}}$, then it follows from (16.31) that

$$\|\varphi_y(x) - x_0\| = \|[\varphi_y(x) - \varphi_y(x_0)] + [\varphi_y(x_0) - x_0]\| \le \|\varphi_y(x) - \varphi_y(x_0)\|$$
$$+ \|\varphi_y(x_0) - x_0\| < \frac{1}{2}\|x - x_0\| + r/2 \le r$$

Hence, $\varphi_y(x) \in \mathcal{B}$. Note that (16.31) also holds if $x', x'' \in \bar{\mathcal{B}}$. Thus $\varphi_y(x)$ is a contraction of $\bar{\mathcal{B}}$ into $\mathcal{B}$. Being a closed subset of $\mathbb{R}^n$, $\bar{\mathcal{B}}$ is complete. Then by the fixed-point theorem 14.17 we conclude that $\varphi_y(x)$ has a fixed point $x \in \bar{\mathcal{B}}$. For this x it follows that $f(x) = y$. Thus $y \in f(\bar{\mathcal{B}}) \subset f(\mathcal{U}) = \mathcal{V}$ which proves (a).

(b) Pick $y \in \mathcal{V}$ and $y + z \in \mathcal{V}$. Then there exist $x \in \mathcal{U}$ and $x + h \in \mathcal{U}$ so that $y = f(x)$ and $y + z = f(x + h)$. So,

$$\varphi_y(x + h) - \varphi_y(x) = h + A^{-1}[f(x) - f(x + h)] = h - A^{-1}z$$

By (16.31) we have

$$\|h - A^{-1}z\| < \frac{1}{2}\|h\|$$

which, by the inequality $\|a - b\| \ge \|a\| - \|b\|$, implies

$$\|h\| - \|A^{-1}z\| \le \|h - A^{-1}z\| < \frac{1}{2}\|h\|$$

and, therefore,

$$\frac{1}{2}\|h\| < \|A^{-1}z\|$$

or, equivalently,

$$\|h\| < 2\|A^{-1}z\| \le 2\|A^{-1}\| \, \|z\| = \|z\|/\lambda \tag{16.32}$$

Since $\frac{\partial}{\partial x} f(x)$ has an inverse on $\mathcal{U}$, say T, we get

$$g(y + z) - g(y) - Tz = h - Tz = -T\left[f(x + h) - f(x) - \frac{\partial}{\partial x} f(x)\, h\right]$$

which in view of (16.32) implies

$$\frac{\|g(y+z)-g(y)-Tz\|}{\|z\|} \leq \frac{\|T\|}{\lambda} \frac{\left\| f(x+h)-f(x)-\frac{\partial}{\partial x} f(x)\, h \right\|}{\|h\|}$$

As $z \to 0$, (16.32) shows that $h \to 0$ and therefore the right-hand side of the last inequality tends to zero which is true of the left. This proves that $\frac{\partial}{\partial x} g(y) = T$. But T is chosen to be the inverse of $\frac{\partial}{\partial x} f(x) = \frac{\partial}{\partial x} f(g(y))$. Thus

$$\frac{\partial}{\partial x} g(y) = \left[\frac{\partial}{\partial x} f(g(y))\right]^{-1} \tag{16.33}$$

for $y \in \mathcal{V}$. But both g and $\frac{\partial}{\partial x} f(x)$ are locally continuous which together with (16.33) implies that g is continuously differentiable on $\mathcal{V}$.
Theorem is proven. □

Summary 16.1. *The inverse function theorem 16.7 states, roughly speaking, that a* ***continuous differentiable mapping*** $f(x)$ ***is invertible*** *in a neighborhood of any point x at which the linear transformation $\frac{\partial}{\partial x} f(x)$ is invertible.*

Corollary 16.3. *The system of n equations*

$$y_i = f_i(x_1, \ldots, x_n), \quad i = 1, \ldots, n$$

can be solved for $(x_1, \ldots, x_n)$ in terms of $(y_1, \ldots, y_n)$ if $\frac{\partial}{\partial x} f(x)$ is invertible in a neighborhood of the point $x = (x_1, \ldots, x_n)$.

16.1.4 Implicit function theorem

For $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ let us consider the extended vector $z := (x^\tau, y^\tau)^\tau \in \mathbb{R}^{n+m}$. Then any linear transformation $A : \mathbb{R}^{n+m} \to \mathbb{R}^n$ can be represented as

$$Az = \begin{bmatrix} A_x & A_y \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = A_x x + A_y y$$

Then the following result seems to be obvious.

Lemma 16.3. (A linear version) *If A_x is invertible, then for every $y \in \mathbb{R}^m$ there exists a unique $x \in \mathbb{R}^n$ such that*

$$\boxed{A_x x + A_y y = 0} \tag{16.34}$$

This x can be calculated as

$$\boxed{x = -(A_x)^{-1} A_y y} \tag{16.35}$$

The theorem given below represents the, so-called, *implicit function theorem* for nonlinear mappings.

Theorem 16.8. (The implicit function theorem) *Let f be a continuously differentiable mapping of an open set $\mathcal{E} \subset \mathbb{R}^{n+m}$ into $\mathbb{R}^n$ such that*

$$\boxed{f(\mathring{x}, \mathring{y}) = 0} \tag{16.36}$$

for some point $\mathring{z} := (\mathring{x}^\top, \mathring{y}^\top)^\top \in \mathcal{E}$. Denote $A := \dfrac{\partial}{\partial z} f(\mathring{z})$ and assume that $A_x := \dfrac{\partial}{\partial x} f(\mathring{z})$ is invertible. Then there exist open sets $\mathcal{U} \subset \mathbb{R}^{n+m}$ and $\mathcal{W} \subset \mathbb{R}^m$ with $\mathring{z} \in \mathcal{U}$ and $\mathring{y} \in \mathcal{W}$, having the following properties:

1. *To every $y \in \mathcal{W}$ there exists a unique x such that*

$$\boxed{z = \begin{pmatrix} x \\ y \end{pmatrix} \in \mathcal{U} \quad \text{and} \quad f(x, y) = 0} \tag{16.37}$$

2. *If x is defined to be $g(y)$, then g is continuously differentiable on $\mathcal{W}$, $g(y) = x$, for any $y \in \mathcal{W}$*

$$\boxed{f(g(y), y) = 0} \tag{16.38}$$

and

$$\boxed{\frac{\partial}{\partial y} g(\mathring{y}) = -(A_x)^{-1} A_y} \tag{16.39}$$

where $A_y := \dfrac{\partial}{\partial y} f(\mathring{z})$.

Proof. For $z \in \mathcal{E}$ define

$$F(x, y) := \begin{pmatrix} f(x) \\ y \end{pmatrix}$$

Then F is a continuously differentiable mapping of $\mathcal{E}$ into $\mathbb{R}^{n+m}$. Show now that $\dfrac{\partial}{\partial z} F(\mathring{z})$ is an invertible element in $\mathbb{R}^{n+m}$. Indeed, since $f(\mathring{z}) = 0$, we have

$$f(\mathring{x} + h, \mathring{y} + k) = A \begin{pmatrix} h \\ k \end{pmatrix} + r(h, k)$$

where $r(h, k)$ is a remainder that occurs in the definition of $A = \dfrac{\partial}{\partial z} f(\mathring{z})$. Again, since

$$F(\mathring{x}+h, \mathring{y}+k) - F(\mathring{x}, \mathring{y}) = \begin{pmatrix} f(\mathring{x}+h, \mathring{y}+k) \\ k \end{pmatrix}$$
$$= \begin{pmatrix} A \begin{pmatrix} h \\ k \end{pmatrix} \\ k \end{pmatrix} + \begin{pmatrix} r(h,k) \\ 0 \end{pmatrix}$$

it follows that $\dfrac{\partial}{\partial z} F(\mathring{z})$ is the linear operator on $\mathbb{R}^{n+m}$ that maps $\begin{pmatrix} h \\ k \end{pmatrix}$ to $\begin{pmatrix} A \begin{pmatrix} h \\ k \end{pmatrix} \\ k \end{pmatrix}$. It is seen that $\dfrac{\partial}{\partial z} F(\mathring{z})$ is one-to-one and hence it is invertible. So, the inverse function theorem can therefore be applied to F that proves (1). To prove (2) define $g(y)$ for $y \in \mathcal{W}$ so that $\Phi(y) := \begin{pmatrix} g(y) \\ y \end{pmatrix} \in \mathcal{U}$ and (16.38) holds. Then $F(g(y), y) = \begin{pmatrix} 0 \\ y \end{pmatrix}$ and $\dfrac{\partial}{\partial y} \Phi(y) k = \begin{pmatrix} \dfrac{\partial}{\partial y} g(y) k \\ k \end{pmatrix}$. In view of the identity $f(\Phi(y)) = 0$ the chain rule shows that

$$\frac{\partial}{\partial z} f(\Phi(y)) \frac{\partial}{\partial y} \Phi(y) = 0$$

Thus

$$A \frac{\partial}{\partial y} \Phi(\mathring{y}) = 0$$

which gives

$$A_x \frac{\partial}{\partial y} g(\mathring{y}) k + A_y k = A \frac{\partial}{\partial y} \Phi(\mathring{y}) = 0$$

This completes the proof. □

Example 16.2. *Let*

$$f_1(x_1, x_2, y_1, y_2, y_5) = 2e^{x_1} + x_2 y_1 - 4y_2 + 3 = 0$$

$$f_2(x_1, x_2, y_1, y_2, y_5) = x_2 \cos x_1 - 6x_1 + 2y_1 - 4y_3 = 0$$

and $\mathring{x} = \begin{pmatrix} 0 & 1 \end{pmatrix}^{\mathsf{T}}$, $\mathring{y} = \begin{pmatrix} 3 & 2 & 7 \end{pmatrix}$. *Then*

$$A_x = \begin{bmatrix} 2 & 3 \\ -6 & 1 \end{bmatrix}, \quad A_y = \begin{bmatrix} 1 & -4 & 0 \\ 2 & 0 & -1 \end{bmatrix}$$

Notice that $\det A_x = 20 \neq 0$ *and hence* A_x *is invertible and* x *in a neighborhood of* $\mathring{x}$, $\mathring{y}$ *can be represented as a function of* y, *that is,* $x = g(y)$.

16.1.5 Vector and matrix differential calculus

16.1.5.1 Differentiation of scalar functions with respect to a vector

Assuming that $x, y \in \mathbb{R}^n$, $P \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{m\times n}$, the direct calculation shows that

$$\frac{\partial}{\partial x}(x^{\mathsf{T}}y) = \frac{\partial}{\partial x}(y^{\mathsf{T}}x) = y \tag{16.40}$$

$$\frac{\partial}{\partial x}(Px) = P^{\mathsf{T}}, \quad \frac{\partial}{\partial x}(x^{\mathsf{T}}Py) = Py \tag{16.41}$$

$$\left.\begin{aligned} &\frac{\partial}{\partial x}(y^{\mathsf{T}}Px) = P^{\mathsf{T}}y, \quad \frac{\partial}{\partial x}(x^{\mathsf{T}}Px) = (P + P^{\mathsf{T}})x \\ &\text{and} \quad \frac{\partial}{\partial x}(x^{\mathsf{T}}Px) = 2Px \quad \text{only when} \quad P = P^{\mathsf{T}} \end{aligned}\right\} \tag{16.42}$$

$$\frac{\partial}{\partial x^{\mathsf{T}}}(Ax) = A \tag{16.43}$$

$$\frac{\partial}{\partial x}\|x\|_2 = \frac{x}{\|x\|_2}, \quad x \neq 0 \tag{16.44}$$

$$\left.\begin{aligned} &\frac{\partial}{\partial x}(x \otimes y) = e \otimes y = \operatorname{col}\{I_{n\times n} \otimes y\} \\ &e := (1, 1, \ldots, 1)^{\mathsf{T}} \end{aligned}\right\} \tag{16.45}$$

16.1.5.2 Differentiation of scalar fun!ctions with respect to a matrix

For the matrices A, B and C the direct calculation implies

$$\frac{\partial}{\partial A}\operatorname{tr}(A) = A \tag{16.46}$$

$$\frac{\partial}{\partial A}\operatorname{tr}(BAC) = B^{\mathsf{T}}C^{\mathsf{T}}, \quad \frac{\partial}{\partial A}\operatorname{tr}(BA^{\mathsf{T}}C) = CB \tag{16.47}$$

$$\frac{\partial}{\partial A}\operatorname{tr}(ABA^{\mathsf{T}}) = AB^{\mathsf{T}} + AB, \quad \frac{\partial}{\partial A}\operatorname{tr}(ABA) = A^{\mathsf{T}}B^{\mathsf{T}} + B^{\mathsf{T}}A^{\mathsf{T}} \tag{16.48}$$

$$\left.\begin{aligned} &\frac{\partial}{\partial A}\operatorname{tr}(BACA) = C^{\mathsf{T}}A^{\mathsf{T}}B^{\mathsf{T}} + B^{\mathsf{T}}A^{\mathsf{T}}C^{\mathsf{T}} \\ &\frac{\partial}{\partial A}\operatorname{tr}(BACA^{\mathsf{T}}) = BAC + B^{\mathsf{T}}AC^{\mathsf{T}} \end{aligned}\right\} \tag{16.49}$$

$$\frac{\partial}{\partial A}\mathrm{tr}\,(A^{\mathsf{T}}A) = 2A \tag{16.50}$$

$$\left.\begin{aligned}\frac{\partial}{\partial A}\mathrm{tr}\,(BA^{\mathsf{T}}AC) &= ACB + AB^{\mathsf{T}}C^{\mathsf{T}}\\ \frac{\partial}{\partial A}\mathrm{tr}\,(BAA^{\mathsf{T}}C) &= CBA + B^{\mathsf{T}}C^{\mathsf{T}}A\end{aligned}\right\} \tag{16.51}$$

$$\frac{\partial}{\partial A}\mathrm{tr}\,(BA^{\mathsf{T}}AB^{\mathsf{T}}) = 2AB^{\mathsf{T}}B, \quad \frac{\partial}{\partial A}\mathrm{tr}\,(BAA^{\mathsf{T}}B^{\mathsf{T}}) = 2B^{\mathsf{T}}BA \tag{16.52}$$

$$\left.\begin{aligned}\frac{\partial}{\partial A}\mathrm{tr}\left(B^{\mathsf{T}}(A^{\mathsf{T}}A)^2B\right) &= \frac{\partial}{\partial A}\mathrm{tr}\,(B^{\mathsf{T}}A^{\mathsf{T}}(AA^{\mathsf{T}})AB)\\ &= 2A(A^{\mathsf{T}}A)B^{\mathsf{T}}B + 2AB^{\mathsf{T}}B(A^{\mathsf{T}}A)\end{aligned}\right\} \tag{16.53}$$

$$\frac{\partial}{\partial A}\mathrm{tr}\,(\exp(A)) = \exp(A) \tag{16.54}$$

$$\frac{\partial}{\partial A}\det(BAC) = \det(BAC)\left(A^{-1}\right)^{\mathsf{T}} \tag{16.55}$$

$$\frac{\partial}{\partial A}\mathrm{tr}\left(A^k\right) = k\left(A^{k-1}\right)^{\mathsf{T}}, \quad \frac{\partial}{\partial A}\mathrm{tr}\left(BA^k\right) = \sum_{i=0}^{k-1}\left(A^iBA^{k-i-1}\right)^{\mathsf{T}} \tag{16.56}$$

$$\frac{\partial}{\partial A}\mathrm{tr}\left(BA^{-1}C\right) = -\left(A^{-1}CBA^{-1}\right)^{\mathsf{T}} \tag{16.57}$$

$$\frac{\partial}{\partial A}\log\det(A) = -(A^{\mathsf{T}})^{-1} \tag{16.58}$$

$$\frac{\partial}{\partial A}\det(A^{\mathsf{T}}) = \frac{\partial}{\partial A}\det(A) = (A^{\mathsf{T}})^{-1}\det(A) \tag{16.59}$$

$$\left.\begin{aligned}\frac{\partial}{\partial A}\det\left(A^k\right) &= \frac{\partial}{\partial A}[\det(A)]^k\\ &= k[\det(A)]^{k-1}\frac{\partial}{\partial A}\det(A)\\ &= k[\det(A)]^{k-1}(A^{\mathsf{T}})^{-1}\det(A) = k(A^{\mathsf{T}})^{-1}\det\left(A^k\right)\end{aligned}\right\} \tag{16.60}$$

16.1.6 Nabla operator in three-dimensional space

Definition 16.5. *Define*

1. *the* ***differential nabla operator*** ∇ *or* ***gradient****, acting to a differentiable function* $\varphi : \mathbb{R}^3 \to \mathbb{R}^1$*, by the following formula*

$$\boxed{\nabla\varphi(x, y, z) := \begin{pmatrix} \dfrac{\partial}{\partial x}\varphi(x, y, z) \\ \dfrac{\partial}{\partial y}\varphi(x, y, z) \\ \dfrac{\partial}{\partial z}\varphi(x, y, z) \end{pmatrix}} \tag{16.61}$$

2. *the differentiable operator* div (***divergence***)*, acting to a differentiable function* $f : \mathbb{R}^3 \to \mathbb{R}^3$*, by the following formula*

$$\boxed{\operatorname{div} f(x, y, z) := \frac{\partial}{\partial x} f_x(x, y, z) + \frac{\partial}{\partial y} f_y(x, y, z) + \frac{\partial}{\partial z} f_z(x, y, z)} \tag{16.62}$$

3. *the differentiable operator* rot (***rotor***)*, acting to a differentiable function* $f : \mathbb{R}^3 \to \mathbb{R}^3$*, by the following formula*

$$\boxed{\begin{aligned} \operatorname{rot} f(x, y, z) &:= \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ f_1(x, y, z) & f_2(x, y, z) & f_3(x, y, z) \end{bmatrix} \\ &= \mathbf{i}\left(\frac{\partial}{\partial y} f_3(x, y, z) - \frac{\partial}{\partial z} f_2(x, y, z)\right) \\ &\quad + \mathbf{j}\left(\frac{\partial}{\partial z} f_1(x, y, z) - \frac{\partial}{\partial x} f_3(x, y, z)\right) \\ &\quad + \mathbf{k}\left(\frac{\partial}{\partial x} f_2(x, y, z) - \frac{\partial}{\partial y} f_1(x, y, z)\right) \end{aligned}}$$

where $(\mathbf{i}, \mathbf{j}, \mathbf{k})$ *is the orthogonal basis in* $\mathbb{R}^3$*.*

Remember some important properties of the scalar (a, b) and the vector product $[a, b]$ of the vectors $a = (a_x, a_y, a_z)^\mathsf{T}$ and $b = (b_x, b_y, b_z)^\mathsf{T}$ in $\mathbb{R}^3$ which are defined by

$$\boxed{(a, b) := a_x b_x + a_y b_y + a_z b_z} \tag{16.63}$$

and

$$\boxed{\begin{aligned}[a,b] &:= \begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_x & a_y & a_z \\ b_x & b_y & b_z \end{bmatrix} \\ &= \mathbf{i}\left(a_y b_z - a_z b_y\right) + \mathbf{j}\left(a_z b_x - a_x b_z\right) + \mathbf{k}\left(a_x b_y - a_y b_x\right)\end{aligned}} \tag{16.64}$$

It is necessary to check that

$$\boxed{\begin{aligned}(a,a) &= \|a\|^2 \\ (a,b) &= (b,a) \\ (a,b) &= 0 \text{ if } a \perp b \\ (a,b+c) &= (a,b) + (a,c)\end{aligned}} \tag{16.65}$$

and

$$\boxed{\begin{aligned}[a,b] &= -[b,a] \\ [a,(b+c)] &= [a,b] + [a,c] \\ [a,b] &= 0 \quad \text{if} \quad a = \lambda b, \quad \lambda \in \mathbb{R} \\ (a,[b,c]) &= (b,[c,a]) = (c,[a,b]) \\ [a,[b,c]] &= b\,(a,c) - c\,(a,b) \\ [a,[b,c]] + [b,[c,a]] + [c,[a,b]] &= 0\end{aligned}} \tag{16.66}$$

Notice also that the operators div and rot, using the definitions above, can be represented in the following manner

$$\boxed{\begin{aligned}\operatorname{div}\varphi &= (\nabla,\varphi) \\ \operatorname{rot}\varphi &= [\nabla,\varphi]\end{aligned}} \tag{16.67}$$

Applying rules (16.65) and (16.66) one can prove that

$$\boxed{\begin{gathered}\operatorname{div}(\alpha f + \beta g) = \alpha \operatorname{div} f + \beta \operatorname{div} g \\ \alpha, \beta \in \mathbb{R}\end{gathered}} \tag{16.68}$$

$$\boxed{\begin{gathered}\operatorname{div}\operatorname{grad}\varphi = (\nabla, \nabla\varphi) = \Delta\varphi \\ \Delta\varphi := \frac{\partial^2}{\partial x^2} f_x(x,y,z) + \frac{\partial^2}{\partial y^2} f_y(x,y,z) + \frac{\partial^2}{\partial y^2} f_x(x,y,z) \\ \text{where } \Delta \text{ is the Laplace operator}\end{gathered}} \tag{16.69}$$

$$\boxed{\text{div rot } f = \big(\nabla, [\nabla, f]\big) = ([\nabla, \nabla], f) = 0} \tag{16.70}$$

$$\boxed{\begin{aligned} &\text{rot rot } f = [\nabla, [\nabla, f]] \\ &= \nabla(\nabla, f) - (\nabla, \nabla) f = \text{grad div } f - \Delta f \end{aligned}} \tag{16.71}$$

16.2 On Riemann–Stieltjes integrals

16.2.1 The necessary condition for existence of Riemann–Stieltjes integrals

Here we will examine the following statement: *when α is of bounded variation on $[a, b]$, continuity of f is sufficient for the existence of the Riemann–Stieltjes integral $\int_{x=a}^{b} f(x)\, d\alpha(x)$?* We conclude that: **continuity** of f throughout $[a, b]$ is by **no means necessary**, however! The next theorem shows that **common discontinuities** (from the right or from the left) **should be avoided** if the integral $\int_{x=a}^{b} f(x)\, d\alpha(x)$ is to exist. Define

$$\boxed{\begin{aligned} U(P_n, f, \alpha) &:= \sum_{i=1}^{n} M_i \Delta\alpha_i \\ M_i &:= \sup\{f(x) : x \in [x_{i-1}, x_i]\} \\ L(P_n, f, \alpha) &:= \sum_{i=1}^{n} M_i \Delta\alpha_i \\ m_i &:= \inf\{f(x) : x \in [x_{i-1}, x_i]\} \end{aligned}} \tag{16.72}$$

which coincide with the upper and lower Darboux sums, respectively, (see (15.4) and (15.5)) for the case $\alpha(x) = x$.

Theorem 16.9. (The necessary condition) *Assume that $\alpha \uparrow$ on $[a, b]$ and $c \in (a, b)$. Assume further that both f and α are discontinuous simultaneously from the right at $x = c$, that is, assume that there exists $\varepsilon > 0$ such that for every $\delta > 0$ there are values of x and y within the interval $(c, c + \delta)$ for which*

$$|f(x) - f(y)| \geq \varepsilon \quad \text{and} \quad |\alpha(x) - \alpha(y)| \geq \varepsilon$$

Then the integral $\int_{x=a}^{b} f(x)\, d\alpha(x)$ cannot exist. The integral also fails to exist if f and α are discontinuous simultaneously from the left at $x = c$.

Proof. Let P_n be a partition of $[a, b]$ containing the point c as a point of subdivision. Then one has

$$U(P_n, f, \alpha) - L(P_n, f, \alpha) := \sum_{i=1}^{n} (M_i - m_i)\,\Delta\alpha_i \tag{16.73}$$

If the ith interval has c as its left endpoint then

$$U(P_n, f, \alpha) - L(P_n, f, \alpha) \geq (M_i - m_i)\,[\alpha_i(x_i) - \alpha_i(c)]$$

since each term in (16.73) is nonnegative. If c is a common discontinuity point from the right, we may assume that the point x_i is chosen in such a way that $[\alpha_i(x_i) - \alpha_i(c)] \geq \varepsilon$. Moreover, by the assumptions of the theorem $(M_i - m_i) \geq \varepsilon$. So,

$$U(P_n, f, \alpha) - L(P_n, f, \alpha) \geq \varepsilon^2 \tag{16.74}$$

But by the definition (15.11) of the Riemann–Stieltjes integral there exists n such that

$$|U(P_n, f, \alpha) - S(P_n, f, \alpha)| < \frac{\varepsilon^2}{2}, \quad |L(P_n, f, \alpha) - S(P_n, f, \alpha)| < \frac{\varepsilon^2}{2}$$

and, hence,

$$\begin{aligned} U(P_n, f, \alpha) - L(P_n, f, \alpha) &= [U(P_n, f, \alpha) - S(P_n, f, \alpha)] \\ &+ [S(P_n, f, \alpha) - L(P_n, f, \alpha)] \leq |U(P_n, f, \alpha) - S(P_n, f, \alpha)| \\ &+ |S(P_n, f, \alpha) - L(P_n, f, \alpha)| < \varepsilon^2 \end{aligned}$$

which is in contradiction with (16.74). If c is a common discontinuity from the left the argument is similar. Theorem is proven. □

16.2.2 The sufficient conditions for existence of Riemann–Stieltjes integrals

Theorem 16.10. (First sufficient (Riemann's) condition) *Assume that $\alpha \uparrow$ on $[a, b]$. If for any $\varepsilon > 0$ there exists a partition P_ε of $[a, b]$ such that P_n is finer than P_ε implies*

$$\boxed{0 \leq U(P_n, f, \alpha) - L(P_n, f, \alpha) < \varepsilon} \tag{16.75}$$

then $f \in \mathcal{R}_{[a,b]}(\alpha)$.

Proof. Since by $\alpha \uparrow$ on $[a, b]$ we have

$$U(P_n, f, \alpha) \leq S(P_n, f, \alpha) \leq L(P_n, f, \alpha)$$

In view of (16.75) this means that $S(P_n, f, \alpha)$ has a limit when $n \to \infty$ which, by the definition (15.11), is the Riemann–Stieltjes integral. Theorem is proven. □

Theorem 16.11. (Second sufficient condition) *If f is continuous on $[a, b]$ and α is of bounded variation on $[a, b]$, then $f \in \mathcal{R}_{[a,b]}(\alpha)$.*

Proof. Since by (15.55) any α of bounded variation can be represented as $\alpha(x) = \alpha^+(x) - \alpha^-(x)$ (where $\alpha^+ \uparrow$ on $[a, b]$ and $\alpha^- \uparrow$ on $[a, b]$), it suffices to prove the theorem when $\alpha \uparrow$ on $[a, b]$ with $\alpha(a) < \alpha(b)$. Continuity of f on $[a, b]$ implies uniform continuity, so that if $\varepsilon > 0$ is given, we can find $\delta = \delta(\varepsilon) > 0$ such that $|x - y| < \delta$ implies $|f(x) - f(y)| < \varepsilon/A$ where $A = 2[\alpha(b) - \alpha(a)]$. If P_ε is a partition with the biggest interval less than δ, then any partition P_n finer than P_ε gives

$$M_i - m_i \leq \varepsilon/A \tag{16.76}$$

since

$$M_i - m_i = \sup\{f(x) - f(y) : x, y \in [x_{i-1}, x_i]\}$$

Multiplying (16.76) by $\Delta\alpha_i$ and summing, we obtain

$$U(P_n, f, \alpha) - L(P_n, f, \alpha) \leq \varepsilon/A \sum_{i=1}^{n} \Delta\alpha_i = \frac{\varepsilon}{2} < \varepsilon$$

So, Riemann's condition (16.75) holds. Theorem is proven. □

Corollary 16.4. *For the special case of the Riemann integral when $\alpha(x) = x$ Theorem 16.11 together with (15.23) state that each of the following conditions is sufficient for* ***the existence of the Riemann integral*** $\int_{x=a}^{b} f(x)\,dx$*:*

1. *f is continuous on $[a, b]$;*
2. *f is of bounded variation on $[a, b]$.*

The following theorem represents the criterion (the necessary and sufficient condition) for the Riemann integrability.

Theorem 16.12. (Lebesgue's criterion for integrability) *Let f be defined and bounded on $[a, b]$. Then it is the Riemann integrable on $[a, b]$, which is $f \in \mathcal{R}_{[a,b]}(x)$, if and only if f is continuous almost everywhere on $[a, b]$.*

Proof. Necessity can be proven by contradiction assuming that the set of discontinuity has a nonzero measure and demonstrating that in this case f is not integrable. *Sufficiency* can be proven by demonstrating that Riemann's condition (16.75) (when $\alpha(x) = x$) is satisfied assuming that the discontinuity points have measure zero. The detailed proof can be found in Apostol (1974). □

16.2.3 Mean-value theorems

Although integrals occur in a wide variety of problems (including control), there are relatively few cases when the explicit value of the integral can be obtained. However, it is often sufficient to have an estimate for the integral rather than its exact value. The *mean-value theorems* of this subsection are especially useful in making such estimates.

Theorem 16.13. (First mean-value theorem) *Assume that* $f \in \mathcal{R}_{[a,b]}(\alpha)$ *with* $\alpha \uparrow$ *on* $[a, b]$. *Denote*

$$M := \sup_{x\in[a,b]} f(x), \quad m := \inf_{x\in[a,b]} f(x)$$

Then there exists a real number $c \in [m, M]$ *such that*

$$\boxed{\int_{x=a}^{b} f(x)\, d\alpha(x) = c \int_{x=a}^{b} d\alpha(x) = c\,[\alpha(b) - \alpha(a)]} \tag{16.77}$$

Proof. If $a = b$ both sides of (16.77) are zero and the result holds trivially. Assume that $\alpha(a) < \alpha(b)$. By (16.72) we have

$$m\,[\alpha(b) - \alpha(a)] \le L(P_n, f, \alpha) \le \int_{x=a}^{b} f(x)\, d\alpha(x)$$

$$\le U(P_n, f, \alpha) \le M\,[\alpha(b) - \alpha(a)]$$

which proves (16.77). Theorem is proven. □

Remark 16.4. *Evidently, if* f *is continuous on* $[a, b]$ *then there exists* $x_0 \in [a, b]$ *such that* $c = f(x_0)$.

Theorem 16.14. (Second mean-value theorem) *Assume that* $\alpha(x)$ *is continuous on* $[a, b]$ *and* $f \uparrow$ *on* $[a, b]$. *Then there exists a point* $x_0 \in [a, b]$ *such that*

$$\boxed{\int_{x=a}^{b} f(x)\, d\alpha(x) = f(a) \int_{x=a}^{x_0} d\alpha(x) + f(b) \int_{x=x_0}^{a} d\alpha(x)} \tag{16.78}$$

Proof. Integrating by parts (see (15.23)) implies

$$\int_{x=a}^{b} f(x)\, d\alpha(x) = f(b)\,\alpha(b) - f(a)\,\alpha(a) - \int_{x=a}^{b} \alpha(x)\, df(x)$$

Applying (16.77) to the integral on the right-hand side of the last identity we have

$$\begin{aligned}\int_{x=a}^{b} f(x)\,d\alpha(x) &= f(b)\,\alpha(b) - f(a)\,\alpha(a) - c\int_{x=a}^{b} df(x)\\ &= f(b)\,\alpha(b) - f(a)\,\alpha(a) - \alpha(x_0)\,[f(b) - f(a)]\\ &= f(b)\,[\alpha(b) - \alpha(x_0)] + f(a)\,[\alpha(x_0) - \alpha(a)]\end{aligned}$$

which is the statement we set out to prove. □

Corollary 16.5. (The Riemann integrals case) *Let g be continuous on $[a,b]$ and $f \uparrow$ on $[a,b]$. Then*

1. there exists a point $x_0 \in [a,b]$ such that

$$\boxed{\int_{x=a}^{b} f(x)\,g(x)\,dx = A\int_{x=a}^{x_0} g(x)\,dx + B\int_{x=x_0}^{b} g(x)\,dx} \tag{16.79}$$

where $A \leq f(a+0)$ and $B \geq f(b-0)$;

*2. **Bonnet's theorem** holds, namely, if, in addition, $f(x) \geq 0$ on $[a,b]$ then $A = 0$ in (16.79) which gives*

$$\boxed{\int_{x=a}^{b} f(x)\,g(x)\,dx = B\int_{x=x_0}^{b} g(x)\,dx} \tag{16.80}$$

16.2.4 The integral as a function of the interval

Theorem 16.15. *Let $\alpha : \mathbb{R} \to \mathbb{R}$ be of bounded variation on $[a,b]$ and $f \in \mathcal{R}_{[a,b]}(\alpha)$. For any $x \in [a,b]$ define*

$$\boxed{F(x) := \int_{s=a}^{x} f(s)\,d\alpha(s)} \tag{16.81}$$

Then

(a) F is of bounded variation on $[a,b]$;
(b) Every point of continuity of α is also a point of continuity of F;
(c) If $f \uparrow$ on $[a,b]$ then the derivative $F'(x)$ exists at each point $x \in (a,b)$ where $\alpha'(x)$ exists and where f is continuous. For such x

$$\boxed{F'(x) = f(x)\,\alpha'(x)} \tag{16.82}$$

Proof. It is sufficient to assume that $\alpha \uparrow$ on $[a, b]$. If $x \neq y$ by (16.77) it follows that

$$F(y) - F(x) = \int_{s=x}^{y} f(s)\, d\alpha(s) = c\,[\alpha(y) - \alpha(x)]$$

where $c \in [m, M]$. So, statements (a) and (b) follow at once from this equation. To prove (c) it is sufficient to divide both sides by $(y - x)$ and observe that $c \to f(x)$ when $y \to x$. Theorem is proven. □

Corollary 16.6.

1. If $f \in \mathcal{R}_{[a,b]}(\alpha)$ then for any $x \in [a, b]$ and the functions F and G defined as

$$F(x) := \int_{s=a}^{x} f(s)\, ds \quad \text{and} \quad G(x) := \int_{s=a}^{x} g(s)\, ds$$

we have

$$\boxed{\int_{s=a}^{b} f(s)\, g(s)\, ds = \int_{s=a}^{b} f(s)\, dG(s) = \int_{s=a}^{b} g(s)\, dF(s)} \tag{16.83}$$

*2. In the Riemann case, when $\alpha(x) = x$, from (16.82) we obtain the, so-called, **first fundamental theorem** of integral calculus:*

$$\boxed{F'(x) = f(x)} \tag{16.84}$$

16.2.5 Derivative integration

Theorem 16.16. *Assume $f \in \mathcal{R}_{[a,b]}(\alpha)$ and g, defined on $[a, b]$, has the derivative g' in (a, b) such that for each $x \in (a, b)$*

$$g'(x) = f(x) \tag{16.85}$$

If in the endpoints

$$g(a) - g(a + 0) = g(b) - g(b - 0)$$

*then the **Newton–Leibniz formula** (the **second fundamental theorem** of integral calculus) holds, namely,*

$$\boxed{\int_{x=a}^{b} f(x)\, dx = \int_{x=a}^{b} g'(x)\, dx = g(b) - g(a)} \tag{16.86}$$

Proof. For every partition P_n of $[a, b]$ and in view of the mean-value theorem (16.77) we have

$$g(b) - g(a) = \sum_{i=1}^{n} [g(x_i) - g(x_{i-1})] = \sum_{i=1}^{n} g'(t_i)\,\Delta x_i = \sum_{i=1}^{n} f(t_i)\,\Delta x_i$$

where $t_i \in [x_{i-1}, x_i]$. But, since f is integrable, for any $\varepsilon > 0$ the partition P_n can be selected so fine that

$$\left| g(b) - g(a) - \int_{x=a}^{b} f(x)\,dx \right| = \left| \sum_{i=1}^{n} f(t_i)\,\Delta x_i - \int_{x=a}^{b} f(x)\,dx \right| < \varepsilon$$

which proves the theorem. □

16.2.6 Integrals depending on parameters and differentiation under integral sign

Theorem 16.17. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be continuous at each point $(x, y) \in Q$ where*

$$Q := \{(x, y) \mid a \le x \le b,\ c \le y \le d\} \tag{16.87}$$

Assume that α is of bounded variation on $[a, b]$ and F is the function defined on $[c, d]$ by the equation

$$\boxed{F(y) = \int_{x=a}^{b} f(x, y)\,d\alpha(x)} \tag{16.88}$$

Then F is continuous on $[c, d]$, or, in other words, if $y_0 \in [c, d]$ then

$$\boxed{\begin{aligned} \lim_{y \to y_0} \int_{x=a}^{b} f(x, y)\,d\alpha(x) &= \int_{x=a}^{b} \lim_{y \to y_0} f(x, y)\,d\alpha(x) \\ &= \int_{x=a}^{b} f(x, y_0)\,d\alpha(x) \end{aligned}} \tag{16.89}$$

Proof. Assume $\alpha \uparrow$ on $[a, b]$. Since Q is a compact then f is uniformly continuous on Q. Hence, for any $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon)$ such that for any pair of points $z := (x, y)$ and $z' := (x', y')$ such that $\|z - z'\| < \delta$ we have $|f(x, y) - f(x', y')| \le \varepsilon$. So, if $|y - y'| < \delta$ we have

$$\left| F(y) - F\left(y'\right) \right| \le \int_{x=a}^{b} \left| f(x, y) - f\left(x', y'\right) \right| d\alpha(x) \le \varepsilon\,[\alpha(b) - \alpha(a)]$$

which establishes the continuity of $F(y)$. □

Corollary 16.7. (The Riemann integral case) *If f is continuous on Q and $g \in \mathcal{R}_{[a,b]}(x)$ then function $F(y)$, defined by (16.88), is continuous on $[c,d]$, that is, if $y_0 \in [c,d]$ then*

$$\boxed{\begin{aligned}\lim_{y\to y_0}\int_{x=a}^{b} g(x)f(x,y)\,dx &= \int_{x=a}^{b} g(x)\lim_{y\to y_0} f(x,y)\,dx \\ &= \int_{x=a}^{b} g(x)f(x,y_0)\,dx\end{aligned}} \tag{16.90}$$

Proof. Define $G(x) := \int_{s=a}^{x} g(s)ds$. Then by (16.83) $F(y)$ may be represented as $F(y) = \int_{x=a}^{b} f(x,y)\,dG(x)$. Now, applying Theorem 16.17, we obtained the desired result. □

Theorem 16.17 permits to establish the following important result.

Theorem 16.18. *Assume that α is of bounded variation on $[a,b]$ and F defined on $[c,d]$ by the equation $F(y) = \int_{x=a}^{b} f(x,y)\,d\alpha(x)$ exists for every $y \in [c,d]$. If the partial derivative $\frac{\partial}{\partial y} f(x,y)$ is continuous on Q (16.87) then $F'(y)$ exists on $[c,d]$ and it is given by the formula*

$$\boxed{F'(y) = \int_{x=a}^{b} \frac{\partial}{\partial y} f(x,y)\,d\alpha(x)} \tag{16.91}$$

Proof. Assuming that $y_0 \in (c,d)$ then we have

$$\frac{F(y)-F(y_0)}{y-y_0} = \int_{x=a}^{b} \frac{f(x,y)-f(x,y_0)}{y-y_0}\,d\alpha(x)$$

$$= \int_{x=a}^{b} \frac{\partial}{\partial y} f(x,\bar{y})\,d\alpha(x), \quad \bar{y} \in [y_0, y]$$

Since $\frac{\partial}{\partial y} f(x,y)$ is continuous on Q, taking $y_0, y \to y_0$ we obtain the validity of (16.91) in the point $y = y_0$. Theorem is proven. □

The following statement can be checked directly.

Proposition 16.1. *If $\varphi_1(t)$ and $\varphi_2(t)$ are differentiable on $[a,b]$, the function $f(t,\tau) \in \mathcal{R}_{[a,b]}(\alpha)$ is differentiable on t and continuous on τ for any fixed $t \in [a,b]$, then*

$$\begin{aligned}&\frac{d}{dt}\int_{\tau=\varphi_1(t)}^{\varphi_2(t)} f(t,\tau)\,d\alpha(\tau)\\&=\varphi_2'(t)\,f(t,\varphi_2(t))\,\alpha'(\varphi_2(t))\\&\quad-\varphi_1'(t)\,f(t,\varphi_1(t))\,\alpha'(\varphi_1(t))+\int_{\tau=\varphi_1(t)}^{\varphi_2(t)}\frac{\partial}{\partial t}f(t,\tau)\,d\alpha(\tau)\end{aligned} \tag{16.92}$$

Particularly,

$$\begin{aligned}&\frac{d}{dt}\int_{\tau=\varphi_1(t)}^{\varphi_2(t)} f(t,\tau)\,d\tau\\&=\varphi_2'(t)\,f(t,\varphi_2(t))\\&\quad-\varphi_1'(t)\,f(t,\varphi_1(t))+\int_{\tau=\varphi_1(t)}^{\varphi_2(t)}\frac{\partial}{\partial t}f(t,\tau)\,d\alpha(\tau)\end{aligned} \tag{16.93}$$

16.3 On Lebesgue integrals

16.3.1 Lebesgue's monotone convergence theorem

Theorem 16.19. (The monotone convergence theorem) *Suppose $\mathcal{E} \in \mathfrak{M}$ and let $\{f_n\}$ be a sequence of measurable nonnegative functions such that for all $x \in \mathcal{E}$*

$$0 \le f_1(x) \le f_2(x) \le \cdots \le \tag{16.94}$$

Let f be defined by

$$f_n(x) \underset{n\to\infty}{\to} f(x), x \in \mathcal{E} \tag{16.95}$$

Then

$$\int_{\mathcal{E}} f_n\,d\mu \underset{n\to\infty}{\to} \int_{\mathcal{E}} f\,d\mu \tag{16.96}$$

Proof. By (16.94) it evidently follows that

$$\int_{\mathcal{E}} f_n \, d\mu \underset{n\to\infty}{\to} \alpha \leq \int_{\mathcal{E}} f \, d\mu \tag{16.97}$$

for some $\alpha \geq 0$ since $\int_{\mathcal{E}} f_n \, d\mu \leq \int_{\mathcal{E}} f \, d\mu$. Choose $c \in (0, 1)$, and let s be a simple measurable function (15.104) such that $0 \leq s \leq f$. Put

$$\mathcal{E}_n := \{x \mid f_n(x) \geq cs(x)\}, \quad n = 1, 2, \ldots$$

By (16.94) $\mathcal{E}_1 \subset \mathcal{E}_2 \subset \cdots$ and by (16.95) it follows that $\mathcal{E} = \bigcup_{n=1}^{\infty} \mathcal{E}_n$. For every n we have

$$\int_{\mathcal{E}} f_n \, d\mu \geq \int_{\mathcal{E}_n} f_n \, d\mu \geq c \int_{\mathcal{E}_n} s \, d\mu$$

Let now $n \to \infty$. By Theorem 15.15 we obtain $\alpha \geq c \int_{\mathcal{E}_n} s \, d\mu$. Letting $c \to 1$ we see that $\alpha \geq \int_{\mathcal{E}_n} s \, d\mu$ and (15.109) implies $\alpha \geq \int_{\mathcal{E}_n} f \, d\mu$ which together with (16.97) proves the theorem. □

Corollary 16.8. *Suppose f_i $(i = 1, 2)$ are Lebesgue measurable, that is, $f_i \in \mathcal{L}(\mu)$ on $\mathcal{E}$. Then $f = (f_1 + f_2) \in \mathcal{L}(\mu)$ on $\mathcal{E}$ and*

$$\boxed{\int_{\mathcal{E}} f \, d\mu = \int_{\mathcal{E}} f_1 \, d\mu + \int_{\mathcal{E}} f_2 \, d\mu} \tag{16.98}$$

Proof.

(a) Suppose, first, that $f_1 \geq 0$ and $f_2 \geq 0$. Choose monotonically increasing sequences $\{s_n'\}$ and $\{s_n''\}$ of nonnegative measurable simple functions which converge to f_1 and f_2, respectively. Since for simple functions (16.98) follows trivially, then for $s_n = s_n' + s_n''$ it follows that

$$\int_{\mathcal{E}} s_n \, d\mu = \int_{\mathcal{E}} s_n' \, d\mu + \int_{\mathcal{E}} s_n'' \, d\mu$$

Taking $n \to \infty$ and applying Theorem 16.19 we obtain (16.98).

(b) If $f_1 \geq 0$ and $f_2 \leq 0$ let us put

$$\mathcal{A} := \{x \mid f(x) \geq 0\}, \quad \mathcal{B} := \{x \mid f(x) < 0\}$$

Then it follows that f, f_1 and $(-f_2)$ are nonnegative on $\mathcal{A}$. Hence, by the previous consideration,

$$\int_{\mathcal{A}} f_1 \, d\mu = \int_{\mathcal{A}} f \, d\mu + \int_{\mathcal{A}} (-f_2) \, d\mu = \int_{\mathcal{A}} f \, d\mu - \int_{\mathcal{A}} f_2 \, d\mu \tag{16.99}$$

Similarly, $-f$, f_1 and $(-f_2)$ are nonnegative on $\mathcal{B}$, so that

$$\int_{\mathcal{B}} (-f_2)\, d\mu = \int_{\mathcal{B}} f_1\, d\mu + \int_{\mathcal{B}} (-f)\, d\mu = \int_{\mathcal{B}} f_1\, d\mu - \int_{\mathcal{B}} (f)\, d\mu \tag{16.100}$$

Adding (16.99) and (16.100) implies (16.98).

(c) In the general case, $\mathcal{E}$ can be discomposed into four sets $\mathcal{E}_i$ on each of which f_1 and f_2 are of constant sign. By previous considerations we have proved that

$$\int_{\mathcal{E}_i} f\, d\mu = \int_{\mathcal{E}_i} f_1\, d\mu + \int_{\mathcal{E}_i} f_2\, d\mu \left(i = \overline{1,4}\right)$$

and (16.98) follows by adding these four equations. □

Corollary 16.9. *Suppose $\mathcal{E} \in \mathfrak{M}$ and let $\{f_n\}$ be a sequence of measurable nonnegative functions such that*

$$\boxed{f(x) = \sum_{n=1}^{\infty} f_n(x), x \in \mathcal{E}} \tag{16.101}$$

Then

$$\boxed{\int_{\mathcal{E}} f\, d\mu = \sum_{n=1}^{\infty} \int_{\mathcal{E}} f_n\, d\mu} \tag{16.102}$$

Proof. The partial sum of (16.101) forms a monotonically increasing sequence that implies (16.102). □

16.3.2 Comparison with the Riemann integral

Let the measurable space $\mathcal{X}$ be the interval $[a, b]$ of the real line with the measure $\mu = m$ (the Lebesgue measure) and $\mathfrak{M}$ be the family of Lebesgue-measurable subsets of $[a, b]$, that is, the Borel σ-algebra.

Theorem 16.20.

(a) If $f \in \mathcal{R}_{[a,b]}(m)$ then $f \in \mathcal{L}(\mu)$ on $[a, b]$, that is, ***each function which is Riemann integrable on an interval is also Lebesgue integrable****, and also both integrals are equal, i.e.,*

$$\boxed{\int_{x=a}^{b} f(x)\, dx = \int_{\mathcal{E}=[a,b]} f\, d\mu} \tag{16.103}$$

(b) Suppose f is bounded on $[a, b]$. Then $f \in \mathcal{R}_{[a,b]}(m)$ if and only if f is ***continuous almost everywhere*** *on $[a, b]$.*

Proof.

(a) follows from Theorem 16.12. To prove (b) suppose that $\{P_k\}$ is a sequence of partitions of $[a, b]$ such that P_{k+1} is a refinement of P_k. Using the definition (16.72) for $\alpha(x) = x$, namely,

$$L(P_n, f, x) := \sum_{i=1}^{n} m_i \Delta x_i, \qquad U(P_n, f, x) := \sum_{i=1}^{n} M_i \Delta x_i$$

$$m_i := \inf_{x \in (x_{i-1}, x_i]} f(x), \qquad M_i := \sup_{x \in (x_{i-1}, x_i]} f(x)$$

and defining

$$U_n(x) = \sum_{i=1}^{n} M_i \chi\left(x \in (x_{i-1}, x_i]\right)$$

$$L_n(x) = \sum_{i=1}^{n} m_i \chi\left(x \in (x_{i-1}, x_i]\right)$$

such that $U_i(a) = L_i(a) = f(a)$ we obtain

$$L_1(x) \leq L_2(x) \leq \cdots \leq f(x) \leq \cdots \leq U_2(x) \leq U_1(x) \tag{16.104}$$

which leads to the existence of the limits

$$L(x) := \lim_{k \to \infty} L_k(x), \qquad U(x) := \lim_{k \to \infty} U_k(x)$$

for which for any $x \in [a, b]$

$$L(x) \leq f(x) \leq U(x)$$

By (16.104) and Theorem 16.19 it follows that there exist the integrals

$$I_L := \lim_{k \to \infty} L(P_n, f, x) = \int_{x=a}^{b} L(x)\, dx,$$

$$I_U := \lim_{k \to \infty} U(P_n, f, x) = \int_{x=a}^{b} U(x)\, dx$$

So far, nothing has been assumed about f except that it is bounded on $[a, b]$. To complete the proof note that $f \in \mathcal{R}_{[a,b]}(x)$ if and only if $I_L = I_U$, or equivalently, if and only if $\int_{x=a}^{b} L(x)\, dx = \int_{x=a}^{b} U(x)\, dx$. But, in view of the fact that $L(x) \leq U(x)$, this happens if and only if $L(x) = U(x)$ for almost all $x \in [a, b]$. This implies that

$L(x) = f(x) = U(x)$ for almost all $x \in [a, b]$, so far as f is measurable and (16.103) follows. □

16.3.3 Fatou's lemma

Lemma 16.4. *Suppose $\mathcal{E} \in \mathfrak{M}$ and let $\{f_n\}$ be a sequence of measurable nonnegative functions. Then*

$$\begin{aligned} &\int_{\mathcal{E}} \liminf_{n\to\infty} f_n \, d\mu \le \liminf_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu \\ &\le \limsup_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu \le \int_{\mathcal{E}} \limsup_{n\to\infty} f_n \, d\mu \end{aligned} \tag{16.105}$$

Proof. The intermediate inequality

$$\liminf_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu \le \limsup_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu$$

trivially follows from the definitions of the upper and lower limits. Denote for all $x \in \mathcal{E}$

$$\begin{aligned} g_n^-(x) &:= \inf_{k\ge n} f_k(x), \quad g_n^+(x) := \sup_{k\ge n} f_k(x) \\ f^-(x) &:= \liminf_{n\to\infty} f_n(x) = \lim_{n\to\infty} g_n^-(x) \\ f^+(x) &:= \limsup_{n\to\infty} f_n(x) = \lim_{n\to\infty} g_n^+(x) \end{aligned} \tag{16.106}$$

Then $g_n^-(x)$ and $g_n^+(x)$ are measurable on $\mathcal{E}$ (see (15.12)) and

$$\begin{aligned} &0 \le g_1^-(x) \le g_2^-(x) \le \cdots \\ &g_n^-(x) \le f_n(x), \quad g_n^-(x) \to f^-(x) \\ &\cdots \ge g_2^+(x) \ge g_1^+(x) \ge 0 \\ &g_n^+(x) \ge f_n(x), \quad g_n^+(x) \to f^+(x) \end{aligned}$$

The integration of the inequality $g_n^-(x) \le f_n(x)$ and the direct application of Theorem 16.19 for $\{g_1^-(x)\}$ leads to

$$\int_{\mathcal{E}} f^- \, d\mu \le \liminf_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu$$

Analogously, the integration of the inequality $g_n^+(x) \ge f_n(x)$ in view of Theorem 16.19 gives

$$\limsup_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu \le \int_{\mathcal{E}} f^+ \, d\mu$$

Lemma is proven. □

Strict inequalities may hold in (16.105) (see the example in Exercise 5 in Chapter 11 of Rudin (1976)).

16.3.4 Lebesgue's dominated convergence

Theorem 16.21. (on a dominate convergence) *Suppose* $\mathcal{E} \in \mathfrak{M}$ *and let* $\{f_n\}$ *be a sequence of measurable functions such that for all* $x \in \mathcal{E}$

$$\boxed{f_n(x) \underset{n\to\infty}{\to} f(x)} \tag{16.107}$$

If there exists a function $g \in \mathcal{L}(\mu)$ *on* $\mathcal{E}$ *such that for* $n = 1, 2, \ldots$

$$\boxed{|f_n(x)| \leq g(x)} \tag{16.108}$$

almost everywhere on $\mathcal{E}$*, then*

$$\boxed{\lim_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu = \int_{\mathcal{E}} f \, d\mu} \tag{16.109}$$

that is, the operation of $\lim_{n\to\infty}$ *and the Lebesgue integration can be interchanged if (16.107) and (16.108) are fulfilled.*

Proof. The inequality (16.108) and Theorem 15.7 imply that $f_n \in \mathcal{L}(\mu)$ and $f \in \mathcal{L}(\mu)$ on $\mathcal{E}$. Since $f_n + g \geq 0$ the Fatou's lemma 16.4 shows that

$$\begin{aligned} \int_{\mathcal{E}} (f+g) \, d\mu &\leq \liminf_{n\to\infty} \int_{\mathcal{E}} (f_n + g) \, d\mu \\ &\leq \liminf_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu + \liminf_{n\to\infty} \int_{\mathcal{E}} (g) \, d\mu \end{aligned}$$

or, equivalently,

$$\int_{\mathcal{E}} f \, d\mu \leq \liminf_{n\to\infty} \int_{\mathcal{E}} f_n \, d\mu \tag{16.110}$$

Similarly, since $g - f_n \geq 0$ we have

$$\begin{aligned} \int_{\mathcal{E}} (g - f) \, d\mu &\leq \liminf_{n\to\infty} \int_{\mathcal{E}} (g - f_n) \, d\mu \\ &\leq \liminf_{n\to\infty} \int_{\mathcal{E}} g \, d\mu + \liminf_{n\to\infty} \int_{\mathcal{E}} (-f_n) \, d\mu \end{aligned}$$

so that

$$-\int_{\mathcal{E}} f\, d\mu \leq \liminf_{n\to\infty} \int_{\mathcal{E}} (-f_n)\, d\mu$$

which is the same as

$$\int_{\mathcal{E}} f\, d\mu \geq \liminf_{n\to\infty} \int_{\mathcal{E}} f_n\, d\mu \tag{16.111}$$

Hence, (16.109) follows from (16.110) and (16.111). □

One important application of Theorem 16.21 refers to a bounded interval.

Theorem 16.22. (Apostol 1974) *Let $\mathcal{I}$ be a bounded interval. Assume that $\{f_n\}$ is a sequence of measurable functions in $\mathcal{L}(\mu)$ on $\mathcal{I}$ which is boundedly convergent almost everywhere on $\mathcal{I}$. That is, assume there is a limit function f and positive constant M such that*

$$\boxed{\lim_{n\to\infty} f_n(x) = f(x) \quad \text{and} \quad |f_n(x)| \leq M \quad \text{almost everywhere on } \mathcal{I}}$$

Then $f \in \mathcal{L}(\mu)$ and

$$\boxed{\lim_{n\to\infty} \int_{\mathcal{I}} f_n(x)\, d\mu = \int_{\mathcal{I}} f(x)\, d\mu}$$

Proof. It follows from Theorem 16.21 if we take $g(x) := M$ for all $x \in \mathcal{I}$. Then $g \in \mathcal{L}(\mu)$ on $\mathcal{I}$, since $\mathcal{I}$ is a bounded interval. Theorem is proven. □

16.3.5 Fubini's reduction theorem

The Lebesgue integral defined on subsets in $\mathbb{R}$ and described in Chapter 15 can be generalized to provide a theory of Lebesgue integration for the function defined on subsets of n-dimensional space $\mathbb{R}^n$. A multiple integral in $\mathbb{R}^n$ can be evaluated by calculating a succession of n one-dimensional integrals. This result is referred to as the *Fubini's theorem.*

Definition 16.6.

(a) If $I := I_1 \times I_2 \times \cdots \times I_n$ is a bounded interval in $\mathbb{R}^n$, where $I_k := [a_k, b_k]$, then ***the n-measure*** *$\mu(I)$ of I may be defined by the equation*

$$\boxed{\mu(I) := \mu(I_1) \cdots \mu(I_n)} \tag{16.112}$$

where $\mu(I_k)$ is the one-dimensional measure, or length, of I_k.

(b) Analogously to the single-dimensional case, a property is said to be of zero n-measure or, to hold ***almost everywhere*** *on a set $S \subset \mathbb{R}^n$, if it holds everywhere on S except for a subset of zero n-measure.*

If P_k is a partition of I_k, then the Cartesian product $P := P_1 \times \cdots \times P_n$ is called a partition of I. So that, if P_k decomposes I_k into m_k one-dimensional subintervals, then P decomposes I into $m = m_1 \cdots m_n$ n-dimensional subintervals, say $J_1, \ldots, J_m$.

Definition 16.7.

(a) A function s defined on I is called a ***step function if a partition*** *P of I exists such that s is constant on the interior of each subinterval J_k, say,*

$$s(x) = c_k \quad \text{if} \quad x \in J_k$$

(b) The ***n-dimensional Lebesgue integral*** *of s over I is defined by the relation*

$$\boxed{\int_I s \, d\mu := \sum_{k=1}^{n} c_k \mu(J_k)} \tag{16.113}$$

Definition 16.8. *A real-valued function f on $I \in \mathbb{R}^n$ is called an* ***upper function*** *on I, and we write $f \in U(I)$, if there exists an increasing (nondecreasing) sequence $\{s_n\}$ of step functions s_n such that*

(a) $s_n \to f$ almost everywhere on I,
(b) $\lim_{n\to\infty} \int_I s_n \, d\mu$ exists.

The sequence $\{s_n\}$ is said to generate f and the ***integral*** *f over I is defined by the equation*

$$\boxed{\int_I f \, d\mu := \lim_{n\to\infty} \int_I s_n \, d\mu} \tag{16.114}$$

The integral $\int_I f \, d\mu$ is also denoted by

$$\boxed{\int_I f(x) \, dx \overset{\text{or}}{=} \int_I f(x_1, \ldots, x_n) \, d(x_1, \ldots, x_n)} \tag{16.115}$$

The notation

$$\int_I f(x_1, \ldots, x_n) \, dx_1 \cdots dx_n$$

is also used. Double integrals are often written with two integral signs, namely,

$$\iint_I f(x, y) \, dx \, dy$$

Theorem 16.23. (Fubini's theorem for step functions) *Let s be a step function on $\mathbb{R}^2$. Then for each fixed $y \in \mathbb{R}$ the integral $\int_{\mathbb{R}} f(x, y)\, dx$ exists and, as a function of x, it is Lebesgue integrable on $\mathbb{R}$. Similarly, for each fixed $x \in \mathbb{R}$ the integral $\int_{\mathbb{R}} f(x, y)\, dy$ exists and, as a function of y, it is Lebesgue integrable on $\mathbb{R}$. Moreover, we have*

$$\boxed{\begin{aligned}&\iint_{\mathbb{R}^2} f(x, y)\, dx\, dy \\ &= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(x, y)\, dx \right] dy = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f(x, y)\, dy \right] dy\end{aligned}} \tag{16.116}$$

Proof. There is a compact interval $I = [a, b] \times [c, d]$ such that s is a step function on I and $s(x, y) = 0$ if $(x, y) \notin I$. There is a partition P of I into mn subrectangles $I_{ij} = [x_{i-1}, x_i] \times [y_{i-1}, y_i]$ such that s is constant on I_{ij}, say,

$$s(x, y) = c_{ij} \quad \text{if} \quad (x, y) \in \operatorname{int} I_{ij}$$

Then

$$\begin{aligned}&\iint_{I_{ij}} f(x, y)\, dx\, dy = c_{ij} (x_i - x_{i-1}) (y_i - y_{i-1}) \\ &= \int_{y_{i-1}}^{y_i} \left[\int_{x_{i-1}}^{x_i} f(x, y)\, dx \right] dy = \int_{x_{i-1}}^{x_i} \left[\int_{y_{i-1}}^{y_i} f(x, y)\, dy \right] dx\end{aligned}$$

Summing on i and j we find

$$\begin{aligned}&\iint_I f(x, y)\, dx\, dy \\ &= \int_c^d \left[\int_a^b f(x, y)\, dx \right] dy = \int_a^b \left[\int_c^d f(x, y)\, dy \right] dx\end{aligned}$$

Since s vanishes outside I this proves (16.116). □

The next theorem is the extension of the previous result to the general class of Lebesgue integrable functions.

Theorem 16.24. (Fubini's theorem for double integrals) *Assume f is Lebesgue integrable on $\mathbb{R}^2$. Then (16.116) holds.*

Proof.

(a) First, let us prove this result for upper functions. If $f \in U\left(\mathbb{R}^2\right)$ then there exists an increasing (nondecreasing) sequence $\{s_n\}$ of step functions s_n such that $s_n(x, y) \to f(x, y)$ for all $(x, y) \in \mathbb{R}^2 - S$ (S is a set of measure zero). Hence, by (16.114)

$$\iint_{\mathbb{R}^2} f(x, y)\, dx\, dy = \lim_{n\to\infty} \iint_{\mathbb{R}^2} s_n(x, y)\, dx\, dy$$

and (16.116) results from Theorem 16.23.

(b) To prove (16.116) for Lebesgue functions it is sufficient to notice that any $f \in \mathcal{L}(\mu)$ can be represented as $f = u - v$ where $u \in U\left(\mathbb{R}^2\right)$ and $v \in U\left(\mathbb{R}^2\right)$. Theorem is proven.

□

Corollary 16.10. *Assume that f is defined and* ***bounded*** *on a compact rectangle $I = [a, b] \times [c, d]$, and also that f is* ***continuous almost everywhere*** *on I. Then $f \in \mathcal{L}(\mu)$ on I and*

$$\iint_I f(x, y)\, dx\, dy = \int_c^d \left[\int_a^b f(x, y)\, dx\right] dy = \int_a^b \left[\int_c^d f(x, y)\, dy\right] dx \tag{16.117}$$

Corollary 16.11. *If f is Lebesgue integrable on $\mathbb{R}^{m+k}$ then the following extension of the Fubini's theorem 16.24 to high-dimensional integrals holds:*

$$\int_{\mathbb{R}^{m+k}} f d\mu = \int_{\mathbb{R}^k} \left[\int_{\mathbb{R}^m} f(x, \mathbf{y})\, dx\right] d\mathbf{y} = \int_{\mathbb{R}^m} \left[\int_{\mathbb{R}^k} f(x, \mathbf{y})\, d\mathbf{y}\right] d\mathbf{y} \tag{16.118}$$

16.3.5.1 Tonelli–Hobson test for integrability in $\mathbb{R}^2$

Theorem 16.25. (The Tonelli–Hobson theorem) *Assume that f is measurable on $\mathbb{R}^2$ and that at least one of two iterated integrals*

$$\int_{\mathbb{R}} \left[\int_{\mathbb{R}} |f(x, y)|\, dx\right] dy \quad or \quad \int_{\mathbb{R}} \left[\int_{\mathbb{R}} |f(x, y)|\, dy\right] dy$$

exists. Then

(a) $f \in \mathcal{L}(\mu)$ on $\mathbb{R}^2$;
(b) The formula (16.116) holds.

Proof. Part (b) follows from (a) because of the Fubini's theorem (16.24). To prove (a) assume that the iterated integral

$$\int_{\mathbb{R}} \left[\int_{\mathbb{R}} |f(x, y)| \, dx \right] dy$$

exists. Let $\{s_n\}$ be an increasing (nondecreasing) sequence of nonnegative step functions defined by the formula

$$s_n(x, y) = \begin{cases} n & \text{if} \quad |x| \leq n \quad \text{and} \quad |y| \leq n \\ 0 & \text{if} \qquad \text{otherwise} \end{cases}$$

Let also $f_n(x, y) := \min\{s_n(x, y), |f(x, y)|\}$. Notice that both s_n and $|f|$ are measurable on $\mathbb{R}^2$. So, f_n is measurable and, since

$$0 \leq |f_n(x, y)| \leq s_n(x, y)$$

so f_n is dominated by a Lebesgue integrable function. Therefore, by Theorem 16.21 $f_n \in \mathcal{L}(\mu)$ on $\mathbb{R}^2$. Hence, we can apply Fubini's theorem 16.24 to f_n along with the inequality

$$0 \leq f_n(x, y) \leq |f_n(x, y)|$$

to obtain

$$\int_{\mathbb{R}^2} f_n \, d\mu = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} f_n(x, y) \, dx \right] dy \leq \int_{\mathbb{R}} \left[\int_{\mathbb{R}} |f_n(x, y)| \, dx \right] dy$$

Since $\{f_n\}$ is increasing this shows that $\lim_{n\to\infty} \int_{\mathbb{R}^2} f_n \, d\mu$ exists. But $\{f_n(x, y)\} \to |f_n(x, y)|$ almost everywhere on $\mathbb{R}^2$. So, $|f| \in \mathcal{L}(\mu)$ on $\mathbb{R}^2$. Since f is measurable, it follows that $f \in \mathcal{L}(\mu)$ on $\mathbb{R}^2$ which proves (a). The proof is similar if the other integral exists. Theorem is proven. □

16.3.6 Coordinate transformation in an integral

Definition 16.9. *Let T be an open set of $\mathbb{R}^n$. A vector function $\mathbf{g} : T \to \mathbb{R}^n$ is called* ***a coordinate transformation*** *(or* ***deffeomorphism****) on T if it has the following three properties:*

(a) $\mathbf{g} \in C^1$ on T, that is, $\mathbf{g}$ is continuously differentiable ($\mathbf{g}$ has the first-order partials which are continuous) on T;
(b) $\mathbf{g}$ is globally one-to-one on T;

(c) for all $\mathbf{t} \in T$ *the Jacobian determinant* $J_{\mathbf{g}}(\mathbf{t})$ *of the transformation* $\mathbf{g}$ *is not equal to zero, that is,*

$$\boxed{J_{\mathbf{g}}(\mathbf{t}) := \det\left[\frac{\partial}{\partial t_k}\mathbf{g}_i(\mathbf{t})\right]_{i,k=1,n} \neq 0} \tag{16.119}$$

Remark 16.5. *The properties of the coordinate transformation* g *mentioned above provide the existence of* $\mathbf{g}^{-1}$ *which is also continuously differentiable on* $\mathbf{g}(T)$.

Remark 16.6. (The Jacobian chain rule) Assume that $\mathbf{g}$ is a coordinate transformation on T and that $\mathbf{h}$ is a coordinate transformation on the image $\mathbf{g}(T)$. Then the composition

$$\boxed{\mathbf{k} = \mathbf{h} \circ \mathbf{g} := \mathbf{h}(\mathbf{g}(\mathbf{t}))} \tag{16.120}$$

is also a deffeomorphism on T with the Jacobian determinant $J_{\mathbf{k}}(\mathbf{t})$ satisfying the equation

$$\boxed{J_{\mathbf{k}}(\mathbf{t}) = J_{\mathbf{h}}(\mathbf{g}(\mathbf{t}))\, J_{\mathbf{g}}(\mathbf{t})} \tag{16.121}$$

Proof. It follows from the relations

$$\left[\frac{\partial}{\partial t_k}\mathbf{h}_i(\mathbf{t})\right]_{i,k=1,n} = \left[\frac{\partial}{\partial g_s}\mathbf{h}_i(\mathbf{t})\right]_{i,s=1,n}\left[\frac{\partial}{\partial t_k}\mathbf{g}_s(\mathbf{t})\right]_{s,k=1,n}$$

$$\det\left[\frac{\partial}{\partial t_k}\mathbf{h}_i(\mathbf{t})\right]_{i,k=1,n} = \det\left[\frac{\partial}{\partial g_s}\mathbf{h}_i(\mathbf{g}(\mathbf{t}))\right]_{i,s=1,n} \det\left[\frac{\partial}{\partial t_k}\mathbf{g}_s(\mathbf{t})\right]_{s,k=1,n}$$

□

Theorem 16.26. *Let* T *be an open subset of* $\mathbb{R}^n$, $\mathbf{g}$ *be a coordinate transformation on* T *and* f *be a real-valued function defined on the image* $\mathbf{g}(T)$ *such that the Lebesgue integral* $\int_{\mathbf{g}(T)} f(x)\, dx$ *exists. Then*

$$\boxed{\int_{\mathbf{g}(T)} f(x)\, dx = \int_T f(\mathbf{g}(\mathbf{t}))\, |J_{\mathbf{g}}(\mathbf{t})|\, d\mathbf{t}} \tag{16.122}$$

Proof. The proof is divided into three parts:

(a) Part 1 shows that (16.122) holds for every linear coordinate transformation $\alpha : T \to \mathbb{R}^n$ with the corollary that

$$\boxed{\mu(\alpha(\mathcal{A})) = |\det \alpha|\, \mu(\mathcal{A})} \tag{16.123}$$

for any subset $\mathcal{A} \subset T$.

(b) In Part 2 one needs to consider a general coordinate transformation $\mathbf{g} : T \to \mathbb{R}^n$ and show that (16.122) is valid when f is the characteristic function of a compact cube $\mathcal{K} \subset \mathbf{g}(T)$ that gives

$$\boxed{\mu(\mathcal{K}) = \int\limits_{g^{-1}(\mathcal{K})} \left|J_{\mathbf{g}}(\mathbf{t})\right| d\mathbf{t}} \tag{16.124}$$

(c) In Part 3 equation (16.124) is used to deduce (16.122) in the general form. The details of this proof can be found in Chapter 15.10 of Apostol (1974).

□

Example 16.3. (The spherical coordinates transformation) *Let us take* $\mathbf{t} := (\rho, \theta, \varphi)$ *and*

$$T := \left\{\mathbf{t} : \rho > 0, \quad \theta \in [0, 2\pi], \quad \varphi \in [0, \pi]\right\}$$

The coordinate transformation $\mathbf{g}$ *maps each point* $\mathbf{t} = (\rho, \theta, \varphi) \in T$ *onto the point* (x, y, z) *in* $\mathbf{g}(T)$ *given by the equations*

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \rho \cos\theta \sin\varphi \\ \rho \sin\theta \sin\varphi \\ \rho \cos\varphi \end{pmatrix}$$

The Jacobian determinant is

$$J_{\mathbf{g}}(\mathbf{t}) = \det \begin{bmatrix} \dfrac{\partial x}{\partial \rho} & \dfrac{\partial x}{\partial \theta} & \dfrac{\partial x}{\partial \varphi} \\ \dfrac{\partial y}{\partial \rho} & \dfrac{\partial y}{\partial \theta} & \dfrac{\partial y}{\partial \varphi} \\ \dfrac{\partial z}{\partial \rho} & \dfrac{\partial z}{\partial \theta} & \dfrac{\partial z}{\partial \varphi} \end{bmatrix}$$

$$= \det \begin{bmatrix} \cos\theta \sin\varphi & \sin\theta \sin\varphi & \cos\varphi \\ -\rho \sin\theta \sin\varphi & \rho \cos\theta \sin\varphi & 0 \\ \rho \cos\theta \cos\varphi & \rho \sin\theta \cos\varphi & -\rho \sin\theta \end{bmatrix} = -\rho^2 \sin\theta$$

So, for any f *to be a real-valued function defined on the image* $\mathbf{g}(T)$ *we have*

$$\iiint\limits_{\mathbf{g}(T)} f(x, y, z)\, dx\, dy\, dz$$
$$= \iiint\limits_{(\rho,\theta,\varphi)\in T} f(\rho \cos\theta \sin\varphi, \rho \sin\theta \sin\varphi, \rho \cos\varphi)\, \rho^2 |\sin\theta|\, d\rho\, d\theta\, d\varphi$$

16.4 Integral inequalities

16.4.1 Generalized Chebyshev inequality

Theorem 16.27. (The generalized Chebyshev inequality) *Let* $g : \mathbb{R} \to \mathbb{R}$ *be a nonnegative nondecreasing function defined on the interval* $[0, \infty)$, *i.e.,*

$$\boxed{g(x) \geq 0 \ \forall x \in [0, \infty), \quad g(x_1) \geq g(x_2) \ \forall x_1 \geq x_2} \tag{16.125}$$

and $\varphi \in \mathcal{L}(\mu)$ *on* $\mathcal{E} \subset \mathbb{R}$ *such that* $g(|\varphi|) \in \mathcal{L}(\mu)$ *on* $\mathcal{E}$, *that is,*

$$\boxed{\int_{\mathcal{E}} g(|\varphi|) \, d\mu < \infty} \tag{16.126}$$

Then for any nonnegative value $a \geq 0$ *the following inequality holds:*

$$\boxed{\int_{\mathcal{E}} g(|\varphi|) \, d\mu \geq g(a)\mu(\{x \mid |\varphi(x)| \geq a\})} \tag{16.127}$$

Proof. By the additivity property of the Lebesgue integral for $\mathcal{E} = \mathcal{E}_1 \cup \mathcal{E}_2$ where

$$\mathcal{E}_1 := \{x \mid |\varphi(x)| \geq a\}$$
$$\mathcal{E}_2 := \{x \mid |\varphi(x)| < a\}$$

and in view of the assumptions of this theorem it follows that

$$\int_{\mathcal{E}} g(|\varphi|) \, d\mu = \int_{\mathcal{E}_1} g(|\varphi|) \, d\mu + \int_{\mathcal{E}_2} g(|\varphi|) \, d\mu$$
$$\geq \int_{\mathcal{E}_1} g(|\varphi|) \, d\mu \geq \int_{\mathcal{E}_1} g(a) \, d\mu = g(a)\,\mu(\mathcal{E}_1)$$

which completes the proof. □

16.4.2 Markov and Chebyshev inequalities

Using the generalized Chebyshev inequality (16.127) one can obtain the following important and commonly used integral relations known as *Markov* and *Chebyshev inequalities*.

Theorem 16.28. (The Markov inequality) *Put in (16.127)*

$$\boxed{g(x) = x^r, \quad x \in [0, \infty), \quad r > 0} \tag{16.128}$$

Then for any $a > 0$ the inequality (16.127) becomes

$$\boxed{\mu\left(\{x \mid |\varphi(x)| \geq a\}\right) \leq a^{-r} \int\limits_{\mathcal{E}} |\varphi|^r \, d\mu} \tag{16.129}$$

Two partial cases corresponding to $r = 1, 2$ present a special interest.

Corollary 16.12. (The first Chebyshev inequality) *For $r = 1$ the Markov inequality (16.129) becomes*

$$\boxed{\mu\left(\{x \mid |\varphi(x)| \geq a\}\right) \leq \frac{1}{a} \int\limits_{\mathcal{E}} |\varphi| \, d\mu} \tag{16.130}$$

Corollary 16.13. (The second Chebyshev inequality) *For $r = 2$ the Markov inequality (16.129) becomes*

$$\boxed{\mu\left(\{x \mid |\varphi(x)| \geq a\}\right) \leq \frac{1}{a^2} \int\limits_{\mathcal{E}} \varphi^2 \, d\mu} \tag{16.131}$$

16.4.3 Hölder inequality

Theorem 16.29. (The Hölder inequality) *Let p and q be positive values such that*

$$\boxed{p > 1, \quad q > 1, \quad p^{-1} + q^{-1} = 1} \tag{16.132}$$

and $\varphi, \eta \in \mathcal{L}(\mu)$ on $\mathcal{E} \subset \mathbb{R}$ such that

$$\boxed{|\varphi|^p \in \mathcal{L}(\mu), \quad \{|\eta|^q \in \mathcal{L}(\mu)\}} \tag{16.133}$$

Then the following inequality holds:

$$\boxed{\int\limits_{\mathcal{E}} |\varphi\eta| \, d\mu \leq \left(\int\limits_{\mathcal{E}} |\varphi|^p \, d\mu\right)^{1/p} \left(\int\limits_{\mathcal{E}} |\eta|^q \, d\mu\right)^{1/q}} \tag{16.134}$$

Proof. If $\int_{\mathcal{E}} |\varphi|^p \, d\mu = \int_{\mathcal{E}} |\eta|^q \, d\mu = 0$ on $\mathcal{E}$ then $\varphi(x) = \eta(x) = 0$ almost everywhere on $\mathcal{E}$ and (16.134) looks trivial. Suppose that $\int_{\mathcal{E}} |\varphi|^p \, d\mu > 0$ and $\int_{\mathcal{E}} |\eta|^q \, d\mu > 0$. Since the function $\ln(x)$ is concave for any $x, y, a, b > 0$ the following inequality holds:

$$\boxed{\ln(ax + by) \geq a \ln(x) + b \ln(y)} \tag{16.135}$$

or, equivalently,

$$\boxed{ax + by \geq x^a y^b} \tag{16.136}$$

Taking $a := 1/p$, $b := 1/q$ and

$$x \triangleq \frac{|\varphi|^p}{\int\limits_{\mathcal{E}} |\varphi|^p \, d\mu}, \quad y \triangleq \frac{|\eta|^p}{\int\limits_{\mathcal{E}} |\eta|^p \, d\mu}$$

implies

$$1/p \frac{|\varphi|^p}{\int\limits_{\mathcal{E}} |\varphi|^p \, d\mu} + 1/q \frac{|\eta|^p}{\int\limits_{\mathcal{E}} |\eta|^p \, d\mu} \geq \frac{|\varphi|}{\left(\int\limits_{\mathcal{E}} |\varphi|^p \, d\mu\right)^{1/p}} \frac{|\eta|}{\left(\int\limits_{\mathcal{E}} |\eta|^p \, d\mu\right)^{1/q}}$$

Integrating both sides of this inequality and using the assumption that $p^{-1} + q^{-1} = 1$ proves (16.134). □

Corollary 16.14. *In the case of the* ***Borel measure*** *when*

$$\mu(\{x \mid x < c \in (a, b)\}) = c - a$$

on $\mathcal{E} = [a, b]$ *we have*

$$\int\limits_{\mathcal{E}} \varphi \, d\mu = \int\limits_{x=a}^{b} \varphi(x) \, dx, \quad \int\limits_{\mathcal{E}} \eta \, d\mu = \int\limits_{x=a}^{b} \eta(x) \, dx$$

and (16.134) becomes

$$\boxed{\begin{aligned} &\int\limits_{x=a}^{b} |\varphi(x) \, \eta(x)| \, dx \\ &\leq \left(\int\limits_{x=a}^{b} |\varphi(x)|^p \, dx\right)^{1/p} \left(\int\limits_{x=a}^{b} |\eta(x)|^q \, dx\right)^{1/q} \end{aligned}} \tag{16.137}$$

Corollary 16.15. *In the vector case when*

$$\varphi := (\varphi_1, \ldots, \varphi_n) \in R^n, \quad \eta := (\eta_1, \ldots, \varphi_n) \in R^n$$

which corresponds to the ***"atomic" measure*** *concentrated in the points* $x = \{x_1, \ldots, x_n\}$ *with the weights* $\mu := (\mu_1, \ldots, \mu_n)$, $\mu_i \geq 0$ $(i = 1, \ldots, n)$ *we have*

$$d\mu = \sum_{i=1}^{n} \delta(x - x_i)\, \mu_i\, dx, \quad \varphi_i := \varphi(x_i), \quad \eta_i := \eta(x_i)$$

$$\int_{\mathcal{E}} \varphi\, d\mu = \sum_{i=1}^{n} \varphi_i \mu_i, \quad \int_{\mathcal{E}} \eta\, d\mu = \sum_{i=1}^{n} \eta_i \mu_i \tag{16.138}$$

and (16.134) becomes

$$\boxed{\sum_{i=1}^{n} |\varphi_i \eta_i|\, \mu_i \leq \left(\sum_{i=1}^{n} |\varphi_i|^p\, \mu_i\right)^{1/p} \left(\sum_{i=1}^{n} |\eta_i|^q\, \mu_i\right)^{1/q}} \tag{16.139}$$

For the ***"atomic" uniform measure*** *when* $\mu_i := \mu_0/n$ $(\mu_0 > 0)$ *we have*

$$\boxed{\frac{\mu_0}{n} \sum_{i=1}^{n} |\varphi_i \eta_i| \leq \left(\frac{\mu_0}{n} \sum_{i=1}^{n} |\varphi_i|^p\right)^{1/p} \left(\frac{\mu_0}{n} \sum_{i=1}^{n} |\eta_i|^q\right)^{1/q}} \tag{16.140}$$

or, equivalently,

$$\boxed{\sum_{i=1}^{n} |\varphi_i \eta_i| \leq \left(\sum_{i=1}^{n} |\varphi_i|^p\right)^{1/p} \left(\sum_{i=1}^{n} |\eta_i|^q\right)^{1/q}} \tag{16.141}$$

16.4.4 Cauchy–Bounyakovski–Schwarz inequality

The following particular case $p = q = 2$ of (16.134) is the most common in use.

Corollary 16.16. (The CBS inequality)

$$\boxed{\int_{\mathcal{E}} |\varphi\eta|\, d\mu \leq \sqrt{\int_{\mathcal{E}} |\varphi|^2\, d\mu} \sqrt{\int_{\mathcal{E}} |\eta|^2\, d\mu}} \tag{16.142}$$

and the equality in (16.142) is reached if

$$\boxed{\varphi(x) = k\eta(x) \quad \textit{for any real} \quad k} \tag{16.143}$$

and almost all $x \in \mathcal{E}$.

Proof. To prove (16.143) it is sufficient to substitute $\varphi(x) = k\eta(x)$ into (16.142). □

Corollary 16.17. *1. In the* ***Borel measure case*** *we have*

$$\int_{x=a}^{b} |\varphi(x)\,\eta(x)|\,dx \le \sqrt{\int_{x=a}^{b} |\varphi(x)|^2\,dx}\sqrt{\int_{x=a}^{b} |\eta(x)|^2\,dx} \tag{16.144}$$

2. In the case of the ***"atomic" measure*** *for any nonnegative* $\mu_i \ge 0$ $(i = 1, \ldots, n)$ *we have*

$$\sum_{i=1}^{n} |\varphi_i \eta_i|\,\mu_i \le \sqrt{\sum_{i=1}^{n} |\varphi_i|^2\,\mu_i}\sqrt{\sum_{i=1}^{n} |\eta_i|^2\,\mu_i} \tag{16.145}$$

which for the uniform measure when

$$\mu_i := \mu_0/n$$

becomes

$$\frac{\mu_0}{n}\sum_{i=1}^{n} |\varphi_i \eta_i| \le \sqrt{\frac{\mu_0}{n}\sum_{i=1}^{n} |\varphi_i|^2}\sqrt{\frac{\mu_0}{n}\sum_{i=1}^{n} |\eta_i|^2} \tag{16.146}$$

or, equivalently,

$$\sum_{i=1}^{n} |\varphi_i \eta_i| \le \sqrt{\sum_{i=1}^{n} |\varphi_i|^2}\sqrt{\sum_{i=1}^{n} |\eta_i|^2} \tag{16.147}$$

16.4.5 Jensen inequality

Theorem 16.30. (The Jensen inequality) *Let* $g_\cup : \mathbb{R} \to \mathbb{R}$ *and* $g_\cap : \mathbb{R} \to \mathbb{R}$ *be convex downward (or, simply,* ***convex****) and convex upward (or, simply,* ***concave****), respectively (see Fig. 16.1) and* $\varphi : \mathbb{R} \to \mathbb{R}$ *be a measurable function such that* $\varphi \in \mathcal{L}(\mu)$ *on* $\mathcal{E} \subset \mathbb{R}$. *Let also* $\int_{\mathcal{E}} d\mu = 1$. *Then*

$$g_\cup\left(\int_{\mathcal{E}} \varphi\,d\mu\right) \le \int_{\mathcal{E}} g_\cup(\varphi)\,d\mu \tag{16.148}$$

and

$$g_\cap\left(\int_{\mathcal{E}} \varphi\,d\mu\right) \ge \int_{\mathcal{E}} g_\cap(\varphi)\,d\mu \tag{16.149}$$

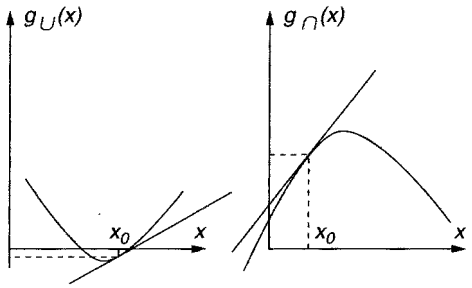

Fig. 16.1. The convex $g_{\cup}(x)$ and concave $g_{\cap}(x)$ functions.

Proof. By the convexity (concavity) definition (see Fig. 16.1) we may conclude that in both convexity and concavity cases there exists a number $\lambda(x_0)$ such that for any $x, x_0 \in \mathbb{R}$ the following inequalities are fulfilled:

$$\begin{aligned} g_{\cup}(x) &\geq g_{\cup}(x_0) + \lambda(x_0)(x - x_0) \\ g_{\cap}(x) &\leq g_{\cap}(x_0) + \lambda(x_0)(x - x_0) \end{aligned} \tag{16.150}$$

Taking $x := \varphi(x)$, $x_0 := \int_{\mathcal{E}} \varphi \, d\mu$ in (16.150) we obtain

$$g_{\cup}(\varphi(x)) \geq g_{\cup}\left(\int_{\mathcal{E}} \varphi \, d\mu\right) + \lambda\left(\int_{\mathcal{E}} \varphi \, d\mu\right)\left(\varphi(x) - \int_{\mathcal{E}} \varphi \, d\mu\right)$$

$$g_{\cap}(\varphi(x)) \leq g_{\cap}\left(\int_{\mathcal{E}} \varphi \, d\mu\right) + \lambda\left(\int_{\mathcal{E}} \varphi \, d\mu r\right)\left(\varphi(x) - \int_{\mathcal{E}} \varphi \, d\mu\right)$$

The application of the Lebesgue integration to both sides of these inequalities leads to (16.148) and (16.149), respectively. Theorem is proven. □

Corollary 16.18. *1. In the* ***Borel measure case*** $\left(\text{when } d\mu = \dfrac{dx}{b-a}\right)$ *we have*

$$\boxed{\begin{aligned} g_{\cup}\left(\frac{1}{b-a}\int_{x=a}^{b} \varphi(x)\, dx\right) &\leq \frac{1}{b-a}\int_{x=a}^{b} g_{\cup}(\varphi(x))\, dx \\ g_{\cap}\left(\frac{1}{b-a}\int_{x=a}^{b} \varphi(x)\, dx\right) &\geq \frac{1}{b-a}\int_{x=a}^{b} g_{\cap}(\varphi(x))\, dx \end{aligned}} \tag{16.151}$$

2. *In the case of the* ***"atomic" measure*** *for any nonnegative* $\mu_i \geq 0$ $(i = 1, \ldots, n)$ *such that* $\sum_{i=1}^{n} \mu_i = 1$ *we have*

$$\boxed{\begin{aligned} g_{\cup}\left(\sum_{i=1}^{n} \varphi_i \mu_i\right) &\leq \sum_{i=1}^{n} g_{\cup}(\varphi_i)\, \mu_i \\ g_{\cap}\left(\sum_{i=1}^{n} \varphi_i \mu_i\right) &\geq \sum_{i=1}^{n} g_{\cap}(\varphi_i)\, \mu_i \end{aligned}} \tag{16.152}$$

which for the uniform measure when $\mu_i := 1/n$ *becomes*

$$\boxed{\begin{aligned} g_{\cup}\left(\frac{1}{n}\sum_{i=1}^{n} \varphi_i\right) &\leq \frac{1}{n}\sum_{i=1}^{n} g_{\cup}(\varphi_i) \\ g_{\cap}\left(\frac{1}{n}\sum_{i=1}^{n} \varphi_i\right) &\geq \frac{1}{n}\sum_{i=1}^{n} g_{\cap}(\varphi_i) \end{aligned}} \tag{16.153}$$

3. *For any* $n = 1, 2, \ldots,$
   (a) *any even* $k = 2s$ $(s = 1, 2, \ldots)$ *and any* $\varphi_i \in \mathbb{R}$ $(i = 1, \ldots n)$ *it follows that*

$$\boxed{\left(\sum_{i=1}^{n} \varphi_i\right)^k \leq n^{k-1} \sum_{i=1}^{n} (\varphi_i)^k} \tag{16.154}$$

   (b) *any odd* $k = 2s - 1$ $(s = 1, 2, \ldots)$ *and any* $\varphi_i \geq 0$ $(i = 1, \ldots n)$ *the inequality (16.154) also holds.*

Proof. Indeed, by (16.153) we have

$$\left(\frac{1}{n}\sum_{i=1}^{n} \varphi_i\right)^k \leq \frac{1}{n}\sum_{i=1}^{n} (\varphi_i)^k$$

which implies (16.154) for an even k, since the function x^k is convex in all axis $\mathbb{R}$. For an even k this function is convex only at the semi-axis $[0, \infty]$ which permits to use the inequality (16.153) only within this region. □

Example 16.4. *For* $g_{\cap}(x) := \ln(|x|)$ *we have*

$$\boxed{\ln\left(\int_{\mathcal{E}} |\varphi|\ d\mu\right) \geq \int_{\mathcal{E}} \ln(|\varphi|)\ d\mu} \tag{16.155}$$

and

$$\boxed{\begin{aligned} &\ln\left(\sum_{i=1}^{n}|\varphi_i|\,\mu_i\right) \geq \sum_{i=1}^{n}\ln\left(|\varphi_i|\right)\mu_i \\ &\ln\left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i|\right) \geq \frac{1}{n}\sum_{i=1}^{n}\ln\left(|\varphi_i|\right) \end{aligned}} \tag{16.156}$$

valid for any $\mu_i > 0$ $(i = 0, 1, \ldots, n)$ *such that* $\sum_{i=1}^{n}\mu_i = 1$.

Corollary 16.19. (The weighted norm case) *If* $\varphi : R \to R^n$ *and* $P = P^{\mathsf{T}} \geq 0$ *then*

$$\boxed{\left\|\frac{1}{b-a}\int_{x=a}^{b}\varphi(x)\,dx\right\|_P^2 \leq \frac{1}{b-a}\int_{x=a}^{b}\|\varphi(x)\|_P^2\,dx} \tag{16.157}$$

Proof. By the definition of the weighted norm and using the matrix-root representation we have

$$\begin{aligned} \left\|\int_{x=a}^{b}\varphi(x)\,dx\right\|_P^2 &= \left(\int_{x=a}^{b}\varphi(x)\,dx, P\int_{x=a}^{b}\varphi(x)\,dx\right) \\ &= \left(P^{1/2}\int_{x=a}^{b}\varphi(x)\,dx,\ P^{1/2}\int_{x=a}^{b}\varphi(x)\,dx\right) \\ &= \left(\int_{x=a}^{b}z(x)\,dx,\ \int_{x=a}^{b}z(x)\,dx\right) \end{aligned}$$

where $z(x) := P^{1/2}\varphi(x)$. Hence, it follows that

$$\left\|\frac{1}{b-a}\int_{x=a}^{b}\varphi(x)\,dx\right\|_P^2 = \sum_{i=1}^{n}\left(\frac{1}{b-a}\int_{x=a}^{b}z_i(x)\,dx\right)^2$$

Applying (16.151) for $g_{\cup}(s) = s^2$ to each term in the sum on the right-hand side we have

$$\sum_{i=1}^{n}\left(\frac{1}{b-a}\int_{x=a}^{b}z_i(x)\,dx\right)^2 \leq \sum_{i=1}^{n}\frac{1}{b-a}\int_{x=a}^{b}z_i^2(x)\,dx = \frac{1}{b-a}\int_{x=a}^{b}\sum_{i=1}^{n}z_i^2(x)\,dx$$

$$= \frac{1}{b-a} \int_{x=a}^{b} \|z(x)\|^2 \, dx = \frac{1}{b-a} \int_{x=a}^{b} \left\|P^{1/2}\varphi(x)\right\|^2 \, dx$$

$$= \frac{1}{b-a} \int_{x=a}^{b} (\varphi(x), \; P\varphi(x)) \, dx = \frac{1}{b-a} \int_{x=a}^{b} \|\varphi(x)\|_P^2 \, dx$$

which proves (16.157). □

16.4.6 Lyapunov inequality

The inequality below is a particular case of the Jensen inequality (16.148).

Corollary 16.20. (The Lyapunov inequality) *For any measurable function $\varphi : \mathbb{R} \to \mathbb{R}$ such that $|\varphi|^t \in \mathcal{L}(\mu)$ on $\mathcal{E} \subset \mathbb{R}$ $(t > 0)$ and when $\int_{\mathcal{E}} d\mu = 1$ the following inequality holds*

$$\boxed{\left(\int_{\mathcal{E}} |\varphi|^s \, d\mu\right)^{1/s} \leq \left(\int_{\mathcal{E}} |\varphi|^t \, d\mu\right)^{1/t}} \tag{16.158}$$

where $0 < s \leq t$.

Proof. Define $r := \dfrac{t}{s}$. Taking in (16.148) $\varphi := |\varphi|^s$ and $g_{\cup}(x) := |x|^r$ implies

$$\left|\int_{\mathcal{E}} |\varphi|^s \, d\mu\right|^{t/s} = \left|\int_{\mathcal{E}} |\varphi|^s \, d\mu\right|^r \leq \int_{\mathcal{E}} (|\varphi|^s)^r \, d\mu = \int_{\mathcal{E}} |\varphi|^t \, d\mu$$

which completes the proof. □

Corollary 16.21. *For any measurable function $\varphi : \mathbb{R} \to \mathbb{R}$ such that $|\varphi|^k \in \mathcal{L}(\mu)$ $(k > 2$ is an integer) on $\mathcal{E} \subset \mathbb{R}$ the following inequalities hold*

$$\boxed{\int_{\mathcal{E}} |\varphi| \, d\mu \leq \left(\int_{\mathcal{E}} |\varphi|^2 \, d\mu\right)^{1/2} \leq \cdots \leq \left(\int_{\mathcal{E}} |\varphi|^k \, d\mu\right)^{1/k}} \tag{16.159}$$

Corollary 16.22.

*1. In the **Borel measure case** we have*

$$\boxed{\left(\frac{1}{b-a} \int_{x=a}^{b} |\varphi(x)|^s \, dx\right)^{1/s} \leq \left(\frac{1}{b-a} \int_{x=a}^{b} |\varphi(x)|^t \, dx\right)^{1/t}} \tag{16.160}$$

and

$$\frac{1}{b-a}\int_{x=a}^{b}|\varphi(x)|\,dx \le \left(\frac{1}{b-a}\int_{x=a}^{b}|\varphi(x)|^2\,dx\right)^{1/2}$$

$$\le \cdots \le \left(\frac{1}{b-a}\int_{x=a}^{b}|\varphi(x)|^k\,dx\right)^{1/k}$$

2. *In the case of the* ***"atomic" measure*** *for any nonnegative* $\mu_i \ge 0$ $(i = 1, \ldots, n)$ *such that* $\sum_{i=1}^{n}\mu_i = 1$ *we have*

$$\boxed{\left(\sum_{i=1}^{n}|\varphi_i|^s\,\mu_i\right)^{1/s} \le \left(\sum_{i=1}^{n}|\varphi_i|^t\,\mu_i\right)^{1/t}} \tag{16.161}$$

and

$$\sum_{i=1}^{n}|\varphi_i|\,\mu_i \le \left(\sum_{i=1}^{n}|\varphi_i|^2\,\mu_i\right)^{1/2} \le \cdots \le \left(\sum_{i=1}^{n}|\varphi_i|^k\,\mu_i\right)^{1/k}$$

which for the uniform measure when $\mu_i := 1/n$ *becomes*

$$\boxed{\left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i|^s\right)^{1/s} \le \left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i|^t\right)^{1/t}} \tag{16.162}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}|\varphi_i| \le \left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i|^2\right)^{1/2} \le \cdots \le \left(\frac{1}{n}\sum_{i=1}^{n}|\varphi_i|^k\right)^{1/k}$$

16.4.7 Kulbac inequality

Theorem 16.31. (The continuous version) *Suppose* $p : \mathbb{R} \to \mathbb{R}$ *and* $q : \mathbb{R} \to \mathbb{R}$ *are any positive function on* $\mathcal{E} \subset \mathbb{R}$ *such that the Lebesgue integral*

$$\boxed{I_{\mathcal{E}}(p,q) := \int_{\mathcal{E}} \ln\left(\frac{p(x)}{q(x)}\right)p(x)dx} \tag{16.163}$$

is finite, that is, $I_{\mathcal{E}}(p,q) < \infty$ *and the following normalizing condition holds*

$$\boxed{\int_{\mathcal{E}} q(x)\,dx = 1, \quad \int_{\mathcal{E}} p(x)\,dx = 1} \tag{16.164}$$

Then

$$\boxed{I_{\mathcal{E}}(p,q) \geq 0} \tag{16.165}$$

and $I_{\mathcal{E}}(p,q) = 0$ *if and only if* $p(x) = q(x)$ *almost everywhere on* $\mathcal{E}$.

Proof. Notice that $(-\ln(x))$ is a convex function on $(0, \infty)$, i.e., $-ln(x) = g_{\cup}(x)$. Hence, by the Jensen inequality (16.151) we have

$$I_{\mathcal{E}}(p,q) = \int_{\mathcal{E}} \ln\left(\frac{p(x)}{q(x)}\right) p(x)dx = \int_{\mathcal{E}} \ln\left(-\left(\frac{q(x)}{p(x)}\right)\right) p(x)dx$$

$$\geq -\ln \int_{\mathcal{E}} \left(\frac{q(x)}{p(x)}\right) p(x)dx = -\ln \int_{\mathcal{E}} q(x)dx = -\ln 1 = 0$$

which proves (16.165). Evidently, $I_{\mathcal{E}}(p,q) = 0$ if $p(x) = q(x)$ almost everywhere on $\mathcal{E}$. Suppose $I_{\mathcal{E}}(p,q) = 0$ and $p(x) \neq q(x)$ for some $x \in \mathcal{E}_0 \subset \mathcal{E}$ such that $\mu(\mathcal{E}_0) = \int_{\mathcal{E}_0} dx > 0$. Then the Jensen inequality (16.151) implies

$$0 = I_{\mathcal{E}}(p,q) = -\int_{\mathcal{E}_0} \ln\left(\frac{q(x)}{p(x)}\right) p(x)dx \geq -\ln\left(\int_{\mathcal{E}_0} \left(\frac{q(x)}{p(x)}\right) p(x)\right) dx$$

$$= -\ln\left(\int_{\mathcal{E}_0} q(x)dx\right) = -\ln\alpha > 0$$

where $\alpha := \int_{\mathcal{E}_0} q(x)dx < 1$ which can always be done selecting $\mathcal{E}_0$ small enough. The last inequality represents a contradiction. So, $\mu(\mathcal{E}_0) = 0$. Theorem is proven. □

Theorem 16.32. (The discrete version) *Suppose* $p = (p_1, \ldots, p_n)$ *and* $q = (q_1, \ldots, q_n)$ *are any vectors with positive components such that*

$$\boxed{\sum_{i=1}^{n} q_i = 1, \quad \sum_{i=1}^{n} p_i = 1} \tag{16.166}$$

Define

$$\boxed{I(p,q) := \sum_{i=1}^{n} p_i \ln\left(\frac{p_i}{q_i}\right)} \tag{16.167}$$

Then

$$\boxed{I(p,q) \geq 0} \tag{16.168}$$

and $I(p,q) = 0$ *if and only if* $p_i = q_i$ *for all* $i = 1, \ldots, n$.

Proof. It practically repeats the proof of the previous theorem where instead of (16.151) we have the inequality (16.152) where $g_{\cup}(\cdot) := -\ln(\cdot)$, $\varphi_i := \dfrac{q_i}{p_i}$ and $\mu_i = p_i$. Theorem is proven. □

16.4.8 Minkovski inequality

Theorem 16.33. (The Minkovski inequality) *Suppose* $\varphi : \mathbb{R} \to \mathbb{R}$ *and* $\eta : \mathbb{R} \to \mathbb{R}$ *are measurable functions such that* $|\varphi|^p \in \mathcal{L}(\mu)$ *and* $|\eta|^p \in \mathcal{L}(\mu)$ *on* $\mathcal{E} \subset \mathbb{R}$ *for some* $p \in [1, \infty)$. *Then the following inequality holds:*

$$\boxed{\left(\int_{\mathcal{E}} |\varphi + \eta|^p \, d\mu\right)^{1/p} \leq \left(\int_{\mathcal{E}} |\varphi|^p \, d\mu\right)^{1/p} + \left(\int_{\mathcal{E}} |\eta|^p \, d\mu\right)^{1/p}} \tag{16.169}$$

Proof. Consider the following inequality

$$|\varphi + \eta|^p = |\varphi + \eta| \, |\varphi + \eta|^{p-1} \leq |\varphi| \, |\varphi + \eta|^{p-1} + |\eta| \, |\varphi + \eta|^{p-1}$$

which after integration becomes

$$\int_{\mathcal{E}} |\varphi + \eta|^p \, d\mu \leq \int_{\mathcal{E}} |\varphi| \, |\varphi + \eta|^{p-1} \, d\mu + \int_{\mathcal{E}} |\eta| \, |\varphi + \eta|^{p-1} \, d\mu \tag{16.170}$$

Applying the Hölder inequality (16.134) to each term in the right-hand side of (16.170) we derive:

$$\int_{\mathcal{E}} |\varphi| \, |\varphi + \eta|^{p-1} \, d\mu \leq \left(\int_{\mathcal{E}} |\varphi|^p \, d\mu\right)^{1/p} \left(\int_{\mathcal{E}} |\varphi + \eta|^{(p-1)q} \, d\mu\right)^{1/q}$$

$$= \left(\int_{\mathcal{E}} |\varphi|^p \, d\mu\right)^{1/p} \left(\int_{\mathcal{E}} |\varphi + \eta|^p \, d\mu\right)^{1/q}$$

since $p = (p-1)q$, and

$$\int_{\mathcal{E}} |\eta| \, |\varphi + \eta|^{p-1} \, d\mu \leq \left(\int_{\mathcal{E}} |\eta|^p \, d\mu\right)^{1/p} \left(\int_{\mathcal{E}} |\varphi + \eta|^{(p-1)q} \, d\mu\right)^{1/q}$$

$$= \left(\int_{\mathcal{E}} |\eta|^p \, d\mu\right)^{1/p} \left(\int_{\mathcal{E}} |\varphi + \eta|^p \, d\mu\right)^{1/q}$$

Using these inequalities for the right-hand side estimation in (16.170) we get

$$\int_{\varepsilon} |\varphi+\eta|^p \, d\mu \le \left[\left(\int_{\varepsilon} |\varphi|^p d\mu \right)^{1/p} + \left(\int_{\varepsilon} |\eta|^p d\mu \right)^{1/p} \right] \left(\int_{\varepsilon} |\varphi+\eta|^p \, d\mu \right)^{1/q}$$

which implies

$$\left(\int_{\varepsilon} |\varphi+\eta|^p \, d\mu \right)^{1-1/q} = \left(\int_{\varepsilon} |\varphi+\eta|^p \, d\mu \right)^{1/p} \le \left[\left(\int_{\varepsilon} |\varphi|^p \, d\mu \right)^{1/p} + \left(\int_{\varepsilon} |\eta|^p \, d\mu \right)^{1/p} \right]$$

Theorem is proven. □

Corollary 16.23.

1. *In the* ***Borel measure case*** *the inequality (16.134) becomes*

$$\boxed{\left(\int_{x=a}^{b} |\varphi(x)+\eta(x)|^p \, dx \right)^{1/p} \le \left(\int_{x=a}^{b} |\varphi(x)|^p \, dx \right)^{1/p} + \left(\int_{x=a}^{b} |\eta(x)|^p \, dx \right)^{1/p}} \tag{16.171}$$

2. *In the case of the* ***"atomic" measure*** *for any nonnegative* $\mu_i \ge 0$ $(i = 1, \ldots, n)$ *we have*

$$\boxed{\left(\sum_{i=1}^{n} |\varphi_i+\eta_i|^p \, \mu_i \right)^{1/p} \le \left(\sum_{i=1}^{n} |\varphi_i|^p \, \mu_i \right)^{1/p} + \left(\sum_{i=1}^{n} |\eta_i|^p \, \mu_i \right)^{1/p}} \tag{16.172}$$

which for the uniform measure when $\mu_i := \mu_0/n$ $(\mu_0 > 0)$ *becomes*

$$\boxed{\begin{aligned}&\left(\frac{\mu_0}{n}\sum_{i=1}^{n}|\varphi_i+\eta_i|^p\right)^{1/p}\\&\leq\left(\frac{\mu_0}{n}\sum_{i=1}^{n}|\varphi_i|^p\right)^{1/p}+\left(\frac{\mu_0}{n}\sum_{i=1}^{n}|\eta_i|^p\right)^{1/p}\end{aligned}} \tag{16.173}$$

or, equivalently,

$$\boxed{\left(\sum_{i=1}^{n}|\varphi_i+\eta_i|^p\right)^{1/p}\leq\left(\sum_{i=1}^{n}|\varphi_i|^p\right)^{1/p}+\left(\sum_{i=1}^{n}|\eta_i|^p\right)^{1/p}} \tag{16.174}$$

16.5 Numerical sequences

16.5.1 Infinite series

16.5.1.1 Partial sums and sums

Let $\{a_n\}$ be a sequence of real numbers. Form a new sequence $\{s_n\}$ where each term is defined as follows:

$$\boxed{s_n := \sum_{t=1}^{n} a_t} \tag{16.175}$$

Definition 16.10. *The number* s_n *is called* ***the nth partial sum of the series****. The series is said to converge (or to diverge) accordingly as* $\{s_n\}$ *is convergent or divergent. If* $\{s_n\}$ *converges to* s*, that is, there exists the limit* $\lim_{n\to\infty} s_n = s$*, then* s *is called* ***the sum of series****.*

It is clear that every theorem about sequences $\{a_n\}$ can be stated in terms of series putting $a_1 := s_1$ and $a_n := s_n - s_{n-1}$ (for $n > 1$), and vice versa. But it is nevertheless useful to consider both concepts. So, the Cauchy criterion 14.8 can be restated as follows.

16.5.1.2 Criterion for series convergence

Criterion 16.1. (The Cauchy criterion for series) *The series* $s_n = \sum_{t=1}^{n} a_t$ *(16.175) converges if and only if for every* $\varepsilon > 0$ *there is an integer* $n_0(\varepsilon)$ *such that*

$$\boxed{\left|\sum_{t=n}^{m} a_t\right| \leq \varepsilon} \tag{16.176}$$

if $m \geq n \geq n_0(\varepsilon)$.

Corollary 16.24. (The necessary condition of convergent) *If the series* $s_n = \sum_{t=1}^{n} a_t$ *(16.175) converges then*

$$\boxed{a_n \underset{n\to\infty}{\to} 0} \tag{16.177}$$

Proof. Taking in (16.176) $m := n + 1$ we obtain (16.177). □

Theorem 16.34. (The criterion for monotonic sequences) *Suppose* $\{a_n\}$ *is* ***monotonic****. Then* $\{a_n\}$ *converges if and only if it is bounded.*

Proof.

(a) *Necessity.* Let $\{a_n\}$ converges. Then it is bounded by Theorem 14.5.

(b) *Sufficiency.* Let $\{a_n\}$ be bounded and suppose that $a_n \le a_{n+1}$ (the proof is analogous in the other case). Let $\mathcal{A}$ be the range of $\{a_n\}$. If $\{a_n\}$ is bounded, then there exist the least upper bound a^+ on $\mathcal{A}$ and for all $n = 1, 2, \ldots$ it follows $a_n \le a^+$. For every $\varepsilon > 0$ there exists an integer $n_0(\varepsilon)$ such that

$$a^+ - \varepsilon \le a_{n_0(\varepsilon)} \le a^+$$

for otherwise $\left(a^+ - \varepsilon\right)$ would be an upper bound. By monotonicity it follows that

$$a^+ - \varepsilon \le a_n \le a^+$$

for all $n \ge n_0(\varepsilon)$ which shows that $\{a_n\}$ converges to a^+. □

16.5.1.3 Sum of series and telescopic series

Lemma 16.5.

1. Let $\left\{s_n^a\right\}$ *and* $\left\{s_n^b\right\}$ *be convergent series, namely,*

$$s_n^a := \sum_{t=1}^{n} a_t \underset{n\to\infty}{\to} s^a, \quad s_n^b := \sum_{t=1}^{n} b_t \underset{n\to\infty}{\to} s^b$$

Then for every pair of constants α *and* β

$$\boxed{\sum_{t=1}^{\infty} (\alpha a_t + \beta b_t) = \alpha \sum_{t=1}^{\infty} a_t + \beta \sum_{t=1}^{\infty} b_t = \alpha s^a + \beta s^b} \tag{16.178}$$

2. If $\{a_n\}$ *and* $\{b_n\}$ *are two* ***telescopic series*** *such that*

$$a_n := b_{n+1} - b_n$$

then $\sum_{t=1}^{n} a_t$ *converges if and only if* $\lim_{n\to\infty} b_n$ *exists, in which case we have*

$$\boxed{\sum_{t=1}^{\infty} a_t = \lim_{n\to\infty} b_n - b_1} \tag{16.179}$$

Proof. The identity (16.178) results from

$$\sum_{t=1}^{n} (\alpha a_t + \beta b_t) = \alpha \sum_{t=1}^{n} a_t + \beta \sum_{t=1}^{n} b_t$$

and (16.179) follows from the identity

$$\sum_{t=1}^{n} a_t = \sum_{t=1}^{n} (b_{t+1} - b_t) = b_{n+1} - b_1$$

□

16.5.1.4 Series of nonnegative terms

Theorem 16.35. (The "partial sum" criterion) *A series of nonnegative terms converges if and only if its partial sums form a bounded sequence.*

Proof. It follows directly from Theorem 16.34. □

Theorem 16.36. ("2^k-criterion") *Suppose* $a_1 \geq a_2 \geq \cdots \geq 0$. *The series* $\sum_{n=1}^{\infty} a_n$ *converges if and only if the series* $\sum_{k=1}^{\infty} 2^k a_{2^k}$ *converges.*

Proof. According to Theorem 16.34 it is sufficient to consider boundedness of the following partial sums

$$s_n := a_1 + a_2 + \ldots + a_n$$
$$t_k := a_1 + 2a_2 + \ldots + 2^k a_{2^k}$$

For $n < 2^k$ we have

$$s_n \leq a_1 + (a_2 + a_3) + \ldots + (a_{2^k} + \ldots a_{2^{k+1}-1})$$
$$\leq a_1 + 2a_2 + \ldots + 2^k a_{2^k} = t_k$$

On the other hand, for $n \geq 2^k$

$$s_n \geq a_1 + (a_2 + a_3) + \ldots + (a_{2^k} + \ldots a_{2^{k+1}-1})$$
$$\geq \frac{1}{2} a_1 + a_2 + \ldots + 2^{k-1} a_{2^k} = \frac{1}{2} t_k$$

So that

$$\frac{1}{2}t_k \leq s_n \leq t_k$$

This means that the sequences $\{s_n\}$ and $\{t_n\}$ are either both bounded or unbounded. This completes the proof. □

Corollary 16.25.

$$\boxed{\sum_{n=1}^{\infty} \frac{1}{n\,(\log n)^p} \begin{cases} < \infty & if \quad p > 1 \\ = \infty & if \quad p \leq 1 \end{cases}} \tag{16.180}$$

Proof. Indeed, since the function $\log n$ is monotonically increasing it follows that the function $\dfrac{1}{n\,(\log n)^p}$ is monotonically decreasing. Applying Theorem 16.36 we obtain

$$\sum_{k=1}^{\infty} 2^k \frac{1}{2^k\,(\log 2^k)^p} = \sum_{k=1}^{\infty} \frac{1}{(k \log 2)^p} = \frac{1}{(\log 2)^p} \sum_{k=1}^{\infty} \frac{1}{k^p}$$

which, in view of Corollary 16.30 to Lemma 16.12 (see below), implies the desired result. □

Corollary 16.26. *Continuing the same procedure one can prove that*

$$\boxed{\sum_{n=1}^{\infty} \frac{1}{n \log n\,(\log\log n)^p} \begin{cases} < \infty & if \quad p > 1 \\ = \infty & if \quad p \leq 1 \end{cases}} \tag{16.181}$$

16.5.1.5 Alternating series

Definition 16.11. *If $a_n > 0$ for all n, then the series $\sum_{n=1}^{\infty} (-1)^{n+1} a_n$ is called an* ***alternating series****.*

Theorem 16.37. (on the convegence of alternating series) *If $\{a_n\}$ is a* ***non-increasing sequence*** *$(a_n > 0)$* ***converging to zero****, then the alternating series $\sum_{n=1}^{\infty} (-1)^{n+1} a_n$ converges, that is,*

$$\boxed{s_n := \sum_{k=1}^{n} (-1)^{k+1} a_k \underset{n\to\infty}{\to} s} \tag{16.182}$$

and for all $n = 1, 2, \ldots$

$$\boxed{0 < (-1)^n (s - s_n) < a_{n+1}} \tag{16.183}$$

Proof. Define $b_k := a_{2k-1} - a_{2k} \geq 0$. Then

$$s_n = (a_1 - a_2) + (a_3 - a_4) + \ldots + \left((-1)^n a_{n-1} + (-1)^{n+1} a_n\right)$$

$$= \begin{cases} \sum_{k=1}^{n} b_k & \text{if} \quad n = 2l \\ \sum_{k=1}^{n-2} b_k + a_n & \text{if} \quad n = 2l-1 \end{cases} \qquad (l = 1, 2, \ldots)$$

And, since $a_n \to 0$, the series s_n converges if and only if $\beta_n := \sum_{k=1}^{n} b_k$ converges, that is, $s_n \to s$ if and only if $\beta_n \to s$. But β_n is the series with nonnegative terms. So, it is monotonically nondecreasing and bounded because of the inequality

$$\beta_n = a_1 - (a_2 - a_3) - \ldots - (a_{2n-2} - a_{2n-1}) - a_{2n} < a_1$$

Hence, β_n converges. The inequality (16.183) is a consequence of the following relations:

$$(-1)^n (s - s_n) = \sum_{k=1}^{\infty} (-1)^{k+1} a_{n+k} = \sum_{k=1}^{\infty} (a_{n+2k-1} - a_{n+2k}) > 0$$

$$(-1)^n (s - s_n) = a_{n+1} - \sum_{k=1}^{\infty} (a_{n+2k} - a_{n+2k+1}) < a_{n+1}$$

□

16.5.1.6 Absolutely convergent series

Definition 16.12. *A series $\sum_{t=1}^{\infty} a_t$ is called **absolutely convergent** if*

$$\boxed{\sum_{t=1}^{\infty} |a_t| < \infty} \qquad (16.184)$$

Lemma 16.6. *Absolute convergence of $\sum_{t=1}^{\infty} a_t$ implies convergence.*

Proof. It is sufficient to apply the Cauchy criterion (16.1) to the inequality

$$\left| \sum_{s=n}^{n+p} a_t \right| \leq \sum_{s=n}^{n+p} |a_t|$$

□

16.5.1.7 The geometric series

Lemma 16.7. *If $|x| < 1$ then the partial geometric series*

$$\boxed{S_n := 1 + x + x^2 + x^3 + \ldots + x_n} \qquad (16.185)$$

converges:

$$\boxed{S_n \to (1-x)^{-1}} \tag{16.186}$$

If $|x| \geq 1$, *the series disconverges.*

Proof. The result (16.186) follows from the identity

$$(1-x)\, S_n = \sum_{k=0}^{n} \left(x^k - x^{k+1}\right) = 1 - x^{n+1}$$

If $|x| \geq 1$, the general term does not tend to zero and, hence, series cannot converge. □

16.5.1.8 Some tests of convergence

Theorem 16.38. (Integral test) *Let* f *be a positive non-increasing function defined on* $[1, \infty)$ *such that* $f(x) \underset{x\to\infty}{\to} 0$. *For integers* $n = 1, 2, \ldots$ *define*

$$\boxed{s_n := \sum_{k=1}^{n} f(k), \quad t_n := \int_{x=1}^{n} f(x)\, dx, \quad d_n := s_n - t_n} \tag{16.187}$$

Then

1.

$$\boxed{0 < f(n+1) \leq d_{n+1} \leq d_n \leq f(1)} \tag{16.188}$$

2. there exists the limit

$$\boxed{d := \lim_{n\to\infty} d_n} \tag{16.189}$$

3. the sequence $\{s_n\}$ *converges if and only if the sequence* $\{t_n\}$ *converges;*

4.

$$\boxed{0 \leq d_n - d \leq f(n)} \tag{16.190}$$

Proof.

1. By monotonicity, one has

$$t_{n+1} = \sum_{k=1}^{n} \int_{k}^{k+1} f(x)\,dx \le \sum_{k=1}^{n} \int_{k}^{k+1} f(k)\,dx = \sum_{k=1}^{n} f(k) = s_n$$

This implies

$$f(n+1) = s_{n+1} - s_n \le s_{n+1} - t_{n+1} = d_{n+1}$$

In addition, we have

$$d_n - d_{n+1} = (t_{n+1} - t_n) - (s_{n+1} - s_n)$$

$$= \int_{n}^{n+1} f(x)\,dx - f(n+1) \ge \int_{n}^{n+1} f(n+1)\,dx - f(n+1) = 0$$

which proves (1) since

$$d_{n+1} \le d_n \le \cdots \le d_1 = f(1)$$

2. But (1) implies (2);
3. And (2) implies (3) since by (1) $\{s_n\}$ dominates $\{t_n\}$ and

$$\lim_{n\to\infty} s_n = d + \lim_{n\to\infty} t_n$$

4. To prove (16.190) notice that

$$d_n - d_{n+1} = \int_{n}^{n+1} f(x)\,dx - f(n+1) \le \int_{n}^{n+1} dx - f(n+1)$$

$$= f(n) - f(n+1)$$

Then summing these inequalities leads to the following inequality

$$d_k - d_{n+1} = \sum_{r=k}^{n} (d_r - d_{r+1}) \le \sum_{r=k}^{n} [f(r) - f(r+1)]$$

$$= f(r) - f(n+1) \le f(k)$$

and, hence, when $n \to \infty$ we get

$$d_k - d \le f(k)$$

Theorem is proven. □

Theorem 16.39. (Dirichlet's test) *Let* $A_n := \sum_{k=1}^{n} a_k$ *be a partial sum of a bounded series, namely, for any* $n = 1, 2, \ldots$ *let*

$$|A_n| \leq M < \infty$$

and let $\{b_n\}$ *be a non-increasing sequence of positive numbers converging to zero, i.e.,* $b_n \underset{n\to\infty}{\downarrow} 0$. *Then the series* $\sum_{k=1}^{\infty} a_k b_k$ *converges, that is,*

$$\boxed{s_n := \sum_{k=1}^{n} a_k b_k \underset{n\to\infty}{\to} s} \tag{16.191}$$

Proof. Notice that by the Abel formula (12.4)

$$\sum_{k=n}^{n+p} a_k b_k = b_{n+p} \sum_{k=n}^{n+p} a_k + \sum_{k=n}^{n+p} (b_{k-1} - b_k) A_{k-1}$$

and, hence,

$$\left|\sum_{k=n}^{n+p} a_k b_k\right| \leq b_{n+p} \left|\sum_{k=n}^{n+p} a_k\right| + \sum_{k=n}^{n+p} (b_{k-1} - b_k) |A_{k-1}|$$

$$\leq b_{n+p} M + \sum_{k=n}^{n+p} (b_{k-1} - b_k) M = M b_{n-1}$$

Since $b_n \underset{n\to\infty}{\downarrow} 0$ for any $\varepsilon > 0$ there exists an integer $n_0(\varepsilon)$ such that for all $n \geq n_0(\varepsilon)$ we have $0 \leq b_n \leq \varepsilon$. Taking in the previous inequality $n := n_0(\varepsilon) + 1$ we obtain

$$\left|\sum_{k=n}^{n+p} a_k b_k\right| \leq M b_{n_0(\varepsilon)} \leq M\varepsilon$$

This means that the Cauchy criterion (16.1) holds which proves the theorem. □

Corollary 16.27. (Abel's test) *The series* $\sum_{k=1}^{\infty} a_k b_k$ *converges if* $\sum_{k=1}^{\infty} a_k$ *converges and if* $\{b_n\}$ *is a monotonically convergent sequence.*

Proof. Denote $b := \lim_{n\to\infty} b_n$ and $A := \lim_{n\to\infty} A_n$. Assume that $\{b_n\}$ is monotonically nonincreasing. Then we have

$$s_n := \sum_{k=1}^{n} a_k b_k = \sum_{k=1}^{n} a_k (b_k - b) + b \sum_{k=1}^{n} a_k$$

So, $\sum_{k=1}^{n} a_k b_k$ if $\sum_{k=1}^{n} a_k (b_k - b)$ converges. But the last satisfies the condition of Theorem 16.39 since $\sum_{k=1}^{n} a_k$ is bounded (it is convergent) and $(b_n - b) \underset{n\to\infty}{\downarrow} 0$. If $\{b_n\}$ is monotonically nondecreasing then

$$s_n := \sum_{k=1}^{n} a_k b_k = -\sum_{k=1}^{n} a_k (b - b_k) + b \sum_{k=1}^{n} a_k$$

and by the same arguments $\left\{-\sum_{k=1}^{n} a_k (b - b_k)\right\}$ converges. Corollary is proven. □

16.5.1.9 Multiplication of series

Definition 16.13. *Given two series $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$. Define*

$$\boxed{c_n := \sum_{k=0}^{n} a_k b_{n-k}, \quad n = 0, 1, 2, \ldots} \tag{16.192}$$

*The series $\sum_{k=0}^{\infty} c_k$ is referred to as the **Cauchy** (or **convoluting**) **product**.*

Lemma 16.8. (Mertens) *If the series $\sum_{k=0}^{\infty} a_k$ converges absolutely and the series $\sum_{k=0}^{\infty} b_k$ converges then $\sum_{k=0}^{\infty} c_k$ also converges and*

$$\boxed{\sum_{k=0}^{\infty} c_k = \left(\sum_{k=0}^{\infty} a_k\right)\left(\sum_{k=0}^{\infty} b_k\right)} \tag{16.193}$$

Proof. Define the partial sums

$$C_n := \sum_{k=0}^{n} c_k, \quad A_n := \sum_{k=0}^{n} a_k, \quad B_n := \sum_{k=0}^{n} b_k$$

and the series sums

$$A := \sum_{k=0}^{\infty} a_k, \quad B := \sum_{k=0}^{\infty} b_k$$

Then

$$C_n = \sum_{k=0}^{n} \sum_{s=0}^{k} a_s b_{k-s} = \sum_{k=0}^{n} \sum_{s=0}^{n} a_s b_{k-s} \chi_{s\le k} = \sum_{s=0}^{n} a_s \sum_{k=0}^{n} b_{k-s} \chi_{s\le k}$$

$$= \sum_{s=0}^{n} a_s \sum_{k=s}^{n} b_{k-s} = \sum_{s=0}^{n} a_s \sum_{\tau=0}^{n-s} b_\tau = \sum_{s=0}^{n} a_s B_{n-s}$$

$$= \sum_{s=0}^{n} a_s B + \sum_{s=0}^{n} a_s (B_{n-s} - B) = A_n B + \sum_{s=0}^{n} a_s (B_{n-s} - B)$$

To complete the proof it is sufficient to show that

$$e_n := \sum_{s=0}^{n} a_s (B_{n-s} - B) \underset{n\to\infty}{\to} 0$$

Define $S := \sum_{s=0}^{\infty} |a_s|$ and select $\varepsilon > 0$. Let for all $n \geq n_0(\varepsilon)$ we have

$$|B_n - B| \leq \frac{\varepsilon}{2S} \quad \text{and} \quad \sum_{s=n_0(\varepsilon)+1}^{\infty} |a_s| \leq \frac{\varepsilon}{2M}$$

where $|B_n - B| \leq M$ for all n. Then for $n > 2n_0(\varepsilon)$ we have

$$|e_n| \leq \sum_{s=0}^{n} |a_s| \, |B_{n-s} - B| = \sum_{s=0}^{n_0(\varepsilon)} |a_s| \, |B_{n-s} - B|$$

$$+ \sum_{s=n_0(\varepsilon)+1}^{n} |a_s| \, |B_{n-s} - B| \leq \frac{\varepsilon}{2S} \sum_{s=0}^{n_0(\varepsilon)} |a_s| + M \sum_{s=n_0(\varepsilon)+1}^{n} |a_s|$$

$$\leq \frac{\varepsilon}{2S} \sum_{s=0}^{\infty} |a_s| + M \frac{\varepsilon}{2M} = \varepsilon$$

Hence, $C_n \to AB$ as $n \to \infty$ which proves the lemma. □

Remark 16.7. (Abel) *The statement of this lemma remains valid if the series* $\sum_{k=0}^{\infty} a_k$ *converges (**not obligatory absolutely**).*

16.5.1.10 Cesàro summability

Definition 16.14. (Cesàro sum) *Let* $A_n := \sum_{k=0}^{n} a_k$ *be a partial sum of the series* $\sum_{k=0}^{\infty} a_k$ *and* $\{s_n\}$ *be the sequence of arithmetic means defined by*

$$s_n := \frac{A_1 + \ldots + A_n}{n} = \frac{1}{n} \sum_{k=0}^{n} A_k = \frac{1}{n} \sum_{k=0}^{n} \sum_{s=0}^{k} a_s \tag{16.194}$$

The series $\sum_{k=0}^{\infty} a_k$ *is said to be* ***Cesàro summable*** *if* $\{s_n\}$ *converges. If* $\lim_{n\to\infty} s_n = S$ *then* S *is called the* ***Cesàro sum*** *(or* $(C, 1)$*-sum) of* $\sum_{k=0}^{\infty} a_k$ *and we write*

$$\boxed{\sum_{k=0}^{\infty} a_k = S \quad (C, 1)} \tag{16.195}$$

Example 16.5. *Let* $a_n = x^n$, $0 < x < 1$. *Then*

$$A_n = \frac{1 - x^{n+1}}{1 - x}$$

$$s_n = \frac{1}{1-x} - \frac{x}{n(1-x)} \sum_{k=1}^{n} x^k = \frac{1}{1-x} - \frac{x\left(1 - x^{n+1}\right)}{n(1-x)^2}$$

Therefore,

$$s_n \to \frac{1}{1-x}$$

and, hence,

$$\sum_{k=0}^{\infty} x^k = \frac{1}{1-x} \quad (C, 1)$$

Claim 16.3. *If* $\sum_{k=0}^{\infty} a_k$ *is summable then it is also* $(C, 1)$*-summable.*

Proof. It follows directly from (16.210) (see below). □

Example 16.6. *Let* $a_n = (-1)^{n+1} n$. *Then*

$$A_n = 1 - 2 + 3 - 4 + \ldots (-1)^{n+1} n$$

$$= \begin{cases} -\dfrac{n}{2} & \text{if} \quad n = 2k \\ 1 + \dfrac{n}{2} & \text{if} \quad n = 2k+1 \end{cases} \qquad (k = 1, 2, \ldots)$$

and, therefore,

$$\limsup_{n\to\infty} s_n = \frac{1}{2}, \quad \liminf_{n\to\infty} s_n = 0$$

and, hence, $\sum_{n=0}^{\infty} (-1)^{n+1} n$ *is not* $(C, 1)$*-summable.*

16.5.2 Infinite products

Definition 16.15. *We say that an infinite product* $\prod_{n=1}^{\infty} u_n$ *converges if there exists a limit*

$$\boxed{\pi := \lim_{n\to\infty} \prod_{n=1}^{\infty} u_n} \tag{16.196}$$

finitely, that is, $|\pi| < \infty$.

16.5.2.1 Cauchy criterion for a product

Theorem 16.40. (The Cauchy criterion for a product) *The infinite product* $\prod_{n=1}^{\infty} u_n$ *converges if and only if for every* $\varepsilon > 0$ *there exists an integer* $n_0(\varepsilon)$ *such that* $n \geq n_0(\varepsilon)$ *implies for all* $k = 1, 2, \ldots$

$$\boxed{\left| \prod_{t=n+1}^{n+k} u_t - 1 \right| < \varepsilon} \tag{16.197}$$

Proof.

(a) *Necessity.* Denote $\pi_n := \prod_{t=1}^{n} u_t$ and suppose that there exists the limit $\lim_{n\to\infty} \pi_n = \pi$. Sure, we can assume that no $u_n = 0$ (since if they are, the results become trivial) and, hence, we can assume that $\pi \neq 0$. Therefore there exists $M > 0$ such that $|\pi_n| > M$. This means that $\{\pi_n\}$ satisfies the Cauchy criterion for convergence (see Theorem 14.8). Hence, given $\varepsilon > 0$, there is $n_0(\varepsilon)$ such that $n \geq n_0(\varepsilon)$ implies for all $k = 1, 2, \ldots$

$$. \; |\pi_{n+k} - \pi_n| < \varepsilon M$$

Dividing by $|\pi_n|$ we obtain (16.197).

(b) *Sufficiency.* Now assume that (16.197) holds. Denote $q_n := \prod_{t=n_0(\varepsilon)+1}^{n} u_t$ and take $\varepsilon = 1/2$ in (16.197). Then evidently $\frac{1}{2} < |q_n| < \frac{3}{2}$. So, if $\{q_n\}$ converges, it cannot converge to 0. Let ε now be arbitrary. Then we can rewrite (16.197) as follows: $\left| \frac{q_{n+k}}{q_n} - 1 \right| < \varepsilon$, which gives

$$|q_{n+k} - q_n| < \varepsilon |q_n| < \varepsilon \frac{3}{2}$$

Therefore, $\{q_n\}$ satisfies the Cauchy criterion (see Theorem 14.8) and, hence, is convergent. This means that π_n converges too. Theorem is proven. □

16.5.2.2 Relation between a product and a sum

Theorem 16.41. *Assume* $u_n \geq 0$ $(n = 1, 2, \ldots)$. *Then the product* $\prod_{n=1}^{\infty}(1 + u_n)$ *converges, that is,*

$$\boxed{\prod_{n=1}^{\infty}(1 + u_n) < \infty} \tag{16.198}$$

if and only if the series $\sum_{n=1}^{\infty} u_n$ *converges, that is,*

$$\boxed{\sum_{n=1}^{\infty} u_n < \infty} \tag{16.199}$$

Proof. Denote $\pi_n := \prod_{t=1}^{n}(1 + u_t)$ and $s_n := \sum_{t=1}^{n} u_t$. Based on the inequality

$$1 + x \leq e^x \tag{16.200}$$

valid for any x, we have $\pi_n \leq \exp(s_n)$. So, if $\{s_n\}$ converges, then $\{\pi_n\}$ converges too. But, on the other hand, the following inequality seems to be obvious: $\pi_n \geq s_n$ implies the convergence of $\{s_n\}$ if $\{\pi_n\}$ is convergent. Theorem is proven. □

Theorem 16.42. *Assume* $u_n > 0$ $(n = 1, 2, \ldots)$. *Then*

$$\boxed{\prod_{n=1}^{\infty}(1 - u_n) = 0} \tag{16.201}$$

if and only if the series

$$\boxed{\sum_{n=1}^{\infty} u_n = \infty} \tag{16.202}$$

Proof. Again, by the inequality (16.200), the convergence of $\pi_n := \prod_{t=1}^{n}(1 - u_t)$ to zero follows from the fact that $s_n \to \infty$. On the other hand, by the inequality

$$\ln(1 - u_t) \geq -u_t$$

valid for $u_t > 0$, we have

$$\pi_n = \exp(\ln \pi_n) = \exp\left(\sum_{t=1}^{n} \ln(1 - u_t)\right)$$

$$\geq \exp\left(-\sum_{t=1}^{n} u_t\right) = \exp(-s_n)$$

So, if $\pi_n \to 0$, then $s_n \to \infty$ which proves the theorem. □

16.5.2.3 Some low estimates for a product

Lemma 16.9. (Nazin & Poznyak 1986) *Let $a \in [0, 1)$, $\beta \in (0, 1)$. Then the following low estimates hold:*

(a)

$$\boxed{\prod_{k=1}^{\infty}\left(1+a\beta^{k}\right) \geq \exp\left(\frac{a\beta}{1-\beta}\left[1-\frac{a\beta}{2(1+\beta)}\right]\right)} \tag{16.203}$$

(b)

$$\boxed{\prod_{k=1}^{\infty}\left(1-a\beta^{k}\right) \geq (1-a)^{1-1/\ln\beta}} \tag{16.204}$$

(c)

$$\boxed{\prod_{k=1}^{\infty}\left(1-a\beta^{k-1}\right)^{k} \geq (1-a)^{1-1/\ln\beta+1/\ln^{2}\beta}} \tag{16.205}$$

Proof.

(a) By the inequality

$$1+x \geq \exp\left(x-x^{2}/2\right)$$

valid for any $x \geq 0$, we have

$$\prod_{k=1}^{\infty}\left(1+a\beta^{k}\right) \geq \exp\left(\sum_{k=1}^{\infty}\left[a\beta^{k}-a^{2}\beta^{2k}/2\right]\right)$$

which implies (16.203).

(b) Applying the inequality

$$1-x \geq \exp\left(\frac{x}{x-1}\right) \tag{16.206}$$

valid for any $x \in [0, 1)$, we have

$$\prod_{k=1}^{\infty}\left(1-a\beta^{k}\right) \geq \exp\left(\sum_{k=1}^{\infty}\frac{a\beta^{k}}{a\beta^{k}-1}\right)$$

$$\geq \exp\left(\int_{x=0}^{\infty}\frac{a\beta^{x}}{a\beta^{x}-1}dx\right) = (1-a)^{-1/\ln\beta}$$

which gives (16.204).

(c) In (16.206) take $x := a\beta^{k-1}$ which implies

$$\prod_{k=1}^{\infty}\left(1-a\beta^{k-1}\right)^k = (1-a)\exp\left(\sum_{k=2}^{\infty} k\ln\left(1-a\beta^{k-1}\right)\right)$$

$$\geq (1-a)\exp\left(-\sum_{k=2}^{\infty} k\frac{a\beta^{k-1}}{1-a\beta^{k-1}}\right)$$

$$\geq (1-a)\exp\left(-\int_{x=1}^{\infty} k\frac{a\beta^{x-1}}{1-a\beta^{x-1}}dx\right)$$

$$= (1-a)\exp\left(-\int_{x=0}^{\infty}\frac{a\beta^{x}}{1-a\beta^{x}}dx - \int_{x=0}^{\infty} x\frac{a\beta^{x}}{1-a\beta^{x}}dx\right)$$

Since

$$\int_{x=0}^{\infty} x\frac{a\beta^{x}}{1-a\beta^{x}}dx = \frac{1}{\ln\beta}\int_{x=0}^{\infty}\ln\left(1-a\beta^{x}\right)dx \leq \frac{1}{\ln\beta}\int_{x=0}^{\infty}\frac{a\beta^{x}}{1-a\beta^{x}}dx$$

we derive

$$\prod_{k=1}^{\infty}\left(1-a\beta^{k-1}\right)^k \geq (1-a)\exp\left(-\left(1-\frac{1}{\ln\beta}\right)\int_{x=0}^{\infty}\frac{a\beta^{x}}{1-a\beta^{x}}dx\right)$$

$$= (1-a)\exp\left(-\left(1-\frac{1}{\ln\beta}\right)\frac{\ln(1-a)}{\ln\beta}\right) = (1-a)^{1-1/\ln\beta+1/\ln^2\beta}$$

which proves (16.205). Lemma is proven. □

16.5.3 Teöplitz lemma

Lemma 16.10. (Teöplitz) *Let* $\{a_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of nonnegative real numbers such that*

$$\boxed{b_n := \sum_{t=1}^{n} a_t \to \infty \quad when \quad n \to \infty} \tag{16.207}$$

and $\{x_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of real numbers which converges to* x^*, *that is,*

$$\boxed{x_n \underset{n\to\infty}{\to} x^*} \tag{16.208}$$

Then

(a) there exists an integer n_0 *such that* $b_n > 0$ *for all* $n \geq n_0$;
(b)

$$\boxed{\frac{1}{b_n}\sum_{t=1}^{n} a_t x_t \to x^* \quad \text{when} \quad n_0 \leq n \to \infty} \tag{16.209}$$

Proof. The claim (a) results from (16.207). To prove (b) let us select $\varepsilon > 0$ and $n_0'(\varepsilon) \geq n_0$ such that for all $n \geq n_0'(\varepsilon)$ we have (in view of (16.208)) $|x_n - x^*| \leq \varepsilon$. Then it follows that

$$\begin{aligned}
\left|\frac{1}{b_n}\sum_{t=1}^{n} a_t x_t - x^*\right| &= \left|\frac{1}{b_n}\sum_{t=1}^{n} a_t (x_t - x^*)\right| \leq \frac{1}{b_n}\sum_{t=1}^{n} a_t |x_t - x^*| \\
&= \frac{1}{b_n}\sum_{t=1}^{n_0'(\varepsilon)-1} a_t |x_t - x^*| + \frac{1}{b_n}\sum_{n_0'(\varepsilon)}^{n} a_t |x_t - x^*| \\
&\leq \frac{1}{b_n}\sum_{t=1}^{n_0'(\varepsilon)-1} a_t |x_t - x^*| + \frac{\varepsilon}{b_n}\sum_{n_0'(\varepsilon)}^{n} a_t \\
&\leq \frac{\text{const}}{b_n} + \varepsilon \to \varepsilon \quad \text{when} \quad b_n \to \infty
\end{aligned}$$

Since this is true for any $\varepsilon > 0$ we obtain the proof of the lemma. □

Corollary 16.28. *If* $x_n \underset{n\to\infty}{\to} x^*$ *then*

$$\boxed{\frac{1}{n}\sum_{t=1}^{n} x_t \underset{n\to\infty}{\to} x^*} \tag{16.210}$$

Proof. To prove (16.210) it is sufficient in (16.209) to take $a_n = 1$ for all $n = 1, 2, \ldots$. □

Corollary 16.29. *Let* $\{a_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of nonnegative real numbers such that*

$$\boxed{\sum_{t=1}^{n} a_t \to \infty \quad \text{when} \quad n \to \infty} \tag{16.211}$$

and for some numerical nonzero sequence $\{b_n\}$ *of real numbers there exists the limit*

$$\boxed{\lim_{n\to\infty} b_n^{-1} \sum_{t=1}^{n} a_t = \alpha} \tag{16.212}$$

Let also $\{x_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of real numbers which converges to* x^*, *that is,*

$$\boxed{x_n \underset{n\to\infty}{\to} x^*} \tag{16.213}$$

Then

$$\boxed{\lim_{n\to\infty} b_n^{-1} \sum_{t=1}^{n} a_t x_n = \alpha x^*} \tag{16.214}$$

Proof. Directly applying the Teöplitz lemma (16.10) we derive

$$b_n^{-1} \sum_{t=1}^{n} a_t x_n = \left[b_n^{-1} \sum_{t=1}^{n} a_t \right] \left[\left(\sum_{t=1}^{n} a_t \right)^{-1} \sum_{t=1}^{n} a_t x_n \right] \to \alpha x^*$$

□

16.5.4 Kronecker lemma

Lemma 16.11. (Kronecker) *Let* $\{a_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of nonnegative nondecreasing real numbers such that*

$$\boxed{0 \le b_n \le b_n \to \infty \quad \textit{when} \quad n \to \infty} \tag{16.215}$$

and $\{x_n\}$ $(n = 1, 2, \ldots)$ *be a sequence of real numbers such that the series* $\sum_{t=1}^{n} x_t$ *converges, that is,*

$$\boxed{s_n := \sum_{t=n_0}^{n} x_t \underset{n\to\infty}{\to} s^*, \quad |s^*| < \infty} \tag{16.216}$$

Then
(a) there exists an integer n_0 *such that* $b_n > 0$ *for all* $n \ge n_0$;
(b)

$$\boxed{\frac{1}{b_n} \sum_{t=1}^{n} b_t x_t \to 0 \quad \textit{when} \quad n_0 \le n \to \infty} \tag{16.217}$$

Proof. Applying the Abel identity (12.4) for the scalar case, namely, using the identity

$$\sum_{t=n_0}^{n} \alpha_t \beta_t = \alpha_n \sum_{t=n_0}^{n} \beta_t - \sum_{t=n_0}^{n} (\alpha_t - \alpha_{t-1}) \sum_{s=n_0}^{t-1} \beta_s$$

we derive

$$\frac{1}{b_n}\sum_{t=n_0}^{n} b_t x_t = \frac{1}{b_n}\left[b_n \sum_{t=n_0}^{n} x_t - \sum_{t=n_0}^{n}(b_t - b_{t-1})\sum_{s=n_0}^{t-1} x_s\right]$$

$$= s_n - \frac{1}{b_n}\sum_{t=n_0}^{n}(b_t - b_{t-1})\, s_{t-1}$$

Denote $a_t := b_t - b_{t-1}$. Then

$$b_n = \sum_{t=n_0}^{n} a_t + b_{n_0} = \sum_{t=n_0}^{n} a_t \left[1 + b_{n_0} / \sum_{t=n_0}^{n} a_t\right]$$

and hence, by the Teöplitz Lemma 16.10, we have

$$\frac{1}{b_n}\sum_{t=n_0}^{n} b_t x_t = s_n - \left[1 + b_{n_0} / \sum_{t=n_0}^{n} a_t\right]^{-1} \left(\sum_{t=n_0}^{n} a_t\right)^{-1} \sum_{t=n_0}^{n} a_t s_{t-1}$$
$$\to s^* - s^* = 0$$

which proves (16.217). □

16.5.5 Abel–Dini lemma

Lemma 16.12. (Abel–Dini) *For any nonnegative number sequence* $\{u_n\}_{n=1,2,\ldots}$ *such that*

$$\boxed{S_n := \sum_{t=1}^{n} u_t \to \infty \quad \textit{when} \quad n \to \infty} \tag{16.218}$$

the following properties hold:

(a) There exists an integer n_0 *such that* $S_n > 0$ *for all* $n \geq n_0$;

(b) The series $\sum_{t=n_0}^{n} \frac{u_t}{S_t^{1+\rho}}$ *converges if* $\rho > 0$ *and it disconverges if* $\rho = 0$, *that is,*

$$\boxed{\sum_{t=n_0}^{\infty} \frac{u_t}{S_t^{1+\rho}} \begin{cases} < \infty & \textit{if} \quad \rho > 0 \\ = \infty & \textit{if} \quad \rho = 0 \end{cases}} \tag{16.219}$$

Proof. (a) follows from (16.218). Evidently $u_t = S_t - S_{t-1}$, so

$$V_\rho(n) := \sum_{t=n_0}^{n} \frac{u_t}{S_t^{1+\rho}} = \sum_{t=n_0}^{n} \frac{S_t - S_{t-1}}{S_t^{1+\rho}}$$

To prove (b) for all positive S define the function $R_\rho(S) := S^{-(1+\rho)}$ (see Fig. 16.2). The dashed area corresponds exactly to the function $V_\rho(n)$.

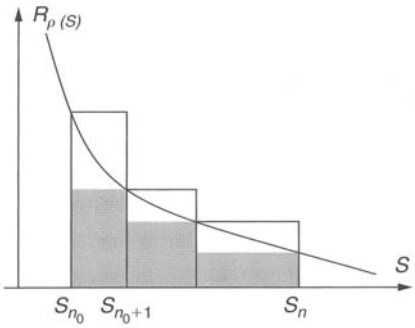

Fig. 16.2. The function $R_\rho(S)$.

For $\rho > 0$ we have

$$V_\rho(n) \le \int_{S_{n_0}}^{S_n} R_\rho(S)\,dS \le \int_{S_{n_0}}^{\infty} S^{-(1+\rho)}dS$$

$$= -\rho^{-1}S^{-\rho}\,|_{S=S_{n_0}}^{\infty} = \rho^{-1}S_{n_0}^{-\rho} < \infty$$

Taking $n \to \infty$ in the left-hand side of this inequality we obtain the result of this lemma for the case $\rho > 0$. Consider now the case $\rho = 0$ and suppose that

$$\sum_{t=n_0}^{\infty} \frac{u_t}{S_t} < \infty \tag{16.220}$$

Then $\frac{u_t}{S_t} \to 0$, or, equivalently, $\frac{S_t - S_{t-1}}{S_t} = 1 - \frac{S_{t-1}}{S_t} \to 0$. This means that for any $\varepsilon > 0$ there exists an integer $k(\varepsilon)$ such that for all $t \ge k(\varepsilon)$ we have $\frac{S_{t-1}}{S_t} > 1-\varepsilon$. In view of this we have

$$V_\rho(n) = \sum_{t=n_0}^{n} \left(\frac{S_t - S_{t-1}}{S_{t-1}}\right)\left(\frac{S_{t-1}}{S_t}\right) \ge \sum_{t=k(\varepsilon)}^{n} \left(\frac{S_t - S_{t-1}}{S_{t-1}}\right)\left(\frac{S_{t-1}}{S_t}\right)$$

$$\ge (1-\varepsilon)\sum_{t=k(\varepsilon)}^{n}\left(\frac{S_t - S_{t-1}}{S_{t-1}}\right) \ge (1-\varepsilon)\int_{S_{k(\varepsilon)}}^{S_n} S^{-1}dS$$

$$= (1-\varepsilon)\ln\frac{S_n}{S_{k(\varepsilon)}} \to \infty \quad \text{since} \quad S_n \to \infty$$

which contradicts (16.220). So, $\sum_{t=n_0}^{\infty}\frac{u_t}{S_t} = \infty$. Lemma is proven. □

Corollary 16.30.

$$\boxed{\sum_{t=n_0}^{\infty} \frac{1}{n^{1+\rho}} \begin{cases} < \infty & \text{if} \quad \rho > 0 \\ = \infty & \text{if} \quad \rho = 0 \end{cases}} \tag{16.221}$$

Proof. It follows from Lemma 16.12 if we take $u_t \equiv 1$. □

16.6 Recurrent inequalities

16.6.1 On the sum of a series estimation

Lemma 16.13. (Nazin & Poznyak 1986)[1] *Let the numerical sequences* $\{\gamma_n\}$ *and* $\{g_n\}$ *satisfy the conditions*

$$\boxed{0 < \gamma_{n+1} \leq \gamma_n, \quad G := \sup_n \left| \sum_{t=1}^{n} g_t \right| < \infty} \tag{16.222}$$

Then the following upper estimate holds

$$\boxed{\left| \sum_{t=1}^{n} g_t \gamma_t^{-1} \right| \leq 2G\gamma_n^{-1}} \tag{16.223}$$

and, hence, for $n \to \infty$

$$\boxed{\frac{1}{h_n} \sum_{t=1}^{n} g_t \gamma_t^{-1} \to 0} \tag{16.224}$$

where $\{h_n\}$ *is any sequence of positive real numbers such that*

$$\boxed{h_n \gamma_n \to 0, \quad h_n > 0} \tag{16.225}$$

Proof. Using the Abel identity (12.4) of the summation by part in the scalar form, namely,

$$\sum_{t=1}^{n} \alpha_t \beta_t = \alpha_n \sum_{t=1}^{n} \beta_t - \sum_{t=1}^{n} (\alpha_t - \alpha_{t-1}) \sum_{s=1}^{t-1} \beta_s$$

[1] This lemma as well as Lemma 16.16 given below were first proven by A.V. Nazin (see the citations in Nazin & Poznyak 1986).

for $\alpha_n := \gamma_n^{-1}$ and $\beta_n := g_n$, we have

$$\left|\sum_{t=1}^{n} g_t \gamma_t^{-1}\right| = \left|\gamma_n^{-1}\sum_{t=1}^{n} g_t - \sum_{t=1}^{n}\left(\gamma_t^{-1} - \gamma_{t-1}^{-1}\right)\sum_{s=1}^{t-1} g_s\right| \le \gamma_n^{-1}\sup_n\left|\sum_{t=1}^{n} g_t\right|$$

$$+\sum_{t=1}^{n}\left(\gamma_t^{-1} - \gamma_{t-1}^{-1}\right)\sup_t\left|\sum_{s=1}^{t-1} g_s\right| \le 2\gamma_n^{-1}\sup_n\left|\sum_{t=1}^{n} g_t\right|$$

which implies (16.223). Lemma is proven. □

16.6.2 Linear recurrent inequalities

Lemma 16.14. (on linear recurrent inequalities) *Let us consider a real numerical non-negative sequence* $\{u_n\}$ *which satisfies the following recurrent inequality*

$$\boxed{0 \le u_{n+1} \le u_n(1-\alpha_n) + \beta_n} \tag{16.226}$$

where $\{\alpha_n\}$ *and* $\{\beta_n\}$ *are numerical sequences such that*

$$\boxed{\begin{array}{l} 0 < \alpha_n \le 1, \quad \sum_{n=1}^{\infty}\alpha_n = \infty \\ 0 \le \beta_n, \quad \limsup_{n\to\infty}(\beta_n/\alpha_n) = p < \infty \end{array}} \tag{16.227}$$

Then

$$\boxed{\limsup_{n\to\infty} u_n \le p} \tag{16.228}$$

Proof. By the definition of lim sup it follows that for any $\varepsilon > 0$ there exists an integer $n_0(\varepsilon)$ such that for any $n \ge n_0(\varepsilon)$ we have $\left|\sup_{t\ge n}(\beta_t/\alpha_t) - p\right| \le \varepsilon$, which implies the inequality

$$\sup_{t\ge n}(\beta_t/\alpha_t) \le p + \varepsilon \tag{16.229}$$

Taking $n > n_0(\varepsilon)$ and making the recursion back up to $n_0(\varepsilon)$ we obtain

$$u_{n+1} \le u_n(1-\alpha_n) + \beta_n \le u_{n-1}(1-\alpha_n)(1-\alpha_{n-1})$$

$$+[\beta_n + \beta_{n-1}(1-\alpha_n)] \le \cdots \le u_{n_0(\varepsilon)}\prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t)$$

$$+\sum_{t=n_0(\varepsilon)}^{n}\beta_t\prod_{t=t+1}^{n}(1-\alpha_t)$$

(here we accept that $\prod_{t=m}^{n}(1-\alpha_t) := 1$ if $m > n$). Applying the inequality (16.229) to the right-hand side we derive

$$u_{n+1} \le u_{n_0(\varepsilon)} \prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t) + \sum_{t=n_0(\varepsilon)}^{n} \beta_t \prod_{t=t+1}^{n}(1-\alpha_t)$$

$$= u_{n_0(\varepsilon)} \prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t) + \sum_{t=n_0(\varepsilon)}^{n} \left(\frac{\beta_t}{\alpha_t}\right)\alpha_t \prod_{t=t+1}^{n}(1-\alpha_t)$$

$$\le u_{n_0(\varepsilon)} \prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t) + (p+\varepsilon)\sum_{t=n_0(\varepsilon)}^{n} \alpha_t \prod_{t=t+1}^{n}(1-\alpha_t)$$

Using the Abel identity (12.5) (the scalar version)

$$\prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t) + \sum_{t=n_0(\varepsilon)}^{n} \alpha_t \prod_{t=t+1}^{n}(1-\alpha_t) = 1$$

and the inequality $1 - x \le e^{-x}$, valid for any x, we get

$$u_{n+1} \le u_{n_0(\varepsilon)} \prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t) + (p+\varepsilon)\left[1 - \prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t)\right]$$

$$= (p+\varepsilon) + \left(u_{n_0(\varepsilon)} - p - \varepsilon\right)\prod_{t=n_0(\varepsilon)}^{n}(1-\alpha_t)$$

$$\le (p+\varepsilon) + \left(u_{n_0(\varepsilon)} - p - \varepsilon\right)\prod_{t=n_0(\varepsilon)}^{n}\exp(-\alpha_t)$$

$$= (p+\varepsilon) + \left(u_{n_0(\varepsilon)} - p - \varepsilon\right)\exp\left(-\sum_{t=n_0(\varepsilon)}^{n}\alpha_t\right) \to (p+\varepsilon)$$

since $\exp\left(-\sum_{t=n_0(\varepsilon)}^{n}\alpha_t\right) \to 0$ by the property (16.227). Since $\varepsilon > 0$ may be selected arbitrarily small the statement (16.228) follows. Lemma is proven. □

Lemma 16.15. (Nazin & Poznyak 1986) *Suppose that sequences* $\{u_n\}$ *and* $\{w_n\}$ *for all* $n \ge n_0$ *satisfy the following recurrent inequalities*

$$\boxed{\begin{array}{l} u_{n+1} \le u_n\left(1 - cn^{-1}\right) + dn^{-(p+1)} \\ w_{n+1} \ge w_n\left(1 - cn^{-1}\right) + dn^{-(p+1)} \end{array}} \tag{16.230}$$

where $c > p > 0$. *Then*

$$\boxed{\limsup_{n\to\infty}(n^p u_n) \le \frac{d}{c-p} \le \liminf_{n\to\infty}(n^p w_n)} \tag{16.231}$$

Moreover, there exist the sequences $\{u_n\}$ *and* $\{w_n\}$ *for which the identities in (16.231) are attained.*

Proof. Without loss of generality we may accept that $n_0 \geq c$. Let us introduce the sequence $\{y_n\}$ generated by the recurrent relation

$$y_{n+1} = y_n \left(1 - cn^{-1}\right) + dn^{-(p+1)} \tag{16.232}$$

Show that

$$\lim_{n\to\infty} (n^p y_n) = \frac{d}{c-p} \tag{16.233}$$

For all $n \geq n_0$ from (16.232) we have

$$y_{n+1} = y_{n_0} \prod_{k=n_0}^{n} \left(1 - \frac{c}{k}\right) + \sum_{k=n_0}^{n} \frac{d}{k^{p+1}} \prod_{m=k+1}^{n} \left(1 - \frac{c}{m}\right) \tag{16.234}$$

Since

$$n^{-p} - (n+1)^{-p} = n^{-p} - n^{-p}\left(1 + n^{-1}\right)^{-p} = n^{-p} - n^{-p}\left[1 - pn^{-1} + O\left(n^{-2}\right)\right]$$

$$= pn^{-(p+1)} + O\left(n^{-(p+2)}\right)$$

we have

$$\frac{d}{n^{p+1}} = \frac{d}{c-p}\left[(c-p)\frac{1}{p}\left(\frac{1}{n^p} - \frac{1}{(n+1)^p}\right) + O\left(n^{-(p+2)}\right)\right]$$

$$= \frac{d}{c-p}\left[\frac{1}{(n+1)^p} - \frac{1}{n^p} + \frac{c}{pn^p}\left(1 - \frac{n^p}{(n+1)^p}\right) + O\left(n^{-(p+2)}\right)\right]$$

$$= \frac{d}{c-p}\left[\frac{1}{(n+1)^p} - \frac{1}{n^p} + \frac{c}{pn^p}\left(1 - p\frac{1}{n} + O\left(n^{-2}\right)\right) + O\left(n^{-(p+2)}\right)\right]$$

$$= \frac{d}{c-p}\left[\frac{1}{(n+1)^p} - \left(1 - \frac{c}{n}\right)\frac{1}{n^p} + O\left(n^{-(p+2)}\right)\right]$$

Substitution of this identity into (16.234) implies

$$y_{n+1} = y_{n_0} \prod_{k=n_0}^{n} \left(1 - \frac{c}{k}\right) + \frac{d}{c-p}\left[\frac{1}{(n+1)^p} - \frac{1}{n_0^p} \prod_{m=n_0}^{n} \left(1 - \frac{c}{m}\right)\right.$$

$$\left. + \sum_{k=n_0}^{n} O\left(k^{-(p+2)}\right) \prod_{m=k+1}^{n} \left(1 - \frac{c}{m}\right)\right] \tag{16.235}$$

Taking into account that

$$\prod_{k=n_0}^{n} \left(1 - \frac{c}{k}\right) \leq \exp\left(-\sum_{k=n_0}^{n} \frac{c}{k}\right) \leq \text{const} \exp\left(-c \ln n\right) = O\left(n^{-c}\right)$$

and

$$\sum_{k=n_0}^{n} O\left(k^{-(p+2)}\right) \prod_{m=k+1}^{n} \left(1-\frac{c}{m}\right)$$

$$= \prod_{m=n_0}^{n} \left(1-\frac{c}{m}\right) \sum_{k=n_0}^{n} O\left(k^{-2}\right) k^{-p} \prod_{m=n_0}^{k} \left(1-\frac{c}{m}\right)^{-1}$$

$$O\left(n^{-c}\right) \sum_{k=n_0}^{n} O\left(k^{-2}\right) k^{-p} = \frac{\text{const}}{n^{p+\varepsilon}} \sum_{k=n_0}^{n} O\left(k^{-(2-\varepsilon)}\right) \frac{k^{c-p-\varepsilon}}{n^{c-p-\varepsilon}} = o\left(n^{-p}\right)$$

(here $\varepsilon \in (0, c-p)$) from (16.235) it follows that

$$y_{n+1} = \frac{d}{c-p}(n+1)^{-p} + o\left(n^{-p}\right) \tag{16.236}$$

We also have $y_n \geq u_n$ if $y_{n_0} = u_{n_0}$ and $y_n \leq w_n$ if $y_{n_0} = w_{n_0}$ which together with (16.236) leads to (16.231). □

Corollary 16.31. (Chung 1954) *Let the sequence $\{u_n\}$ of nonnegative real numbers satisfy the following recurrent equation:*

$$\boxed{u_{n+1} = u_n\left(1 - c_n n^{-1}\right) + d_n n^{-(p+1)}, \quad n \geq n_0} \tag{16.237}$$

where $\{c_n\}$ and $\{d_n\}$ are the sequences or real numbers such that

$$\boxed{\lim_{n\to\infty} c_n = c > p > 0, \quad \lim_{n\to\infty} d_n = d > 0} \tag{16.238}$$

then

$$\boxed{\lim_{n\to\infty} n^p u_n = \frac{d}{c-p}} \tag{16.239}$$

Proof. By lim definition for any $\varepsilon > 0$ there exists a number n_0 such that for all $n \geq n_0$

$$|c_n - c| \leq \varepsilon, \quad |d_n - d| \leq \varepsilon$$

which implies

$$c - \varepsilon \leq c_n \leq c + \varepsilon, \quad d - \varepsilon \leq d_n \leq d + \varepsilon$$

Using these inequalities in (16.237) gives

$$u_{n+1} \le u_n \left(1 - \frac{c-\varepsilon}{n}\right) + \frac{d+\varepsilon}{n^{p+1}}$$

$$u_{n+1} \ge u_n \left(1 - \frac{c+\varepsilon}{n}\right) + \frac{d-\varepsilon}{n^{p+1}}$$

Applying Lemma 16.15 we obtain

$$\frac{d-\varepsilon}{c+\varepsilon-p} \le \liminf_{n\to\infty} n^p u_n \le \limsup_{n\to\infty} n^p u_n \le \frac{d+\varepsilon}{c-\varepsilon-p}$$

Taking $\varepsilon \to +0$ proves (16.239). □

16.6.3 Recurrent inequalities with root terms

The lemmas below seem to be extremely important for the Lyapunov-like stability analysis for discrete time nonlinear systems.

Lemma 16.16. (Nazin & Poznyak 1986) *Let the sequence $\{u_n\}$ of nonnegative real numbers satisfy the following recurrent equation:*

$$\boxed{u_{n+1} \le u_n (1-\alpha_n) + \beta_n + \delta_n u_n^r, \quad n \ge n_0} \tag{16.240}$$

where $r \in (0, 1)$ and $\{\alpha_n\}, \{\beta_n\}, \{\delta_n\}$ are sequences of real numbers such that

$$\boxed{\liminf_{n\to\infty} \left(n_n^{\alpha} \alpha_n\right) \ge c, \quad \limsup_{n\to\infty} (n^t \beta_n) \le d, \quad \limsup_{n\to\infty} (n^s \delta_n) \le a} \tag{16.241}$$

for some c, d and a satisfying

$$\boxed{a \ge 0, \quad c > 0, \quad d \ge 0, \quad \alpha \in (0, 1), \quad t > \alpha, \quad s > \alpha} \tag{16.242}$$

Then

(a) if $s > r\alpha + (1-r)t$ and under $\alpha = 1$, $c > t - 1$, it follows that

$$\boxed{\begin{aligned} &\limsup_{n\to\infty} \left(n^{t-\alpha} u_n\right) \le \frac{d}{\hat{c}(\alpha)} \\ &\hat{c}(\alpha) = c - (t-1)\chi(\alpha = 1) \\ &\chi(\alpha = 1) = \begin{cases} 1 & \text{if} \quad \alpha = 1 \\ 0 & \text{if} \quad \alpha \ne 1 \end{cases} \end{aligned}} \tag{16.243}$$

(b) if $s = r\alpha + (1-r)\,t$ *and under* $\alpha = 1$, $c > t - 1$, *it follows that*

$$\boxed{\limsup_{n\to\infty}\left(n^{t-\alpha}u_n\right) \le f} \tag{16.244}$$

where $f > 0$ *is the root of the nonlinear equation*

$$\boxed{\hat{c}\,(\alpha)\,f = d + af^r} \tag{16.245}$$

(for $r = 1/2$ *we have* $f = \left(\dfrac{a + \sqrt{a^2 + 4\hat{c}\,(\alpha)\,d}}{2\hat{c}\,(\alpha)}\right)^2$*);*

(c) if $s < r\alpha + (1-r)\,t$ *and under* $\alpha = 1$, $c > \lambda = \dfrac{s-\alpha}{1-r}$, *it follows that*

$$\boxed{\limsup_{n\to\infty}\left(n^{\lambda}u_n\right) \le \left(\frac{a}{c - \lambda\chi\,(\alpha = 1)}\right)^{1/(1-r)}} \tag{16.246}$$

Proof. Let us use the inequality

$$x^r \le (1-r)\,x_0^r + \frac{r}{x_0^{1-r}}x \tag{16.247}$$

valid for any $x, x_0 > 0$. Indeed (see Fig. 16.3),

$$x^r \le l\,(x) = c + kx$$

where the parameters of the linear function $l\,(x)$ can be found from the following system of linear equations

$$c + kx_0 = x_0^r,\ (x^r)'\,|_{x=x_0} = k$$

which gives $k = rx_0^{r-1}$, $c = (1-r)\,x_0^r$ and, hence, $x^r \le c + kx = (1-r)\,x_0^r + rx_0^{r-1}x$. Taking

$$x := u_n, \quad x_0 := fn^{-\rho}\ \ (\rho := \min\{t - \alpha, \lambda\})$$

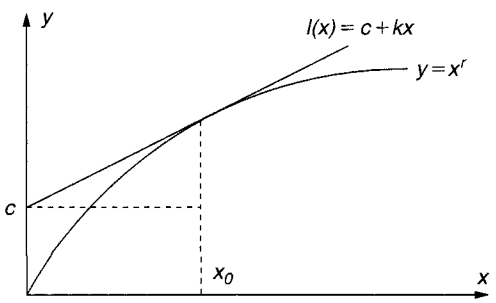

Fig. 16.3. Illustration of the inequality (16.247).

and using the obtained inequality for u_n^r in (16.240), for all $n \geq n_0$ we obtain

$$u_{n+1} \leq u_n \left(1 - \frac{c + o(1)}{n} + \frac{r\,(a + o\,(1))}{f^{1-r} n^{s-(1-r)/\rho}}\right) + \frac{d + o(1)}{n^t} + \frac{(1-r)\,(a + o\,(1))\,f^r}{n^{s+r\rho}}$$

Now the results (16.243), (16.244) and (16.246) follow immediately if applied to the last inequality Lemma 16.15 for the case $\alpha = 1$ and Corollary 16.31 for the case $\alpha < 1$. □

Lemma 16.17. (Nazin & Poznyak 1986) *Let the sequence $\{u_n\}$ of nonnegative real numbers satisfy the following recurrent equation:*

$$\boxed{u_{n+1} \leq u_n\,(1 - \alpha_n) + \beta_n + \bar{\delta}_n u_n^r, \quad n \geq n_0} \tag{16.248}$$

where $r \in (0, 1)$ and $\{\alpha_n\}$, $\{\beta_n\}$, $\{\delta_n\}$ are sequences of nonnegative real numbers such that

$$\boxed{\sum_{n=n_0}^{\infty} \alpha_n \geq c, \quad \limsup_{n\to\infty} (\beta_n/\alpha_n) = b, \quad \limsup_{n\to\infty} \left(\bar{\delta}_n/\alpha_n\right) = d \qquad \alpha_n \in (0, 1], \quad \beta_n \geq 0, \quad \bar{\delta}_n \geq 0} \tag{16.249}$$

Then

$$\boxed{\limsup_{n\to\infty} u_n \leq \inf_{c>r} u\,(c)} \tag{16.250}$$

where

$$\boxed{u\,(c) := \left(1 - \frac{r}{c}\right)^{-1} \left[b + (1-r)\,c^{r/(1-r)} d^{1/(1-r)}\right]} \tag{16.251}$$

Proof. Using the inequality (16.247) for $x = u_n$ and $x_0 = c\bar{\delta}_n/\alpha_n$, $c > r$ in (16.248), we derive

$$u_{n+1} \leq u_n \left[1 - \alpha_n \left(1 - \frac{r}{c}\right)\right] + \beta_n + (1-r)\left(c\bar{\delta}_n/\alpha_n\right)_n^{r/(1-r)} \bar{\delta}_n$$

or, equivalently,

$$u_{n+1} \leq u_n [1 - \tilde{\alpha}_n] + \tilde{\beta}_n$$

$$\tilde{\alpha}_n := \alpha_n \left(1 - \frac{r}{c}\right), \quad \tilde{\beta}_n := \beta_n + (1 - r) \left(c\bar{\delta}_n/\alpha_n\right)_n^{r/(1-r)} \bar{\delta}_n$$

Then applying Lemma 16.14 we obtain

$$\limsup_{n\to\infty} u_n \leq \limsup_{n\to\infty} \left(\tilde{\beta}_n/\tilde{\alpha}_n\right) := u\,(c)$$

Taking $\inf_{c>r}$ of the right-hand side we get (16.250) and (16.251). Lemma is proven. □

17 Complex Analysis

Contents

17.1 Differentiation

17.1.1 Differentiability

Let f be a complex-valued function of a complex variable z. Any complex function can be represented as follows

$$\boxed{\begin{array}{l} f(z) = u(x, y) + iv(x, y) \\ z = x + iy \end{array}} \tag{17.1}$$

We refer to u and v as the *real* and *imaginary* parts of f and write

$$\boxed{u = \mathrm{Re} f, \quad v = \mathrm{Im} f} \tag{17.2}$$

Example 17.1.

$$f(z) = z^2 = (x + iy)^2 = x^2 - y^2 + i2xy$$

$$u(x, y) = x^2 - y^2, \quad v(x, y) = xy$$

Definition 17.1. *If f is defined in some neighborhood of a finite point z and $\lim_{\Delta z \to 0} \dfrac{f(z + \Delta z) - f(z)}{\Delta z}$ exists finitely we say that the function $f(z)$ is* ***differentiable*** *at the point z. This limiting value is called the* ***derivative*** *of $f(z)$ at z and we write*

$$\boxed{f'(z) := \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}} \tag{17.3}$$

17.1.2 Cauchy–Riemann conditions

The next theorem gives conditions which must be imposed on the functions $u(x, y)$ and $v(x, y)$ in order that $f(z)$ should be differentiable at the point z.

Theorem 17.1. (The necessary conditions of differentiability) *Let f be defined in a neighborhood of the point $z \in \mathbb{C}$ and be differentiable at z. Then*

(a) the partial derivatives

$$\frac{\partial}{\partial x}u(x, y), \quad \frac{\partial}{\partial y}u(x, y), \quad \frac{\partial}{\partial x}v(x, y), \quad \frac{\partial}{\partial y}v(x, y)$$

exist;

(b) and the following relations hold:

$$\boxed{\frac{\partial}{\partial x}u(x, y) = \frac{\partial}{\partial y}v(x, y), \quad \frac{\partial}{\partial y}u(x, y) = -\frac{\partial}{\partial x}v(x, y)} \tag{17.4}$$

Proof. Since $f'(z)$ exists then for any given $\varepsilon > 0$ there exists $\delta = \delta(\varepsilon) > 0$ such that

$$\left|\frac{f(z + \Delta z) - f(z)}{\Delta z} - f'(z)\right| < \varepsilon \tag{17.5}$$

whenever $0 < |\Delta z| < \delta$. Representing Δz as $\Delta z = te^{i\alpha}$, where $t = |\Delta z|$ and $\alpha = \arg z$ (see (13.40)), we can see that (17.5) is fulfilled independently of α when $0 < t < \delta$. Let us take $\alpha = 0$. This means that $\Delta z = t = \Delta x$. This implies

$$\begin{aligned} f'(z) &= \lim_{\Delta x \to 0}\left[\frac{u(x + \Delta x, y) - u(x, y)}{\Delta x} + i\frac{v(x + \Delta x, y) - v(x, y)}{\Delta x}\right] \\ &= \frac{\partial}{\partial x}u(x, y) + i\frac{\partial}{\partial x}v(x, y) \end{aligned} \tag{17.6}$$

Taking $\alpha = \pi/2$ we find that $\Delta z = it = i\Delta y$ and, therefore,

$$\begin{aligned} f'(z) &= \lim_{\Delta y \to 0}\left[\frac{u(x, y + \Delta y) - u(x, y)}{i\Delta y} + i\frac{v(x, y + \Delta y) - v(x, y)}{i\Delta y}\right] \\ &= \frac{1}{i}\frac{\partial}{\partial y}u(x, y) + \frac{\partial}{\partial y}v(x, y) = -i\frac{\partial}{\partial y}u(x, y) + \frac{\partial}{\partial y}v(x, y) \end{aligned} \tag{17.7}$$

Comparing (17.6) with (17.7) we obtain (17.4). Theorem is proven. □

The conditions (17.4) are called the **Cauchy–Riemann conditions**. They are also known as the **d'Alembert–Euler conditions**. The theorem given below shows that these conditions are also sufficient to provide the differentiability.

Theorem 17.2. (The sufficient conditions of differentiability) *The Cauchy–Riemann conditions (17.4) are also sufficient for the differentiability of $f(z)$ provided the functions $u(x, y)$ and $v(x, y)$ are totally differentiable (all partial derivatives exist) at the considered point. The derivative $f'(z)$ can be calculated as*

$$\boxed{f'(z) = \frac{\partial}{\partial x}u(x, y) + i\frac{\partial}{\partial x}v(x, y) = \frac{\partial}{\partial y}v(x, y) - i\frac{\partial}{\partial y}u(x, y)} \tag{17.8}$$

Proof. By the total differentiability it follows that

$$\Delta u := u(x + \Delta x, y + \Delta y) - u(x, y) = \frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y + o(|\Delta z|)$$

$$\Delta v := v(x + \Delta x, y + \Delta y) - v(x, y) = \frac{\partial v}{\partial x}\Delta x + \frac{\partial v}{\partial y}\Delta y + o(|\Delta z|)$$

Therefore

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{\Delta u + i\Delta v}{\Delta x + i\Delta y}$$
$$= \frac{\left(\frac{\partial u}{\partial x}\Delta x + \frac{\partial u}{\partial y}\Delta y\right) + i\left(\frac{\partial v}{\partial x}\Delta x + \frac{\partial v}{\partial y}\Delta y\right) + o(|\Delta z|)}{\Delta x + i\Delta y}$$

Using now the Cauchy–Riemann conditions (17.4), the simple rearrangement gives

$$\frac{f(z + \Delta z) - f(z)}{\Delta z} = \frac{\left(\left[\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right]\Delta x + \left[\frac{\partial u}{\partial y} + i\frac{\partial v}{\partial y}\right]\Delta y\right) + o(|\Delta z|)}{\Delta x + i\Delta y}$$
$$= \frac{\left(\left[\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right]\Delta x + \left[-i\frac{\partial u}{\partial y} + \frac{\partial v}{\partial y}\right]i\Delta y\right) + o(|\Delta z|)}{\Delta x + i\Delta y}$$
$$= \frac{\left(\left[\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right]\Delta x + \left[i\frac{\partial v}{\partial x} + \frac{\partial u}{\partial x}\right]i\Delta y\right) + o(|\Delta z|)}{\Delta x + i\Delta y}$$
$$= \left[\frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}\right] + o(1)$$

where $o(1) \to 0$ whenever $|\Delta z| \to 0$. So that $f'(z)$ exists and is given by

$$f'(z) = \frac{\partial u}{\partial x} + i\frac{\partial v}{\partial x}$$

which completes the proof. □

Example 17.2. *For the same function* $f(z) = z^2$ *as in Example 17.1 we have*

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial x} = 2x, \quad \frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} = -2y$$
$$f'(z) = 2x - i2y$$

Definition 17.2. *A function* $f(z)$*, differentiable at each point of an open set* $\mathcal{D} \subset \mathbb{C}$*, is called* ***regular*** *(or* ***holomorphic****) on* $\mathcal{D}$*. Sure, here we assume that we deal with a single-valued (or uniform) function since the notion of differentiability (17.3) has been introduced only for single-valued functions. If a regular function* $f(z)$ *possesses a continuous derivative on* $\mathcal{D}$ *then it is called an* ***analytic function****.*[1]

Below these definitions will be extensively used.

Example 17.3. *It is easy to check that, as in real analysis,*

$$\boxed{\begin{aligned} \frac{d}{dz}e^z &= e^z, \quad \frac{d}{dz}\ln z = \frac{1}{z} \\ \frac{d}{dz}\sin z &= \cos z, \quad \frac{d}{dz}\cos z = -\sin z \\ \frac{d}{dz}\tan z &= \frac{1}{\cos^2 z}, \quad \frac{d}{dz}\cot z = -\frac{1}{\sin^2 z} \end{aligned}} \tag{17.9}$$

17.1.3 Theorem on a constant complex function

An application of the Cauchy–Riemann equations (17.4) is given in the next lemma.

Lemma 17.1. *Let* $f = u + iv$ *be a function with a derivative everywhere in an open disc* $\mathcal{D} \subset \mathbb{C}$ *centered at the point* $z = (a, b)$*.*

1. *If any of* u, v *or* $|f|^2 := u^2 + v^2$ *is constant on* $\mathcal{D}$*, then* f *is constant on* $\mathcal{D}$*.*
2. *Also,* f *is constant on* $\mathcal{D}$ *if* $f'(z) = 0$ *for all* $z \in \mathcal{D}$*.*

Proof. Suppose u is a constant on $\mathcal{D}$. By (17.4) it follows that $\dfrac{\partial v}{\partial x} = \dfrac{\partial v}{\partial y} = 0$. Therefore

$$v(x, y) = v(a, b)$$

So, v is a constant on $\mathcal{D}$. By the same argument we show that u is a constant on $\mathcal{D}$ if v is a constant. Now suppose that $|f|^2 := u^2 + v^2$ is a constant. This, in view of (17.4), implies

$$0 = u\frac{\partial u}{\partial x} + v\frac{\partial v}{\partial x} = u\frac{\partial u}{\partial y} + v\frac{\partial v}{\partial y} = -u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x}$$

[1] It can be shown that the existence of $f'(z)$ on D automatically implies continuity of $f'(z)$ on D (Goursat 1900). So, **regularity** and **analyticity** can be considered as two definitions having an identical mathematical sense.

and, hence,

$$\left(u^2 + v^2\right) \frac{\partial u}{\partial x} = |f|^2 \frac{\partial u}{\partial x} = 0$$

If $|f|^2 = 0$, then $u = v = 0$ and so $f = 0$. If $|f|^2 \neq 0$, then $\dfrac{\partial u}{\partial x} = 0$ and so u is a constant. Hence, by the arguments above, v is also a constant that shows that f is a constant too. Finally, if $f' = 0$ on $\mathcal{D}$, then both $\dfrac{\partial v}{\partial x}$ and $\dfrac{\partial v}{\partial y}$ are equal zero. So, v is a constant, and hence, u is a constant too. Lemma is proven. □

17.2 Integration

17.2.1 Paths and curves

Definition 17.3. *A **path** (**contour**) in a complex plane is a complex-valued function $C = C(t)$, continuous on a compact interval $[a, b] \in \mathbb{R}$. The image of $[a, b]$ under C (the graph of C) is said to be a **curve** described by C and it is said to join the points $C(a)$ and $C(b)$. If*

*(a) $C(a) \neq C(b)$, the curve is called an **arc** with the endpoints $C(a)$ and $C(b)$;*
*(b) $C(t)$ is one-to-one on $[a, b]$, the curve is called a **simple** (or **Jordan**) **arc**;*
*(c) $C(a) = C(b)$, the curve is a **closed curve**;*
*(d) $C(a) = C(b)$ and $C(t)$ is one-to-one on $[a, b]$, the curve is called a **simple** (or **Jordan**) **closed curve**.*

These types of curves are shown in Fig. 17.1.

Definition 17.4.

*(a) A path C is called **rectifiable** if it has a finite arc length.*

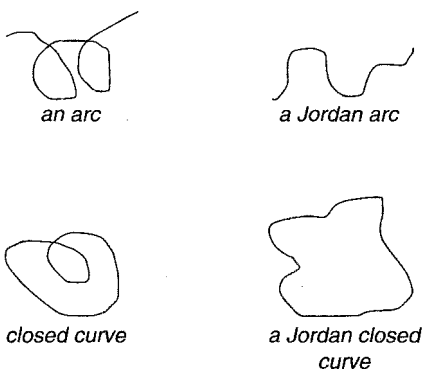

Fig. 17.1. Types of curves in the complex plane.

(b) *A path C is called* ***piecewise smooth*** *if it has a bounded derivative C' which is continuous everywhere on $[a, b]$ except (possibly) a finite number of points for which it is required that both right- and left-hand derivatives exist.*
(c) *A piecewise smooth closed path is called a* ***circuit.***
(d) *The* ***arc length*** *$\Lambda_C(a, b)$ of a path C on $[a, b]$ is defined by*

$$\boxed{\begin{aligned}\Lambda_C(a,b) &:= \sup_{P_m}\left\{\Lambda_m^C(a,b) : P_m \in \mathcal{P}(a,b)\right\}\\ \Lambda_{P_m}^C(a,b) &:= \sum_{i=1}^{m} |C(t_i) - C(t_{i-1})|\end{aligned}} \tag{17.10}$$

where $\mathcal{P}(a, b)$ is the set of all possible partitions of $[a, b]$.

Lemma 17.2. *A path C is rectifiable if and only if $C(t)$ is of bounded variation on $[a, b]$.*

Proof. If $P_m := \{t_0 = a, t_1, t_2, \ldots, t_m = b\}$ is a partition of $[a, b]$, and if $C(t)$ is a function of bounded variation on $[a, b]$, that is, for all $a \le t_{i-1} \le t_i \le b$

$$|C(t_i) - C(t_{i-1})| \le M(t_i - t_{i-1})$$

then

$$\begin{aligned}\Lambda_C(a,b) &= \sup_{P_m}\left\{\sum_{i=1}^{m} |C(t_i) - C(t_{i-1})|\right\}\\ &\le \sup_{P_m}\left\{M\sum_{i=1}^{m}(t_i - t_{i-1})\right\} = M(b-a) < \infty\end{aligned}$$

which proves this lemma. □

Corollary 17.1. *The arc length $\Lambda_C(a, b)$ may be calculated as the Lebesgue integral*

$$\boxed{\Lambda_C(a,b) = \int_{t=a}^{b} |C'(t)|\, dt} \tag{17.11}$$

and

$$\boxed{\Lambda_C(a,b) = \Lambda_C(a,c) + \Lambda_C(c,b)} \tag{17.12}$$

Definition 17.5. *If $a \in \mathbb{C}$, $r > 0$ and the path C is defined by the equation*

$$\boxed{C(t) := a + re^{it}, \quad t \in [0, 2\pi]} \tag{17.13}$$

then this path is called a ***positively*** *(counterclockwise)* ***oriented circle*** *(or a* ***sphere in*** $\mathbb{C}$*) with the center at a and the radius r. It is denoted by* $B(a, r)$ *and is referred to as* $C = B(a, r)$.

17.2.2 Contour integrals

Let C be a path in the complex plane $\mathbb{C}$ with domain $[a, b]$, and $f : \mathbb{C} \to \mathbb{C}$ be a complex-valued function

$$f(z) = u(x, y) + iv(x, y), \quad z = x + iy$$

defined on the graph of C.

Definition 17.6. *The* ***contour integral*** *of* f *along* C*, denoted by* $\int_C f(z)dz$*, is defined by*

$$\boxed{\begin{gathered}\int_C f(z)dz := \int_{t=a}^{b} f(C(t))dC(t) = \lim_{m\to\infty} \sup_{P_m} s\left(P_m, \zeta_k^{(m)}\right) \\ s\left(P_m, \zeta_k^{(m)}\right) := \sum_{k=1}^{m} f(\zeta_k)(z_k - z_{k-1}) \\ z_k \in C,\ z_0 = C(a),\ z_k = C(t_k),\ z_m = C(b) \\ \Delta(P_m) := \max_{k=1,\ldots,m} |z_k - z_{k-1}| \underset{m\to\infty}{\to} 0\end{gathered}} \tag{17.14}$$

whenever the Riemann–Stieltjes integral $\int_{t=a}^{b} f(C(t))dC(t)$ *on the right-hand side of (17.14) exists. If the contour C is closed, that is,* $C(a) = C(b)$ *(see Fig. 17.2), then the integral (17.14) is denoted by*

$$\boxed{\oint_C f(z)\,dz := \int_{t=a}^{b=a} f(C(t))dC(t)} \tag{17.15}$$

Remark 17.1. *If* $f(z)$ *is a partially continuous bounded function, the integral (17.14) always exists.*

Lemma 17.3. *The calculation of the contour integral* $\int_C f(z)\,dz$ *(17.14) can be realized by the calculation of four real integrals according to the following formula:*

$$\boxed{\int_C f(z)\,dz = \int_C [u(x,y)\,dx - v(x,y)\,dy] + i\int_C [u(x,y)\,dy + v(x,y)\,dx]} \tag{17.16}$$

Proof. It follows from the presentation

$$\sum_{k=1}^{m} f(\zeta_k)(z_k - z_{k-1}) = \sum_{k=1}^{m} [u_k \Delta x_k - v_k \Delta y_k] + i \sum_{k=1}^{m} [u_k \Delta y_k + v_k \Delta x_k]$$

$$u_k := u(x_k, y_k), \quad v_k := v(x_k, y_k)$$

$$\Delta x_k := x_k - x_{k-1}, \quad \Delta y_k := y_k - y_{k-1}, \quad z_k = x_k + i y_k = C(t_k)$$

□

Denote by C^- the same contour C but passed in the clockwise direction (see Fig. 17.2). Then the following properties seem to be evident.

Proposition 17.1.

1.

$$\boxed{\int_C f(z)\,dz = -\int_{C^-} f(z)\,dz} \tag{17.17}$$

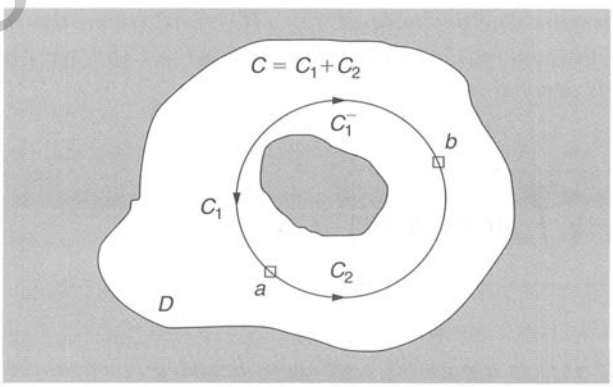

Fig. 17.2. The closed contour $C = C_1 + C_2$ within the region $\mathcal{D}$.

2.

$$\boxed{\int_C f(z)\,dz = \int_{t=a}^{b} f(C(t))C'(t)\,dt} \tag{17.18}$$

3. *For any* $\alpha, \beta \in \mathbb{C}$

$$\boxed{\int_C [\alpha f(z) + \beta g(z)]\,dz = \alpha \int_C f(z)\,dz + \beta \int_C g(z)\,dz} \tag{17.19}$$

4.

$$\boxed{\int_{C_1+C_2} f(z)\,dz = \int_{C_1} f(z)\,dz + \int_{C_2} f(z)\,dz} \tag{17.20}$$

17.2.3 Cauchy's integral law

We investigate the conditions under which the integral $\int_C f(z)\,dz$ along a path C joining any two given points a and b in a domain $\mathcal{D}$ is independent of the particular path C (in $\mathcal{D}$), but depends on a and b only.

Lemma 17.4. *It is necessary and sufficient that the integral of* $f(z)$ *along a path joining any two given points* a *and* b *in a domain* $\mathcal{D}$ *is independent of the particular path* C *(in* $\mathcal{D}$*) in that the integral of the same function* $f(z)$ *around a closed path* C *in* $\mathcal{D}$ *(see Fig. 17.2) should vanish, that is,*

$$\boxed{\oint_C f(z)\,dz = 0} \tag{17.21}$$

Proof.

(a) *Necessity.* Suppose that the integral $\int_C f(z)\,dz$ along any path C in $\mathcal{D}$ depends only on the endpoints a and b, i.e., $\int_C f(z)\,dz = \phi(a, b)$. Let us choose two distinct arcs C_1 and C_2 of C joining a and b (see Fig. 17.2). Then we have

$$\oint_C f(z)\,dz = \int_{C_2} f(z)\,dz + \int_{C_1} f(z)\,dz$$

$$= \int_{C_2} f(z)\,dz - \int_{C_1^-} f(z)\,dz$$

$$= \phi(a, b) - \phi(a, b) = 0$$

(b) *Sufficiency.* Suppose now that (17.21) holds for a domain $\mathcal{D}$. Suppose $a, b \in \mathcal{D}$ and select any two paths C_1^- and C_2 in $\mathcal{D}$ joining a to b. Since $C = C_1 + C_2$ and (17.21) holds, we get

$$0 = \oint_C f(z)\, dz = \int_{C_2} f(z)\, dz + \int_{C_1} f(z)\, dz$$

$$= \int_{C_2} f(z)\, dz - \int_{C_1^-} f(z)\, dz$$

Hence, $\int_{C_2} f(z)\, dz = \int_{C_1^-} f(z)\, dz$ for any paths C_2 and C_1^-. This means that $\int_C f(z)\, dz = \phi(a, b)$ exactly. The lemma is proven. □

17.2.3.1 Simply-connected domains

Now we are ready to formulate the following fundamental integral theorem.

Theorem 17.3. (Cauchy's integral law) *If $f(z)$ is a **regular function** in a simply-connected domain[2] $\mathcal{D}$, then the integral of $f(z)$ along any path C in $\mathcal{D}$ depends only on the endpoints of this path, or in other words, if C is any closed contour in $\mathcal{D}$ then*

$$\oint_C f(z)\, dz = 0$$

Proof. By (17.16) it suffices to show that each of the real line integrals

$$\oint_C [u(x, y)\, dx - v(x, y)\, dy], \quad \oint_C [u(x, y)\, dy + v(x, y)\, dx]$$

vanishes. By Corollary 16.2 (from the Real Analysis chapter) it follows that

$$\oint_C [P(x, y)\, dx + Q(x, y)\, dy] = 0$$

along any closed contour in a simply-connected domain $\mathcal{D}$ if and only if the partial derivative of the real functions $P(x, y)$ and $Q(x, y)$ exist and are continuous in $\mathcal{D}$ and

$$\frac{\partial}{\partial y} P(x, y) = \frac{\partial}{\partial x} Q(x, y) \tag{17.22}$$

[2] We are reminded that a domain D in an open plane is *simply connected* if and only if any closed Jordan contour C in D is *reducible* in D, that is, can be continuously shrunk to a point in D without leaving D.

at each point of $\mathcal{D}$. Indeed,

$$P(x, y)\,dx + Q(x, y)\,dy = d\varphi(x, y)$$

and

$$\oint_C [P(x, y)\,dx + Q(x, y)\,dy] = \oint_C d\varphi(x, y)$$
$$= \varphi(x(a), y(a)) - \varphi(x(b), y(b)) = 0$$

since $a = b$. But, in our case,

$$u(x, y) = P(x, y), \quad v(x, y) = -Q(x, y)$$

for the first integral, and

$$v(x, y) = P(x, y), \quad u(x, y) = Q(x, y)$$

So, (17.22) coincides exactly with the Cauchy–Riemann conditions (17.4) which proves the theorem. □

Remark 17.2. *The converse of Theorem 17.3 is also true, namely, if $f(z)$ is continuous in a simply-connected domain D and $\oint_C f(z)\,dz = 0$ for every closed contour C in D, then $f(z)$ is regular in D. The proof can be found in Fuchs & Shabat (1964).*

As it follows from the consideration above, Theorem 17.3 enables us to give an equivalent alternative definition of a regular function: **a single-valued function *f(z)* is regular in *D* if it is continuous in *D* and its integral around any closed contour *C* in *D* is equal to zero.**

17.2.3.2 Multiply-connected domains

The Cauchy theorem 17.3 can be generalized so as to apply to multiply-connected domains. Let D be an $(n + 1)$-ply connected bounded domain whose frontier consists of $(n + 1)$ disjoint contours C_0 (the external boundary component), $C_1, \ldots, C_n$, and let $f(z)$ be regular at each point of the closed region $\bar{D}$ (see Fig. 17.3 showing a case in which $n = 3$).

By taking suitable (disjoint) cuts $\gamma_1, \gamma_2, \ldots \gamma_n$ we form from D a simply-connected domain D' whose boundary we denote by C. We will consider each cut γ_i $(i = 1, \ldots, n)$ as two-edged as in Fig. 17.3.

Theorem 17.4. *If $f(z)$ is regular at each point of the closed region $\bar{D}$ whose frontier C consists of a finite number of disjoint contours C_i $(i = 0, 1, \ldots, n)$, that is, $C = \bigcup_{i=0}^{n} C_i$, then the integral $\oint_C f(z)\,dz$ of $f(z)$ around the boundary of $\bar{D}$ (taken so that each component of the boundary is traversed in a sense such that the interior D of $\bar{D}$ remains on the left) is equal to zero.*

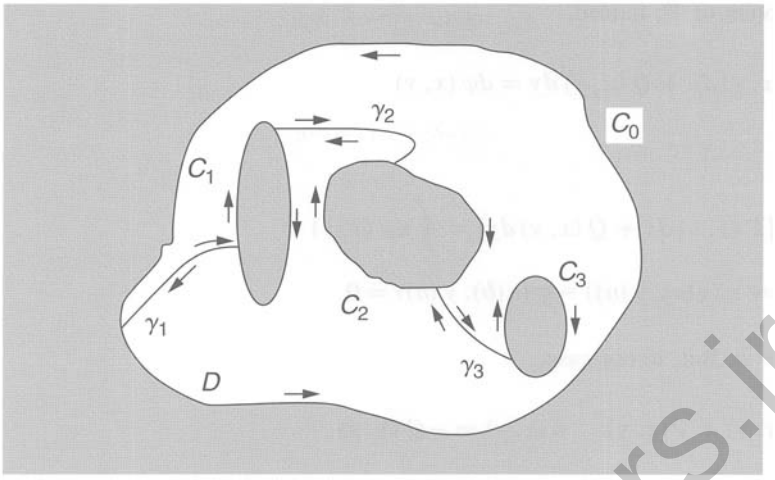

Fig. 17.3. Multiply-connected domains ($n = 3$).

Proof. As C is piecewise smooth and $f(z)$ is regular in the domain D^* containing $\bar{D}$ it follows by Theorem 17.3 that $\oint_C f(z)\,dz = 0$. So, C_0 is traversed in the anti-clockwise (positive) sense and C_i $(i = 1, \ldots, n)$ in the clockwise (negative) sense. Each of cuts $\gamma_1, \gamma_2, \ldots \gamma_n$ is traversed twice: first, in one sense and then in the opposite sense. As the integrals along the two edges of each cut conceal each other, it follows by the standard properties of the contour integral (17.1) that

$$\begin{aligned} \oint_C f(z)\,dz &= \oint_{C_0} f(z)\,dz - \sum_{i=1}^{n} \oint_{C_k} f(z)\,dz \\ &= \oint_{C_0} f(z)\,dz + \sum_{i=1}^{n} x \oint_{C_k^-} f(z)\,dz = 0 \end{aligned} \tag{17.23}$$

Theorem is proven. □

Corollary 17.2. *Assuming that the conditions under which (17.23) holds are satisfied, we have*

$$\boxed{\oint_{C_0} f(z)\,dz = \sum_{i=1}^{n} \oint_{C_k} f(z)\,dz} \tag{17.24}$$

which for $n = 1$ *gives*

$$\boxed{\oint_{C_0} f(z)\,dz = \oint_{C_1} f(z)\,dz} \tag{17.25}$$

17.2.4 Singular points and Cauchy's residue theorem

17.2.4.1 Types of singularities

Definition 17.7. *Consider a function $f(z)$ which is regular (analytic) everywhere on an open set D bounded by a closed contour C (here we are writing C for C_0), except at a finite number of isolated points $a_1, a_2, \ldots, a_n$. These exceptional points are called* ***singular points*** *(singularities) of $f(z)$.*

Isolated singularities are divided into **three types** according to the behavior of the function in a deleted neighborhood of the point concerned.

Definition 17.8. *An isolated singularity a of the function $f(z)$ is said to be*

1. *a* ***removable singularity****, if $\lim\limits_{z\to a} f(z)$ exists finitely;*
2. *a* ***pole****, if $\lim\limits_{z\to a} f(z) = \pm\infty$;*
3. *an* ***essential singularity****, $f(z)$ does not tend to a limit (finite or infinite) as $z \to a$.*

Remark 17.3. *All of these notions are closely connected with the, so-called, Laurent expansion of the function $f(z)$ which will be discussed below. There will be shown that*

- *a removable singularity cannot contain the term $\dfrac{c}{(z-a)^n}$ for any finite $n \geq 1$ (for example, the function $\dfrac{\sin z}{z}$ at the point $z = a = 0$ has a removable singularity);*
- *evidently, that a function $f(z)$ defined in some deleted neighborhood of $z = a$ has a pole at $z = a$ if and only if the function $g(z) := \dfrac{1}{f(z)}$ is regular at a and has zero at $z = a$, i.e., $g(a) = 0$ (while $g(z)$ is not identically equal to zero);*
- *in the case of isolated essential singularity there exist (the Sokhotsky–Cazoratti theorem, 1868) at least two sequences $\{z'_n\}$ and $\{z''_n\}$, each converging to a, such that the corresponding sequences $\{f(z'_n)\}$ and $\{f(z''_n)\}$ tend to different limits as $n \to \infty$ (for example, the function $e^{1/z}$ at the point $z = a = 0$ has an essential singularity and is regular for all other z).*

Definition 17.9. *A function $f(z)$ is called* ***meromorphic*** *(ratio type) if its singularities are only poles.*

From this definition it immediately follows that in any bounded closed domain of the complex plane a meromorphic function may have only a finite number of poles: for, otherwise, there would exist a sequence of distinct poles converging to a (finite) point in the region; such point would necessarily be a nonisolated singularity that contradicts our hypothesis that any finite singular point of this function must be a pole.

Example 17.4. *Meromorphic functions are* $1/\sin z$, $\tan z$, $\cot z$.

17.2.4.2 Cauchy's residue theorem

We enclose the a_k by mutually disjoint circles C_k in D such that each circle C_k enclosing no singular points other than the corresponding point a_k. It follows readily

from (17.25) that the integral of $f(z)$ around C_k is equal to the integral around any other contour C_k' in D which also encloses a_k, but does not enclose or pass through any other singular point of $f(z)$. Thus the value of this integral is a characteristic of $f(z)$ and the singular point a_k.

Definition 17.10. *The **residue** of $f(z)$ at the singular point a_k is denoted by* $\operatorname{res} f(a_k)$ *and is defined by*

$$\boxed{\operatorname{res} f(a_k) := \frac{1}{2\pi i} \oint_{C_k} f(z)\, dz} \tag{17.26}$$

Formula (17.24) leads immediately to the following result.

Theorem 17.5. (Cauchy's residue theorem) *Let D be an open domain bounded by a closed contour C and let $f(z)$ be regular (analytic) at all points of $\bar{D}$ with the exception of a finite number of singular points $a_1, a_2, \ldots, a_n$ contained in the domain D. Then the integral of $f(z)$ around C is $2\pi i$ times the sum of its residues at the singular points, that is,*

$$\boxed{\oint_C f(z)\, dz = 2\pi i \sum_{k=1}^{n} \operatorname{res} f(a_k)} \tag{17.27}$$

Corollary 17.3. *The residue of $f(z)$ at a removable singularity is equal to zero.*

The next subsection deals with method of residues calculating without integration.

17.2.5 Cauchy's integral formula

17.2.5.1 Representation of an analytic function through its contour integral

The theorem below reveals a remarkable property of analytical functions: it relates the value of an analytical function at a point with the value on a closed curve not containing the point.

Theorem 17.6. (Cauchy's integral formula) *Assume f is regular (analytic) on an open set D, and let C be any contour (circuit) in D which encloses a point $z \in D$ but does not cross it. Then*

$$\boxed{\oint_C \frac{f(w)}{w - z} dw = f(z) \oint_C \frac{1}{w - z} dw} \tag{17.28}$$

Proof. Define a new function g on D as follows:

$$g(w) := \begin{cases} \dfrac{f(w) - f(z)}{w - z} & \text{if} \quad w \neq z \\ f'(z) & \text{if} \quad w = z \end{cases}$$

Then $g(w)$ is regular (analytic) at each point $w \neq z$ in D and at the point z itself it is continuous. Applying the Cauchy theorem 17.3 to g gives

$$0 = \oint_C g(w)\,dw = \oint_C \frac{f(w) - f(z)}{w - z} dw = \oint_C \frac{f(w)}{w - z} dw - f(z) \oint_C \frac{1}{w - z} dw$$

which proves (17.28). □

Example 17.5. *If $C = B(z, r)$ is a* ***positively*** *(counterclockwise)* ***oriented circle*** *(17.13) with the center at a and the radius r, that is,*

$$C = C(t) := z + re^{it}, \quad t \in [0, 2\pi]$$

then

$$C'(t) = ire^{it} = i[C(t) - z]$$

and by (17.18) we derive that

$$\boxed{\oint_C \frac{1}{w - z} dw = \int_{t=0}^{2\pi} \frac{C'(t)}{C(t) - z} dt = \int_{t=0}^{2\pi} i\, dt = 2\pi i} \tag{17.29}$$

In this case (17.28) becomes

$$\boxed{\oint_C \frac{f(w)}{w - z} dw = 2\pi i\, f(z)} \tag{17.30}$$

Corollary 17.4. (Mean-value theorem) *For $C = B(z, r)$ it follows that*

$$\boxed{f(z) = \frac{1}{2\pi} \int_{t=0}^{2\pi} f\left(z + re^{it}\right) dt} \tag{17.31}$$

Proof. By (17.18) and (17.30) we have

$$2\pi i f(z) = \oint_C \frac{f(w)}{w-z} dw$$

$$= \int_{t=0}^{2\pi} f\left(z + re^{it}\right) \frac{C'(t)}{C(t)-z} dt$$

$$= \int_{t=0}^{2\pi} f\left(z + re^{it}\right) i\, dt$$

which proves (17.31). □

The next theorem shows that formula (17.30) holds not only for positively oriented circles (17.13) but for any circuit containing z as an internal point.

Theorem 17.7. *If f is regular (analytic) on an open set D, and let C be any contour (circuit) in D (not obligatory $C = B(z, r)$) which encloses a point $z \in D$ but does not cross it, then*

$$\boxed{\oint_C \frac{1}{w-z} dw = 2\pi i n(C, z)} \tag{17.32}$$

where $n(C, z)$ is an integer called the ***winding number*** *(or index) of C with respect to z which is the number of times the point $C(t)$ "winds around" the point z as t varies over the interval $[a, b]$.*

Proof. By (17.18) it follows that

$$\oint_C \frac{1}{w-z} dw = \int_{t=a}^{b} \frac{C'(t)}{C(t)-z} dt$$

Define the complex-valued function $F(x)$ by the equation

$$F(x) := \int_{t=a}^{x} \frac{C'(t)}{C(t)-z} dt, \quad t \in [a, b]$$

To prove the theorem we must show that $F(b) = 2\pi i n$ for some integer n. Notice that $F(x)$ is continuous and $F'(x) = \dfrac{C'(x)}{C(x)-z}$ at each point where $C'(t)$ exists, and, hence, the function $G(x)$ defined by

$$G(x) := e^{-F(x)} [C(x) - z]$$

is also continuous on $[a, b]$ and, moreover, at each point, where $C'(t)$ exists, we have

$$G'(x) = e^{-F(x)}C'(x) - F'(x)e^{-F(x)}[C(x) - z]$$

$$= e^{-F(x)}C'(x) - \frac{C'(x)}{C(x) - z}e^{-F(x)}[C(x) - z] = 0$$

Therefore, $G'(x) = 0$ for each t in $[a, b]$ except (possibly) for a finite number of points. By continuity of $G(x)$ on $[a, b]$ we conclude that $G(x)$ is a constant throughout $[a, b]$ that implies $G(a) = G(b)$ or, equivalently,

$$G(b) = e^{-F(b)}[C(b) - z] = G(a) = e^{-F(a)}[C(a) - z] = C(a) - z$$

Since $C(a) = C(b) \neq z$ we find $e^{-F(b)} = 1$ that gives $F(b) = 2\pi i n$ exactly where n is an integer and corresponds to the number of times the point $C(t)$ "winds around" the point z as t varies over the interval $[a, b]$. This completes the proof. □

Corollary 17.5. *Cauchy's integral formula (17.28) can now be restated in the form*

$$\boxed{\frac{1}{2\pi i}\oint_C \frac{f(w)}{w - z}dw = n(C, z)\ f(z)} \tag{17.33}$$

Example 17.6. *Let C be any contour enclosing the point $a \in \mathbb{C}$. We need to calculate $J := \oint_C (z - a)^n\, dz$ for every integer $n = \ldots, -1, 0, 1, \ldots$ By (17.25) it follows that*

$$J = \oint_C (z - a)^n\, dz = \oint_{B(a,r)} (z - a)^n\, dz$$

Letting $z - a = re^{i\varphi}$ we get $dz = re^{i\varphi}i\, d\varphi$, and hence,

$$J = \oint_{B(a,r)} (z - a)^n\, dz = \int_{\varphi=0}^{2\pi} r^n e^{in\varphi} re^{i\varphi} i\, d\varphi = ir^{n+1}\int_{\varphi=0}^{2\pi} e^{i(n+1)\varphi}\, d\varphi$$

Since

$$\int e^{k\varphi}\, d\varphi = \frac{1}{k}e^{k\varphi} + K, \quad k \neq 0, \quad K = const$$

one gets

$$J = ir^{n+1}\int_{\varphi=0}^{2\pi} e^{i(n+1)\varphi}\, d\varphi = \frac{r^{n+1}}{n+1}e^{i(n+1)\varphi}\Big|_{\varphi=0}^{\varphi=2\pi}$$

$$= \frac{r^{n+1}}{n+1}\left(e^{i2\pi(n+1)} - 1\right) \quad \text{if} \quad n \neq -1$$

and $J = 2\pi i$ *if* $n = -1$. *So,*

$$\boxed{J = \oint_C (z-a)^n \, dz = \begin{cases} 0 & for \quad n \neq -1 \\ 2\pi i & for \quad n = -1 \end{cases}} \tag{17.34}$$

Notice that case $n = -1$ *follows directly from (17.30) if we take* $f(z) = 1$.

17.2.5.2 High-order derivatives integral representation

By definition, the analytical function is a function of a complex-variable differentiable at any point of a domain D. The next theorem shows that an analytical function automatically has a derivative of any order which in turn is analytical too.

Theorem 17.8. (The bounds for high-order derivatives) *If* $f(z)$ *is regular (analytical) in an open domain* D *and is continuous in* $\bar{D}$, *then* ***it possesses the derivatives of all orders*** *at each point* $z \in D$ *and the derivative of the order* n *can be calculated as*

$$\boxed{f^{(n)}(z) = \frac{n!}{2\pi i} \oint_C \frac{f(w)}{(w-z)^{n+1}} dw} \tag{17.35}$$

where C *is the boundary of* D, *that is,* $C = \bar{D} \setminus D$.

Proof. By the derivative definition and using (17.32) for $n(C, z) = 1$, we have

$$\begin{aligned} f'(z) &= \lim_{h \to 0} \frac{f(z+h) - f(z)}{h} \\ &= \frac{1}{2\pi i} \lim_{h \to 0} \frac{1}{h} \oint_C f(w) \left[\frac{1}{(w-z-h)} - \frac{1}{(w-z)} \right] dw \\ &= \frac{1}{2\pi i} \lim_{h \to 0} \oint_C \frac{f(w)}{(w-z-h)(w-z)} dw = \frac{1}{2\pi i} \oint_C \frac{f(w)}{(w-z)^2} dw \end{aligned}$$

So, for $n = 1$ the theorem is proven. Let us use now the induction method, namely, supposing that it is true for some fixed n and using the same calculations as before, we can easily show that it is true also for $(n+1)$ which completes the proof. □

Remark 17.4. *Formula (17.35) may be obtained by the formal direct differentiation of Cauchy's formula (17.28) by* z.

Remark 17.5. *If the function* $\varphi(z)$ *is continuous on the boundary* C *of an open domain* D, *then the function*

$$\Phi(z) := \frac{1}{2\pi i} \oint_C \frac{\varphi(w)}{(w-z)} dw$$

is regular (analytical) in D.

Example 17.7.

$$\frac{1}{\pi i}\oint_C \frac{\sin(w)}{(w-z)^3}dw = \frac{2!}{2\pi i}\oint_C \frac{\sin(w)}{(w-z)^{2+1}}dw = \sin^{(2)}(z) = -\sin z$$

17.2.5.3 Cauchy's inequalities

Formula (17.35) leads directly to the following important inequalities known as the Cauchy's inequalities for the module of the nth derivative. Indeed, from (17.35) it follows that

$$\boxed{\left|f^{(n)}(z)\right| = \frac{n!}{2\pi}\left|\oint_C \frac{f(w)}{(w-z)^{n+1}}dw\right| \leq \frac{n!}{2\pi}\frac{Ml}{r^{n+1}}} \tag{17.36}$$

where r is the distance between the point z and the boundary C, i.e., $r := \inf_{w\in C}\|z-w\|$, M is the supremum of the module of $f(w)$ in D, i.e., $M := \sup_{z\in D}|f(z)|$ and l is the length of C, i.e., $l := |\oint_C dw|$. In particular, if $f(w)$ is analytical in the disc $D = \{w \in \mathbb{C} : |w-z| < r\}$, then $l = 2\pi r$ and we obtain

$$\boxed{\left|f^{(n)}(z)\right| \leq \frac{Mn!}{r^n}} \tag{17.37}$$

17.2.5.4 Liouville's theorem

Definition 17.11. *A function analytical everywhere on $\mathbb{C}$ is called an* ***entire*** *function.*

Example 17.8. *Entire functions are polynomials,* $\sin z$ *and* $\cos z$*, and* e^z*.*

Theorem 17.9. (Liouville) *Every bounded entire function is constant.*

Proof. Suppose $|f(z)| \leq M$ for all $z \in \mathbb{C}$. Then by (17.37) applied for $n = 1$ it follows that $\left|f^{(n)}(z)\right| \leq \frac{M}{r}$ for every $r > 0$. Letting $r \to \infty$ implies $f'(z) = 0$ for every $z \in \mathbb{C}$ which completes the proof. □

17.2.6 Maximum modulus principle and Schwarz's lemma

Theorem 17.10. (Maximum modulus principle) *If a function $f(z)$ is analytic and not constant on an open region D and is continuous on $\bar{D}$, then its module $|f(z)|$ cannot achieve its maxima in any point D, that is, every contour*

$$C = B(a,r) := \{z \in \mathbb{C} \,|\, |z-a| = r\} \subset D$$

contains points z such that

$$\boxed{|f(z)| > |f(a)|} \tag{17.38}$$

Proof. By continuity of $|f(z)|$ it achieves its maximum M on $\bar{D}$. Denote by $\mathcal{E}$ the set of all extrema points, i.e.,

$$\mathcal{E} := \left\{ z \in \bar{D} \,|\, |f(z)| = M \right\}$$

Suppose that $\mathcal{E} = D$. This means that $|f(z)| = M$ for all points $z \in D$ and, by Lemma 17.1, it follows that $f(z) = \text{const}$ on D that contradicts with the assumptions of the theorem. Suppose now that $\mathcal{E} \subset D$, namely, there exists a boundary point $z_0 \in \mathcal{E}$ such that it is an internal point of D. Let us construct a circuit $C = B(z_0, r)$ which contains a point $z_1 \in D$ such that $z_1 \notin \mathcal{E}$ (this can always be done since z_0 is a boundary point). Then $|f(z)| < M$, and for any small enough $\varepsilon > 0$, by continuity of $f(z)$, there exists a set C_1, which is a part of C where $|f(z)| < M - \varepsilon$. Denote $C_2 := C \setminus C_1$. Evidently that for any $z \in C_2$ one has $|f(z)| \leq M$. Then by Theorem 17.4 it follows that

$$f(z_0) = \frac{1}{2\pi} \int_{=0}^{2\pi} f(z)\, d\varphi \overset{ds = r d\varphi}{=} \frac{1}{2\pi r} \left(\int_{C_1} f(z)\, dz + \int_{C_2} f(z)\, dz \right)$$

which implies

$$|f(z_0)| = M \leq \frac{1}{2\pi r} \left([M - \varepsilon] l_1 + M l_2 \right) = M - \frac{\varepsilon l_1}{2\pi r}$$

$$l_1 := \int_{C_1} dz, \quad l_2 := \int_{C_2} f(z)\, dz$$

But the last inequality is impossible which leads to the contradiction. Theorem is proven. □

Corollary 17.6. (Minimum modulus principle) *If a function $f(z)$ is analytical and not constant on D, and it is continuous and nonequal to zero on $\bar{D}$, then the minimum of $|f(z)|$ cannot be achieved on D.*

Proof. It can be easily done if we apply Theorem 17.10 to the function $g(z) = 1/f(z)$. □

Using the maximum modulus principle it is possibly easy to state the following useful result.

Lemma 17.5. (Schwartz, around 1875) *If function $f(z)$ is analytical in the open domain $|z| < 1$, it is continuous on $|z| \leq 1$, and, in the addition,*

$$f(0) = 0, \quad |f(z)| \leq 1$$

then

$$\boxed{|f(z)| \leq |z|} \tag{17.39}$$

If at least in one internal point of the domain $|z| < 1$ the exact equality $|f(z)| = |z|$ holds, then this equality takes place at any point of this domain and, besides,

$$\boxed{f(z) = e^{i\alpha} z} \tag{17.40}$$

where α is a real constant.

Proof. To prove this result it is sufficient to consider the function

$$\varphi(z) := \begin{cases} \dfrac{f(z)}{z} & \text{if} \quad z \neq 0 \\ f'(0) & \text{if} \quad z = 0 \end{cases}$$

which is analytical on the set $0 < |z| < 1$ and continuous in $|z| \leq 1$. Applying to this function the maximum modulus principle 17.10 we derive that on the circle $|z| = 1$ we have

$$|\varphi(z)| = \left| \frac{f(z)}{z} \right| \leq 1$$

and by this principle, $|\varphi(z)| \leq 1$ everywhere on $|z| \leq 1$, which gives $|f(z)| \leq |z|$. So, the first part of the lemma is proven. If in some internal point z_0 we have $|f(z_0)| = |z_0|$, then in this point $|\varphi(z)| = 1$ and, again by maximum modulus principle 17.10, it follows that $|\varphi(z)| \equiv 1$ everywhere on $|z| \leq 1$. By Lemma 17.1 we have that $\varphi(z) = \text{const}$ which can be represented as $e^{i\alpha}$ which implies (17.40). □

17.2.7 Calculation of integrals and Jordan lemma

17.2.7.1 Real integral calculation using the Cauchy's residue theorem

The main idea of integral calculus using Cauchy's residue theorem consists of the following. Assume we must calculate the usual (Riemann) integral $\int_{x=a}^{b} f(x)\,dx$ of the real function $f(x)$ over the given interval (finite or infinite) $(a, b) \in \mathbb{R}$. Let us complete this interval with some curve C' which together with (a, b) contains a domain $\mathcal{D}$. Let us then extend (analytically) our given function $f(x)$ up to a function $f(z)$ defined on $\bar{\mathcal{D}}$. Hence, by Cauchy's residue theorem (17.27)

$$\oint_{C' \cup (a,b)} f(z)\,dz = \int_{x=a}^{b} f(x)\,dx + \oint_{C'} f(z)\,dz = 2\pi i \sum_{k=1}^{n} \operatorname{res} f(a_k)$$

which gives

$$\boxed{\int_{x=a}^{b} f(x)\,dx = 2\pi i \sum_{k=1}^{n} \operatorname{res} f(a_k) - \oint_{C'} f(z)\,dz} \tag{17.41}$$

If the integral over C' can be calculated or expressed as a function of the integral $\int_{x=a}^{b} f(x)\,dx$, then the problem of the integral calculus might be solved! This technique clearly shows that the integral of a real function $f(x)$ can be calculated as the sum of residues in its singular points that is significantly simpler especially in the case of poles.

Remark 17.6. *Usually to simplify calculations, the extended function $f(z)$ is selected in such a manner that on (a, b) it would be its real or imaginary part that permits to calculate $\int_{x=a}^{b} f(x)\,dx$ by simple separation of real and imaginary parts.*

17.2.7.2 Improper integrals and Jordan lemma

If the interval is infinite, then one may be considered an extended family of the contours $C'_k \cup (a_k, b_k) \subset C'_{k+1} \cup (a_{k+1}, b_{k+1})$ such that $(a_k, b_k) \to (a, b)$ as $k \to \infty$. In this case it is not obligatory to calculate the integral $\oint_{C'_k} f(z)\,dz$ but it is sufficient only to find its limit. Very often it turns out that this limit is equal to zero. This fact may be shown using the lemma given below.

Lemma 17.6. (Jordan) *If on some sequence $\{C'_k\}$ of contours cuts*

$$C'_k := \left\{ z \in \mathbb{C} \mid |z| = R_k,\ Im\, z > -a,\ R_k \underset{k\to\infty}{\to} \infty,\ a \text{ is fixed} \right\}$$

the function $g(z)$ tends to zero uniformly on $\arg z$, *then for any $\lambda > 0$*

$$\boxed{\lim_{k\to\infty} \oint_{C'_k} g(z)\, e^{i\lambda z}\, dz = 0} \tag{17.42}$$

Proof. Denote $z = x + iy = re^{i\varphi}$, $M_k := \max_{C'_k} |g(z)|$ and $\alpha_k := \arcsin \frac{a}{R_k}$. By the lemma assumption, $M_k \underset{k\to\infty}{\to} 0$ and $\alpha_k \underset{k\to\infty}{\to} 0$ such that $\alpha_k R_k \underset{k\to\infty}{\to} a$. Let $a > 0$ (see Fig. 17.4).

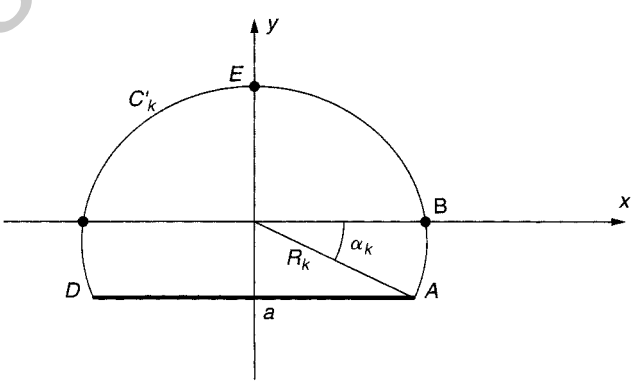

Fig. 17.4. The contour C'_k with cuts AB and CD.

On the cuts AB and CD we have $\left|e^{i\lambda z}\right| = e^{-\lambda y} \le e^{\alpha\lambda}$, therefore

$$\left| \int\limits_{AB\cup CD} g(z)\, e^{i\lambda z}\, dz \right| \le M_k e^{\alpha\lambda} \alpha_k R_k \underset{k\to\infty}{\to} 0$$

Applying the inequality $\sin\varphi \ge \frac{2}{\pi}\varphi$ valid[3] for $\varphi \in [0, 2\pi]$ we get

$$\left| e^{i\lambda z} \right| = e^{-\lambda R_k \sin\varphi} \le e^{-\frac{2\lambda R_k}{\pi}\varphi}$$

at the cut BE which implies

$$\left| \int\limits_{BE} g(z)\, e^{i\lambda z} dz \right| \le M_k R_k \int\limits_{\varphi=0}^{\pi/2} e^{-\frac{2\lambda R_k}{\pi}\varphi} d\varphi = M_k \frac{\pi}{2\lambda}\left(1 - e^{-\lambda R_k}\right) \underset{k\to\infty}{\to} 0$$

Analogously, $\left|\int_{EC} g(z)\, e^{i\lambda z}\, dz\right| \underset{k\to\infty}{\to} 0$ which proves the lemma for the case $a > 0$. If $a < 0$, the proof is significantly simpler since there does not need to calculate integrals over the cuts AB and CD. Lemma is proven. □

Example 17.9. *Let us calculate the, so-called,* ***Laplace integral*** $\int_{t=0}^{\infty} \frac{\cos t}{t^2 + b^2} dt$. *Select the auxiliary function* $f(z) = \frac{e^{iz}}{z^2 + b^2}$ *and the contour* C'_{R_k} *as in Fig. 17.4 with* $a = 0$. *Since the function* $g(z) := \frac{1}{z^2 + b^2}$ *satisfies on* C'_{R_k} *the inequality* $|g(z)| < \frac{1}{R_k^2 - b^2}$, *then it converges uniformly to zero as* $R_k \to \infty$ *and hence, by the Jordan lemma (17.6),* $\int_{C'_{R_k}} f(z)\, dz = \int_{C'_{R_k}} g(z)\, e^{iz}\, dz \underset{k\to\infty}{\to} 0$. *Then, for any* $R_k > |b|$ *by Cauchy's residue theorem (17.27) it follows that*

$$\int\limits_{t=-R_k}^{R_k} \frac{e^{it}}{t^2 + b^2} dt + \int\limits_{C'_{R_k}} f(z)\, dz = 2\pi i \frac{e^{-|b|}}{2|b|}$$

---

[3] To prove this inequality it is sufficient to notice that $\left(\frac{\sin\varphi}{\varphi}\right)' = \frac{\cos\varphi}{\varphi^2}(\varphi - \tan\varphi) < 0$ at $(0, 2\pi)$ and, hence, the function $\frac{\sin\varphi}{\varphi}$ decreases at this interval.

since $f(z)$ inside the joint contour has the unique singular point (pole of the multiplicity one) $z = |b|\,i$. Separating the real and imaginary parts and using that the function $f(z)$ is even, we finally conclude that

$$\boxed{\int_{t=0}^{\infty} \frac{\cos t}{t^2 + b^2} dt = \pi \frac{e^{-|b|}}{|b|}} \tag{17.43}$$

17.3 Series expansions

In this section we will consider the problem of the representation of analytical functions by power-series expansions and their generalizations ("negative power"). More exactly, we will deal with a series given by

$$\sum_{n=-\infty}^{\infty} c_n (z-a)^n = \ldots + c_{-n}(z-a)^{-n} \ldots + c_{-1}(z-a)^{-1} + c_0 + c_1(z-a) + \ldots + c_n (z-a)^n + \ldots \tag{17.44}$$

where z is a complex variable, and c_n and a are constants named coefficients and the center of the series, respectively.

17.3.1 Taylor (power) series

Theorem 17.11. (O. Cauchy, 1831) *A function $f(z)$ can be represented by the corresponding* ***Taylor series***

$$\boxed{\begin{aligned} &f(z) = f(a) + \frac{f'(a)}{1!}(z-a) + \ldots + \frac{f^{(n)}(a)}{n!}(z-a)^n + R_n \\ &f^{(n)}(a) \quad \text{are given by (17.35)} \\ &R_n = \frac{(z-a)^{n+1}}{2\pi i} \oint_C \frac{f(w)}{(w-z)(w-a)^{n+1}} dw \end{aligned}} \tag{17.45}$$

in any open domain circle with a boundary $C = B(a, r)$ where this function is analytical. In any closed domain $\bar{\mathcal{R}}$, belonging to this circle, this Taylor series converges uniformly, that is, $R_n \to 0$ when $n \to \infty$ independently of $z \in \bar{\mathcal{R}}$.

Proof. Let us use the known formula of the geometric progression

$$\frac{1 - q^{n+1}}{1 - q} = 1 + q + q^2 + \ldots + q^n$$

valid not only for real, but for complex variables $q \in \mathbb{C}$ $(q \neq 1)$, rewriting it as

$$\frac{1}{1-q} = 1 + q + q^2 + \ldots + q^n + \frac{q^{n+1}}{1-q} \tag{17.46}$$

Fixing some point $a \in D$ (D is the open domain where $f(z)$ is analytical) and using (17.46) we may write

$$\frac{1}{(w-z)} = \frac{1}{(w-a)}\left[\frac{1}{1-\dfrac{z-a}{w-a}}\right]$$

$$= \frac{1}{(w-a)}\left[1+\frac{z-a}{w-a}+\ldots\left(\frac{z-a}{w-a}\right)^{n}+\frac{1}{1-\dfrac{z-a}{w-a}}\left(\frac{z-a}{w-a}\right)^{n+1}\right]$$

Multiplying both sides by $\dfrac{1}{2\pi i} f(w)$ and integrating along the contour $C = B(a, r)$, lying in D and containing both points z and a, and applying Cauchy's formula (17.28), we obtain (17.45). Let us now consider any positive r' such that $0 < r' < r$, where r is the radius of the circle $C = B(a, r)$, and the circle $|z-a| \le kr'$ with any k satisfying $0 < k < 1$. Let z belong to this last circle and $C' = B(a, r')$. Then $|w-a| = r'$, and hence,

$$|w-z| = |(w-a)+(a-z)|$$
$$\ge |w-a| - |z-a| \ge r' - kr' = (1-k)r'$$

Applying this inequality to (17.45) we have

$$|R_n| = \left|\frac{(z-a)^{n+1}}{2\pi i}\oint_C \frac{f(w)}{(w-z)(w-a)^{n+1}}dw\right|$$

$$\le \frac{k^{n+1}(r')^{n+1}}{2\pi}\cdot\frac{M2\pi r'}{(1-k)(r')^{n+2}} = \frac{Mk^{n+1}}{1-k}$$

where $M = \sup_{z:|z-a|\le r'} |f(z)|$ (the function $f(z)$ is analytical within this circle and, hence, it is bounded). Since $k < 1$, we obtain $R_n \to 0$ when $n \to \infty$ for every z satisfying $|z-a| \le kr'$. Theorem is proven. □

Claim 17.1. (The Cauchy–Adhamar formula) *Every power (Taylor) series has a definite radius of convergence*

$$R := \left\{ z \in \mathbb{C} \mid |z-a| < R, \quad \left|\sum_{n=0}^{\infty} c_n (z-a)^n\right| < \infty \right\}$$

which is finite or $+\infty$ *and may be calculated as*

$$\boxed{R = \frac{1}{\limsup\limits_{n\to\infty}\sqrt[n]{|c_n|}}} \tag{17.47}$$

Proof. To prove this result it is sufficient to show that for any z, for which $|z-a| \leq kR$, $0 < k < 1$, the series $\sum_{n=0}^{\infty} c_n (z-a)^n$ converges, and for any z, for which $|z-a| > R$, this series disconverges. By the upper limit definition, for any $\varepsilon > 0$ there exists $n_0(\varepsilon)$ such that $\sqrt[n]{|c_n|} < \dfrac{1}{R} + \varepsilon$ for all $n \geq n_0(\varepsilon)$. Selecting ε such that $\dfrac{1}{R} + \varepsilon < \dfrac{1}{R\left(\dfrac{k+1}{2}\right)}$,

we obtain

$$|c_n (z-a)^n| < \frac{k^n R^n}{R^n \left(\dfrac{k+1}{2}\right)^n} = \left(\frac{2k}{k+1}\right)^n = q^n$$

$$0 < q = 1 - \frac{1-k}{1+k} < 1$$

for any $n \geq n_0(\varepsilon)$ and z satisfying $|z-a| \leq kR$. So, we get

$$\left| \sum_{n=0}^{\infty} c_n (z-a)^n \right| \leq \sum_{n=0}^{\infty} |c_n (z-a)^n| \leq \sum_{n=0}^{\infty} q^n = \frac{1}{1-q} < \infty$$

To prove the rest of the theorem, again notice that by the definition of the upper limit, for any $\varepsilon > 0$ there exists a subsequence $n = n_k$ such that $\sqrt[n_k]{|c_{n_k}|} > \dfrac{1}{R} - \varepsilon$, or, equivalently,

$$\left| c_{n_k} (z-a)^{n_k} \right| > \left[\left(\frac{1}{R} - \varepsilon \right) |z-a| \right]^{n_k}$$

But if $|z-a| > R$ we can always select ε such that $\left(\dfrac{1}{R} - \varepsilon\right)|z-a| > 1$. This means that the term $c_{n_k} (z-a)^{n_k}$ will tend to ∞ and, hence, $\sum_{n=0}^{\infty} c_n (z-a)^n$ disconverges. Claim is proven. □

Example 17.10.

(a) The following series converge for any $z \in \mathbb{C}$

$$e^z = 1 + z + \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \ldots;$$

$$\sin z = 1 - \frac{1}{3!}z^3 + \frac{1}{5!}z^5 - \ldots; \quad \cos z = 1 - \frac{1}{2!}z^2 + \frac{1}{4!}z^4 - \ldots$$

$$\sinh z = z + \frac{1}{3!}z^3 + \frac{1}{5!}z^5 + \ldots; \quad \cosh z = 1 + \frac{1}{2!}z^2 + \frac{1}{4!}z^4 + \ldots$$

(b) The following series converge for any z such that $|z| < 1$

$$\ln(1+z) = z - \frac{1}{2!}z^2 + \frac{1}{3!}z^3 + \dots$$

$$(1+z)^a = 1 + az + \frac{a(a-1)}{2!}z^2 + \frac{a(a-1)(a-2)}{3!}z^3 + \dots$$

Remark 17.7. *If $f(z)$ is regular at $z = a$ and $f(a) = 0$ ($f(z)$ is not zero identically), then by Taylor's theorem it may be presented by the power series (17.44) as follows*

$$\boxed{f(z) = \sum_{n=N(f,a)}^{\infty} c_n (z-a)^n = c_N (z-a)^N + \dots} \tag{17.48}$$

where $c_N \neq 0$ and $N(f,a) \geq 1$. The number $N(f,a)$, appearing in (17.48), is called ***the order of the zero*** *of $f(z)$ at $z = a$.*

Lemma 17.7. (Parseval's identity) *For any function analytical on the disk $|z-a| \leq R$ and any $r \in [0, R)$ the following identity holds*

$$\boxed{\frac{1}{2\pi}\int_{\theta=0}^{2\pi} \left|f\left(a + re^{i\theta}\right)\right|^2 d\theta = \sum_{n=0}^{\infty} |c_n|^2 r^{2n}} \tag{17.49}$$

Proof. By the direct calculation of the circuit integral of $|f(z)|^2$ over the $|z-a| = r$ using the Taylor expansion $f(z) = \sum_{n=0}^{\infty} c_n (z-a)^n$ gives (17.49). □

17.3.2 Laurent series

Suppose $f(z)$ is regular in the annulus K defined by $r < |z-a| < R$, $0 \leq r < R \leq \infty$. We construct the annular domains K' and K'' defined by $r' < |z-a| < R'$ and $r'' < |z-a| < R''$ where $r < r' < r'' < R'' < R' < R$ so that K contains $\bar{K}'$ and K' contains $\bar{K}''$ (see Fig. 17.5).

As $f(z)$ is regular on $\bar{K}'$, by the Cauchy integral formula (17.30) it can be represented as

$$\boxed{\begin{aligned} &f(z) = f_1(z) + f_2(z) \\ &f_1(z) = \frac{1}{2\pi i}\oint_{C_{R'}} \frac{f(\zeta)}{\zeta - z} d\zeta, \quad f_2(z) = -\frac{1}{2\pi i}\oint_{C_{r'}} \frac{f(\zeta)}{\zeta - z} d\zeta \end{aligned}} \tag{17.50}$$

where $C_{R'}$ and $C_{r'}$ denote circles of the respective radii R' and r' with centers at the point a (see Fig. 17.5). Then for all points ζ on $C_{R'}$ it follows that $\left|\dfrac{z-a}{\zeta-a}\right| < \dfrac{R''}{R'} := q_1 < 1$. So, the fraction $1/(\zeta - z)$ can be expended as a geometric series uniformly convergent on ζ on $C_{R'}$, that is,

$$\frac{1}{(\zeta - z)} = \left(\frac{1}{\zeta - z}\right) / \left(1 - \frac{z-a}{\zeta - a}\right)$$

$$= \frac{1}{\zeta - z} + \frac{z-a}{(\zeta - z)^2} + \ldots + \frac{(z-a)^n}{(\zeta - z)^{n+1}} + \ldots$$

Substitution of this expansion in (17.50) gives

$$\boxed{{}_1(z) = \frac{1}{2\pi i} \oint_{C_{R'}} \frac{f(\zeta)}{\zeta - z} d\zeta = \sum_{n=0}^{\infty} c_n (z-a)^n} \tag{17.51}$$

where

$$\boxed{c_n = \frac{1}{2\pi i} \oint_{C_{R'}} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta} \tag{17.52}$$

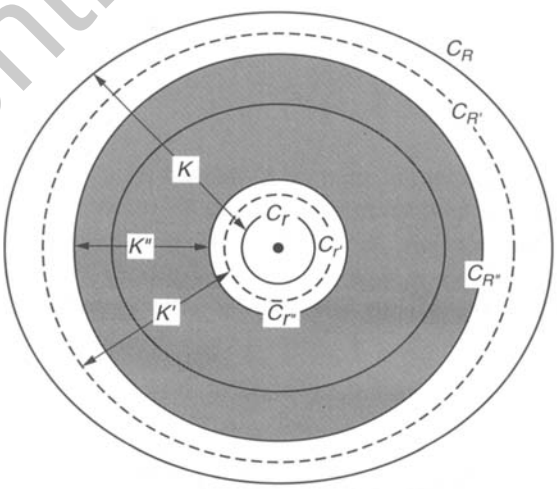

Fig. 17.5. The annular domains.

Notice that, in general, c_n cannot be represented in the form $\dfrac{f^{(n)}(a)}{n!}$ since $f(z)$ is not regular at $z = a$. Analogously, for any ζ on $C_{r'}$ we have $\left|\dfrac{\zeta - a}{z - a}\right| < \dfrac{r'}{r''} := q_2 < 1$. Hence,

$$\frac{1}{(\zeta - z)} = -\left(\frac{1}{z - a}\right) / \left(1 - \frac{\zeta - a}{z - a}\right)$$

$$= -\frac{1}{z - a} - \frac{\zeta - a}{(z - a)^2} - \ldots - \frac{(\zeta - a)^{n-1}}{(z - a)^n} - \ldots$$

and, as the result,

$$\boxed{f_2(z) = -\frac{1}{2\pi i}\oint_{C_{r'}} \frac{f(\zeta)}{\zeta - z} d\zeta = \sum_{n=0}^{\infty} c_{-n}(z - a)^{-n}} \tag{17.53}$$

with

$$\boxed{c_{-n} = \frac{1}{2\pi i}\oint_{C_{r'}} f(\zeta)(\zeta - a)^{n-1} d\zeta} \tag{17.54}$$

Combining (17.52) and (17.54) in (17.50) we obtain the following result.

Theorem 17.12. (Laurent, 1843) *Every function $f(z)$ which is regular in the annulus $K := \{z \in C \mid r < |z - a| < R\}$ can be represented in this annulus by its* ***Laurent series***

$$\boxed{\begin{aligned} f(z) &= \sum_{n=-\infty}^{\infty} c_n (z - a)^n = f_1(z) + f_2(z) \\ f_1(z) &:= \sum_{n=0}^{\infty} c_n (z - a)^n, \quad f_2(z) := \sum_{n=-1}^{-\infty} c_n (z - a)^n \\ c_n &= \frac{1}{2\pi i}\oint_{C_{R'}} \frac{f(\zeta)}{(\zeta - z)^{n+1}} d\zeta \quad (n = 0, \pm 1, \pm 2, \ldots) \end{aligned}} \tag{17.55}$$

The term $f_1(z)$ is called the ***regular part*** *of the Laurent series and the term $f_2(z)$ is called the* ***principal part*** *of the Laurent series, respectively.*

Corollary 17.7. ***Cauchy's inequalities for the Laurent series*** *are as*

$$\boxed{|c_n| < M/\rho^n} \tag{17.56}$$

if the function $f(z)$ is bounded on the circle $|z - a| = \rho \in (r, R)$, i.e., $|f(z)| \leq M$.

Example 17.11. *The function*

$$f(z) = \frac{1}{(z-1)(z-2)} = \frac{1}{z-2} - \frac{1}{z-1}$$

is regular in the annulus ("rings")

$$K_1 := \{z \in C \mid 0 \leq |z| < 1\}$$
$$K_2 := \{z \in C \mid 1 < |z| < 2\}$$
$$K_3 := \{z \in C \mid 2 < |z|\}$$

So, in K_1

$$\frac{1}{z-2} = -\frac{1}{2}\left(\frac{1}{1-z/2}\right) = -\frac{1}{2}\left(1 + \frac{z}{2} + \frac{z^2}{4} + \ldots\right)$$
$$\frac{1}{z-1} = -\frac{1}{1-z} = -\left(1 + z + z^2 \ldots\right)$$
$$f(z) = -\frac{1}{2}\left(1 + \frac{z}{2} + \frac{z^2}{4} + \ldots\right) + \left(1 + z + z^2 \ldots\right)$$
$$= \frac{1}{2} + \frac{3}{4}z + \frac{7}{8}z^2 + \ldots$$

In K_2

$$\frac{1}{z-2} = -\frac{1}{2}\left(\frac{1}{1-z/2}\right) = -\frac{1}{2}\left(1 + \frac{z}{2} + \frac{z^2}{4} + \ldots\right)$$
$$\frac{1}{z-1} = \frac{1}{z}\frac{1}{1-1/z} = \frac{1}{z}\left(1 + z^{-1} + z^{-2} \ldots\right)$$
$$f(z) = -\frac{1}{2}\left(1 + \frac{z}{2} + \frac{z^2}{4} + \ldots\right) - \left(z^{-1} + z^{-2} \ldots\right)$$

In K_3

$$\frac{1}{z-2} = \frac{1}{z}\left(\frac{1}{1-2/z}\right) = \frac{1}{z}\left(1 + \frac{2}{z} + \frac{4}{z^2} + \ldots\right)$$
$$\frac{1}{z-1} = \frac{1}{z}\frac{1}{1-1/z} = \frac{1}{z}\left(1 + z^{-1} + z^{-2} \ldots\right)$$
$$f(z) = \left(\frac{1}{z} + \frac{2}{z^2} + \frac{4}{z^3} + \ldots\right) - \left(z^{-1} + z^{-2} \ldots\right) = z^{-2} + 3z^{-3} + \ldots$$

Example 17.12. *Let us calculate* ***Euler's integral*** $\int_{x=-\infty}^{\infty} \frac{\sin x}{x} dx$. *Evidently,*

$$I_E = \int_{x=0}^{\infty} \frac{\sin x}{x} dt + \int_{x=-\infty}^{0} \frac{\sin t}{x} dt = 2\int_{x=0}^{\infty} \frac{\sin x}{x} dx$$

Introduce the auxiliary function $f(z) = \dfrac{e^{iz}}{z}$ *and select the contour*

$$C := C_r \cup [r, R] \cup C_R \cup [-R, -r]$$

as it is shown in Fig. 17.6. Within this contour the function $f(z)$ *is regular and, hence, by Cauchy's residue theorem (17.27) it follows that*

$$\begin{aligned} 0 &= \oint_C f(z)\,dz \\ &= \int_{C_r} f(z)\,dz + \int_r^R f(z)\,dz + \int_{C_R} f(z)\,dz + \int_{-R}^{-r} f(z)\,dz \end{aligned} \tag{17.57}$$

By Jordan's lemma (17.6) $\lim\limits_{R\to\infty} \int_{C_R} f(z)\,dz = 0$. *To estimate* $\int_{C_r} f(z)\,dz$ *let us consider the Laurent expansion (17.55) of* $f(z)$ *in the neighborhood of the point* $z = 0$*:*

$$f(z) = \frac{1 + iz + \dfrac{(iz)^2}{2} + \dots}{z} = \frac{1}{z} + P(z)$$

where $P(z)$ *is a function continuous in* $z = 0$. *Thus, using representation* $z = re^{i\varphi}$, *we also have*

$$\begin{aligned} \int_{C_r} f(z)\,dz &= \int_{C_r} \frac{e^{iz}}{z}\,dz \\ &= \int_{C_r} \frac{1}{z}\,dz + \int_{C_r} P(z)\,dz \\ &= \int_{\varphi=\pi}^{0} \frac{1}{re^{i\varphi}} \left(re^{i\varphi} i\,d\varphi\right) + O(r) \\ &= -i\pi + O(r) \end{aligned}$$

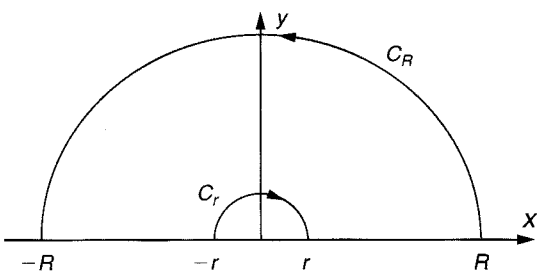

Fig. 17.6. The contour C.

So, (17.57) can be rewritten as

$$0 = -i\pi + \int_{r}^{R} f(z)\,dz + \int_{-R}^{-r} f(z)\,dz + O\left(\frac{1}{R}\right) + O(r) \tag{17.58}$$

where, in view of $z = x + iy$,

$$\int_{r}^{R} f(z)\,dz + \int_{-R}^{-r} f(z)\,dz = \int_{r}^{R} \frac{e^{ix}}{x}dx + \int_{-R}^{-r} \frac{e^{ix}}{x}dx$$

$$= \int_{r}^{R} \frac{e^{ix} - e^{-ix}}{x}dx = \int_{r}^{R} 2i\frac{\sin x}{x}dx$$

Taking $R \to \infty$ *and* $r \to 0$ *from (17.58) we obtain* $\int_0^\infty \frac{\sin x}{x}dx = \frac{\pi}{2}$ *and, finally*

$$\boxed{\int_{x=-\infty}^{\infty} \frac{\sin x}{x}dx = \pi} \tag{17.59}$$

17.3.3 Fourier series

Let function $f(z)$ be analytical in annulus

$$K := \{z \in C \mid 1 - \varepsilon < |z| < 1 + \varepsilon\}$$

Thus within this annulus it may be represented by the Laurent expansion (17.55)

$$\boxed{\begin{aligned} f(z) &= \sum_{n=-\infty}^{\infty} c_n z^n \\ c_n &= \frac{1}{2\pi i}\oint_{|z|=1} \frac{f(\zeta)}{\zeta^{n+1}}d\zeta = \frac{1}{2\pi}\int_{\theta=0}^{2\pi} f\left(e^{i\theta}\right)e^{-in\theta}\,d\theta \end{aligned}} \tag{17.60}$$

In particular, for the points $z = e^{it}$ of the unitary circle we obtain

$$\boxed{\begin{aligned} \varphi(t) := f\left(e^{it}\right) &= \sum_{n=-\infty}^{\infty} c_n e^{int} = c_0 + \sum_{n=1}^{\infty}\left(c_n e^{int} + c_{-n}e^{-int}\right) \\ &= \frac{a_0}{2} + \sum_{n=1}^{\infty}\left[a_n \cos(nt) + b_n \sin(nt)\right] \end{aligned}} \tag{17.61}$$

where $a_0 := 2c_0$, $a_n := c_n + c_{-n}$, $b_n := i\,(c_n - c_{-n})$, and, hence, by (17.60),

$$\boxed{\begin{gathered} a_0 = \frac{1}{\pi}\int\limits_{\theta=0}^{2\pi} f\left(e^{i\theta}\right) d\theta,\ a_n = \frac{1}{\pi}\int\limits_{\theta=0}^{2\pi} f\left(e^{i\theta}\right)\cos(nt)\, d\theta \\ b_n = \frac{1}{\pi}\int\limits_{\theta=0}^{2\pi} f\left(e^{i\theta}\right)\sin(nt)\, d\theta \end{gathered}} \tag{17.62}$$

The series (17.60) is known as the *Fourier series* of the function $\varphi(t)$ written in complex form.

17.3.4 Principle of argument

Theorem 17.13. (The principle of the argument) *Let $\mathcal{D}$ be the interior domain bounded by a contour C and $f(z)$ be a function having a finite number of multi-poles b_1, $b_2, \ldots, b_P$ with the respective orders $p_1, p_2, \ldots, p_P$ and a finite number of multi-zeros $a_1, a_2, \ldots, a_N$ with the respective orders $n_1, n_2, \ldots, n_N$. Then the logarithmic derivative $f'(z)/f(z)$ is regular on C, and has in $\mathcal{D}$ at most a finite number of singularities such that the following identity holds*

$$\boxed{\begin{gathered} \oint\limits_C \frac{f'(z)}{f(z)} dz = 2\pi i\left(N_f - P_f\right) = i\,\Delta_C \arg f(z) \\ N_f := n_1 + n_2 + \ldots + n_N,\ P_f := p_1 + p_2 + \ldots + p_P \\ \Delta_C \arg f(z) := \frac{1}{2\pi}\left(N_f - P_f\right) \end{gathered}} \tag{17.63}$$

Proof. Considering the multi-connected domain $\mathcal{D}_{\varepsilon,\mu}$ (with the "joint" boundary $C_{\varepsilon,\mu}$), obtained from $\mathcal{D}$ by excluding (deleting) all singularity points (in this case zeros and poles), we conclude that the logarithmic derivative $f'(z)/f(z)$ is regular on $\mathcal{D}_{\varepsilon,\mu}$, and, hence, by (17.23) the integral $\oint_C \dfrac{f'(z)}{f(z)} dz$ can be represented as a finite sum of the individual integrals taken over the contours

$$\begin{aligned} C_k^{zero} &:= \{z \in C \mid |z - a_k| = \varepsilon_k > 0\},\ k = 1, \ldots, N \\ C_s^{pole} &:= \{z \in C \mid |z - b_s| = \mu_s > 0\},\ s = 1, \ldots, P \end{aligned}$$

Indeed, by (17.23)

$$\begin{aligned} 0 &= \oint\limits_{C_{\varepsilon,\mu}} \frac{f'(z)}{f(z)} dz \\ &= \oint\limits_C \frac{f'(z)}{f(z)} dz - \sum_{k=1}^{N} \oint\limits_{C_k^{zero}} \frac{f'(z)}{f(z)} dz - \sum_{s=1}^{P} \oint\limits_{C_s^{pole}} \frac{f'(z)}{f(z)} dz \end{aligned}$$

which implies

$$\oint_C \frac{f'(z)}{f(z)}dz = \sum_{k=1}^{N} \oint_{C_k^{zero}} \frac{f'(z)}{f(z)}dz + \sum_{s=1}^{P} \oint_{C_s^{pole}} \frac{f'(z)}{f(z)}dz \tag{17.64}$$

So, it is sufficient to consider the individual integrals $\oint_{C_k^{zero}} \frac{f'(z)}{f(z)}dz$ and $\oint_{C_s^{pole}} \frac{f'(z)}{f(z)}dz$ for each fixed k and s. For the zero a_k by (17.48) it follows that

$$f(z) = c_{n_k}(z-a)^{n_k} + c_{n_k+1}(z-a)^{n_k+1} + \ldots \qquad (c_{n_k} \neq 0)$$

and hence,

$$\begin{aligned} f'(z) &= n_k c_{n_k}(z-a)^{n_k-1} + (n_k+1)c_{n_k+1}(z-a)^{n_k} + \ldots \\ \frac{f'(z)}{f(z)} &= \frac{n_k c_{n_k}(z-a)^{n_k-1} + (n_k+1)c_{n_k+1}(z-a)^{n_k} + \ldots}{c_{n_k}(z-a)^{n_k} + c_{n_k+1}(z-a)^{n_k+1} + \ldots} \\ &= \frac{1}{z-a}\frac{n_k c_{n_k} + (n_k+1)c_{n_k+1}(z-a) + \ldots}{c_{n_k} + c_{n_k+1}(z-a) + \ldots} \\ &= \frac{1}{z-a}\left[n_k + \tilde{c}_0(z-a) + \tilde{c}_1(z-a)^2 + \ldots\right] \\ &= \frac{n_k}{z-a} + \tilde{c}_0 + \tilde{c}_1(z-a) + \ldots \end{aligned}$$

in some neighborhood of the point a. Thus, the logarithmic residue of a regular function at a zero is equal to the order n_k of that zero. Analogously, for the pole b_k in some of its deleted neighborhood we have

$$f(z) = \frac{c_{-p_k}}{(z-a)^{p_k}} + \frac{c_{-p_k+1}}{(z-a)^{p_k-1}} + \ldots \qquad (c_{p_k} \neq 0)$$

and, hence,

$$\begin{aligned} f'(z) &= -\frac{p_k c_{-p_k}}{(z-a)^{p_k+1}} - (p_k-1)\frac{c_{-p_k+1}}{(z-a)^{p_k}} + \ldots \\ \frac{f'(z)}{f(z)} &= -\frac{1}{z-b_k}\frac{p_k c_{-p_k} + (p_k-1)c_{-p_k+1}(z-b_k) + \ldots}{c_{-p_k} + c_{-p_k+1}(z-b_k) + \ldots} \\ &= -\frac{p_k}{z-a} + \check{c}_0(z-b_k) + \check{c}_1(z-b_k)^2 + \ldots \end{aligned}$$

Thus, the logarithmic residue of $f(z)$ at a pole is equal to the order p_k of that pole with the sign reversed. Combining these two logarithmic residues in (17.64) we derive

the relation $\oint_C \frac{f'(z)}{f(z)} dz = 2\pi i \left(N_f - P_f\right)$. To complete the proof it is sufficient to notice that $\oint_C \frac{f'(z)}{f(z)} dz = \Delta_C \ln f(z)$ where $\ln f(z)$ denotes a value of the logarithm which varies continuously as z makes one complete circuit of C (starting from some fixed point, z_0, say) and $\Delta_C \ln f(z)$ denotes the corresponding variation of $\ln f(z)$. Since

$$\Delta_C \ln f(z) = \Delta_C \ln |f(z)| + i \Delta_C \arg f(z)$$

$$\Delta_C \ln |f(z)| = \ln |f(z_0)| - \ln |f(z_0)| = 0$$

we have $\oint_C \frac{f'(z)}{f(z)} dz = i \Delta_C \arg f(z)$ which completes the proof. □

17.3.5 Rouché theorem

One of the important applications of the principle of the argument is the following theorem.

Theorem 17.14. (Rouché) *Let $\mathcal{D}$ be the interior domain bounded by a contour C. If functions $f(z)$ and $g(z)$ are analytical on $\mathcal{D}$, continuous on $\bar{\mathcal{D}}$ and satisfy the inequality*

$$\boxed{|f(z)| > |g(z)|} \tag{17.65}$$

at each point z on C, then the functions $f(z)$ and $[f(z) + g(z)]$ have the same number of zeros in $\mathcal{D}$, each zero being counted according to its multiplicity.

Proof. Notice that by the assumption of this theorem $|f(z)| > 0$ on C and

$$|f(z) + g(z)| \geq |f(z)| - |g(z)| > 0$$

Hence, the functions $f(z)$ and $[f(z) + g(z)]$ have no zeros on C and the principle of the arguments is applicable to both. Based on the identity

$$\arg [f(z) + g(z)] = \arg f(z) + \arg \left(1 + \frac{g(z)}{f(z)}\right)$$

we derive

$$\Delta_C \arg [f(z) + g(z)] = \Delta_C \arg f(z) + \Delta_C \arg \left(1 + \frac{g(z)}{f(z)}\right)$$

But $\Delta_C \arg\left(1+\dfrac{g(z)}{f(z)}\right)=0$ since the point $w=1+\dfrac{g(z)}{f(z)}$ always remains inside the circle $|w-1|<1$. So, $\Delta_C \arg[f(z)+g(z)]=\Delta_C \arg f(z)$ and to complete the proof one needs to apply the formula (17.63), namely, since $P_f=P_{f+g}=0$ we have

$$N_f=N_f-P_f=\frac{1}{2\pi i}\oint\limits_C \frac{f'(z)}{f(z)}dz$$

$$=\frac{1}{2\pi}\Delta_C \arg f(z)$$

$$=\frac{1}{2\pi}\Delta_C \arg[f(z)+g(z)]$$

$$=\frac{1}{2\pi i}\oint\limits_C \frac{f'(z)+g'(z)}{f(z)+g(z)}dz$$

$$=N_{f+g}-P_{f+g}=N_{f+g}$$

Theorem is proven. □

17.3.6 Fundamental algebra theorem

Theorem 17.15. (The fundamental theorem of algebra) *Any polynomial of degree $n\geq 1$ has a zero (root), that is, for any $n\geq 1$ there exists a point $z_0\in\mathbb{C}$ such that*

$$\boxed{p_a(z_0)=0} \tag{17.66}$$

where

$$\boxed{p_a(z):=a_0z^n+a_1z^{n-1}+\ldots+a_{n-1}z+a_n,\quad a_0\neq 0} \tag{17.67}$$

Corollary 17.8. *Every polynomial $p_a(z)$ (17.67) of degree $n\geq 1$ has exactly n zeros (roots), that is, for any $z\in\mathbb{C}$ the polynomial $p_a(z)$ can be represented as*

$$\boxed{p_a(z)=a_0\prod_{i=1}^{n}(z-z_i)} \tag{17.68}$$

First proof. **(based on Liouville's theorem)** Assume that $p_a(z)$ (17.67) has no zero and prove that $p_a(z)$ is a constant. Let $f(z)=1/p_a(z)$. Then f is analytic everywhere on $\mathbb{C}$ since, by the assumption, $p_a(z)\neq 0$ in $\mathbb{C}$. Since

$$p_a(z)=z^n\left[a_0+a_1z^{-1}+\ldots+a_{n-1}z^{-n+1}+a_nz^{-n}\right]\to\infty$$

as $|z| \to \infty$, so $|f(z)| \to 0$ as $|z| \to \infty$. Therefore, $f(z)$ is bounded on $\mathbb{C}$, and so, by Liouville's theorem 17.9, $f(z)$ and hence $p_a(z)$ is a constant on $\mathbb{C}$ that is possible if and only if $a_i = 0$ for $i = 0, 1, \ldots, n-1$. This contradicts with the condition $a_0 \neq 0$. Theorem is proven. □

Second proof. **(based on the Rouché theorem)** Let us put $g(z) := a_1 z^{n-1} + \ldots + a_{n-1} z + a_n$ and select R large enough such that on the circle $|z| = R$ there would be $|f(z)| > |g(z)|$ (this always may be done since $|f(z)| = |a_0| R^n$ and $|g(z)| \leq |a_1| R^{n-1} + \ldots + |a_{n-1}| R + |a_n|$). Then by the Rouché theorem (17.14) these two functions $f(z)$ and $[f(z) + g(z)]$ have the same number of roots. But $f(z) := a_0 z^n$ has exactly n roots which completes the proof. □

Example 17.13. *Let us define how many roots the polynomial*

$$p_a(z) = z^8 - 4z^5 + z^2 - 1$$

has in the disc $|z| < 1$. *Define* $f(z) := z^8 - 4z^5$ *and* $g(z) := z^2 - 1$. *Notice that on the circle* $|z| = 1$ *we have* $|f(z)| = |z^3 - 4| \geq 4 - |z^3| = 3$ *and* $|g(z)| \leq |z^2| + 1 = 2$. *Thus, by the Rouché theorem (17.14) the number of roots of* $p_a(z)$ *is equal to the number of roots of* $f(z) := z^8 - 4z^5 = z^5(z^3 - 4)$ *in the disc* $|z| < 1$ *which is equal to* 5 *(since* $z^3 - 4 \neq 0$ *within the disc).*

17.4 Integral transformations

In this section we will consider the class of the, so-called, *integral transformations* of an original complex function $f(t)$ of a real argument (defined on $\mathbb{R}^+$) into the corresponding function $F(p)$, called the *image*, defined on the complex plane $\mathbb{C}$. This class of transformations is given by the relation

$$\boxed{F(p) := \int_{t=0}^{\infty} f(t) K(t, p) dt} \tag{17.69}$$

where the function $K : \mathbb{R}^+ \times \mathbb{C} \to \mathbb{C}$ is called the *kernel* of the integral transformation (17.69). Such sorts of transformations are actively applied in theory of differential equations and many other fields of physics and engineering practice. Let us briefly present the most important of them. In any case, we will assume that the original function $f(t)$ satisfies the following conditions:

A1 It satisfies *Hölder's condition*, i.e., for any $t \in \mathbb{R}^+$ (maybe with the exception of some exclusive points) there exists positive constants L, h_0 and $\alpha \leq 1$ such that for all $h : |h| \leq h_0$

$$\boxed{|f(t+h) - f(t)| \leq L |h|^{\alpha}} \tag{17.70}$$

A2

$$\boxed{f(t) \equiv 0 \text{ if } t < 0} \tag{17.71}$$

A3 There exist such constants $M > 0$ and $s_0 \geq 0$ such that

$$\boxed{|f(t)| \leq Me^{s_0 t}} \tag{17.72}$$

(the constant s_0 is called the *increment index* of the function f).[4]

17.4.1 Laplace transformation $\left(K(t, p) = e^{-pt}\right)$

17.4.1.1 Direct Laplace transformation

Definition 17.12. *The **Laplace image** of the function $f(t)$ satisfying assumptions A1–A3 is called the complex function $F : \mathbb{C} \to \mathbb{C}$ of the complex variable $p := s + i\sigma$ defined by the relation*

$$\boxed{F(p) := \int_{t=0}^{\infty} f(t)\, e^{-pt} dt} \tag{17.73}$$

where the integral is taken over the positive semi-axis.[5] *We will write*

$$\boxed{F(p) = \mathcal{L}\{f\}} \tag{17.74}$$

Theorem 17.16. *For any original function $f(t)$ satisfying assumptions A1–A3 its Laplace image $F(p)$ is **correctly defined** within the semi-plane $\mathrm{Re}\, p = s > s_0$, where s_0 is the increment index of f, and $F(p)$ is **analytical** (regular) within this semi-plane.*

Proof. Indeed, for any p such that $\mathrm{Re}\, p = s > s_0$ the integral (17.73) converges absolutely since by A3 (17.72) it is estimated from above by a convergent integral, that is,

$$\left| \int_{t=0}^{\infty} f(t)\, e^{-pt}\, dt \right| \leq \int_{t=0}^{\infty} |f(t)|\, e^{-pt}\, dt \leq \int_{t=0}^{\infty} Me^{-(s-s_0)t}\, dt = \frac{M}{s - s_0} < \infty \tag{17.75}$$

---

[4] More exactly,

$$s_0 := \inf \left\{ s \geq 0 : \limsup_{t \to \infty} |f(t)|\, e^{-s_0 t} \leq M \right\}$$

[5] It is known also as the, so-called, **double-side Laplace transformation** defined by

$$F(p) := \int_{t=-\infty}^{\infty} f(t)\, e^{-pt}\, dt$$

Then, for any p within the semi-plane $\mathrm{Re}p \geq s_1 > s_0$ we have

$$|F'(p)| = \left| \frac{d}{dp} \int_{t=0}^{\infty} f(t)\, e^{-pt}\, dt \right|$$

$$= \left| \int_{t=0}^{\infty} f(t)\, te^{-pt}\, dt \right| \leq \int_{t=0}^{\infty} |f(t)|\, te^{-pt}\, dt \leq \int_{t=0}^{\infty} Mte^{-(s-s_0)t}\, dt$$

$$= \frac{M}{(s_1 - s_0)^2} < \infty \tag{17.76}$$

which exactly means that the function $F(p)$ possesses its derivative and, hence, is analytical in any point of the semi-plane $\mathrm{Re}p > s_0$. Theorem is proven. □

Corollary 17.9. *If* $p \to \infty$ *such that* $\mathrm{Re}p = s \to \infty$, *then* $F(p)$ *tends to zero, i.e.,*

$$\boxed{\lim_{s \to \infty} F(p) = 0} \tag{17.77}$$

Proof. It follows directly from (17.75). □

17.4.1.2 Inverse Laplace transformation

To obtain the main result on the inverse *Laplace transformation* we need the following simple lemma.

Lemma 17.8. *For any function* $\phi(x)$ *integrable (in Riemann sense) on the interval* $[\alpha, \beta]$, *we have*

$$\boxed{\lim_{b \to \infty} \int_{x=\alpha}^{\beta} \phi(x) \sin(bx)\, dx = 0} \tag{17.78}$$

Proof. If $\phi(x)$ is continuously differentiable then the integration by parts implies

$$\int_{x=\alpha}^{\beta} \phi(x) \sin(bx)\, dx = -\phi(x) \frac{\cos(bx)}{b} \Big|_{\alpha}^{\beta} + \int_{x=\alpha}^{\beta} \phi'(x) \frac{\cos(bx)}{b} dx \underset{b \to \infty}{\to} 0$$

If $\phi(x)$ is an integrable function then for any $\varepsilon > 0$ there exists a continuously differentiable function $\phi_\varepsilon(x)$ and the constant $b_\varepsilon > 0$ such that

$\int_{x=\alpha}^{\beta} |\phi(x) - \phi_\varepsilon(x)|\, dx \leq \frac{\varepsilon}{2}$ and $\left|\int_{x=\alpha}^{\beta} \phi_\varepsilon(x) \sin(bx)\, dx\right| \leq \frac{\varepsilon}{2}$. Therefore, $\left|\int_{x=\alpha}^{\beta} \phi(x) \sin(bx)\, dx\right| \leq \int_{x=\alpha}^{\beta} |\phi(x) - \phi_\varepsilon(x)|\, dx + \left|\int_{x=\alpha}^{\beta} \phi_\varepsilon(x) \sin(bx)\, dx\right| \leq \varepsilon$. Lemma is proven. □

The next theorem presents the main result on the inverse Laplace transformation.

Theorem 17.17. (on the inverse transformation) *If $f(t)$ is an original function satisfying assumptions A1–A3 and $F(p)$ is its image, then in any point $t \geq 0$ where it satisfies Hölder's condition (17.70) the following representation holds*

$$\boxed{f(t) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} F(p)\, e^{pt}\, dp := \mathcal{L}^{-1}\{F\}} \tag{17.79}$$

Here the integral is taken over any line $\mathrm{Re}\, p = a > s_0$ and is understood in the main-valued sense, that is, as the limit of the integral along the interval $[a - ib, a + ib]$ when $b \to \infty$.

Proof. Let us consider the integral

$$f_b(t) := \frac{1}{2\pi i} \int_{a-ib}^{a+ib} F(p)\, e^{pt}\, dp = \frac{1}{2\pi i} \int_{a-ib}^{a+ib} e^{pt} \left(\int_{\tau=0}^{\infty} f(\tau)\, e^{-p\tau}\, d\tau \right) dp$$

Since by (17.75) the integral $\int_{\tau=0}^{\infty} f(\tau)\, e^{-p\tau}\, d\tau$ converges uniformly on p in the semi-plane $\mathrm{Re}\, p \geq a$, we may change the order of the integration which gives

$$f_b(t) = \frac{1}{2\pi i} \int_{\tau=0}^{\infty} f(\tau) \left(\int_{a-ib}^{a+ib} e^{p(t-\tau)}\, dp \right) d\tau$$

$$= \frac{1}{\pi} \int_{\tau=0}^{\infty} f(\tau)\, e^{a(t-\tau)} \frac{\sin b(t-\tau)}{t-\tau} d\tau$$

$$= \frac{1}{\pi} e^{at} \int_{\tau=-t}^{\infty} f(x+t)\, e^{-a(x+t)} \frac{\sin bx}{x} dx$$

Denote $g(t) := f(t)e^{-at}$ and notice that by A2 $g(t) \equiv 0$ for $t < 0$. Therefore,

$$f_b(t) = \frac{e^{at}}{\pi} \int_{\tau=-\infty}^{\infty} \frac{g(x+t) - g(t)}{x} e^{-a(x+t)} \sin bx\, dx + \frac{f(t)}{\pi} \int_{\tau=-\infty}^{\infty} \frac{\sin bx}{x}\, dx$$

The second integral for any $b > 0$ is exactly the Euler's integral (17.59) and, hence, it is equal to π which leads to the following expression

$$f_b(t) = \frac{1}{\pi} e^{at} \int_{\tau=-\infty}^{\infty} \frac{g(x+t) - g(t)}{x} e^{-a(x+t)} \sin bx \, dx + f(t)$$

The first integral by Lemma 17.8 tends to zero as $b \to \infty$ which completes the proof. □

Corollary 17.10. *The original function $f(t)$ is completely defined by its image $F(p)$ (see formula (17.79)) with the exception of the points of discontinuity.*

17.4.1.3 Some properties of the Laplace transformation

1. By direct calculation using (17.73) it follows that

$$\boxed{\mathcal{L}\{1\} = \frac{1}{p}, \quad \mathcal{L}\{e^{p_0 t}\} = \frac{1}{p - p_0}, \quad \mathcal{L}\{\delta(t-\tau)\} = e^{-p\tau}} \tag{17.80}$$

2. Denoting $G(p) := \int_{t=0}^{\infty} g(t) e^{-pt} \, dt$, for any complex numbers α and β we have

$$\boxed{\mathcal{L}\{\alpha f(t) + \beta g(t)\} = \alpha F(p) + \beta G(p)} \tag{17.81}$$

3.

$$\boxed{\begin{aligned} &\mathcal{L}\{\sin(\omega t)\} = \mathcal{L}\left\{\frac{e^{i\omega t} - e^{-i\omega t}}{2i}\right\} \\ &= \frac{1}{2i}\left(\frac{1}{p - i\omega} - \frac{1}{p + i\omega}\right) = \frac{\omega}{p^2 + \omega^2} \end{aligned}} \tag{17.82}$$

4.

$$\boxed{\begin{aligned} &\mathcal{L}\{\cos(\omega t)\} = \mathcal{L}\left\{\frac{e^{i\omega t} + e^{-i\omega t}}{2}\right\} \\ &= \frac{1}{2}\left(\frac{1}{p - i\omega} + \frac{1}{p + i\omega}\right) = \frac{p}{p^2 + \omega^2} \end{aligned}} \tag{17.83}$$

5.

$$\boxed{\mathcal{L}\{\sinh(\omega t)\} = \frac{\omega}{p^2 - \omega^2}, \mathcal{L}\{\cosh(\omega t)\} = \frac{p}{p^2 - \omega^2}} \tag{17.84}$$

6. For any $\alpha > 0$

$$\boxed{\mathcal{L}\{f(\alpha t)\} = \frac{1}{\alpha} F\left(\frac{p}{\alpha}\right)} \tag{17.85}$$

7. *Differentiation of original functions*: If a function $f(t)$ is continuous for $t > 0$ and $f'(t)$ or, in general, $f^{(n)}(t)$ is an original function too, then

$$\boxed{\mathcal{L}\{f'(t)\} = pF(p) - f(0)} \tag{17.86}$$

or,

$$\boxed{\begin{array}{l}\mathcal{L}\left\{f^{(n)}(t)\right\} = p^n F(p) - p^{n-1} f(0) \\ -p^{n-2} f'(0) - \ldots - f^{(n-1)}(0)\end{array}} \tag{17.87}$$

Indeed, integrating by part we derive

$$\mathcal{L}\left\{f'(t)\right\} = \int_{t=0}^{\infty} f'(t)\, e^{-pt}\, dt = \left[f(t)\, e^{-pt}\right]_0^{\infty} + p \int_{t=0}^{\infty} f(t)\, e^{-pt}\, dt$$

and, since $\mathrm{Re}\, p = s > s_0$, it follows that $\left|f(t)\, e^{-pt}\right| \leq M e^{-(s-s_0)t}$. Therefore

$$\left[f(t)\, e^{-pt}\right]_0^{\infty} = -f(0)$$

which implies (17.86). Applying (17.86) n times we obtain (17.87).

8. *Differentiation of images*:

$$\boxed{F^{(n)}(p) = \mathcal{L}\{(-1)^n t^n f(t)\}} \tag{17.88}$$

This can be obtained by the direct differentiation (since $F(p)$ is analytical in $\mathrm{Re}\, p = s > s_0$), that is,

$$F'(p) = -\int_{t=0}^{\infty} t f(t)\, e^{-pt}\, dt,\ F''(p) = \int_{t=0}^{\infty} t^2 f(t)\, e^{-pt}\, dt$$

$$F^{(n)}(p) = (-1)^n \int_{t=0}^{\infty} t^n f(t)\, e^{-pt}\, dt$$

Example 17.14.

$$\boxed{\begin{aligned}&\mathcal{L}\{t^n\} = \frac{n!}{p^{n+1}},\ \mathcal{L}\{t^n e^{p_0 t}\} = \frac{n!}{(p-p_0)^{n+1}}\\ &\mathcal{L}\{t\sin(\omega t)\} = \frac{2p\omega}{\left(p^2+\omega^2\right)^2}\\ &\mathcal{L}\{t\cos(\omega t)\} = \frac{p^2-\omega^2}{\left(p^2+\omega^2\right)^2}\end{aligned}} \tag{17.89}$$

9. *Integration of original functions*:

$$\boxed{\mathcal{L}\left\{\int_{\tau=0}^{t} f(\tau)\,d\tau\right\} = \frac{F(p)}{p}} \tag{17.90}$$

It follows from (17.86) that if we take $g(t) := \int_{\tau=0}^{t} f(\tau)\,d\tau$ and calculate $F(p) = \mathcal{L}\{f(t)\} = \mathcal{L}\{g'(t)\} = pG(p)$ that gives $G(p) = \dfrac{F(p)}{p}$.

10. *Integration of images*: If the integral $\int_p^\infty F(p)\,dp$ (the path of integrations completely belongs to the semi-plane $\mathrm{Re}\,p \geq a > s_0$) converges, then

$$\boxed{\int_p^\infty F(p)\,dp = \mathcal{L}\left\{\frac{f(t)}{t}\right\}} \tag{17.91}$$

It follows from changing the order of integration:

$$\int_p^\infty F(p)\,dp = \int_{t=0}^\infty f(t)\left(\int_p^\infty e^{-pt}\,dp\right)dt = \int_{t=0}^\infty \frac{f(t)}{t}e^{-pt}\,dt$$

11. *Theorem on delay effect*: For any positive τ

$$\boxed{\mathcal{L}\{f(t-\tau)\} = e^{-p\tau}F(p)} \tag{17.92}$$

12. *Theorem on shifting effect*:

$$\boxed{\mathcal{L}\{e^{p_0 t}f(t)\} = F(p-p_0)} \tag{17.93}$$

13. *Multiplication (Borel) theorem*: Denoting the convolution $(f * g)$ of two functions by

$$\boxed{(f*g) := \int_{\tau=0}^{t} f(\tau)\,g(t-\tau)\,d\tau} \tag{17.94}$$

we have

$$\boxed{\mathcal{L}\{(f * g)\} = F(p)G(p)} \tag{17.95}$$

Indeed,

$$\begin{aligned}
\mathcal{L}\{(f * g)\} &= \int_{t=0}^{\infty} e^{-pt} \left(\int_{\tau=0}^{t} f(\tau)\, g(t-\tau)\, d\tau \right) dt \\
&= \int_{t=0}^{\infty} e^{-pt} \left(\int_{\tau=0}^{\infty} f(\tau)\, g(t-\tau)\, d\tau \right) dt \\
&= \int_{\tau=0}^{\infty} f(\tau) \left(\int_{t=0}^{\infty} e^{-pt} g(t-\tau)\, dt \right) d\tau \\
&= \int_{\tau=0}^{\infty} f(\tau) \left(\int_{t=\tau}^{\infty} e^{-pt} g(t-\tau)\, dt \right) d\tau \\
&= \int_{\tau=0}^{\infty} f(\tau) \left(\int_{t'=0}^{\infty} e^{-p(t'+\tau)} g(t')\, dt' \right) d\tau \\
&= \int_{\tau=0}^{\infty} f(\tau) e^{-p\tau}\, d\tau \left(\int_{t'=0}^{\infty} e^{-pt'} g(t')\, dt' \right) = F(p)G(p)
\end{aligned}$$

Corollary 17.11. (Duhammel's integral)

$$\boxed{\mathcal{L}\{f(0)g(t) + (f' * g)\} = pF(p)G(p)} \tag{17.96}$$

14. *Theorem on the inverse transformation*: Let $f(t)$ and $g(t)$ have the increment indices s_f and s_g, correspondingly. Then

$$\boxed{\mathcal{L}\{f(t)g(t)\} = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} F(q)\, G(p-q) dq} \tag{17.97}$$

where $a > s_f$ and $\mathrm{Re}\, p > s_g + a$.

Indeed,

$$\mathcal{L}\{f(t)g(t)\} = \int_{t=0}^{\infty} f(t)g(t)\,e^{-pt}\,dt$$

$$= \frac{1}{2\pi i}\int_{t=0}^{\infty}\left[\int_{a-i\infty}^{a+i\infty} F(q)\,e^{qt}dq\right] g(t)\,e^{-pt}\,dt$$

$$= \frac{1}{2\pi i}\left[\int_{a-i\infty}^{a+i\infty} F(q)\int_{t=0}^{\infty} g(t)\,e^{-(p-q)t}dt\right] dq$$

$$= \frac{1}{2\pi i}\int_{a-i\infty}^{a+i\infty} F(q)\,G(p-q)\,dq$$

15. *First theorem on the expansion*: If in a neighborhood of a point p such that $|p| \geq R$ (R is large enough) the function $F(p)$ may be presented by the Laurent series

$$\boxed{F(p) = \sum_{k=1}^{\infty} \frac{c_k}{p^k}} \tag{17.98}$$

then its original $f(t) = \mathcal{L}^{-1}\{F(p)\}$ can be represented as

$$\boxed{f(t) = \sum_{k=1}^{\infty} \frac{c_k}{(k-1)!}t^{k-1}} \tag{17.99}$$

This can be obtained using formulas (17.89).

16. *Second theorem on the expansion*: Let the function $F(p)$ be a meromorphic in a semi-plane $\mathrm{Re}\,p > s_0$, for any $a > s_0$ the integral $\int_{a-i\infty}^{a+i\infty} F(p)\,dp$ converges absolutely and there exists a system of circles C_n ($|p| = R_n \to \infty$, $R_1 < R_2 < \ldots$) such that $F(p) \to 0$ uniformly respectively $\arg p$. Then $F(p)$ is the image of the function

$$\boxed{f(t) = \sum_{(p_k)} \operatorname*{res}_{p_k} F(p)e^{pt}} \tag{17.100}$$

where the sum is taken over all singular points p_k. This result may be proven using Cauchy's residue theorem (17.27) and the Jordan lemma (17.6).

Corollary 17.12. *If* $F(p) = \dfrac{A(p)}{B(p)}$ *is rational such that*

$$\boxed{\deg A(p) < \deg B(p) = \sum_{k=1}^{N_p} n_k}$$

with p_k is a pole of $F(p)$ and n_k is its multiplicity, then

$$\boxed{f(t) = \sum_{k=1}^{N_p} \frac{1}{(n_k - 1)!} \lim_{p \to p_k} \frac{d^{n_k - 1}}{dp^{n_k - 1}} \left[F(p) (p - p_k)^{n_k} e^{pt} \right]} \tag{17.101}$$

17.4.2 Other transformations

17.4.2.1 Heavyside transformation

It is given by (17.69) with

$$\boxed{K(t, p) = pe^{-pt}} \tag{17.102}$$

17.4.2.2 Fourier transformation $K(t, i\omega) = \frac{1}{\sqrt{2\pi}} e^{-i\omega t}$

Main definitions

If in the double-side Laplace transformations [(in the direct one (17.73) put the increment index (17.72) $s_0 = 0$ and in the inverse transformation (17.79)) put $a = 0$] such that the integration is done over the imaginary axis ($p = i\omega$), we obtain the, so-called, *Fourier transformation*:

$$\boxed{\begin{aligned} \mathcal{F}(\omega) &:= \frac{1}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} f(t) e^{-i\omega t} dt \\ f(t) &= \frac{1}{\sqrt{2\pi}} \int_{\omega=-\infty}^{\infty} F(\omega) e^{i\omega t} d\omega \end{aligned}} \tag{17.103}$$

In physics the function $\mathcal{F}(\omega)$ is called the *spectral function* of the "oscillations" $f(t)$.

Remark 17.8. *The range of the application of the Fourier transformation (17.103) is significantly narrower than one for the Laplace transformation (17.73) since the corresponding first nonproper integral in (17.103) converges if the function $f(t)$ is absolutely integrable, i.e., $\int_{t=-\infty}^{\infty} |f(t)| dt < \infty$. In the case of the Laplace transformation (17.73) such condition becomes as $\int_{t=-\infty}^{\infty} |f(t) e^{-st}| dt < \infty$ $(s > s_0)$ which significantly extends the class of the original functions. From a physical point of view the Fourier transformation (17.103) is more natural than the Laplace transformation (17.73) since formulas (17.103) coincide (maybe some constants are different) with those of the representation of the original function $f(t)$ as the Fourier series (17.62)*

$$f(t) = \sum_{n=-\infty}^{\infty} G_n e^{in(2\pi/T)t}, \quad G_n = \frac{1}{T} \int_{t=0}^{T} f(t) e^{-in(2\pi/T)t} dt$$

valid for periodical (with the period T) functions $f(t)$ treated as "oscillations".

The auxiliary function $h_\lambda(\omega)$
According to Rudin (1973) put

$$H(t) := e^{-|t|}, \quad t \in (-\infty, \infty)$$

and notice that

$$0 < H(t) \leq 1, \quad H(\lambda t) \to 1 \quad \text{as} \quad \lambda \to 0$$

Define also the following parametric family of functions

$$h_\lambda(\omega) := \frac{1}{\sqrt{2\pi}} \int_{t=-\infty}^{\infty} H(\lambda t)\, e^{it\omega}\, dt, \quad \lambda > 0, \quad \omega \in \mathbb{R}$$

A simple computation gives

$$h_\lambda(\omega) = \sqrt{\frac{2}{\pi}} \frac{\lambda}{\lambda^2 + \omega^2}$$

and, therefore,

$$\int_{\omega=-\infty}^{\infty} h_\lambda(\omega)\, d\omega = \sqrt{2\pi} \tag{17.104}$$

Proposition 17.2. *If a function $g : \mathbb{R}_+ \to \mathbb{R}$ is absolutely integrable on $[0, \infty)$, that is,*

$$\int_{t=0}^{\infty} |g(t)|\, dt < \infty, \quad g(t) = 0 \quad \textit{for} \quad t < 0$$

and its Laplace transformation is $G(p)$, then

$$\begin{aligned} (g \circledast h_\lambda)(\omega) &:= \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^{\infty} g(\omega - y)\, h_\lambda(y)\, dy \\ &= \frac{1}{2\pi} \int_{t=-\infty}^{\infty} H(\lambda t)\, G(it)\, e^{it\omega}\, dt \end{aligned} \tag{17.105}$$

Proof. The simple application of Fubini's reduction theorem 16.24 gives

$$(g \circledast h_\lambda)(\omega) = \frac{1}{2\pi} \int_{y=-\infty}^{\infty} g(\omega - y) \left[\int_{t=-\infty}^{\infty} H(\lambda t)\, e^{ity}\, dt \right] dy$$

$$= \frac{1}{2\pi} \int_{t=-\infty}^{\infty} H(\lambda t) \left[\int_{y=-\infty}^{\infty} g(\omega - y)\, e^{ity} dy \right] dt$$

$$= \frac{1}{2\pi} \int_{t=-\infty}^{\infty} H(\lambda t) \left[\int_{y=-\infty}^{\infty} g(y)\, e^{it(\omega - y)} dy \right] dt$$

$$= 12\pi \int_{t=-\infty}^{\infty} H(\lambda t)\, e^{it\omega} \left[\int_{y=0}^{\infty} g(y)\, e^{-ity} dy \right] dt$$

$$= \frac{1}{2\pi} \int_{t=-\infty}^{\infty} H(\lambda t)\, e^{it\omega} G(it)\, dt$$

□

Proposition 17.3. *If a function* $g : \mathbb{R} \to \mathbb{R}$ *is bounded almost everywhere, that is,* $ess\sup |g(\omega)| < \infty$, *and it is continuous at a point* ω, *then*

$$(g * h_\lambda)(\omega) := \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^{\infty} g(\omega - y)\, h_\lambda(y)\, dy \to g(\omega) \tag{17.106}$$

as $\lambda \to 0$.

Proof. In view of (17.104) and by the dominated convergence theorem we have

$$(g * h_\lambda)(\omega) - g(\omega) = \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{\infty} [g(\omega - y) - g(\omega)]\, h_\lambda(y)\, dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{x=-\infty}^{\infty} [g(\omega - y) - g(\omega)]\, \lambda^{-1} h_{\lambda=1}\left(\frac{y}{\lambda}\right) dy$$

$$= \frac{1}{\sqrt{2\pi}} \int_{s=-\infty}^{\infty} [g(\omega - \lambda s) - g(\omega)]\, h_{\lambda=1}(s)\, ds \to 0 \text{ as } \lambda \to 0$$

□

Corollary 17.13. *For any* $p \in [1, \infty)$

$$\|(g \circledast h_\lambda) - g\|_{L_p} \to 0 \text{ as } \lambda \to 0$$

The Plancherel theorem

One of the most important results of the Fourier transformation theory is the next theorem.

Theorem 17.18. (Plancherel, around 1800) *If* $f(t) \in L_2[0, \infty)$ *and its* ***Laplace transformation*** *(17.73) is* $F(p) \in \mathbb{H}_2$,[6] *then the following identity (known as* ***Parseval's identity****) holds:*

$$\boxed{\begin{aligned} \|f\|_{L_2} &:= \left(\int_{t=0}^{\infty} |f(t)|^2 \, dt \right)^{1/2} \\ &= \|F\|_{\mathbb{H}_2} := \left(\frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} F(j\omega) \, F^{\sim}(j\omega) \, d\omega \right)^{1/2} \\ &\text{where} \quad F^{\sim}(s) := F(-s) \end{aligned}} \tag{17.107}$$

Proof. Recalling that $f(t) = 0$ for $t < 0$, define the function

$$g(x) := \left(f \circledast \tilde{f} \right)(x)$$

where $\tilde{f}(x) := f(-x)$. Then

$$g(t) = \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^{\infty} f(t-y) \, f(-y) \, dy = \frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f(t+y) \, f(y) \, dy$$

and, in view of (17.95), its Laplace transformation is

$$G(p) = \frac{1}{\sqrt{2\pi}} F(p) \, F(-p)$$

[6] The exact definitions of the functional spaces $L_2[0, \infty)$ and $\mathbb{H}_2$ are given in Chapter 18.

Indeed,

$$\begin{aligned}
\mathcal{L}\{g\} &= \int_{t=0}^{\infty} e^{-pt} \left(\frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f(t+y) f(y)\, dy \right) dt \\
&= \frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f(y) \left(\int_{t=0}^{\infty} e^{-pt} f(t+y)\, dt \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f(y) \left(\int_{t=0}^{\infty} e^{-p(t'-y)} f(t')\, dt' \right) dy \\
&= \frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f(y) e^{py} \left(\int_{t=0}^{\infty} e^{-pt'} f(t')\, dt' \right) dy \\
&= \frac{1}{\sqrt{2\pi}} F(p) F(-p)
\end{aligned}$$

It is easy to see that $g(x)$ is a continuous function. But it is also bounded since

$$\begin{aligned}
|g(x)| &\leq \frac{1}{\sqrt{2\pi}} \sqrt{\int_{y=-\infty}^{\infty} f^2(x+y)\, dy} \sqrt{\int_{y=-\infty}^{\infty} f^2(y)\, dy} \\
&\leq \frac{1}{\sqrt{2\pi}} \int_{y=-\infty}^{\infty} f^2(y)\, dy < \infty
\end{aligned}$$

Therefore, by Propositions 17.2 and 17.3, it follows that

$$(g * h_\lambda)(0) = \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} H(\lambda\omega) G(i\omega)\, d\omega$$

But

$$\lim_{\lambda \to 0} (g * h_\lambda)(0) = g(0) = \frac{1}{\sqrt{2\pi}} \int_{y=0}^{\infty} f^2(y)\, dy = \frac{1}{\sqrt{2\pi}} \|f\|_{L_2}^2$$

and

$$\lim_{\lambda\to 0}\frac{1}{2\pi}\int\limits_{\omega=-\infty}^{\infty} H(\lambda\omega)\,G(i\omega)\,d\omega=\frac{1}{2\pi}\int\limits_{\omega=-\infty}^{\infty} G(i\omega)\,d\omega$$

$$\frac{1}{2\pi}\int\limits_{\omega=-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}F(i\omega)\,F(-i\omega)\right]d\omega=\frac{1}{\sqrt{2\pi}}\|F\|_{\mathbb{H}_2}^2$$

which completes the proof. □

Corollary 17.14. *If* $f(t), g(t) \in L_2[0,\infty)$ *and their* ***Laplace transformation*** *(17.73) are* $F(p), G(p) \in \mathbb{H}_2$, *then the following identity holds:*

$$\begin{aligned}\langle f,g\rangle_{L_2} &:= \left(\int\limits_{t=0}^{\infty} f(t)\,g(t)\,dt\right)^{1/2}\\ &= \langle F,G\rangle_{\mathbb{H}_2} := \left(\frac{1}{2\pi}\int\limits_{\omega=-\infty}^{\infty} F(j\omega)\,G^{\sim}(j\omega)\,d\omega\right)^{1/2}\\ \text{where}\quad G^{\sim}(s) &:= G(-s)\end{aligned} \tag{17.108}$$

Proof. It completely repeats the proof of the previous theorem. □

17.4.2.3 Two-dimensional Fourier transformation

It is given by

$$\begin{aligned}G(\sigma,\tau) &:= \frac{1}{2\pi}\int\limits_{x=-\infty}^{\infty}\int\limits_{y=-\infty}^{\infty} g(x,y)\,e^{-i(\sigma x+\tau y)}\,dx\,dy\\ g(x,y) &= \frac{1}{2\pi}\int\limits_{\sigma=-\infty}^{\infty}\int\limits_{\tau=-\infty}^{\infty} G(\sigma,\tau)\,e^{i(\sigma x+\tau y)}\,d\sigma\,d\tau\end{aligned} \tag{17.109}$$

17.4.2.4 Both-side Laplace transformation

If we refuse assumption A2 (17.71) and in the Fourier transformation (17.103) make the integration in the range $(-\infty, \infty)$, we obtain the, so-called, *both-side Laplace transformation* given by

$$\boxed{F(p) := \int_{t=-\infty}^{\infty} f(t)\, e^{-pt}\, dt, \quad f(t) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} F(p)\, e^{pt}\, dp} \tag{17.110}$$

17.4.2.5 Melline transformation

In (17.110) if we change p with $(-p)$ and t with $\ln \tau$, we get

$$F(-p) := \int_{t=-\infty}^{\infty} f(\ln \tau)\, e^{p \ln \tau} \frac{d\tau}{\tau}, \; f(\ln \tau) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} F(-p)\, e^{-p \ln \tau}\, dp$$

Defining $g(\tau) := f(\ln \tau)$ and $G(p) := F(-p)$ we obtain the, so-called, *Melline transformation*:

$$\boxed{G(p) := \int_{\tau=-\infty}^{\infty} g(\tau)\, \tau^{p-1}\, d\tau, \; g(\tau) = \frac{1}{2\pi i} \int_{a-i\infty}^{a+i\infty} \frac{G(p)}{\tau^p} dp} \tag{17.111}$$

Denote this transformation by $G(p) := \mathcal{M}\{g(\tau)\}$.

Claim 17.2. *It is easy to check that*

1.

$$\boxed{\mathcal{M}\{g(\alpha\tau)\} = \frac{G(p)}{\alpha^p}, \quad \alpha > 0} \tag{17.112}$$

2.

$$\boxed{\mathcal{M}\{\tau^{\alpha} g(\tau)\} = G(p+\alpha)} \tag{17.113}$$

3.

$$\boxed{\mathcal{M}\{f(\tau)\, g(\tau)\} = \int_{a-i\infty}^{a+i\infty} F(q) G(p-q)\, dq} \tag{17.114}$$

4.

$$\boxed{\begin{aligned}\mathcal{M}\{g'(\tau)\} &= -(p-1)\,G(p-1)\\ \mathcal{M}\{\tau g'(\tau)\} &= -pG(p)\\ \mathcal{M}\{\tau^2 g'(\tau)\} &= (p+1)\,pG(p)\end{aligned}} \tag{17.115}$$

This transformation turns out to be very useful for the solution of partial differential equations of a heating type.

17.4.2.6 Hankel (Fourier–Bessel) transformation

Let us in (17.109) make the transformation to the polar coordinates, i.e., $y = r\sin\varphi$, $\sigma = \rho\cos\theta$, $\tau = \rho\sin\theta$ which gives

$$G(\rho,\theta) = \frac{1}{2\pi}\int_{r=-\infty}^{\infty} r\left[\int_{\varphi=-\infty}^{\infty} g(r,\varphi)\,e^{-ir\rho\cos(\varphi-\theta)}\,d\varphi\right]dr$$

$$g(r,\varphi) = \frac{1}{2\pi}\int_{\rho=-\infty}^{\infty} \rho\left[\int_{\theta=0}^{2\pi} G(\rho,\theta)\,e^{ir\rho\cos(\varphi-\theta)}\,d\theta\right]d\rho$$

Representing $g(r,\varphi)$ as $g(r,\varphi) = e^{-in\varphi}\tilde{g}(r)$ (where n is an integer) and $(\varphi-\theta)$ as $\varphi-\theta = \dfrac{\pi}{2}+t$, we derive

$$G(\rho,\theta) = \frac{1}{2\pi}e^{-in(\theta+\pi/2)}\int_{r=0}^{\infty}\tilde{g}(r)\,r\left[\int_{t=0}^{2\pi} e^{i(r\rho\sin t - nt)}\,dt\right]dr$$

Defining $J_n(z) := \dfrac{1}{2\pi}\int_{t=0}^{2\pi}\cos(nt - z\sin t)\,dt$, $G_n(\rho) = e^{in(\theta+\pi/2)}G(\rho,\theta)$, we may write

$$\boxed{G_n(\rho) = \int_{t=0}^{2\pi} g(r)\,J_n(r\rho)\,r\,dr} \tag{17.116}$$

and, hence, substituting $\theta-\varphi = t-\dfrac{\pi}{2}$ we obtain

$$\tilde{g}(r) = \frac{1}{2\pi}\int_{\rho=-\infty}^{\infty} G_n(\rho)\,\rho\left[\int_{\theta=0}^{2\pi} e^{i[n(\varphi-\theta-\pi/2)+r\rho\cos(\varphi-\theta)]}\,d\theta\right]d\rho$$

which, finally, implies

$$\boxed{\tilde{g}(r) = \int_{\rho=-\infty}^{\infty} G_n(\rho)\, J_n(r\rho)\, \rho\, d\rho} \tag{17.117}$$

Formulas (17.116) and (17.117) are called the *Hankel (Fourier–Bessel) transformation*. It is frequently used for the solution of the partial differential equation describing potential electric two-dimensional fields.

18 Topics of Functional Analysis

Contents

In Chapter 14 important concepts were introduced such as

1. *Lineality* of a space of elements;
2. *Metric* (or norm) in a space;
3. *Compactness*, *convergence* of a sequence of elements and *Cauchy sequences*;
4. *Contraction principle*.

As examples we have considered in detail the finite dimensional spaces $\mathbb{R}^n$ and $\mathbb{C}^n$ of real and complex vectors (numbers). But the same definitions of lineality and norms remain true if we consider as another example a functional space (where an element is a function) or a space of sequences (where an element is a sequence of real or complex vectors). The specific feature of such spaces is that they are all *infinite dimensional*. This chapter deals with the analysis of such spaces which is called "functional analysis".

Let us introduce two important additional concepts which we will use below.

Definition 18.1. *The subset $\mathcal{V}$ of a linear normed space $\mathcal{X}$ is said to be **dense** in $\mathcal{X}$ if its closure is equal to $\mathcal{X}$.*

This property means that every element $x \in \mathcal{X}$ may be approximated as closely as we like by some element $v \in \mathcal{V}$, that is, for any $x \in \mathcal{X}$ and any $\varepsilon > 0$ there exists an element $v \in \mathcal{V}$ such that $\|x - v\| < \varepsilon$.

All normed linear spaces have dense subsets, but they need not be obligatory countable subsets.

Definition 18.2. *A normed linear space $\mathcal{X}$ is said to be **separable** if it contains at least one dense subset which is countable.*

The separable spaces have special properties that are important in different applications. In particular, denoting the elements of such countable subsets by $\{e_i\}_{i=1,\ldots}$ it is possible to represent each element $x \in \mathcal{X}$ as the convergent series

$$\boxed{x = \sum_{i=1}^{\infty} \xi_i e_i} \tag{18.1}$$

where the scalars $\xi_i \in \mathbb{R}$ are called the coordinates of the element x in the *basis* $\{e_i\}_{i=1,\ldots}$.

18.1 Linear and normed spaces of functions

Below we will introduce examples of some functional spaces with the corresponding norm within. The lineality and main properties of a norm (metric) can be easily verified that is why we leave this for the reader as an exercise.

18.1.1 Space m_n of all bounded complex numbers

Let us consider a set m of sequences $x := \{x_i\}_{i=1}^{\infty}$ such that

$$\boxed{x_i \in \mathbb{C}^n \quad \text{and} \quad \sup_i \|x_i\| < \infty} \tag{18.2}$$

where $\|x_i\| := \sqrt{\sum_{s=1}^{n} x_{is}\bar{x}_{is}}$ and introduce the norm in m as

$$\boxed{\|x\| := \sup_i \|x_i\|} \tag{18.3}$$

18.1.2 Space l_p^n of all summable complex sequences

By definition

$$\boxed{l_p^n := \left\{ x = \{x_i\}_{i=1}^{\infty} \mid x_i \in \mathbb{C}^n, \quad \|x\|_{l_p^n} := \left(\sum_{i=1}^{\infty} \|x_i\|^p \right)^{1/p} \right\} < \infty} \tag{18.4}$$

18.1.3 Space $C[a, b]$ of continuous functions

It is defined as follows

$$\boxed{\begin{aligned} C[a,b] := \Big\{ f(t) \mid f \quad & \text{is continuous for all} \quad t \in [a,b], \\ & \|f\|_{C[a,b]} := \max_{t\in[a,b]} \left| f(t) \right| < \infty \Big\} \end{aligned}} \tag{18.5}$$

18.1.4 Space $C^k[a, b]$ of continuously differentiable functions

It contains all functions which are k-times differentiable and the kth derivative is continuous, that is,

$$C^k[a,b] := \left\{ f(t) \mid f^{(k)} \quad \text{exists and is continuous} \right.$$
$$\left. \text{for all} \quad t \in [a,b], \quad \|f\|_{C^k[a,b]} := \sum_{i=0}^{k} \max_{t\in[a,b]} \left|f^{(i)}(t)\right| < \infty \right\} \tag{18.6}$$

18.1.5 Lebesgue spaces $L_p[a,b]$ $(1 \le p < \infty)$

For each $1 \le p < \infty$ it is defined by the following way:

$$L_p[a,b] := \left\{ f(t) : [a,b] \to \mathbb{C} \mid \int_{t=a}^{b} |f(t)|^p \, dt < \infty \right.$$
(here the integral is understood in the Lebesgue sense),
$$\left. \|f\|_p := \left(\int_{t=a}^{b} |f(t)|^p \, dt \right)^{1/p} \right\} \tag{18.7}$$

Remark 18.1. *Sure, here functions $f(t)$ are not obligatory continuous.*

18.1.6 Lebesgue spaces $L_\infty[a,b]$

It contains all measurable functions from $[a,b]$ to $\mathbb{C}$, namely,

$$L_\infty[a,b] := \left\{ f(t) : [a,b] \to \mathbb{C} \mid \right.$$
$$\left. \|f\|_\infty := \operatorname*{ess\,sup}_{t\in[a,b]} |f(t)| < \infty \right\} \tag{18.8}$$

18.1.7 Sobolev spaces $S_p^l(G)$

It consists of all functions (for simplicity, real valued) $f(t)$ defined on G which have p-integrable continuous derivatives $f^{(i)}(t)$ $(i = 1, \ldots, l)$, that is,

$$S_p^l(G) := \left\{ f(t) : G \to \mathbb{R} \mid < \infty \ \ (i = 1, \ldots, l) \right.$$
(the integral is understood in the Lebesgue sense),
$$\left. \|f\|_{S_p^l(G)} := \left(\int_{t\in G} |f(t)|^p \, dt + \sum_{i=1}^{l} \int_{t\in G} \left|f^{(i)}(t)\right|^p dt \right)^{1/p} \right\} \tag{18.9}$$

More exactly, the Sobolev space is the completion (see definition below) of (18.9).

18.1.8 Frequency domain spaces $\mathbb{L}_p^{m\times k}$, $\mathbb{RL}_p^{m\times k}$, $\mathbb{L}_\infty^{m\times k}$ and $\mathbb{RL}_\infty^{m\times k}$

By definition

1. The Lebesgue space $\mathbb{L}_p^{m\times k}$ is the space of all p-integrable complex matrices, i.e.,

$$\mathbb{L}_p^{m\times k} := \left\{ F : \mathbb{C} \to \mathbb{C}^{m\times k} \mid \|F\|_{\mathbb{L}_p^{m\times k}} := \left(\frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} (\operatorname{tr}\{F(j\omega)\, F^{\sim}(j\omega)\})^{p-1}\, d\omega \right)^{1/p} < \infty \right\} \tag{18.10}$$

(here $F^{\sim}(j\omega) := F^{\mathsf{T}}(-j\omega)$)

2. The Lebesgue space $\mathbb{RL}_p^{m\times k}$ is the subspace of $\mathbb{L}_p^{m\times k}$ containing only complex matrices with rational elements, i.e., in

$$F = \left\| F_{i,j}(s) \right\|_{i=\overline{1,m};\ j=\overline{1,k}}$$

each element $F_{i,j}(s)$ represents the polynomial ratio

$$F_{i,j}(s) = \frac{a_{i,j}^0 + a_{i,j}^1 s + \ldots + a_{i,j}^{p_{i,j}} s^{p_{i,j}}}{b_{i,j}^0 + b_{i,j}^1 s + \ldots + b_{i,j}^{q_{i,j}} s^{q_{i,j}}} \tag{18.11}$$

$p_{i,j}$ and $q_{i,j}$ are positive integers

Remark 18.2. *If $p_{i,j} \le q_{i,j}$ for each element F_{ij} of (18.11), then $F(s)$ can be interpreted as a matrix **transfer function** of a linear (finite-dimensional) system.*

3. The Lebesgue space $\mathbb{L}_\infty^{m\times k}$ is the space of all complex matrices bounded (almost everywhere) on the imaginary axis elements, i.e.,

$$\mathbb{L}_\infty^{m\times k} := \left\{ F : \mathbb{C} \to \mathbb{C}^{m\times k} \mid \|F\|_{\mathbb{L}_\infty^{m\times k}} := \operatorname*{ess\,sup}_{s:\operatorname{Re} s>0} \lambda_{\max}^{1/2}\{F(s)\, F^{\sim}(s)\} \right\} = \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}^{1/2}\{F(j\omega)\, F^{\sim}(j\omega)\} < \infty \tag{18.12}$$

(the last equality may be regarded as the generalization of the maximum modulus principle 17.10 for matrix functions).

4. The Lebesgue space $\mathbb{RL}_\infty^{m\times k}$ is the subspace of $\mathbb{L}_\infty^{m\times k}$ containing only complex matrices with rational elements given in the form (18.11).

18.1.9 Hardy spaces $\mathbb{H}_p^{m\times k}$, $\mathbb{RH}_p^{m\times k}$, $\mathbb{H}_\infty^{m\times k}$ and $\mathbb{RH}_\infty^{m\times k}$

The Hardy spaces $\mathbb{H}_p^{m\times k}$, $\mathbb{RH}_p^{m\times k}$, $\mathbb{H}_\infty^{m\times k}$ and $\mathbb{RH}_\infty^{m\times k}$ are subspaces of the corresponding Lebesgue spaces $\mathbb{L}_p^{m\times k}$, $\mathbb{RL}_p^{m\times k}$, $\mathbb{L}_\infty^{m\times k}$ and $\mathbb{RL}_\infty^{m\times k}$ containing complex matrices with only regular (holomorphic) (see Definition 17.2) elements on the open half-plane $\operatorname{Re} s > 0$.

Remark 18.3. *If $p_{i,j} \leq q_{i,j}$ for each element F_{ij} of, then $F(s) \in \mathbb{RH}_p^{m\times k}$ can be interpreted as a matrix transfer function of a* ***stable*** *linear (finite-dimensional) system.*

Example 18.1.

$$\begin{array}{ll} \dfrac{1}{2-s} \in \mathbb{RL}_2 := \mathbb{RL}_2^{1\times 1}, & \dfrac{1-s}{2-s} \in \mathbb{RL}_\infty := \mathbb{RL}_\infty^{1\times 1} \\ \dfrac{1}{2+s} \in \mathbb{HL}_2 := \mathbb{HL}_2^{1\times 1}, & \dfrac{1-s}{2+s} \in \mathbb{RH}_\infty := \mathbb{RH}_\infty^{1\times 1} \\ \dfrac{e^{-s}}{2-s} \in \mathbb{L}_2 := \mathbb{L}_2^{1\times 1}, & \dfrac{e^{-s}}{2+s} \in \mathbb{H}_2 := \mathbb{H}_2^{1\times 1} \\ e^{-s}\dfrac{1-s}{2-s} \in \mathbb{L}_\infty := \mathbb{L}_\infty^{1\times 1}, & e^{-s}\dfrac{1-s}{2+s} \in \mathbb{H}_\infty := \mathbb{H}_\infty^{1\times 1} \end{array}$$

18.2 Banach spaces

18.2.1 Basic definition

Remember that a linear normed (topological) space $\mathcal{X}$ is said to be complete (see Definition 14.14) if every Cauchy (fundamental) sequence has a limit in the same space $\mathcal{X}$. The concept of a complete space is very important since even without evaluating the limit one can determine whether a sequence is convergent or not. So, if a metric (topological) space is not complete it is impossible to talk about a convergence, limits, differentiation and so on.

Definition 18.3. *A linear, normed and complete space is called a* ***Banach space****.*

18.2.2 Examples of incomplete metric spaces

Sure, not all linear normed (metric) spaces are complete. The example given below illustrates this fact.

Example 18.2. (of a noncomplete normed space) *Let us consider the space $CL[0, 1]$ of all continuous functions $f : [0, 1] \to R$ which are absolutely integrable (in this case, in the Riemann sense) on $[0, 1]$, that is, for which*

$$\|f\|_{CL[0,1]} := \int_{t=-1}^{1} |f(t)|\, dt < \infty \tag{18.13}$$

Consider the sequence $\{f_n\}$ of the continuous functions

$$f_n := \begin{cases} nt & \text{if } t \in [0, 1/n] \\ 1 & \text{if } t \in [1/n, 1] \end{cases}$$

Then for $n > m$

$$\|f_n - f_m\|_{CL[0,1]} = \int_{t=0}^{1} |f_n(t) - f_m(t)|\, dt$$

$$= \int_{t=0}^{1/n} |nt - mt|\, dt + \int_{t=1/n}^{1/m} |1 - mt|\, dt + \int_{t=1/m}^{1} |1 - 1|\, dt$$

$$\frac{(n-m)}{2n^2} + \frac{(1-m/n)^2}{2m} = \frac{1}{2}\left(\frac{1}{m} - \frac{1}{n}\right) \to 0$$

as $n, m \to \infty$. So, $\{f_n\}$ is a Cauchy sequence. However, its pointwise limit is

$$f_n(t) \to \begin{cases} 1 & \text{if} \quad 0 < t \le 1 \\ 0 & \text{if} \quad t = 0 \end{cases}$$

In other words, the limit is a discontinuous function and, hence, it is not in $CL[0, 1]$. This means that the functional space $CL[0, 1]$ is not complete.

Example 18.3. *By the same reason, the spaces $CL_p[0, 1]$ (the space of continuous and p-integrable functions) are not complete.*

18.2.3 Completion of metric spaces

There exist two possibilities to correct the situation and to provide the completeness property for a linear normed space if initially some is not complete:

- try to change the definition of a norm;
- try to extend the class of considered functions (it was suggested by Cauchy).

18.2.3.1 Changing of a norm

To illustrate the first approach related to changing of a norm let us consider again the space of all functions *continuous* at the interval $[0, 1]$, but instead of the Lebesgue norm (18.13) we consider the Chebyshev type norm $\|f\|_{C[a,b]}$ as in (18.5). This means that instead of the space $CL[0, 1]$ we will consider the space $C[a, b]$ (18.5). Evidently, that this space is complete, since it is known that uniform convergent sequences of continuous functions converge to a continuous function. Hence, $C[a, b]$ is a *Banach space* under this norm.

Claim 18.1. *By the same reasons it is not difficult to show that all spaces $C^k[a, b]$ (18.6) are Banach.*

Claim 18.2. *The spaces $L_p[a, b]$ $(1 \le p < \infty)$ (18.7), $L_\infty[a, b]$ (18.8), $\mathbb{L}_p^{m\times k}$ (23.19) and $\mathbb{L}_\infty^{m\times k}$ (18.12) are Banach too.*

18.2.3.2 Completion

Theorem 18.1. *Any linear normed space $\mathcal{X}$ with a norm $\|x\|_{\mathcal{X}}$ can be considered as a linear manifold which is complete in some Banach space $\hat{\mathcal{X}}$. This space $\hat{\mathcal{X}}$ is called the* ***completion*** *of $\mathcal{X}$.*

Proof. Consider two fundamental sequences $\{x_n\}$ and $\{x'_n\}$ with elements from $\mathcal{X}$. We say that they are *equivalent* if $\|x_n - x'_n\| \to 0$ as $n \to \infty$ and we will write $\{x_n\} \sim \{x'_n\}$. The set of all fundamental sequences may be separated (*factorized*) at noncrossed classes: $\{x_n\}$ and $\{x'_n\}$ are included in the same class if and only if $\{x_n\} \sim \{x'_n\}$. The set of all such classes $\mathcal{X}_i$ we denoted by $\hat{\mathcal{X}}$. So,

$$\hat{\mathcal{X}} := \bigcup_i \mathcal{X}_i, \quad \mathcal{X}_i \bigcap_{i \neq j} \mathcal{X}_j = \varnothing$$

Let us make the space $\hat{\mathcal{X}}$ a normed space. To do that, define the *operation of summing of the classes* $\mathcal{X}_i$ by the following manner: if $\{x_n\} \in \mathcal{X}_i$ and $\{y_n\} \in \mathcal{X}_j$ then class $(\mathcal{X}_i + \mathcal{X}_j)$ may be defined as the class containing $\{x_n + y_n\}$. The *operation of the multiplication by a constant* may be introduced as follows: we denoted by $\lambda\mathcal{X}_i$ the class containing $\{\lambda x_n\}$ if $\{x_n\} \in \mathcal{X}_i$. It is evident that $\hat{\mathcal{X}}$ is *a linear space*. Define now the norm in $\hat{\mathcal{X}}$ as

$$\|\mathcal{X}_i\| := \lim_{n\to\infty} \|x_n\|_{\mathcal{X}} \quad (\{x_n\} \in \mathcal{X}_i)$$

It easy to check the norm axioms for such norm and to show that

(a) $\mathcal{X}$ may be considered as a linear manifold in $\hat{\mathcal{X}}$;
(b) $\mathcal{X}$ is dense in $\hat{\mathcal{X}}$, i.e., there exists $\{x_n\} \in \mathcal{X}$ such that $\|x_n - \mathcal{X}_i\|_{\mathcal{X}} \to 0$ as $n \to \infty$ for some $\mathcal{X}_i \in \mathcal{X}$;
(c) $\hat{\mathcal{X}}$ is complete (Banach).

This completes the proof. □

This theorem can be interpreted as the following statement.

Corollary 18.1. *For any linear norm space $\mathcal{X}$ there exists a Banach space $\hat{\mathcal{X}}$ and a linear, injective map $T : \mathcal{X} \to \hat{\mathcal{X}}$ such that $T(\mathcal{X})$ is dense in $\hat{\mathcal{X}}$ and for all $x \in \mathcal{X}$*

$$\boxed{\|Tx\|_{\hat{\mathcal{X}}} = \|x\|_{\mathcal{X}}}$$

18.3 Hilbert spaces

18.3.1 Definition and examples

Definition 18.4. *A* ***Hilbert space*** *$\mathcal{H}$ is an inner (scalar) product space that is complete as a linear normed space under the* ***induced norm***

$$\boxed{\|z\|_{\mathcal{H}} := \sqrt{\langle z, z\rangle}} \tag{18.14}$$

Example 18.4. *The following spaces are Hilbert*

1. *The* ***space*** l_2^n ***of all summable complex sequences*** *(see (18.4) for $p = 2$) under the inner product*

$$\boxed{\langle x, y\rangle_{l_2^n} := \sum_{i=1}^{\infty} x_i \bar{y}_i} \tag{18.15}$$

2. *The* ***Lebesgue space*** $L_2[a, b]$ *of all integrable (in Lebesgue sense) complex functions (see (18.7) for $p = 2$) under the inner product*

$$\boxed{\langle x, y\rangle_{L_2[a,b]} := \int_{t=a}^{b} x(t)\bar{y}(t)\, dt} \tag{18.16}$$

3. *The* ***Sobolev's space*** $S_2^l(G)$ ***of*** *all l times differentiable on G quadratically integrable (in Lebesgue sense) complex functions (see (18.9) for $p = 2$) under the inner product*

$$\boxed{\langle x, y\rangle_{S_p^l(G)} := \sum_{i=0}^{l} \left\langle \frac{d^i}{dt^i} x, \frac{d^i}{dt^i} y \right\rangle_{L_2[a,b]}} \tag{18.17}$$

4. *The* ***frequency domain space*** $\mathbb{L}_2^{m\times k}$ *of all p-integrable complex matrices (23.19) under the inner product*

$$\boxed{\langle x, y\rangle_{\mathbb{L}_p^{m\times k}} := \int_{\omega=-\infty}^{\infty} tr\{X(j\omega)\, Y^{\sim}(j\omega)\}\, d\omega} \tag{18.18}$$

5. *The* ***Hardy spaces*** $\mathbb{H}_2^{m\times k}$ *(the subspace of $\mathbb{L}_2^{m\times k}$ containing only holomorphic in the right-hand semi-plan $\mathbb{C}^+ := \{s \in \mathbb{C} \mid \mathrm{Re}\, s > 0\}$ functions) under the inner product (18.18).*

18.3.2 Orthogonal complement

Definition 18.5. *Let $\mathcal{M}$ be a subset of a Hilbert space $\mathcal{H}$, i.e., $\mathcal{M} \subset \mathcal{H}$. Then the* ***distance*** *between a point $x \in \mathcal{H}$ and $\mathcal{M}$ is defined by*

$$\boxed{\rho(x, \mathcal{M}) := \inf_{y\in\mathcal{M}} \|x - y\|} \tag{18.19}$$

The following claim seems to be evident.

Claim 18.3. *If $x \in \mathcal{M}$, then $\rho(x, \mathcal{M}) = 0$. If $x \notin \mathcal{M}$ and $\mathcal{M}$ is a closed set (see Definition 14.7), then $\rho(x, \mathcal{M}) > 0$.*

Corollary 18.2. *If $\mathcal{M} \subset \mathcal{H}$ is a closed convex set and $x \notin \mathcal{M}$, then there exists a unique element $y \in \mathcal{M}$ such that $\rho(x, \mathcal{M}) = \|x - y\|$.*

Proof. Indeed, suppose that there exists another element $y^* \in \mathcal{M}$ such that

$$\rho(x, \mathcal{M}) = \|x - y\| = \|x - y^*\| := d$$

Then

$$4d^2 = 2\|x - y\|^2 + 2\|x - y^*\|^2$$

$$= \|x - y^*\|^2 + 4\left\|x - \frac{y + y^*}{2}\right\|^2 \geq \|x - y^*\|^2 + 4 \inf_{y \in \mathcal{M}} \|x - y\|^2$$

$$\geq \|x - y^*\|^2 + 4d^2$$

which gives $\|x - y^*\|^2 \leq 0$, or, equivalently, $y = y^*$. □

Corollary 18.3. *If $\mathcal{M} \subset \mathcal{H}$ is a subspace of $\mathcal{H}$ (this means that it is a closed convex linear manifold in $\mathcal{H}$) then for any $x \in \mathcal{H}$ there exists a unique element $x_{\mathcal{M}} \in \mathcal{M}$ such that*

$$\boxed{\rho(x, \mathcal{M}) := \inf_{y \in \mathcal{M}} \|x - y\| = \|x - x_{\mathcal{M}}\|} \tag{18.20}$$

*This element $x_{\mathcal{M}} \in \mathcal{M}$ is called the **orthogonal projection** of the element $x \in \mathcal{H}$ onto the subspace $\mathcal{M} \subset \mathcal{H}$.*

Lemma 18.1. *Let $\rho(x, \mathcal{M}) = \|x - x_{\mathcal{M}}\|$ where $\mathcal{M}$ is a subspace of a Hilbert space $\mathcal{H}$ with the inner product $\langle x, y \rangle_{\mathcal{H}}$. Then $(x - x_{\mathcal{M}}) \perp \mathcal{M}$, that is, for any $y \in \mathcal{M}$*

$$\boxed{\langle x - x_{\mathcal{M}}, y \rangle_{\mathcal{H}} = 0} \tag{18.21}$$

Proof. By the definition (18.20) for any $\lambda \in \mathbb{C}$ (here $x_{\mathcal{M}} + \lambda y \in \mathcal{M}$) we have

$$\|x - (x_{\mathcal{M}} + \lambda y)\| \geq \|x - x_{\mathcal{M}}\|$$

which implies

$$\lambda \langle x - x_{\mathcal{M}}, y \rangle_{\mathcal{H}} + \bar{\lambda} \langle y, x - x_{\mathcal{M}} \rangle_{\mathcal{H}} + \lambda\bar{\lambda} \|y\|^2 \geq 0$$

Taking $\lambda = -\dfrac{\langle x - x_{\mathcal{M}}, y \rangle_{\mathcal{H}}}{\|y\|^2}$ one has $-\dfrac{|\langle x - x_{\mathcal{M}}, y \rangle|^2}{\|y\|^2} \geq 0$ which leads to the equality $\langle x - x_{\mathcal{M}}, y \rangle_{\mathcal{H}} = 0$. Lemma is proven. □

Definition 18.6. *If $\mathcal{M}$ is a subspace of a Hilbert space $\mathcal{H}$ then the* ***orthogonal complement*** *$\mathcal{M}^{\perp}$ is defined by*

$$\boxed{\mathcal{M}^{\perp} := \{x \in \mathcal{H} \mid \langle x, y\rangle_{\mathcal{H}} = 0 \quad \text{for all} \quad y \in \mathcal{M}\}} \tag{18.22}$$

It is easy to show that $\mathcal{M}^{\perp}$ is a closed linear subspace of $\mathcal{H}$ and that $\mathcal{H}$ can be uniquely decomposed as the direct sum

$$\boxed{\mathcal{H} = \bar{\mathcal{M}} \oplus \mathcal{M}^{\perp}} \tag{18.23}$$

This means that any element $x \in \mathcal{H}$ has the unique representation

$$\boxed{x = x_{\bar{\mathcal{M}}} + x_{\mathcal{M}^{\perp}}} \tag{18.24}$$

where $x_{\mathcal{M}} \in \bar{\mathcal{M}}$ and $x_{\mathcal{M}^{\perp}} \in \mathcal{M}^{\perp}$ such that $\|x\|^2 = \|x_{\mathcal{M}}\|^2 + \|x_{\mathcal{M}^{\perp}}\|^2$.

Theorem 18.2. *Let $\mathcal{M}$ be a subspace of a Hilbert space $\mathcal{H}$. $\mathcal{M}$ is dense in $\mathcal{H}$ if and only if $\mathcal{M}^{\perp} = \{0\}$.*

Proof.

(a) *Necessity.* Let $\mathcal{M}$ be dense in $\mathcal{H}$. This means that $\bar{\mathcal{M}} = \mathcal{H}$. Assume that there exists $x_0 \in \mathcal{H}$ such that $x_0 \perp \mathcal{M}$. Let $\{y_n\} \subset \mathcal{M}$ and $y_n \to y \in \mathcal{H}$. Then $0 = \langle y_n, x_0\rangle \to \langle y, x_0\rangle = 0$ since $\mathcal{M}$ is dense in $\mathcal{H}$. Taking $y = x_0$ we get that $\langle x_0, x_0\rangle = 0$ which gives $x_0 = 0$.

(b) *Sufficiency.* Let $\mathcal{M}^{\perp} = \{0\}$, that is, if $\langle y, x_0\rangle = 0$ for any $y \in \mathcal{M}$, then $x_0 = 0$. Suppose that $\mathcal{M}$ is not dense in $\mathcal{H}$. This means that there exists $x_0 \notin \bar{\mathcal{M}}$. Then by the orthogonal decomposition $x_0 = y_0 + z_0$ where $y_0 \in \bar{\mathcal{M}}$ and $z_0 \in \left(\bar{\mathcal{M}}\right)^{\perp} = \mathcal{M}^{\perp}$. Here $z_0 \neq 0$ for which $\langle z_0, y\rangle_{\mathcal{H}} = 0$ for any $y \in \bar{\mathcal{M}}$. By the assumption such element $z_0 = 0$. We get the contradiction. Theorem is proven. □

18.3.3 Fourier series in Hilbert spaces

Definition 18.7. *An* ***orthonormal system (set)*** *$\{\phi_n\}$ of functions in a Hilbert space $\mathcal{H}$ is a nonempty subset $\{\phi_n \mid n \geq 1\}$ of $\mathcal{H}$ such that*

$$\boxed{\langle \phi_n, \phi_m\rangle_{\mathcal{H}} = \delta_{n,m} = \begin{cases} 1 & \text{if} \quad n = m \\ 0 & \text{if} \quad n \neq m \end{cases}} \tag{18.25}$$

1. *The series $\sum_{n=1}^{\infty} \alpha_k \phi_n$ is called the series in $\mathcal{H}$ with respect to the system $\{\phi_n\}$ (18.25);*
2. *For any $x \in \mathcal{H}$ the representation (if it exists)*

$$\boxed{x(t) = \sum_{n=1}^{\infty} \alpha_n \phi_n(t)} \tag{18.26}$$

is called the ***Fourier expansion*** *of x with respect to $\{\phi_n\}$.*

Lemma 18.2. *In (18.26)*

$$\boxed{\alpha_k = \langle x, \phi_k \rangle_{\mathcal{H}}} \tag{18.27}$$

Proof. Pre-multiplying (18.26) by ϕ_k and using (18.25) we find

$$\langle x, \phi_k \rangle_{\mathcal{H}} = \sum_{n=1}^{\infty} \alpha_n \langle \phi_k, \phi_n \rangle_{\mathcal{H}} = \sum_{k=1}^{\infty} \alpha_k \delta_{k,n} = \alpha_k$$

which proves (18.27). Lemma is proven. □

Corollary 18.4. (The Parseval equality)

$$\boxed{\|x\|^2 = \sum_{n=1}^{\infty} |\langle x, \phi_n \rangle_{\mathcal{H}}|^2} \tag{18.28}$$

Proof. It follows from the relation

$$\begin{aligned} \langle x, x \rangle_{\mathcal{H}} &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \langle x, \phi_n \rangle_{\mathcal{H}} \overline{\langle x, \phi_m \rangle_{\mathcal{H}}} \langle \phi_n, \phi_m \rangle_{\mathcal{H}} \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} \langle x, \phi_n \rangle_{\mathcal{H}} \overline{\langle x, \phi_m \rangle_{\mathcal{H}}} \delta_{n,m} = \sum_{n=1}^{\infty} |\langle x, \phi_n \rangle_{\mathcal{H}}|^2 \end{aligned}$$

□

Example 18.5.

1. ***Classical Fourier expansion.*** *In* $\mathcal{H} = L_2[0, 1]$ *the corresponding orthogonal basis* $\{\phi_n\}$ *is*

$$\boxed{\{\phi_n\} = \left\{1,\ \sqrt{2}\sin(2\pi nt),\ \sqrt{2}\cos(2\pi nt),\ n \geq 1\right\}}$$

which implies

$$x(t) = a_0 + \sqrt{2}\sum_{n=1}^{\infty} a_n \sin(2\pi nt) + \sqrt{2}\sum_{n=1}^{\infty} b_n \cos(2\pi nt)$$

where

$$a_0 = \int_{t=0}^{1} x(t)\, dt,\ a_n = \int_{t=0}^{1} x(t)\sqrt{2}\cos(2\pi nt)\, dt$$

$$b_n = \int_{t=0}^{1} x(t)\sqrt{2}\sin(2\pi nt)\, dt$$

2. ***Legendre expansion.*** *In* $\mathcal{H} = L_2[0, 1]$ *the corresponding orthogonal basis* $\{\phi_n\}$ *is* $\{\phi_n\} = \{p_n\}$ *where*

$$\boxed{p_k := \frac{1}{2^k k!} \frac{d^k}{dt^k} \left[\left(t^2 - 1\right)^k \right], \quad k \geq 1}$$

18.3.4 Linear ***n****-manifold approximation*

Definition 18.8. *The collection of the elements*

$$\boxed{u_n := \sum_{k=1}^{n} c_k \phi_k \in \mathcal{H}, \quad c_k \in \mathbb{C} \ (k \geq 1)} \tag{18.29}$$

is called the ***linear*** *n****-manifold*** *generated by the system of functions* $\{\phi_k\}_{k=\overline{1,n}}$.

Theorem 18.3. *The best* L_2*-approximation of any elements* $x \in \mathcal{H}$ *by the element* u_n *from the n-manifold (18.29) is given by the Fourier coefficients* $c_k = \alpha_k$ *(18.27), namely,*

$$\boxed{\inf_{c_k : k=\overline{1,n}} \left\| x - \sum_{k=1}^{n} c_k \phi_k \right\|_{L_2}^2 = \left\| x - \sum_{k=1}^{n} \alpha_k \phi_k \right\|_{L_2}^2} \tag{18.30}$$

Proof. It follows from the identity

$$\begin{aligned} \|x - u_n\|_{L_2}^2 &= \left\| x - \sum_{k=1}^{n} c_k \phi_k \right\|_{L_2}^2 \\ &= \left\| \sum_{n=1}^{\infty} \alpha_n \phi_n(t) - \sum_{k=1}^{n} c_k \phi_k \right\|_{L_2}^2 \\ &= \sum_{k=n+1}^{\infty} |\alpha_k|^2 \|\phi_k\|_{L_2}^2 + \sum_{k=1}^{n} |\alpha_k - c_k|^2 \|\phi_k\|_{L_2}^2 \end{aligned}$$

which reaches the minimum if $c_k = \alpha_k \ \left(c_k : k = \overline{1, n}\right)$. Theorem is proven. □

18.4 Linear operators and functionals in Banach spaces

18.4.1 Operators and functionals

Definition 18.9.

1. *Let* $\mathcal{X}$ *and* $\mathcal{Y}$ *be linear normed spaces (usually either Banach or Hilbert spaces) and* $T : \mathcal{D} \to \mathcal{Y}$ *be a* ***transformation*** *(or* ***operator****) from a subset* $\mathcal{D} \subset \mathcal{X}$ *to* $\mathcal{Y}$. $\mathcal{D} = \mathcal{D}(T)$ *is called the* ***domain*** *(image) of the operator T and values* $T(\mathcal{D})$ *constitute the* ***range*** *(the set of possible values)* $\mathcal{R}(T)$ *of T. If the range of the operator T is finite-dimensional then we say that the operator has* ***finite range****.*

2. *If $\mathcal{Y}$ is a scalar field $\mathcal{F}$ (usually $\mathbb{R}$) then the transformations T are called* ***functionals***.
3. *A functional T is* ***linear*** *if it is additive, i.e., for any $x, y \in \mathcal{D}$*

$$\boxed{T(x+y) = Tx + Ty}$$

and homogeneous, i.e., for any $x \in \mathcal{D}$ and any $\lambda \in \mathcal{F}$

$$\boxed{T(\lambda x) = \lambda T x}$$

4. *Operators for which the domain $\mathcal{D}$ and the range $T(\mathcal{D})$ are in one-to-one correspondence are called* ***invertible***. *The* ***inverse operator*** *is denoted by $T^{-1} : T(\mathcal{D}) \to \mathcal{D}$, so that*

$$\boxed{\mathcal{D} \supseteq T^{-1}(T(\mathcal{D}))}$$

Example 18.6.

1. *The* ***shift operator*** *$T_{sh} : l_p^n \to l_p^n$ defined by*

$$\boxed{T_{sh} x_i = x_{i+1}}$$

for any $i = 1, 2, \ldots$.

2. *The* ***integral operator*** *$T_g : L_2[a,b] \to \mathbb{R}$ defined by*

$$\boxed{T_g f := \int_{t=a}^{b} f(t)\, g(t)\, dt}$$

for any $f, g \in L_2[a,b]$.

3. *The* ***differential operator*** *$T_d : \mathcal{D}(T) = C^1[a,b] \to C[a,b]$ defined by*

$$\boxed{T_d f := \frac{d}{dt} f(t)}$$

for any $f \in C^1[a,b]$ and any $t \in [a,b]$.

It is evident in the following claim.

Claim 18.4.

1. *T is invertible if and only if it is* ***injective***, *that is, $Tx = 0$ implies $x = 0$. The set $\{x \in \mathcal{D} \mid Tx = 0\}$ is called the kernel of the operator and denoted by*

$$\boxed{\ker T := \{x \in \mathcal{D} \mid Tx = 0\}}$$

So, T is injective if and only if $\ker T = \{0\}$.
2. *If T is linear and invertible then T^{-1} is also linear.*

18.4.2 Continuity and boundedness

18.4.2.1 Continuity

Definition 18.10.

1. *Let $T : \mathcal{D}(T) \to \mathcal{Y}$ be a map (**operator**) between two linear normed spaces $\mathcal{X}$ (with a norm $\|\cdot\|_{\mathcal{X}}$) and $\mathcal{Y}$ (with a norm $\|\cdot\|_{\mathcal{Y}}$). It is said to be **continuous at** $x_0 \in \mathcal{X}$ if, given $\varepsilon > 0$, there exists a $\delta = \delta(\varepsilon) > 0$ such that $\|T(x) - T(x_0)\|_{\mathcal{Y}} < \varepsilon$, whenever $\|x - x_0\|_{\mathcal{X}} < \delta$.*
2. *T is **semi-continuous** at a point $x_0 \in \mathcal{X}$ if it transforms any convergent sequence $\{x_n\} \subset \mathcal{D}(T)$, $x_n \to x_0$, $n \to \infty$ into a sequence $\{T(x_n)\} \subset \mathcal{R}(T)$ weakly convergent to $T(x_0)$, i.e., $\|T(x_n) \to T(x_0)\| \to 0$ when $n \to \infty$.*
3. *T is **continuous** (or **semi-continuous**) **on** $\mathcal{D}(T)$ if it is continuous (or semi-continuous) at every point in $\mathcal{D}(T)$.*

Lemma 18.3. *Let $\mathcal{X}$ and $\mathcal{Y}$ be Banach spaces and A be a linear operator defined at $\mathcal{X}$. If A is **continuous** at the point $0 \in \mathcal{X}$, then A is continuous at any point $x_0 \in \mathcal{X}$.*

Proof. This result follows from the identity $Ax - Ax_0 = A(x - x_0)$. If $x \to x_0$, then $z := x - x_0 \to 0$. By continuity at zero $Az \to 0$ that implies $Ax - Ax_0 \to 0$. Lemma is proven. □

So, a linear operator A may be called *continuous* if it is continuous at the point $x_0 = 0$.

18.4.2.2 Boundedness

Definition 18.11.

1. *A linear operator $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ between two linear normed spaces $\mathcal{X}$ (with a norm $\|\cdot\|_{\mathcal{X}}$) and $\mathcal{Y}$ (with a norm $\|\cdot\|_{\mathcal{Y}}$) is said to be **bounded** if there exists a real number $c > 0$ such that for all $x \in \mathcal{D}(A)$*

$$\boxed{\|Ax\|_{\mathcal{Y}} \le c\,\|x\|_{\mathcal{X}}} \tag{18.31}$$

   *The **set of all bounded linear operators** $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ is usually denoted by $\mathcal{L}(\mathcal{X}, \mathcal{Y})$.*
2. *A linear operator $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ is called a **compact operator** if it maps any bounded subset of $\mathcal{X}$ onto a compact set of $\mathcal{Y}$.*
3. *The **induced norm** of a linear bounded operator $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ may be introduced as follows*

$$\boxed{\|A\| := \sup_{x \in \mathcal{D}(A),\, x \neq 0} \frac{\|Ax\|_{\mathcal{Y}}}{\|x\|_{\mathcal{X}}} = \sup_{x \in \mathcal{D}(A),\, \|x\|_{\mathcal{X}} = 1} \|Ax\|_{\mathcal{Y}}} \tag{18.32}$$

   (here it is assumed that if $\mathcal{D}(A) = \{0\}$ then by definition $\|A\| = 0$ since $A0 = 0$).

It seems to be evident that the continuity and boundedness for linear operators are equivalent concepts.

Claim 18.5. *A linear operator $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ is continuous if and only if it is bounded.*

Example 18.7.

1. *If* $\beta := \left(\sum_{i,j=1}^{\infty} |a_{ij}|^q \right)^{1/q} < \infty$ $(q > 1)$, *then the* ***"weighting" operator*** *A defined by*

$$\boxed{y = Ax_i := \sum_{j=1}^{\infty} a_{ij} x_j} \tag{18.33}$$

making from l_p to l_q $\left(p^{-1} + q^{-1} = 1\right)$ is linear and bounded since by the Hölder inequality (16.134)

$$\|Ax_i\|_{l_q}^q = \sum_{i=1}^{\infty} \left| \sum_{j=1}^{\infty} a_{ij} x_j \right|^q \leq \sum_{i=1}^{\infty} \left(\sum_{j=1}^{\infty} |a_{ij}|^q \right) \left(\sum_{j=1}^{\infty} |x_j|^p \right)^{q/p}$$

$$= \beta \left(\sum_{j=1}^{\infty} |x_j|^p \right)^{q/p} = \beta \|x\|_{l_p}^q$$

2. *If* $\beta := \int_{x=a}^{b} \int_{s=a}^{b} |K(x,s)|^q \, ds \, dx < \infty$, *then the* ***integral operator*** $A : \mathcal{X} = L_p[a,b] \to \mathcal{Y} = L_q[a,b] = \mathcal{Y}$ $\left(p^{-1} + q^{-1} = 1\right)$ *defined by*

$$\boxed{y = Af := \int_{s=a}^{b} K(x,s) f(s) \, ds} \tag{18.34}$$

is linear and bounded since by the Hölder inequality (16.134)

$$\|Af\|_{L_q[a,b]}^q := \int_{x=a}^{b} \left| \int_{s=a}^{b} K(x,s) f(s) \, ds \right|^q dx$$

$$\leq \int_{x=a}^{b} \left(\int_{s=a}^{b} |K(x,s)|^q \, ds \right) \left(\int_{s=a}^{b} |f(s)|^p \, ds \right)^{q/p} dx = \beta \|f\|_{L_p[a,b]}^q$$

3. *If* $\beta := \max_{t \in \mathcal{D}} \sum_{\alpha=0}^{l} |a_\alpha(t)| < \infty$, *then the* ***differential operator*** $A : \mathcal{D} \subset \mathcal{X} = C^k[a,b] \to \mathcal{Y} = C[a,b] = \mathcal{Y}$ *defined by*

$$\boxed{y = Af := \sum_{\alpha=0}^{l} a_\alpha(t) f^{(\alpha)}(t)} \tag{18.35}$$

is linear and bounded since

$$\|Af\|_{C[a,b]} := \max_{t\in\bar{\mathcal{D}}} \left| \sum_{\alpha=0}^{l} a_\alpha(t)\, f^{(\alpha)}(t) \right|$$
$$\leq \max_{t\in\bar{\mathcal{D}}} \left(\sum_{\alpha=0}^{l} |a_\alpha(t)| \sum_{\alpha=0}^{l} \left| f^{(\alpha)}(t) \right| \right) \leq \beta \|f\|_{C^l[a,b]}$$

18.4.2.3 Sequence of linear operators and uniform convergence

It is possible to introduce several different notions of a convergence in the space of linear bounded operators $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ acting from $\mathcal{X}$ to $\mathcal{Y}$.

Definition 18.12. *Let $\{A_n\} \subset \mathcal{L}(\mathcal{X}, \mathcal{Y})$ be a sequence of operators.*

1. *We say that*
   - *A_n **uniformly converges** to $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ if $\|A_n - A\| \to 0$ whenever $n \to \infty$. Here the norm $\|A_n - A\|$ is understood as in (18.32);*
   - *A_n **strongly converges** to $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ if $\|A_n f - Af\|_{\mathcal{Y}} \to 0$ whenever $n \to \infty$ for any $f \in \mathcal{X}$.*
2. *If the operator A is dependent on the parameter $\alpha \in \mathcal{A}$, then*
   - *$A(\alpha)$ is **uniformly continuous** at $\alpha_0 \in \mathcal{A}$, if*

$$\boxed{\|A(\alpha) - A(\alpha_0)\| \to 0 \quad as \quad \alpha \to \alpha_0}$$

   - *$A(\alpha)$ is **strongly continuous** at $\alpha_0 \in \mathcal{A}$, if for all $f \in \mathcal{X}$*

$$\boxed{\|A(\alpha) f - A(\alpha_0) f\|_{\mathcal{Y}} \to 0 \quad as \quad \alpha \to \alpha_0}$$

In view of this definition the following claim seems to be evident.

Claim 18.6. *A_n uniformly converges to $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ if and only if $A_n f \to Af$ uniformly on $f \in \mathcal{X}$ in the ball $\|f\|_{\mathcal{X}} \leq 1$.*

Theorem 18.4. *If $\mathcal{X}$ is a linear normed space and $\mathcal{Y}$ is a Banach space, then $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ is a Banach space too.*

Proof. Let $\{A_n\}$ be a fundamental sequence in the metric of $\mathcal{L}(\mathcal{X}, \mathcal{Y})$, that is, for any $\varepsilon > 0$ there exists a number $n_0 = n_0(\varepsilon)$ such that for any $n > n_0$ and any natural p we have $\|A_{n+p} - A\| < \varepsilon$. Then the sequence $\{A_n f\}$ is also fundamental. But $\mathcal{Y}$ is complete, and, hence, $\{A_n f\}$ converges. Denote $y := \lim_{n\to\infty} A_n f$. By this formula any element $f \in \mathcal{X}$ is mapped into an element of $\mathcal{Y}$, and, hence, it defines the operator $y = Af$. Let us prove that the linear operator A is bounded (continuous). First, notice that $\{\|A_n\|\}$ is also fundamental. This follows from the inequality $|\|A_{n+p}\| - \|A_n\|| \leq \|A_{n+p} - A_n\|$. But it means that $\{\|A_n\|\}$ is bounded, that is, there exists $c > 0$ such that $\|A_n\| \leq c$ for every $n \geq 1$. Hence, $\|A_n f\| \leq c\|f\|$. Taking the limit in the right-hand side we obtain $\|Af\| \leq c\|f\|$ which shows that A is bounded. Theorem is proven. □

18.4.2.4 Extension of linear bounded operators

Bounded linear operators that map into a Banach space always have a unique extension to the closure of their domain without changing of its norm value.

Theorem 18.5. *Let $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ be a linear bounded operator (functional) mapping the linear normed space $\mathcal{X}$ into a Banach space $\mathcal{Y}$. Then it has a unique bounded extension $\tilde{A} : \overline{\mathcal{D}(A)} \to \mathcal{Y}$ such that*

1. $\tilde{A}f = Af$ *for any* $f \in \mathcal{D}(A)$;
2. $\|\tilde{A}\| = \|A\|$.

Proof. If $f \in \mathcal{D}(A)$, put $\tilde{A}f = Af$. Let $f \in \mathcal{X}$, but $x \notin \mathcal{D}(A)$. By the density of $\mathcal{D}(A)$ in $\mathcal{X}$, there exists the sequence $\{f_n\} \subset \mathcal{D}(A)$ converging to x. Put $\tilde{A}f = \lim\limits_{n\to\infty} Af_n$. Let us show that this definition is correct, namely, that the limit exists and it does not depend on the selection of the convergent sequence $\{f_n\}$. The existence follows from the completeness property of $\mathcal{Y}$ since $\|Af_n - Af_m\| \le \|A\| \, \|f_n - f_m\|_{\mathcal{X}}$. Hence, $\lim\limits_{n\to\infty} Af_n$ exists. Supposing that there exists another sequence $\{f_n'\} \subset \mathcal{D}(A)$ converging to f we may denote $a := \lim\limits_{n\to\infty} Af_n$ and $b := \lim\limits_{n\to\infty} Af_n'$. Then we get

$$\|a - b\| \le \|a - Af_n\| + \|Af_n - Af_n'\| + \|Af_n' - b\| \underset{n\to\infty}{\to} 0$$

But $\|Af_n\| \le \|A\| \, \|f_n\|$ that for $n \to \infty$ implies $\|\tilde{A}f\| \le \|A\| \, \|f\|$, or, equivalently, $\|\tilde{A}\| \le \|A\|$. We also have $\|\tilde{A}\| := \sup\limits_{\|f\|_{\mathcal{X}} \le 1} \|\tilde{A}f\| \ge \sup\limits_{f\in\mathcal{D}(A),\ \|f\|_{\mathcal{X}} \le 1} \|\tilde{A}f\| = \|A\|$. So, we have $\|\tilde{A}\| = \|A\|$. The linearity property of $\tilde{A}$ follows from the linearity of A. Theorem is proven. □

Definition 18.13. *The operator $\tilde{A}$ constructed in Theorem 18.5 is called the* ***extension*** *of A to the closure $\overline{\mathcal{D}(A)}$ of its domain $\mathcal{D}(A)$ without increasing its norm.*

The principally more complex case arises when $\overline{\mathcal{D}(A)} = \mathcal{X}$. The following important theorem says that any linear bounded functional (operator) can be extended to the *whole* space $\mathcal{X}$ without increasing into a norm. A consequence of this result is the existence of nontrivial linear bounded functionals on any normed linear space.

Theorem 18.6. (The Hahn–Banach theorem) *Any linear bounded functional $A : \mathcal{D}(A) \subset \mathcal{X} \to \mathcal{Y}$ defined on a linear subspace $\mathcal{D}(A)$ of a linear normed space $\mathcal{X}$ can be extended to a linear bounded functional $\tilde{A}$ defined on the whole $\mathcal{X}$ with the preservation of the norm, i.e., $\tilde{A}f = Af$ for any $f \in \mathcal{D}(A)$ such that $\|\tilde{A}\| = \|A\|$.*

Proof. Here we present only the main idea of the proof.

(a) If $\mathcal{X}$ is separable, then the proof is based on Theorem 18.5 using the following lemma.

Lemma 18.4. *Let $\mathcal{X}$ be a real normed space and $\mathcal{L}$ a linear manifold in $\mathcal{X}$ where there is defined a linear functional A. If $f_0 \notin \mathcal{L}$ and $\mathcal{L}_1 := \{f + tf_0 \mid f \in \mathcal{L},\ t \in \mathbb{R}\}$ is a linear manifold containing all elements $f + tf_0$, then there exists a linear bounded functional*

A_1 defined on $\mathcal{L}_1$ such that it coincides with A on $\mathcal{L}$ and preserving the norm on $\mathcal{L}_1$, namely, $\|A_1\| = \|A\|$.

Then, since $\mathcal{X}$ is separable, there exists a basis $\{f_n\}_{n\geq 1}$ such that we can construct the sequence of s-manifolds

$$\mathcal{L}_{s\geq 1} := \left\{ \sum_{i=1}^{s} \lambda_i f_i \mid f_i \in \mathcal{X},\ \lambda_i \in \mathbb{R} \right\}$$

connected by $\mathcal{L}_{s+1} = \mathcal{L}_s + \{f_{n+1}\}$, $\mathcal{L}_0 := \varnothing$. Then we make the extension of A to each of the subspaces $\mathcal{L}_{s\geq 1}$ based on the lemma above. Finally we apply Theorem 18.5 to the space $\mathcal{X} = \bigcup_{s\geq 1} \mathcal{L}_s$ using the density property of $\mathcal{X}$.

(b) In general, the proof is based on Zorn's lemma (see Yoshida (1979)). □

Corollary 18.5. *Let $\mathcal{X}$ be a normed (topological) space and $x \in \mathcal{X}$, $x \neq 0$. Then there exists a linear bounded functional f, defined on $\mathcal{X}$, such that its value at any point x is equal to*

$$\boxed{f(x) := \langle x, f\rangle = \|x\|} \tag{18.36}$$

and

$$\boxed{\|f\| := \sup_{x\in D(f), \|x\|\leq 1} \langle x, f\rangle = 1} \tag{18.37}$$

Proof. Consider the linear manifold $\mathcal{L} := \{tx\}$, $t \in \mathbb{R}$ where we define f as follows: $\langle tx, f\rangle = t\,\|x\|$. So, we have $\langle x, f\rangle = \|x\|$. Then for any $y = tx$ it follows $|\langle y, f\rangle| = |t| \cdot \|x\| = \|tx\| = \|y\|$. This means that $\|f\| = 1$ and completes the proof. □

Corollary 18.6. *Let in a normed space $\mathcal{X}$ there be defined a linear manifold $\mathcal{L}$ and the element $x_0 \notin \mathcal{L}$ having the distance d up to this manifold, that is, $d := \inf_{x\in\mathcal{L}} \|x - x_0\|$. Then there exists a linear functional f defined on the whole $\mathcal{X}$ such that*

1. *$\langle x, f\rangle = 0$ for any $x \in \mathcal{L}$*
2. *$\langle x_0, f\rangle = 1$*
3. *$\|f\| = 1/d$*

Proof. Take $\mathcal{L}_1 := \mathcal{L} + \{x_0\}$. Then any element $y \in \mathcal{L}_1$ is uniquely defined by $y = x + tx_0$ where $x \in \mathcal{L}$ and $t \in \mathbb{R}$. Define on $\mathcal{L}_1$ the functional $f := t$. Now, if $y \in \mathcal{L}$, then $t = 0$ and $\langle y, f\rangle = 0$. So, statement 1 holds. If $y = x_0$, then $t = 1$ and, hence, $\langle x_0, f\rangle = 1$ which verifies statement 2. Finally,

$$|\langle y, f\rangle| = |t| = \frac{|t| \cdot \|y\|}{\|y\|} = \frac{\|y\|}{\left\|\frac{x}{t} + x_0\right\|} \leq \frac{\|y\|}{d}$$

which gives $\|f\| \leq 1/d$. On the other hand, by the "inf" definition, there exists a sequence $\{x_n\} \in \mathcal{L}$ such that $d = \lim_{n\to\infty} \|x_n - x_0\|$. This implies

$$1 = \langle x_0 - x_n, f\rangle \leq \|x_n - x_0\| \cdot \|f\|$$

Taking limit in the last inequality we obtain $1 \leq d\,\|f\|$ which gives $\|f\| \geq 1/d$. Combining both inequalities we conclude the statement 3. Corollary is proven. □

Corollary 18.7. *A linear manifold $\mathcal{L}$ is not* ***dense*** *in a Banach space $\mathcal{X}$ if and only if there exists a linear bounded functional $f \neq 0$ such that $\langle x, f\rangle = 0$ for any $x \in \mathcal{L}$.*

Proof.

(a) *Necessity.* Let $\bar{\mathcal{L}} \neq \mathcal{X}$. Then there exists a point $x_0 \in \mathcal{X}$ such that the distance between x_0 and $\mathcal{L}$ is positive, namely, $\rho(x_0, \mathcal{L}) = d > 0$. By Corollary 18.6 there exists f such that $\langle x_0, f\rangle = 1$, that is, $f \neq 0$ but $\langle x, f\rangle = 0$ for any $x \in \mathcal{L}$.

(b) *Sufficiency.* Let now $\bar{\mathcal{L}} = \mathcal{X}$. Then for any $x \in \mathcal{X}$, in view of the density property, there exists $\{x_n\} \in \mathcal{L}$ such that $x_n \to x$ when $n \to \infty$. By the condition that there exists $f \neq 0$, $= 0$ for any $x \in \mathcal{L}$, we have $\langle x, f\rangle = \lim_{n\to\infty} \langle x_n, f\rangle = 0$. Since x is arbitrary, it follows that $f = 0$. Contradiction. Corollary is proven.

□

Corollary 18.8. *Let $\{x_k\}_1^n$ be a system of linearly independent elements in a normed space $\mathcal{X}$. Then there exists a system of linear bounded functionals $\{f_l\}_1^n$, defined on the whole $\mathcal{X}$, such that*

$$\boxed{\langle x_k, f_l\rangle = \delta_{kl}\ (k, l = 1, \ldots, n)} \tag{18.38}$$

These two systems $\{x_k\}_1^n$ and $\{f_l\}_1^n$ are called ***bi-orthogonal.***

Proof. Take x_1 and denote by L_1 the linear span of the elements $x_2, \ldots, x_n$. By linear independency, it follows that $\rho(x_1, L_1) > 0$. By Corollary 18.6 we can find the linear bounded functional f_1 such that $\langle x_1, f_1\rangle = 1$, $\langle x, f_1\rangle = 0$ on L_1. Iterating this process we construct the desired system $\{f_l\}_1^n$. □

18.4.3 Compact operators

In this subsection we will consider a special subclass of bounded linear operators having properties rather similar to those enjoyed by operators on finite-dimensional spaces.

Definition 18.14. *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces. An operator $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ is said to be a* ***compact operator*** *if A maps a bounded set of $\mathcal{X}$ onto* ***relative compact sets*** *of $\mathcal{Y}$, that is, A is linear and for any bounded sequence $\{x_n\}$ in $\mathcal{X}$ the sequence $\{Ax_n\}$ has a convergence subsequence in $\mathcal{Y}$.*

Claim 18.7. *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed linear spaces and $A : \mathcal{X} \to \mathcal{Y}$ be a linear operator. Then the following assertions holds:*

(a) If A is bounded, that is, $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ and $\dim(Ax) < \infty$, then the operator A is compact.

(b) If $\dim(\mathcal{X}) < \infty$, then A is compact.

(c) The range of A is separable if A is compact.

(d) If $\{A_n\}$ is a sequence of compact operators from $\mathcal{X}$ to Banach space $\mathcal{Y}$ that converges uniformly to A, then A is a compact operator.

(e) The identity operator I on the Banach space $\mathcal{X}$ is compact if and only if $\dim(\mathcal{X}) < \infty$.

(f) If A is a compact operator in $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ whose range is a closed subspace of $\mathcal{Y}$, then the range of A is finite-dimensional.

Proof. It can be found in Rudin (1976) and Yoshida (1979). □

Example 18.8.

1. *Let $\mathcal{X} = l_2$ and $A : l_2 \to l_2$ be defined by $Ax := \left(x_1, \frac{x_2}{2}, \frac{x_3}{3}, \ldots\right)$. Then A is compact. Indeed, defining A_n by*

$$A_n x := \left(x_1, \frac{x_2}{2}, \frac{x_3}{3}, \ldots, \frac{x_n}{n}, 0, 0, \ldots\right)$$

we have

$$\|Ax - A_n x\|^2 = \sum_{k=n+1}^{\infty} \frac{1}{k^2} |x_i|^2 \leq \frac{\|x\|^2}{(n+1)^2}$$

and, hence, $\|A - A_n\| \leq (n+1)^{-1}$. This means that A_n converges uniformly to A and, by the previous claim (d), A is compact.

2. *Let $k(t,s) \in L_2([a,b] \times [a,b])$. Then the integral operator $K : L_2([a,b]) \to L_2([a,b])$ defined by $(Ku)(t) := \int_{s=a}^{b} k(t,s)\, u(s)\, ds$ is a compact operator (see Yoshida (1979)).*

Theorem 18.7. (Approximation theorem) *Let $\Phi : \mathcal{M} \subset \mathcal{X} \to \mathcal{Y}$ be a compact operator where $\mathcal{X}, \mathcal{Y}$ are Banach spaces and $\mathcal{M}$ is a bounded nonempty subset of $\mathcal{X}$. Then for every $n = 1, 2, \ldots$ there exists a continuous operator $\Phi_n : \mathcal{M} \to \mathcal{Y}$ such that*

$$\boxed{\sup_{x \in \mathcal{M}} \|\Phi(x) - \Phi_n(x)\| \leq n^{-1} \quad \text{and} \quad \dim(\operatorname{span} \Phi_n(\mathcal{M})) < \infty} \tag{18.39}$$

as well as $\Phi_n(\mathcal{M}) \subseteq \operatorname{co} \Phi(\mathcal{M})$ – the convex hull of $\Phi(\mathcal{M})$.

Proof. (see Zeidler (1995)). For every n there exists a finite $(2n)^{-1}$-net for $A(\mathcal{M})$ and elements $u_j \in \Phi(\mathcal{M})$ $(j = 1, \ldots, J)$ such that for all $x \in \mathcal{M}$

$$\min_{1 \leq j \leq J} \left\|\Phi(x) - u_j\right\| \leq (2n)^{-1}$$

Define for all $x \in \mathcal{M}$ the, so-called, *Schauder operator* A_n by

$$\Phi_n(x) := \sum_{j=1}^{J} a_j(x)\, u_j \left(\sum_{j=1}^{J} a_j(x) \right)^{-1}$$

where

$$a_j(x) := \max\left\{ n^{-1} - \left\| \Phi(x) - u_j \right\|; 0 \right\}$$

are continuous functions. In view of this A_n is also continuous and, moreover,

$$\begin{aligned} \|\Phi(x) - \Phi_n(x)\| = \Phi_n(x) &= \left\| \sum_{j=1}^{J} a_j(x) \left(u_j - \Phi_n(x)\right) \right\| \left(\sum_{j=1}^{J} a_j(x) \right)^{-1} \\ &\le \sum_{j=1}^{J} a_j(x) \left\| \left(u_j - \Phi_n(x)\right) \right\| \left(\sum_{j=1}^{J} a_j(x) \right)^{-1} \\ &\le \sum_{j=1}^{J} a_j(x)\, n^{-1} \left(\sum_{j=1}^{J} a_j(x) \right)^{-1} = n^{-1} \end{aligned}$$

Theorem is proven. □

18.4.4 Inverse operators

Many problems in the theory of ordinary and partial differential equations may be presented as a linear equation $Ax = y$ given in functional spaces $\mathcal{X}$ and $\mathcal{Y}$ where $A : \mathcal{X} \to \mathcal{Y}$ is a linear operator. If there exists the inverse operator $A^{-1} : \mathcal{R}(A) \to \mathcal{D}(A)$, then the solution of this linear equation may be formally represented as $x = A^{-1}y$. So, it seems to be very important to notice under which conditions the inverse operator exists.

18.4.4.1 Set of nulls and isomorphic operators

Let $A : \mathcal{X} \to \mathcal{Y}$ be a linear operator where $\mathcal{X}$ and $\mathcal{Y}$ are linear spaces such that $\mathcal{D}(A) \subseteq \mathcal{X}$ and $\mathcal{R}(A) \subseteq \mathcal{Y}$.

Definition 18.15. *The subset* $\mathcal{N}(A) \subseteq \mathcal{D}(A)$ *defined by*

$$\boxed{\mathcal{N}(A) := \{x \in \mathcal{D}(A) \mid Ax = 0\}} \tag{18.40}$$

is called the ***null space*** *of the operator* A.

Notice that

1. $\mathcal{N}(A) \neq \varnothing$ since $0 \in \mathcal{N}(A)$.
2. $\mathcal{N}(A)$ is a linear subspace (manifold).

Theorem 18.8. *An operator* A *is* ***isomorphic*** *(it transforms each point* $x \in \mathcal{D}(A)$ *only into a unique point* $y \in \mathcal{R}(A)$*) if and only if* $\mathcal{N}(A) = \{0\}$*, that is, when the set of nulls consists only of the single* 0*-element.*

Proof.

(a) *Necessity.* Let A be isomorphic. Suppose that $\mathcal{N}(A) \neq \{0\}$. Take $z \in \mathcal{N}(A)$ such that $z \neq 0$. Let also $y \in \mathcal{R}(A)$. Then the equation $Ax = y$ has a solution. Consider a point $x^* + z$. By lineality of A it follows $A(x^* + z) = y$. So, the element y has at least two different images x^* and $x^* + z$. We have obtained the contradiction to isomorphic property assumption.

(b) *Sufficiency.* Let $\mathcal{N}(A) = \{0\}$. But assume that there exist at least two $x_1, x_2 \in \mathcal{D}(A)$ such that $Ax_1 = Ax_2 = y$ and $x_1 \neq x_2$. The last implies $A(x_1 - x_2) = 0$. But this means that $(x_1 - x_2) \in \mathcal{N}(A) = \{0\}$, or, equivalently, $x_1 = x_2$. Contradiction. □

Claim 18.8. *Evidently,*

- *if a linear operator A is isomorphic then there exists the inverse operator A^{-1};*
- *the operator A^{-1} is a linear operator too.*

18.4.4.2 Bounded inverse operators

Theorem 18.9. *An operator A^{-1}* ***exists*** *and, simultaneously, is* ***bounded*** *if and only if the following inequality holds*

$$\boxed{\|Ax\| \geq m \|x\|} \tag{18.41}$$

for all $x \in \mathcal{D}(A)$ and some $m > 0$.

Proof.

(a) *Necessity.* Let A^{-1} exists and be bounded on $\mathcal{D}\left(A^{-1}\right) = \mathcal{R}(A)$. This means that there exists $c > 0$ such that for any $y \in \mathcal{R}(A)$ we have $\left\|A^{-1}y\right\| \leq c \|y\|$. Taking $y = Ax$ in the last inequality, we obtain (18.41).

(b) *Sufficiency.* Let now (18.41) hold. Then if $Ax = 0$ then by (18.41) we find that $x = 0$. This means that $\mathcal{N}(A) = \{0\}$ and by Theorem 18.8 it follows that A^{-1} exists. Then taking in (18.41) $x = A^{-1}y$ we get $\left\|A^{-1}y\right\| \leq m^{-1} \|y\|$ for all $y \in \mathcal{R}(A)$ which proves the boundedness of A^{-1}. □

Definition 18.16. *A linear operator $A : \mathcal{X} \to \mathcal{Y}$ is said to be* ***continuously invertible*** *if $\mathcal{R}(A) = \mathcal{Y}$, A is invertible and $A^{-1} \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ (that is, it is bounded).*

Theorem 18.9 may be reformulated in the following manner.

Theorem 18.10. *An operator A is* ***continuously invertible*** *if and only if $\mathcal{R}(A) = \mathcal{Y}$ and for some constant $m > 0$ the inequality (18.41) holds.*

It is not so difficult to prove the following result.

Theorem 18.11. (Banach) *If $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ (that is, A is linear bounded), $\mathcal{R}(A) = \mathcal{Y}$ and A is invertible, then it is continuously invertible.*

Example 18.9. *Let us consider in $C[0, 1]$ the following simplest integral equation*

$$(Ax)(t) := x(t) - \int_{s=0}^{1} tsx(s)\, ds = y(t) \tag{18.42}$$

The linear operator $A : C[0,1] \to C[0,1]$ *is defined by the left-hand side of (18.42). Notice that* $x(t) = y(t) + ct$, *where* $c = \int_{s=0}^{1} sx(s)\,ds$. *Integrating the equality* $sx(s) = sy(s) + cs^2$ *on* $[0,1]$, *we obtain* $c = \frac{3}{2}\int_{s=0}^{1} sy(s)\,ds$. *Hence, for any* $y(t)$ *in the right-hand side of (18.42) the solution is* $x(t) = y(t) + \frac{3}{2}\int_{s=0}^{1} tsy(s)\,ds := \left(A^{-1}y\right)(t)$. *Notice that* A^{-1} *is bounded, but this means by definition that the operator* A *is continuously invertible.*

Example 18.10. *Let* $y(t)$ *and* $a_i(t)$ $(i = 1, \ldots, n)$ *be continuous on* $[0, T]$. *Consider the following linear ordinary differential equation (ODE)*

$$(Ax)(t) := x^{(n)}(t) + a_1(t)\,x^{(n-1)}(t) + \ldots + a_n(t)\,x(t) = y(t) \tag{18.43}$$

under the initial conditions $x(0) = x'(0) = \ldots = x^{(n-1)}(0) = 0$ *and define the operator* A *as the left-hand side of (18.43) which is, evidently, linear with* $\mathcal{D}(A)$ *consisting of all functions which are n-times continuously differentiable, i.e.,* $x(t) \in C^n[0,T]$. *We will solve the Cauchy problem finding the corresponding* $x(t)$. *Let* $x_1(t), x_2(t), \ldots, x_n(t)$ *be the system of n linearly independent solutions of (18.43) when* $y(t) \equiv 0$. *Construct the, so-called, Wronsky's determinant*

$$W(t) := \begin{vmatrix} x_1(t) & \cdots & x_n(t) \\ x_1'(t) & \cdots & x_n'(t) \\ \vdots & \vdots & \vdots \\ x_1^{(n-1)}(t) & & x_n^{(n-1)}(t) \end{vmatrix}$$

It is well known (see, for example, El'sgol'ts (1961)) that $W(t) \neq 0$ *for all* $t \in [0,T]$. *According to the Lagrange approach dealing with the variation of arbitrary constants we may find the solution of (18.43) for any* $y(t)$ *in the form*

$$x(t) = c_1(t)\,x_1(t) + c_2(t)\,x_2(t) + \ldots + c_n(t)\,x_n(t)$$

which leads to the following ODE-system for $c_i(t)$ $(i = 1, \ldots, n)$:

$$\left.\begin{array}{l} c_1'(t)\,x_1(t) + c_2'(t)\,x_2(t) + \ldots + c_n'(t)\,x_n(t) = 0 \\ c_1'(t)\,x_1'(t) + c_2'(t)\,x_2'(t) + \ldots + c_n'(t)\,x_n'(t) = 0 \\ \ldots \\ c_1'(t)\,x_1^{(n-1)}(t) + c_2'(t)\,x_2^{(n-1)}(t) + \ldots + c_n'(t)\,x_n^{(n-1)}(t) = y(y) \end{array}\right\}$$

Resolving this system by Cramer's rule we derive $c_k'(t) = \dfrac{w_k(t)}{W(t)}\,y(t)$ $(k = 1, \ldots, n)$ *where* $w_k(t)$ *is the algebraic complement of the kth element of the last nth row. Taking into account the initial conditions we conclude that*

$$x(t) = \sum_{k=1}^{n} x_k(t) \int_{s=0}^{t} \frac{w_k(s)}{W(s)}\,y(s)\,ds := \left(A^{-1}y\right)(t)$$

which implies the following estimate $\|x\|_{C[0,T]} \leq c\,\|y\|_{C[0,T]}$ *with* $c = \max\limits_{t\in[0,T]} \sum_{k=1}^{n} |x_k(t)| \int_{s=0}^{t} \left|\frac{w_k(s)}{W(s)}\right| ds$ *that proves that the operator* A *is continuously invertible.*

18.4.4.3 Bounds for $\left\|A^{-1}\right\|$

Theorem 18.12. *Let* $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ *be a linear bounded operator such that* $\|I - A\| < 1$ *where* I *is the identical operator (which is, obviously, continuously invertible). Then* A *is continuously invertible and the following bounds hold:*

$$\boxed{\left\|A^{-1}\right\| \leq \frac{1}{1 - \|I - A\|}} \tag{18.44}$$

$$\boxed{\left\|I - A^{-1}\right\| \leq \frac{\|I - A\|}{1 - \|I - A\|}} \tag{18.45}$$

Proof. Consider in $\mathcal{L}(\mathcal{X}, \mathcal{Y})$ the series $(I + C + C^2 + \ldots)$ where $C := I - A$. Since $\left\|C^k\right\| \leq \|C\|^k$ this series uniformly converges (by the Weierstrass rule), i.e.,

$$S_n := I + C + C^2 + \ldots C^n \underset{n\to\infty}{\to} S$$

It is easy to check that

$$(I - C)\, S_n = I - C^{n+1}$$
$$S_n\,(I - C) = I - C^{n+1}$$
$$C^{n+1} \underset{n\to\infty}{\to} 0$$

Taking the limits in the last identities we obtain

$$(I - C)\, S = I,\ S\,(I - C) = I$$

which shows that the operator S is invertible and $S^{-1} = I - C = A$. So, $S = A^{-1}$ and

$$\|S_n\| \leq \|I\| + \|C\| + \|C\|^2 + \ldots \|C\|^n = \frac{1 - \|C\|^{n+1}}{1 - \|C\|}$$
$$\|I - S_n\| \leq \|C\| + \|C\|^2 + \ldots \|C\|^n = \frac{\|C\| - \|C\|^{n+1}}{1 - \|C\|}$$

Taking $n \to \infty$ we obtain (18.44) and (18.45). □

18.5 Duality

Let $\mathcal{X}$ be a linear normed space and $\mathcal{F}$ be the real axis $\mathbb{R}$, if $\mathcal{X}$ is real, and be the complex plane $\mathbb{C}$, if $\mathcal{X}$ is complex.

18.5.1 Dual spaces

Definition 18.17. *Consider the space $\mathcal{L}(\mathcal{X}, \mathcal{F})$ of all linear bounded functionals defined on $\mathcal{X}$. This space is called* ***dual to*** *$\mathcal{X}$ and is denoted by $\mathcal{X}^*$, so that*

$$\boxed{\mathcal{X}^* := \mathcal{L}(\mathcal{X}, \mathcal{F})} \tag{18.46}$$

The value of the linear functional $f \in \mathcal{X}^*$ on the element $x \in \mathcal{X}$ we will denote by $f(x)$, or $\langle x, f \rangle$, that is,

$$\boxed{f(x) = \langle x, f \rangle} \tag{18.47}$$

The notation $\langle x, f \rangle$ is analogous to the usual scalar product and turns out to be very useful in concrete calculations. In particular, the lineality of $\mathcal{X}$ and $\mathcal{X}^*$ implies the following identities (for any scalars $\alpha_1, \alpha_2, \beta_1, \beta_2$, any elements $x_1, x_2 \in \mathcal{X}$ and any functionals $f, f_1, f_2 \in \mathcal{X}^*$):

$$\boxed{\begin{array}{l} \langle \alpha_1 x_1 + \alpha_2 x_2, f \rangle = \alpha_1 \langle x_1, f \rangle + \alpha_2 \langle x_2, f \rangle \\ \langle x, \beta_1 f_1 + \beta_2 f_2 \rangle = \bar{\beta}_1 \langle x, f_1 \rangle + \bar{\beta}_2 \langle x, f_2 \rangle \end{array}} \tag{18.48}$$

($\bar{\beta}$ means the complex conjugated value to β. In a real case $\bar{\beta} = \beta$). If $\langle x, f \rangle = 0$ for any $x \in \mathcal{X}$, then $f = 0$. This property can be considered as the definition of the "null"-functional. Less trivial seems to be the next property.

Lemma 18.5. *If $\langle x, f \rangle = 0$ for any $f \in \mathcal{X}^*$, then $x = 0$.*

Proof. It is based on Corollary 18.5 of the Hahn–Banach theorem 18.6. Assuming the existence of $x \neq 0$, we can find $f \in \mathcal{X}^*$ such that $f \neq 0$ and $\langle x, f \rangle = \|x\| \neq 0$ which contradicts the identity $\langle x, f \rangle = 0$ valid for any $f \in \mathcal{X}^*$. So, $x = 0$. □

Definition 18.18. *In $\mathcal{X}^*$ one can introduce two types of convergence.*

- ***Strong convergence*** *(on the norm in $\mathcal{X}^*$):*
  $f_n \underset{n\to\infty}{\to} f (f_n, f \in \mathcal{X}^*)$, *if* $\|f_n - f\| \underset{n\to\infty}{\to} 0$.
- ***Weak convergence*** *(in the functional sense):*
  $f_n \overset{*}{\underset{n\to\infty}{\to}} f (f_n, f \in \mathcal{X}^*)$, *if for any $x \in \mathcal{X}$ one has*
  $\langle x, f_n \rangle \underset{n\to\infty}{\to} \langle x, f \rangle$.

Remark 18.4.

1. *Notice that the* ***strong convergence*** *of a functional sequence $\{f_n\}$* ***implies*** *its* ***weak convergence****.*
2. *(**Banach–Shteingauss**): $f_n \overset{*}{\underset{n\to\infty}{\to}} f$ if and only if*
   (a) $\{\|f_n\|\}$ is bounded;
   (b) $\langle x, f_n \rangle \underset{n\to\infty}{\to} \langle x, f \rangle$ on some dense linear manifold in $\mathcal{X}$.

Claim 18.9. *Independently of the fact whether the original topological space $\mathcal{X}$ is Banach or not,* ***the space $\mathcal{X}^* = \mathcal{L}(\mathcal{X}, \mathcal{F})$ of all linear bounded functionals is always Banach****.*

Proof. It can be easily seen from Definition 18.3. □

More exactly this statement can be formulated as follows.

Lemma 18.6. *$\mathcal{X}^*$ is a Banach space with the norm*

$$\boxed{\|f\| = \|f\|_{\mathcal{X}^*} := \sup_{x \in \mathcal{X},\ \|x\|_{\mathcal{X}} \leq 1} |f(x)|} \tag{18.49}$$

Furthermore, the following duality between two norms $\|\cdot\|_{\mathcal{X}}$ and $\|\cdot\|_{\mathcal{X}^*}$ takes place:

$$\boxed{\|x\|_{\mathcal{X}} = \sup_{f \in \mathcal{X}^*,\ \|f\|_{\mathcal{X}^*} \leq 1} |f(x)|} \tag{18.50}$$

Proof. The details of the proof can be found in Yoshida (1979). □

Example 18.11. *The spaces $L_p[a, b]$ and $L_q[a, b]$ are dual, that is,*

$$\boxed{L_p^*[a, b] = L_q[a, b]} \tag{18.51}$$

where $p^{-1} + q^{-1} = 1$, $1 < p < \infty$. Indeed, if $x(t) \in L_p[a, b]$ and $y(t) \in L_q[a, b]$, then the functional

$$\boxed{f(x) = \int_{t=a}^{b} x(t)\, y(t)\, dt} \tag{18.52}$$

is evidently linear, and boundedness follows from the Hölder inequality (16.137).

Since the dual space of a linear normed space is always a Banach space, one can consider the bounded linear functionals on $\mathcal{X}^*$, which we shall denote by $\mathcal{X}^{**}$. Moreover, each element $x \in \mathcal{X}$ gives rise to a bounded linear functional f^* in $\mathcal{X}^*$ by $f^*(f) = f(x)$, $f \in \mathcal{X}^*$. It can be shown that $\mathcal{X} \subset \mathcal{X}^{**}$, called the natural embedding of $\mathcal{X}$ into $\mathcal{X}^{**}$. Sometimes it happens that these spaces coincide. Notice that this is possible if $\mathcal{X}$ is a Banach space (since $\mathcal{X}^{**}$ is always Banach).

Definition 18.19. *If $\mathcal{X} = \mathcal{X}^{**}$, the Banach space $\mathcal{X}$ is called* ***reflexive****.*

Such spaces play an important role in different applications since they possess many properties resembling those in Hilbert spaces.

Claim 18.10. *Reflexive spaces are all Hilbert spaces, $\mathbb{R}^n$, l_p^n, and $L_{p>1}(\bar{G})$.*

Theorem 18.13. *The Banach space $\mathcal{X}$ is reflexive if and only if any bounded (by a norm) sequence of its elements contains a subsequence which weakly converges to some point in $\mathcal{X}$.*

Proof. See Trenogin (1980), Section 17.5, and Yoshida (1979) p. 264, the Eberlein–Shmulyan theorem. □

18.5.2 Adjoint (dual) and self-adjoint operators

Let $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces. Construct the linear functional $\varphi(x) = \langle x, \varphi \rangle := \langle Ax, f \rangle$ where $x \in \mathcal{X}$ and $f \in \mathcal{Y}^*$.

Lemma 18.7. *(1) $\mathcal{D}(\varphi) = \mathcal{X}$, (2) φ is a linear operator, (3) φ is bounded.*

Proof. (1) is evident. (2) is valid since

$$\begin{aligned}\varphi(\alpha_1 x_1 + \alpha_2 x_2) &= \langle A(\alpha_1 x_1 + \alpha_2 x_2), f \rangle \\ &= \alpha_1 \langle A(x_1), f \rangle + \alpha_2 \langle A(x_2), f \rangle \\ &= \alpha_1 \varphi(x_1) + \alpha_2 \varphi(x_2)\end{aligned}$$

And (3) holds since $|\varphi(x)| = |\langle Ax, f \rangle| \leq \|Ax\| \|f\| \leq \|A\| \|f\| \|x\|$. □

From this lemma it follows that $\varphi \in \mathcal{X}^*$. So, the linear continuous operator $\varphi = A^* f$ is correctly defined.

Definition 18.20. *The operator $A^* \in \mathcal{L}(\mathcal{Y}^*, \mathcal{X}^*)$ defined by*

$$\boxed{\langle x, A^* f \rangle := \langle Ax, f \rangle} \tag{18.53}$$

*is called the **adjoint** (or **dual**) **operator** of A.*

Lemma 18.8. *The representation $\langle Ax, f \rangle = \langle x, \varphi \rangle$ is unique ($\varphi \in \mathcal{X}^*$) for any $x \in \mathcal{D}(A)$ if and only if $\overline{\mathcal{D}(A)} = \mathcal{X}$.*

Proof.

(a) *Necessity.* Suppose $\overline{\mathcal{D}(A)} \neq \mathcal{X}$. Then by Corollary 18.7 from the Hahn–Banach theorem 18.6 there exists $\varphi_0 \in \mathcal{X}^*$, $\varphi_0 \neq 0$ such that $\langle x, \varphi_0 \rangle = 0$ for all $x \in \overline{\mathcal{D}(A)}$. But then $\langle Ax, f \rangle = \langle x, \varphi + \varphi_0 \rangle = 0$ for all $x \in \mathcal{D}(A)$ which contradicts the assumption of the uniqueness of the presentation.

(b) *Sufficiency.* Let $\overline{\mathcal{D}(A)} = \mathcal{X}$. If $\langle Ax, f \rangle = \langle x, \varphi_1 \rangle = \langle x, \varphi_2 \rangle$ then $\langle x, \varphi_1 - \varphi_2 \rangle = 0$ and by the same Corollary 18.7 it follows that $\varphi_1 - \varphi_2 = 0$ which means that the representation is unique.

□

Lemma 18.9. *If $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are Banach spaces, then $\|A^*\| = \|A\|$.*

Proof. By property (3) of the previous lemma we have $\|\varphi\| \leq \|A\| \|f\|$, i.e., $\|A^*\| \leq \|A\|$. But, by Corollary 18.5 from the Hahn–Banach theorem 18.6, for any x_0 such that $Ax_0 \neq 0$ there exists a functional $f_0 \in \mathcal{Y}^*$ such that $\|f_0\| = 1$ and $|\langle Ax_0, f_0 \rangle| = \|Ax_0\|$ which leads to the following estimate:

$$\|Ax_0\| = |\langle Ax_0, f_0 \rangle| = |\langle x_0, A^* f_0 \rangle| \leq \|A^*\| \|f_0\| \|x_0\| = \|A^*\| \|x_0\|$$

So, $\|A^*\| \geq \|A\|$ and, hence, $\|A^*\| = \|A\|$ which proves the lemma. □

Example 18.12. *Let $\mathcal{X} = \mathcal{Y} = \mathbb{R}^n$ be n-dimensional Euclidean spaces. Consider the linear operator*

$$\boxed{y = Ax \left(y_i := \sum_{k=1}^{n} a_{ik} x_k,\ i = 1, \ldots, n \right)} \tag{18.54}$$

Let $z \in (\mathbb{R}^n)^* = \mathbb{R}^n$. Since in Euclidean spaces the action of an operator is the corresponding scalar product, then $\langle Ax, z \rangle = (Ax, z) = (x, A^\top z) = \langle x, A^* z \rangle$. So,

$$\boxed{A^* = A^\top} \tag{18.55}$$

Example 18.13. *Let $\mathcal{X} = \mathcal{Y} = L_2[a, b]$. Let us consider the integral operator $y = Kx$ given by*

$$\boxed{y(t) = \int_{s=a}^{b} k(t, s)\, x(s)\, ds} \tag{18.56}$$

with the kernel $k(t, s)$ which is continuous on $[a, b] \times [a, b]$. We will consider the case when all variables are real. Then we have

$$\begin{aligned} \langle Kx, z \rangle &= \int_{t=a}^{b} \left(\int_{s=a}^{b} k(t, s)\, x(s)\, ds \right) z(t)\, dt \\ &= \int_{s=a}^{b} \left(\int_{t=a}^{b} k(t, s)\, z(t)\, dt \right) x(s)\, ds \\ &= \int_{t=a}^{b} \left(\int_{s=a}^{b} k(s, t)\, z(s)\, ds \right) x(t)\, dt = \langle x, K^* z \rangle \end{aligned}$$

which shows that the operator K^* $(\omega = K^* z)$ is defined by

$$\boxed{\omega(t) = \int_{s=a}^{b} k(s, t)\, z(s)\, ds} \tag{18.57}$$

that is, K^* is also integral with the kernel $k(s, t)$ which is inverse to the kernel $k(t, s)$ of K.

Definition 18.21. *The operator $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces, is said to be **self-adjoint** (or **Hermitian**) if $A^* = A$, that is, if it coincides with its adjoint (dual) form.*

Remark 18.5. *Evidently for self-adjoint operators $\mathcal{D}(A) = \mathcal{D}(A^*)$.*

Example 18.14.

1. *In $\mathbb{R}^n$, where any linear operator A is a matrix transformation, it will be self-adjoint if it is symmetric, i.e., $A = A^\top$, or, equivalently, $a_{ij} = a_{ji}$.*
2. *In $\mathbb{C}^n$, where any linear operator A is a complex matrix transformation, it will be self-adjoint if it is Hermitian, i.e., $A = A^*$, or, equivalently, $a_{ij} = \bar{a}_{ji}$.*
3. *The integral operator K in Example 18.13 is self-adjoint in $L_2[a, b]$ if its kernel is symmetric, namely, if $k(t, s) = k(s, t)$.*

It is easy to check the following simple properties of self-adjoint operators.

Proposition 18.1. *Let A and B be self-adjoint operators. Then*

1. *$(\alpha A + \beta B)$ is also self-adjoint for any real α and β.*
2. *(AB) is self-adjoint if and only if these two operators commute, i.e., if $AB = BA$. Indeed, $(ABx, f) = (Bx, Af) = (x, BAf)$.*
3. *The value (Ax, x) is always real for any $x \in \mathcal{F}$ (real or complex).*
4. *For any self-adjoint operator A we have*

$$\boxed{\|A\| = \sup_{\|x\| \le 1} |(Ax, x)|} \tag{18.58}$$

18.5.3 Riesz representation theorem for Hilbert spaces

Theorem 18.14. (F. Riesz) *If $\mathcal{H}$ is a Hilbert space (complex or real) with a scalar product $(\cdot, \cdot)$, then for any linear bounded functional f, defined on $\mathcal{H}$, there exists the unique element $y \in \mathcal{H}$ such that for all $x \in \mathcal{H}$ one has*

$$\boxed{f(x) = \langle x, f\rangle = (x, y)} \tag{18.59}$$

and, furthermore, $\|f\| = \|y\|$.

Proof. Let L be a subspace of $\mathcal{H}$. If $L = \mathcal{H}$, then for $f = 0$ one can take $y = 0$ and the theorem is proven. If $L \neq \mathcal{H}$, there exists $z_0 \perp L$, $z_0 \neq 0$ (it is sufficient to consider the case $f(z_0) = \langle z_0, f\rangle = 1$; if not, instead of z_0 we can consider $z_0/\langle z_0, f\rangle$). Let now $x \in \mathcal{H}$. Then $x - \langle x, f\rangle z_0 \in L$, since

$$\langle x - \langle x, f\rangle z_0, f\rangle = \langle x, f\rangle - \langle x, f\rangle \langle z_0, f\rangle = \langle x, f\rangle - \langle x, f\rangle = 0$$

Hence, $[x - \langle x, f\rangle z_0] \perp z_0$ which implies

$$0 = (x - \langle x, f\rangle z_0, z_0) = (x, z_0) - \langle x, f\rangle \|z_0\|^2$$

or, equivalently, $\langle x, f\rangle = \left(x, z_0/\|z_0\|^2\right)$. So, we can take $y = z_0/\|z_0\|^2$. Show now the uniqueness of y. If $\langle x, f\rangle = (x, y) = (x, \tilde{y})$, then $(x, y - \tilde{y}) = 0$ for any $x \in \mathcal{H}$. Taking $x = y - \tilde{y}$ we obtain $\|y - \tilde{y}\|^2 = 0$ which proves the identity $y = \tilde{y}$. To complete the proof we need to prove that $\|f\| = \|y\|$. By the Cauchy–Bounyakovski–Schwarz inequality $|\langle x, f\rangle| = |(x, y)| \le \|f\| \|y\|$. By the definition of the norm $\|f\|$ it follows that $\|f\| \le \|y\|$. On the other hand, $\langle y, f\rangle = (y, y) \le \|f\| \|y\|$ which leads to the inverse inequality $\|y\| \le \|f\|$. So, $\|f\| = \|y\|$. Theorem is proven. □

A different application of this theorem can be found in Riesz & Nagy (1978 (original in French, 1955)).

18.5.4 Orthogonal projection operators in Hilbert spaces

Let M be a subspace of a Hilbert space $\mathcal{H}$.

Definition 18.22. *The operator $P \in \mathcal{L}(\mathcal{H}, M)$ $(y = Px)$, acting in $\mathcal{H}$ such that*

$$y := \arg\min_{z \in M} \|x - z\| \tag{18.60}$$

*is called the **orthogonal projection operator** to the subspace M.*

Lemma 18.10. *The element $y = Px$ is unique and $(x - y, z) = 0$ for any $z \in M$.*

Proof. See subsection 18.3.2. □

The following evident properties of the projection operator hold.

Proposition 18.2.

1. *$x \in M$ if and only if $Px = x$.*
2. *Let $M^{\perp}$ be the orthogonal complement to M, that is,*

$$\boxed{M^{\perp} := \{z \in \mathcal{H} \mid z \perp M\}} \tag{18.61}$$

Then any $x \in \mathcal{H}$ can be represented as $x = y + z$ where $y \in M$ and $z \perp M$. Then the operator $P^{\perp} \in \mathcal{L}(\mathcal{H}, M^{\perp})$, defining the orthogonal projection any point from $\mathcal{H}$ to $M^{\perp}$, has the following representation:

$$\boxed{P^{\perp} = I - P} \tag{18.62}$$

3. *$x \in M^{\perp}$ if and only if $Px = 0$.*
4. *P is a linear operator, i.e., for any real α and β we have*

$$\boxed{P(\alpha x_1 + \beta x_2) = \alpha P(x_1) + \beta P(x_2)} \tag{18.63}$$

5.

$$\boxed{\|P\| = 1} \tag{18.64}$$

Indeed, $\|x\|^2 = \|Px + (I - P)x\|^2 = \|Px\|^2 + \|(I - P)x\|^2$ which implies $\|Px\|^2 \leq \|x\|^2$ and thus $\|P\| \leq 1$. On the other hand, if $M \neq \{0\}$, take $x_0 \in M$ with $\|x_0\| = 1$. Then $1 = \|x_0\| = \|Px_0\| \leq \|P\| \|x_0\| = \|P\|$. The inequalities $\|P\| \leq 1$ and $\|P\| \geq 1$ give (18.64).

6.

$$\boxed{P^2 = P} \tag{18.65}$$

since for any $x \in M$ *we have* $P^2(Px) = Px$.

7. P *is self-adjoint, that is,*

$$\boxed{P^* = P} \tag{18.66}$$

8. *For any* $x \in \mathcal{H}$

$$\boxed{(Px, x) = \left(P^2 x, x\right) = (Px, Px) = \|Px\|^2} \tag{18.67}$$

which implies

$$\boxed{(Px, x) \geq 0} \tag{18.68}$$

9. $\|Px\| = \|x\|$ *if and only if* $x \in M$.
10. *For any* $x \in \mathcal{H}$

$$\boxed{(Px, x) \leq \|x\|^2} \tag{18.69}$$

which follows from (18.67), the Cauchy–Bounyakovsk–Schwarz inequality and (18.64).

11. *Let* $A = A^* \in \mathcal{L}(\mathcal{H}, \mathcal{H})$ *and* $A^2 = A$. *Then* A *is obligatory an orthogonal projection operator to some subspace* $M = \{x \in \mathcal{H} \mid Ax = x\} \subset \mathcal{H}$. *Indeed, since* $x = Ax + (I - A)x$ *it follows that* $Ax = A^2 x = A(Ax) \in M$ *and* $(I - A)x \in M^{\perp}$.

The following lemma can be easily verified.

Lemma 18.11. *Let* P_1 *be the orthogonal projector to a subspace* M_1 *and* P_2 *be the orthogonal projector to a subspace* M_2. *Then the following four statements are equivalent:*

1.

$$\boxed{M_2 \subset M_1}$$

2.

$$\boxed{P_1 P_2 = P_2 P_1 = P_2}$$

3. $\|P_2 x\| \leq \|P_1 x\|$ *for any* $x \in \mathcal{H}$.
4. $(P_2 x, x) \leq (P_1 x, x)$ *for any* $x \in \mathcal{H}$.

Corollary 18.9.

1. $M_2 \perp M_1$ *if and only if* $P_1 P_2 = 0$.
2. $P_1 P_2$ *is a projector if and only if* $P_1 P_2 = P_2 P_1$.

3. *Let* P_i $(1 = 1, \ldots, N)$ *be a projection operator. Then* $\sum_{i=1}^{N} P_i$ *is a projection operator too if and only if*

$$\boxed{P_i P_k = \delta_{ik} P_i}$$

4. $P_1 - P_2$ *is a projection operator if and only if* $P_1 P_2 = P_2$, *or, equivalently, when* $P_1 \geq P_2$.

18.6 Monotonic, nonnegative and coercive operators

Remember the following elementary lemma from real analysis.

Lemma 18.12. *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a continuous function such that*

$$(x - y)[f(x) - f(y)] \geq 0 \tag{18.70}$$

for any $x, y \in \mathbb{R}$ *and*

$$xf(x) \to \infty \quad \textit{when} \quad |x| \to \infty \tag{18.71}$$

Then the equation $f(x) = 0$ *has a solution. If (17.80) holds in the strong sense, i.e.,*

$$(x - y)[f(x) - f(y)] > 0 \quad \textit{when} \quad x \neq y \tag{18.72}$$

then the equation $f(x) = 0$ *has a unique solution.*

Proof. For $x < y$ from (17.80) it follows that $f(x)$ is a nondecreasing function and, in view of (17.81), there exist numbers a and b such that $a < b$, $f(a) < 0$ and $f(b) < 0$. Then, considering $f(x)$ on $[a, b]$, by the theorem on intermediate values, there exists a point $\xi \in [a, b]$ such that $f(\xi) = 0$. If (18.72) is fulfilled, then $f(x)$ is a monotonically increasing function and the root of the function $f(x)$ is unique. □

The following definitions and theorems represent the generalization of this lemma to functional spaces and nonlinear operators.

18.6.1 Basic definitions and properties

Let $\mathcal{X}$ be a real separable normed space and $\mathcal{X}^*$ be a space dual to $\mathcal{X}$. Consider a nonlinear operator $T : \mathcal{X} \to \mathcal{X}^*$ $(\mathcal{D}(T) = \mathcal{X}, \mathcal{R}(T) \subset \mathcal{X}^*)$ and, as before, denoted by $f(x) = \langle x, f \rangle$ the value of the linear functional $f \in \mathcal{X}^*$ on the element $x \in \mathcal{X}$.

Definition 18.23.

1. *An operator* T *is said to be* ***monotone*** *if for any* $x, y \in \mathcal{D}(T)$

$$\boxed{\langle x - y, T(x) - T(y) \rangle \geq 0} \tag{18.73}$$

2. *It is called* ***strictly monotone*** *if for any* $x \neq y$

$$\boxed{\langle x - y, T(x) - T(y)\rangle > 0} \tag{18.74}$$

and the equality is possible only if $x = y$.

3. *It is called* ***strongly monotone*** *if for any* $x, y \in \mathcal{D}(T)$

$$\boxed{\langle x - y, T(x) - T(y)\rangle \geq \alpha(\|x - y\|)\|x - y\|} \tag{18.75}$$

where the nonnegative function $\alpha(t)$, *defined at* $t \geq 0$, *satisfies the condition* $\alpha(0) = 0$ *and* $\alpha(t) \to \infty$ *when* $t \to \infty$.

4. *An operator* T *is called* ***nonnegative*** *if for all* $x \in \mathcal{D}(T)$

$$\boxed{\langle x, T(x)\rangle \geq 0} \tag{18.76}$$

5. *An operator* T *is* ***positive*** *if for all* $x \in \mathcal{D}(T)$

$$\boxed{\langle x, T(x)\rangle > 0} \tag{18.77}$$

6. *An operator* T *is called* ***coercive*** *(or,* ***strongly positive****) if for all* $x \in \mathcal{D}(T)$

$$\boxed{\langle x, T(x)\rangle \geq \gamma(\|x\|)} \tag{18.78}$$

where function $\gamma(t)$, *defined at* $t \geq 0$, *satisfies the condition* $\gamma(t) \to \infty$ *when* $t \to \infty$.

Example 18.15. *The function* $f(x) = x^3 + x - 1$ *is the strictly monotone operator in* $\mathbb{R}$.

The following lemma installs the relation between monotonicity and coercivity properties.

Lemma 18.13. *If an operator* $T : \mathcal{X} \to \mathcal{X}^*$ *is* ***strongly monotone*** *then it is* ***coercive*** *with*

$$\boxed{\gamma(\|t\|) = \alpha(t) - \|T(0)\|} \tag{18.79}$$

Proof. By the definition (when $y = 0$) it follows that

$$\langle x, T(x) - T(0)\rangle \geq \alpha(\|x\|)\|x\|$$

This implies

$$\langle x, T(x)\rangle \geq \langle x, T(0)\rangle + \alpha(\|x\|)\|x\| \geq \alpha(\|x\|)\|x\| - \|x\|\,\|T(0)\| = [\alpha(\|x\|) - \|T(0)\|]\|x\|$$

which proves the lemma. □

Remark 18.6. *Notice that if an operator $T : \mathcal{X} \to \mathcal{X}^*$ is coercive then $\|T(x)\| \to \infty$ when $\|x\| \to \infty$. This follows from the inequalities*

$$\|T(x)\| \|x\| \geq \langle x, T(x)\rangle \geq \gamma(\|x\|)\|x\|$$

or, equivalently, from $\|T(x)\| \geq \gamma(\|x\|) \to \infty$ when $\|x\| \to \infty$.

The next theorem generalizes Lemma 18.12 to the nonlinear vector-function case.

Theorem 18.15. Trenogin (1980) *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a nonlinear operator (a vector function) which is continuous everywhere in $\mathbb{R}^n$ and such that for any $x, y \in \mathcal{X}$*

$$\boxed{\langle x - y, \quad T(x) - T(y)\rangle \geq c\|x - y\|^2, \quad c > 0} \tag{18.80}$$

(i.e., in (18.75) $\alpha(t) = ct$). Then the system of nonlinear equations

$$\boxed{T(x) = 0} \tag{18.81}$$

has a unique solution $x^* \in \mathbb{R}^n$.

Proof. Let us apply the induction method. For $n = 1$ the result is true by Lemma 18.12. Let it be true in $\mathbb{R}^{n-1}$ $(n \geq 2)$. Show that this result holds in $\mathbb{R}^n$. Consider in $\mathbb{R}^n$ a standard orthonormal basis $\{e_i\}_{i=1}^n$ $\left(e_i = (\delta_{ik})_{k=1}^n\right)$. Then $T(x)$ can be represented as $T(x) = \{T_i(x)\}_{i=1}^n$, $x = \sum_{j=1}^n x_j e_j$. For some fixed $t \in \mathbb{R}$ define the operator T_t by $T_t : \mathbb{R}^{n-1} \to \mathbb{R}^{n-1}$ for all $x = \sum_{j=1}^{n-1} x_j e_j$ acting as $T_t(x) := \{T_i(x + te_n)\}_{i=1}^{n-1}$. Evidently, $T_t(x)$ is continuous on $\mathbb{R}^{n-1}$ and, by the induction supposition, for any $x, y \in \mathbb{R}^{n-1}$ it satisfies the following inequality

$$\langle x - y, T_t(x) - T_t(y)\rangle = (t - t)\left[T_n(x + te_n) - T_n(y + te_n)\right]$$
$$+ \sum_{i=1}^{n-1} (x_i - y_i)\left[T_i(x + te_n) - T_i(y + te_n)\right] \geq c\|x - y\|^2$$

This means that the operator T_t also satisfies (18.80). By the induction supposition the system of nonlinear equations

$$T_i(x + te_n) = 0, \quad i = 1, \ldots, n - 1 \tag{18.82}$$

has a unique solution $\hat{x} \in \mathbb{R}^{n-1}$. This means exactly that there exists a vector-function $\hat{x} = \sum_{j=1}^{n-1} x_j e_j : \mathbb{R} \to \mathbb{R}^{n-1}$ which solves the system of nonlinear equations $T_i(x) = 0$. It is not difficult to check that the function $\hat{x} = \hat{x}(t)$ is continuous. Consider then the function $\psi(t) := T_n(x + te_n)$. It is also not difficult to check that this function satisfies all conditions of Lemma 18.12. Hence, there exists such $\tau \in \mathbb{R}$ that $\psi(\tau) = 0$. This exactly means that the equation (18.81) has a unique solution. □

The following proposition seems useful.

Theorem 18.16. Trenogin (1980) *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a continuous monotonic operator such that for all $x \in \mathbb{R}^n$ with $\|x\| > \lambda$ the following inequality holds:*

$$\boxed{(x, T(x)) \geq 0} \tag{18.83}$$

Then the equation $T(x) = 0$ has a solution x^* such that $\|x^*\| \leq \lambda$.

Proof. Consider the sequence $\{\varepsilon_k\}$, $0 < \varepsilon_k \underset{k\to\infty}{\to} 0$ and the associated sequence $\{T_k\}$, $T : \mathbb{R}^n \to \mathbb{R}^n$ of the operators defined by $T_k(x) := \varepsilon_k x + T(x)$. Then, in view of monotonicity of T, we have for all $x, y \in \mathbb{R}^n$

$$\begin{aligned}\langle x - y, T_k(x) - T_k(y)\rangle &= (x - y, T_k(x) - T_k(y)) \\ &= \varepsilon_k (x - y, (x - y) + (x - y, T(x) - T(y)) \geq \varepsilon_k \|x - y\|^2\end{aligned}$$

Hence, by Theorem 18.15 it follows that the equation $T_k(x) = 0$ has the unique solution x_k^* such that $\left\|x_k^*\right\| \leq \lambda$. Indeed, if not, we obtain the contradiction: $0 = \left(x_k^*, T_k(x_k^*)\right) \geq \varepsilon_k \left\|x_k^*\right\|^2 > 0$. Therefore, the sequence $\left\{x_k^*\right\} \subset \mathbb{R}^n$ is bounded. By the Bolzano–Weierstrass theorem there exists a subsequence $\left\{x_{k_t}^*\right\}$ convergent to some point $\bar{x} \subset \mathbb{R}^n$ when $k_t \to \infty$. This implies $0 = T_{k_t}(x_{k_t}^*) = \varepsilon_{x_{k_t}^*} x_{k_t}^* + T\left(x_{k_t}^*\right)$. Since $T(x)$ is continuous then when $k_t \to \infty$ we obtain $T(\bar{x})$. Theorem is proven. □

18.6.2 Galerkin method for equations with monotone operators

The technique given below presents the constructive method for finding an approximative solution of the operator equation $T(x) = 0$ where $T : \mathcal{X} \to \mathcal{X}^*$ ($\mathcal{D}(T) = \mathcal{X}$, $\mathcal{R}(T) \subset \mathcal{X}^*$). Let $\{\varphi_k\}_{k=1}^{\infty}$ be a complete sequence of linearly independent elements from $\mathcal{X}$, and $\mathcal{X}_n$ be a subspace spanned on $\varphi_1, \ldots, \varphi_n$.

Definition 18.24. *The element $x_n \in \mathcal{X}$ having the construction*

$$\boxed{x_n = \sum_{l=1}^{n} c_l \varphi_l} \tag{18.84}$$

*is said to be **the Galerkin approximation** to the solution of the equation $T(x) = 0$ with the monotone operator T if it satisfies the following system of equations*

$$\boxed{\langle \varphi_k, T(x_n)\rangle = 0, k = 1, \ldots, n} \tag{18.85}$$

or, equivalently,

$$\boxed{\sum_{l=1}^{n} \langle \varphi_l, T(x_n)\rangle \varphi_l = 0} \tag{18.86}$$

Remark 18.7.

1. *It is easy to prove that x_n is a solution of (18.85) if and only if $\langle u, T(x_n)\rangle = 0$ for any $u \in \mathcal{X}_n$.*

2. *The system (18.85) can be represented in the operator form* $J_n\bar{c}_n = x_n$ *where the operator* J_n *is defined by (18.85) with* $\bar{c}_n := (c_1, \ldots, c_n)$. *Notice that* $\|J_n\| \leq \sqrt{\sum_{l=1}^{n} \|\varphi_l\|^2}$. *In view of this, the equation (18.85) can be rewritten in the standard basis as*

$$\boxed{\langle \varphi_k, T(J_n\bar{c}_n)\rangle = 0, k = 1, \ldots, n} \tag{18.87}$$

Lemma 18.14. *If an operator* $T : \mathcal{X} \to \mathcal{X}^*$ $(\mathcal{D}(T) = \mathcal{X}, \mathcal{R}(T) \subset \mathcal{X}^*)$ *is strictly monotone (18.74) then*

1. *The equation (18.81) has a unique solution.*
2. *For any n the system (18.85) has a unique solution.*

Proof. If u and v are two solutions of (18.81), then $T(u) = T(v) = 0$, and, hence, $\langle x - y, T(u) - T(v)\rangle = 0$ which, in view of (18.74), takes place if and only if $u = v$. Again, if x_n' and x_n'' are two solutions of (18.85) then $\langle x_n', T(x_n'')\rangle = \langle x_n'', T(x_n')\rangle = \langle x_n', T(x_n')\rangle = \langle x_n'', T(x_n'')\rangle = 0$, or, equivalently,

$$\langle x_n' - x_n'', T(x_n') - T(x_n'')\rangle = 0$$

which, by (18.74), is possible if and only if $x_n' = x_n''$. □

Lemma 18.15. Trenogin (1980) *Let an operator* $T : \mathcal{X} \to \mathcal{X}^*$ $(\mathcal{D}(T) = \mathcal{X}, \mathcal{R}(T) \subset \mathcal{X}^*)$ *be monotone and semi-continuous, and there exists a constant* $\lambda > 0$ *such that for all* $x \in \mathcal{X}$ *with* $\|x\| > \lambda$ *we have* $\langle x, T(x)\rangle > 0$. *Then for any n the system (18.85) has the solution* $x_n \in \mathcal{X}$ *such that* $\|x_n\| \leq \lambda$.

Proof. It is sufficient to introduce in $\mathbb{R}^n$ the operator T_n defined by

$$T_n(\bar{c}_n) := \{\langle \varphi_k, T(J_n\bar{c}_n)\rangle\}_{k=1}^{n}$$

and to check that it satisfies all condition of Theorem 18.16. □

Based on these two lemmas it is possible to prove the following *main result on the Galerkin approximations.*

Proposition 18.3. Trenogin (1980) *Let the conditions of Lemma 18.15 be fulfilled and* $\{x_n\}$ *is the sequence of solutions of the system (18.85). Then the sequence* $\{T(x_n)\}$ *weakly converges to zero.*

18.6.3 Main theorems on the existence of solutions for equations with monotone operators

Theorem 18.17. Trenogin (1980) *Let* $T : \mathcal{X} \to \mathcal{X}^*$ $(\mathcal{D}(T) = \mathcal{X}, \mathcal{R}(T) \subset \mathcal{X}^*)$ *be an operator, acting from a real separable reflexive Banach space* $\mathcal{X}$ *into its dual space* $\mathcal{X}^*$, *which is monotone and semi-continuous. Let also there exist a constant* $\lambda > 0$ *such that for all* $x \in \mathcal{X}$ *with* $\|x\| > \lambda$ *we have* $\langle x, T(x)\rangle > 0$. *Then the equation* $T(x) = 0$ *has the solution* x^* *such that* $\|x^*\| \leq \lambda$.

Proof. By Lemma 18.15 for any n the Galerkin system (18.85) has the solution x_n such that $\|x_n\| \leq \lambda$. By reflexivity, from any sequence $\{x_n\}$ one can take out the subsequence $\{x_{n'}\}$ weakly convergent to some $x_0 \in \mathcal{X}$ such that $\|x_0\| \leq \lambda$. Then, by monotonicity of T, it follows that

$$S_{n'} := \langle x - x_{n'}, T(x) - T(x_{n'}) \rangle \geq 0$$

But $S_{n'} = \langle x - x_{n'}, T(x) \rangle - \langle x, T(x_{n'}) \rangle$, and, by Proposition 18.3, $\langle x, T(x_{n'}) \rangle \to 0$ weakly if $n' \to \infty$. Hence, $S_{n'} \to \langle x - x_0, T(x) \rangle$, and, therefore for all $x \in \mathcal{X}$

$$\langle x - x_0, T(x) \rangle \geq 0 \tag{18.88}$$

If $T(x_0) = 0$ then the theorem is proven. Let now $T(x_0) \neq 0$. Then, by Corollary 18.5 from the Hahn–Banach theorem 18.6 (for the case $\mathcal{X} = \mathcal{X}^{**}$), it follows the existence of the element $z_0 \in \mathcal{X}$ such that $\langle z_0, T(x_0) \rangle = \|T(x_0)\|$. Substitution of $x := x_0 - tz_0$ $(t > 0)$ into (18.88) implies $\langle z_0, T(x_0 - tz_0) \rangle \leq 0$, which for $t \to +0$ gives $\langle z_0, T(x_0) \rangle = \|T(x_0)\| \leq 0$. This is equivalent to the identity $T(x_0) = 0$. So, the assumption that $T(x_0) \neq 0$ is incorrect. Theorem is proven. □

Corollary 18.10. *Let an operator T be, additionally, coercive. Then the equation*

$$\boxed{T(x) = y} \tag{18.89}$$

has a solution for any $y \in \mathcal{X}^*$.

Proof. For any fixed $y \in \mathcal{X}^*$ define the operator $F(x) : \mathcal{X} \to \mathcal{X}^*$ acting as $F(x) := T(x) - y$. It is monotone and semi-continuous too. So, we have

$$\begin{aligned} \langle x, F(x) \rangle &= \langle x, T(x) \rangle - \langle x, y \rangle \geq \gamma(\|x\|)\|x\| - \|y\|\|x\| \\ &= [\gamma(\|x\|) - \|y\|]\|x\| \end{aligned}$$

and, therefore, there exists $\lambda > 0$ such that for all $x \in \mathcal{X}$ with $\|x\| > \lambda$ one has $\langle x, F(x) \rangle > 0$. Hence, the conditions of Theorem 18.17 hold which implies the existence of the solution for the equation $F(x) = 0$. □

Corollary 18.11. *If in Corollary 18.10 the operator is strictly monotone, then the solution of (18.89) is unique, i.e., there exists the operator T^{-1} inverse to T.*

Example 18.16. (Existence of the unique solution for ODE boundary problem) *Consider the following ODE boundary problem*

$$\left.\begin{aligned} &\mathcal{D}x(t) - f(t, x) = 0,\ t \in (a, b) \\ &\mathcal{D}x(t) := \sum_{l=1}^{m} (-1)^l D^l \{P_l(x) D^l x(t)\}, \\ &D := \frac{d}{dt} \text{ is the differentiation operator} \\ &D^k x(a) = D^k x(b) = 0,\ 0 \leq k \leq m - 1 \end{aligned}\right\} \tag{18.90}$$

in the Sobolev space $S_2^m(a,b)$ (18.9). Suppose that $f(t,x)$ for all x_1 and x_2 satisfies the condition

$$[f(t,x_1) - f(t,x_2)](x_1 - x_2) \geq 0$$

Let for the functions $P_l(x)$ the following additional condition be fulfilled for some $\alpha > 0$:

$$\int_{t=a}^{b} \left(\sum_{l=1}^{m} P_l(x) \left[D^l x(t) \right]^2 \right) dt \geq \alpha \|x\|_{S_2^m(a,b)}$$

Consider now in $S_2^m(a,b)$ the bilinear form

$$b(x,z) := \int_{t=a}^{b} \sum_{l=1}^{m} P_l(x) \left[D^l x(t) \right] \left[D^l z(t) \right] dt + \int_{t=a}^{b} f(t, x(t)) z(t)\, dt$$

defining in $S_2^m(a,b)$ the nonlinear operator

$$(T(x), z)_{S_2^m(a,b)} = b(x,z)$$

which is continuous and strongly monotone since

$$b(x_1, z) - b(x_2, z) \geq \alpha \|x_1 - x_2\|_{S_2^m(a,b)}$$

Then by Theorem 18.17 and Corollary 18.10 it follows that the problem (18.90) has the unique solution.

18.7 Differentiation of nonlinear operators

Consider a nonlinear operator $\Phi : \mathcal{X} \to \mathcal{Y}$ acting from a Banach space $\mathcal{X}$ to another Banach space $\mathcal{Y}$ and having a domain $\mathcal{D}(\Phi) \subset \mathcal{X}$ and a range $\mathcal{R}(T) \subset \mathcal{Y}$.

18.7.1 Fréchet derivative

Definition 18.25. *We say that an operator $\Phi : \mathcal{X} \to \mathcal{Y}$ ($\mathcal{D}(\Phi) \subset \mathcal{X}$, $\mathcal{R}(\Phi) \subset \mathcal{Y}$) acting in Banach spaces is **Fréchet differentiable** in a point $x_0 \in \mathcal{D}(\Phi)$, if there exists a linear bounded operator $\Phi'(x_0) \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$ such that*

$$\boxed{\begin{gathered} \Phi(x) - \Phi(x_0) = \Phi'(x_0)(x - x_0) + \omega(x - x_0) \\ \|\omega(x - x_0)\| = o(\|x - x_0\|) \end{gathered}} \tag{18.91}$$

or, equivalently,

$$\boxed{\lim_{x \to x_0} \frac{\Phi(x) - \Phi(x_0) - \langle x - x_0, \Phi'(x_0) \rangle}{\|x - x_0\|} = 0} \tag{18.92}$$

Definition 18.26. *If the operator* $\Phi : \mathcal{X} \to \mathcal{Y}$ $(\mathcal{D}(\Phi) \subset \mathcal{X}, \mathcal{R}(\Phi) \subset \mathcal{Y})$, *acting in Banach spaces, is* ***Fréchet-differentiable*** *in a point* $x_0 \in \mathcal{D}(\Phi)$ *the expression*

$$\boxed{d\Phi(x_0 \mid h) := \langle h, \Phi'(x_0) \rangle} \tag{18.93}$$

is called the ***Fréchet differential of the operator*** Φ *in the point* $x_0 \in \mathcal{D}(\Phi)$ *under the variation* $h \in \mathcal{X}$, *that is, the Fréchet differential of* Φ *in* x_0 *is nothing more than the value of the operator* $\Phi'(x_0)$ *at the element* $h \in \mathcal{X}$.

Remark 18.8. *If originally* $\Phi(x)$ *is a linear operator, namely, if* $\Phi(x) = Ax$ *where* $A \in \mathcal{L}(\mathcal{X}, \mathcal{Y})$, *then* $\Phi'(x_0) = A$ *in any point* $x_0 \in \mathcal{D}(A)$.

Several simple propositions follow from these definitions.

Proposition 18.4.

1. *If* $F, G : \mathcal{X} \to \mathcal{Y}$ *and both operators are Fréchet differentiable in* $x_0 \in \mathcal{X}$ *then*

$$\boxed{(F+G)'(x_0) = F'(x_0) + G'(x_0)} \tag{18.94}$$

and for any scalar α

$$\boxed{(\alpha F)'(x_0) = \alpha F'(x_0)} \tag{18.95}$$

2. *If* $F : \mathcal{X} \to \mathcal{Y}$ *is Fréchet-differentiable in* $x_0 \in \mathcal{D}(A)$ *and* $G : \mathcal{Z} \to \mathcal{X}$ *is Fréchet-differentiable in* $z_0 \in \mathcal{D}(G)$ *such that* $G(z_0) = x_0$ *then there is well-defined and continuous in the point* z_0 *the superposition* $(F \circ G)$ *of the operators* F *and* G, *namely,*

$$\boxed{F(G(z)) := (F \circ G)(z)} \tag{18.96}$$

and

$$\boxed{(F \circ G)'(z_0) = F'(x_0)\, G'(z_0)} \tag{18.97}$$

Example 18.17. *In finite-dimensional spaces* $F : \mathcal{X} = \mathbb{R}^k \to \mathcal{Y} = \mathbb{R}^l$ *and* $G : \mathcal{Z} = \mathbb{R}^m \to \mathcal{X} = \mathbb{R}^k$ *we have the systems of two algebraic nonlinear equations*

$$y = F(x), \quad x = G(z)$$

and, moreover,

$$F'(x_0) = A := \left\| \frac{\partial f_i(x_0)}{\partial x_j} \right\|_{i=1,\ldots,k;\, j=1,\ldots,l}$$

where A *is called the* ***Jacobi matrix****. Additionally, (18.97) is converted into the following representation:*

$$(F \circ G)'(z_0) = \left\| \sum_{j=1}^{k} \frac{\partial f_i(x_0)}{\partial x_j} \frac{\partial g_j(z_0)}{\partial z_s} \right\|_{i=1,\ldots,l;\, s=1,\ldots,m}$$

Example 18.18. *If F is the nonlinear integral operator acting in $C[a,b]$ and is defined by*

$$F(u) := u(x) - \int_{t=a}^{b} f(x,t,u(t))\,dt$$

then $F'(u_0)$ exists in any point $u_0 \in C[a,b]$ such that

$$F'(u_0)h = h(x) - \int_{t=a}^{b} \frac{\partial f(x,t,u_0(t))}{\partial u} h(t)\,dt$$

18.7.2 Gâteaux derivative

Definition 18.27. *If for any $h \in \mathcal{X}$ there exists the limit*

$$\boxed{\lim_{t\to+0} \frac{\Phi(x_0+th)-\Phi(x_0)}{t} = \delta\Phi(x_0 \mid h)} \tag{18.98}$$

*then the nonlinear operator $\delta\Phi(x_0 \mid h)$ is called the **first variation of the operator** $\Phi(x)$ in the point $x_0 \in \mathcal{X}$ at the direction h.*

Definition 18.28. *If in (18.98)*

$$\boxed{\delta\Phi(x_0 \mid h) = A_{x_0}(h) = \langle h, A_{x_0}\rangle} \tag{18.99}$$

*where $A_{x_0} \in \mathcal{L}(\mathcal{X},\mathcal{Y})$ is a linear bounded operator then Φ is **Gâteaux-differentiable** in a point $x_0 \in \mathcal{D}(\Phi)$ and the operator $A_{x_0} := \Phi'(x_0)$ is called the **Gâteaux derivative** of Φ in the point x_0 (independently on h). Moreover, the value*

$$\boxed{d\Phi(x_0 \mid h) := \langle h, A_{x_0}\rangle} \tag{18.100}$$

*is known as the **Gâteaux differential of** Φ in the point x_0 at the direction h.*

It is easy to check the following connections between the Gâteaux and Fréchet differentiability.

Proposition 18.5.

1. *Fréchet-differentiability **implies** Gâteaux-differentiability.*
2. *Gâteaux-differentiability **does not guarantee** Fréchet-differentiability. Indeed, for the function*

$$f(x,y) = \begin{cases} 1 & \text{if} \quad y = x^2 \\ 0 & \text{if} \quad y \neq x^2 \end{cases}$$

which, evidently, is not differentiable in the point $(0, 0)$ *in the Fréchet sense, the Gâteaux differential in the point* $(0, 0)$ *exists and is equal to zero since, in view of the properties* $f(0, 0) = 0$ *and* $f(th, tg) = 0$ *for any* (h, g), *we have* $\dfrac{f(th, tg) - f(0, 0)}{t} = 0$.

3. *The existence of the first variation* ***does not imply*** *the existence of the Gâteaux differential.*

18.7.3 Relation with "variation principle"

The main justification of the concept of differentiability is related to the optimization (or optimal control) theory in Banach spaces and is closely connected with the, so-called, *variation principle* which allows us to replace a minimization problem by an equivalent problem in which the loss function is linear.

Theorem 18.18. Aubin (1979) *Let* $\Phi : \mathcal{U} \to \mathcal{Y}$ *be a functional Gâteaux-differentiable on a convex subset* $\mathcal{X}$ *of a topological space* $\mathcal{U}$. *If* $x^* \in \mathcal{X}$ *minimizes* $\Phi(x)$ *on* $\mathcal{X}$ *then*

$$\boxed{\langle x^*, \Phi'(x^*)\rangle = \min_{x \in \mathcal{X}} \langle x, \Phi'(x^*)\rangle} \tag{18.101}$$

In particular, if x^* *is an interior point of* $\mathcal{X}$, *i.e.,* $x^* \in \operatorname{int} \mathcal{X}$, *then this condition implies*

$$\boxed{\Phi'(x^*) = 0} \tag{18.102}$$

Proof. Since $\mathcal{X}$ is convex then $\tilde{y} = x^* + \lambda(x - x^*) \in \mathcal{X}$ for any $\lambda \in (0, 1]$ whenever $x \in \mathcal{X}$. Therefore, since x^* is a minimizer of $\Phi(x)$ on $\mathcal{X}$, we have $\dfrac{\Phi(\tilde{y}) - \Phi(x^*)}{\lambda} \geq 0$. Taking the limit $\lambda \to +0$ we deduce from the Gâteaux-differentiability of $\Phi(x)$ on $\mathcal{X}$ that $\langle x - x^*, \Phi'(x^*)\rangle \geq 0$ for any $x \in \mathcal{X}$. In particular if $x^* \in \operatorname{int} \mathcal{X}$ then for any $y \in \mathcal{X}$ there exists $\varepsilon > 0$ such that $x = x^* + \varepsilon y \in \mathcal{X}$, and, hence, $\langle x - x^*, \Phi'(x^*)\rangle = \varepsilon \langle y, \Phi'(x^*)\rangle \geq 0$ which is possible for any $y \in \mathcal{X}$ if $\Phi'(x^*) = 0$. Theorem is proven. □

18.8 Fixed-point theorems

This section deals with the most important topics of functional analysis related with

- The existence principle;
- The convergence analysis.

18.8.1 Fixed points of a nonlinear operator

In this section we follow Trenogin (1980) and Zeidler (1995).

Let an operator $\Phi : \mathcal{X} \to \mathcal{Y}$ $(\mathcal{D}(\Phi) \subset \mathcal{X}, \mathcal{R}(\Phi) \subset \mathcal{Y})$ acts in Banach space $\mathcal{X}$. Suppose that the set $\mathcal{M}_\Phi := \mathcal{D}(\Phi) \cap \mathcal{R}(\Phi)$ is not empty.

Definition 18.29. *The point* $x^* \in \mathcal{M}_\Phi$ *is called a* ***fixed point*** *of the operator* Φ *if it satisfies the equality*

$$\boxed{\Phi(x^*) = x^*} \tag{18.103}$$

Remark 18.9. *Any operator equation (18.81): $T(x) = 0$ can be transformed to the form (18.103). Indeed, one has*

$$\tilde{T}(x) := T(x) + x = x$$

That's why all results, concerning the existence of the solution to the operator equation (18.81), can be considered as ones but with respect to the equation $\tilde{T}(x) = x$. The inverse statement is also true.

Example 18.19. *The fixed points of the operator $\Phi(x) = x^3$ are $\{0, -1, 1\}$ which follows from the relation $0 = x^3 - x = x\left(x^2 - 1\right) = x(x-1)(x+1)$.*

Example 18.20. *Let us try to find the fixed points of the operator*

$$\Phi(x) := \int_{s=0}^{1} x(t)\, x(s)\, ds + f(t) \tag{18.104}$$

assuming that it acts in $C[0, 1]$ (which is real) and that $\int_{t=0}^{1} f(t)\, dt \leq 1/4$. By the definition (18.103) we have $x(t) \int_{s=0}^{1} x(s)\, ds + f(t) = x(t)$. Integrating this equation leads to the following:

$$\left(\int_{t=0}^{1} x(t)\, dt\right)^2 + \int_{t=0}^{1} f(t)\, dt = \int_{t=0}^{1} x(t)\, dt$$

which gives

$$\int_{t=0}^{1} x(t)\, dt = \frac{1}{2} \pm \sqrt{\frac{1}{4} - \int_{t=0}^{1} f(t)\, dt} \tag{18.105}$$

So, any function $x(t) \in C[0, 1]$ satisfying (18.105) is a fixed point of the operator (18.104).

The main results related to the existence of the solution of the operator equation

$$\boxed{\Phi(x) = x} \tag{18.106}$$

are as follows:

- The **contraction principle** (see (14.17)) or **the Banach theorem** (1920) which states that *if the operator $\Phi : X \to X$ (X is a compact) is k-contractive, i.e., for all $x, x' \in X$*

$$\left\|\Phi(x) - \Phi\left(x'\right)\right\| \leq k \left\|x - x'\right\|, k \in [0, 1)$$

then

(a) *the solution of (18.106) exists and is unique;*

(b) *the iterative method* $x_{n+1} = \Phi(x_n)$ *exponentially converges to this solution.*

- The **Brouwer fixed-point theorem** for finite-dimensional Banach space.
- The **Schauder fixed-point theorem** for infinite-dimensional Banach space.
- **The Leray–Schauder principle** which states that *a priori estimates yield existence.*

There are many other versions of these fixed-point theorems such as Kakutani, Ky-Fan etc. related some generalizations of the theorems mentioned above. For details see Aubin (1979) and Zeidler (1986).

18.8.2 Brouwer fixed-point theorem

To deal correctly with the Brouwer fixed-point theorem we need the preparations considered below.

18.8.2.1 The Sperner lemma

Let

$$S_N(x_0, \ldots, x_N) := \left\{ x \in \mathcal{X} \mid x = \sum_{i=0}^{N} \lambda_i x_i, \quad \lambda_i \geq 0, \quad \sum_{i=0}^{N} \lambda_i = 1 \right\} \tag{18.107}$$

be an N-simplex in a finite-dimensional normed space $\mathcal{X}$ and $\{S_1, \ldots, S_J\}$ be a *triangulation* of S_N consisting of N-simplices S_j $(j = 1, \ldots, J)$ (see Fig. 18.1) such that

(a) $S_N = \bigcup_{j=1}^{J} S_j$;

(b) if $j \neq k$, then the intersection $S_j \bigcap_{j \neq k} S_k$ or is empty or a common face of dimension less than N.

Let one of the numbers $(0, 1, \ldots, N)$ be associated with each vertex v of the simplex S_j. So, suppose that if $v \in S_j := S_j(x_{i_0}, \ldots, x_{i_N})$, then one of the numbers $i_0, \ldots, i_N$ is associated with v.

Definition 18.30. *S_j is called a **Sperner simplex** if and only if all of its vertices carry different numbers, i.e., the vertices of S_N carry different numbers* $0, 1, \ldots, N$.

Lemma 18.16. (Sperner) *The number of Sperner simplices is always odd.*

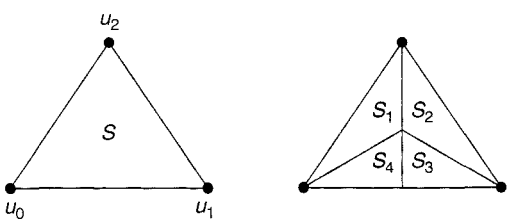

Fig. 18.1. N-simplex and its triangulation.

Proof. It can be easily proven by induction that for $N = 1$ each S_j is a 1-simplex (segment). In this case a 0-face (vertex) of S_j is called distinguished if and only if it carries the number 0. So, one has exactly two possibilities (see Fig. 18.2a): (i) S_j has precisely one distinguished $(N-1)$-faces, i.e., S_j is a Sperner simplex; (ii) S_j has precisely two or more distinguished $(N-1)$-faces, i.e., S_j is not a Sperner simplex. But since the distinguished 0-face occurs twice in the interior and once on the boundary, the total number of distinguished 0-faces is odd. Hence, the number of Sperner simplices is odd. Let now $N = 2$ (see Fig. 18.2b). Then S_j is 2-simplex and a 1-face (segment) of S_j is called distinguished if and only if it carries the numbers 0, 1. Conditions (i) and (ii) given above are satisfied for $N = 2$. The distinguished 1-faces occur twice in the interior and, by the case $N = 1$, it follows that the number of the distinguished 1-faces is odd. Therefore, the number of Sperner simplices is odd. Now let $N \geq 3$. Supposing that the lemma is true for $(N-1)$, as in the case $N = 2$, we easily obtain the result. □

18.8.2.2 The Knaster–Kuratowski–Mazurkiewicz (KKM) lemma

Lemma 18.17. (Knaster–Kuratowski–Mazurkiewicz) *Let $S_N(x_0, \ldots, x_N)$ be an N-simplex in a finite-dimensional normed space $\mathcal{X}$. Suppose we are given closed sets $\{C_i\}_{i=1}^N$ in $\mathcal{X}$ such that*

$$S_N(x_0, \ldots, x_N) \subseteq \bigcup_{m=0}^{k} C_{i_m} \tag{18.108}$$

for all possible systems of indices $\{i_0, \ldots, i_k\}$ and all $k = 0, \ldots, N$. Then there exists a point $v \in S_N(x_0, \ldots, x_N)$ such that $v \in C_j$ for all $j = 0, \ldots, N$.

Proof. Since for $N = 0$ the set $S_0(x_0)$ consists of a single point x_0, and the statement looks trivial. Let $N \geq 1$. Let v be any vertex of S_j $(j = 0, \ldots, N)$ (for a triangulation $S_1, \ldots, S_N$) such that $v \in S_j(x_{i_0}, \ldots, x_{i_N})$. By the assumptions of this lemma there exists a set C_k such that $v \in C_k$. We may associate the index k with the vertex v. By the Sperner lemma 18.16 it follows that there exists a Sperner simplex S_j whose vertices carry the numbers $0, \ldots, N$. Hence the vertices $v_0, \ldots, v_N$ satisfy the condition $v_k \in C_k$ $(k = 0, \ldots, N)$. Consider now a sequence of triangulations of simplex $S_N(x_0, \ldots, x_N)$ such that the diameters of the simplices of the triangulations tend to zero

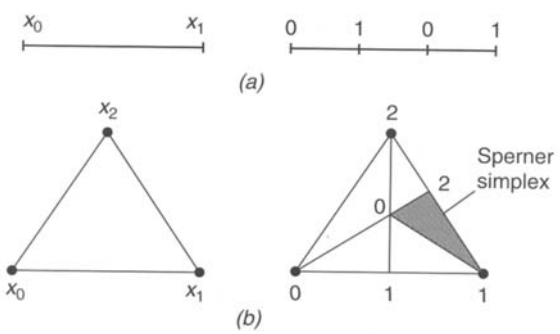

Fig. 18.2. The Sperner simplex.

(selecting, for example, a sequence barycentric subdivision of S). So, there are points $v_k^{(n)} \in C_k$ $(k = 0, \ldots, N;\ n = 1, 2, \ldots)$ such that $\lim\limits_{n\to\infty} \operatorname{diam} S_N\left(v_0^{(n)}, \ldots, v_N^{(n)}\right) = 0$. Since the simplex $S_N(x_0, \ldots, x_N)$ is a compact, there exists a subsequence $\left\{v_k^{(s)}\right\}$ such that $v_k^{(s)} \underset{s\to\infty}{\to} v \in S_N(x_0, \ldots, x_N)$ for all $k = 0, \ldots, N$. And since the set C_k is closed, this implies $v \in C_k$ for all $k = 0, \ldots, N$. Lemma is proven. □

Now we are ready to formulate the main result of this section.

18.8.2.3 The Brouwer theorem

Theorem 18.19. (Brouwer 1912) *The continuous operator $\Phi : \mathcal{M} \to \mathcal{M}$ has at least one fixed point when $\mathcal{M}$ is a compact, convex, nonempty set in a finite-dimensional normed space over the field $\mathcal{F}$ (real or complex).*

Proof.

(a) Consider this operator when $\mathcal{M} = S_N$ and demonstrate that the continuous operator $\Phi : S_N \to S_N$ $(N = 0, 1, \ldots)$ has at least one fixed point when $S_N = S_N(x_0, \ldots, x_N)$ is an N-simplex in a finite-dimensional normed space $\mathcal{X}$. For $N = 0$ the set S_0 consists of a single point and the statement is trivial. For $N = 1$ the statement is also trivial. Let now $N = 2$. Then $S_2 = S_2(x_0, x_1, x_2)$ and any point x in S_2 can be represented as

$$x = \sum_{i=0}^{2} \lambda_i(x)\, x_i,\ \lambda_i \geq 0,\ \sum_{i=0}^{2} \lambda_i = 1 \tag{18.109}$$

We set

$$C_i := \{x \in S_N \mid \lambda_i(\Phi x) \leq \lambda_i(x),\ i = 0, 1, 2\}$$

Since $\lambda_i(x)$ and Φ are continuous on S_N, the sets C_i are closed and the condition (18.108) of Lemma 18.17 is fulfilled, that is, $S_N \in \bigcup\limits_{m=0}^{k} C_{i_m}$ $(k = 0, 1, 2)$. Indeed, if it is not true, then there exists a point $x \in S_2\left(x_{i_0}, x_{i_1}, x_{i_2}\right)$ such that $x \notin \bigcup\limits_{m=0}^{k} C_{i_m}$, i.e., $\lambda_{i_m}(\Phi x) > \lambda_{i_m}(x)$ for all $m = 0, \ldots, k$. But this is in contradiction to the representation (18.109). Then by Lemma 18.17 there is a point $y \in S_2$ such that $y \in C_j$ $(j = 0, 1, 2)$. This implies $\lambda_i(\Phi y) \leq \lambda_i(y)$ for all $j = 0, 1, 2$. Since also $\Phi y \in S_2$ we have

$$\sum_{i=0}^{2} \lambda_i(y) = \sum_{i=0}^{2} \lambda_i(\Phi y) = 1$$

and, hence, $\lambda_i(\Phi y) = \lambda_i(y)$ for all $j = 0, 1, 2$ which is equivalent to the expression $\Phi y = y$. So, y is the desired fixed point of Φ in the case $N = 2$. In $N \geq 3$ one can use the same arguments as for $N = 2$.

(b) Now, when $\mathcal{M}$ is a compact, convex, nonempty set in a finite-dimensional normed space, it is easy to show that $\mathcal{M}$ is homeomorphic to some N-simplex

$(N = 0, 1, 2, \ldots)$. This means that there exist homeomorphisms $\Phi : \mathcal{M} \to \mathcal{B}$ and $C : S_N \to \mathcal{B}$ such that the map

$$C^{-1} \circ \Phi : \mathcal{M} \overset{\Phi}{\to} \mathcal{B} \overset{C^{-1}}{\to} S_N$$

is the desired homeomorphism from the given set $\mathcal{M}$ onto the simplex S_N. Using now this fact shows that each continuous operator $\Phi : \mathcal{M} \to \mathcal{M}$ has at least one fixed point. This completes the proof. □

Corollary 18.12. *The continuous operator $B : \mathcal{K} \to \mathcal{K}$ has at least a fixed point when $\mathcal{K}$ is a subset of a normed space that is homeomorphic to a set $\mathcal{M}$ as it is considered in Theorem 18.19.*

Proof. Let $C : \mathcal{M} \to \mathcal{K}$ be a homeomorphism. Then the operator

$$C^{-1} \circ B \circ C : \mathcal{M} \overset{C}{\to} \mathcal{K} \overset{B}{\to} \mathcal{K} \overset{C^{-1}}{\to} \mathcal{M}$$

is continuous. By Theorem 18.19 there exists a fixed point x^* of the operator $\Phi := C^{-1} \circ B \circ C$, i.e., $C^{-1}(B(Cx^*)) = x^*$. Let $y = Cx$. Then $By = y$, $y \in \mathcal{K}$. Therefore B has a fixed point. Corollary is proven. □

18.8.3 Schauder fixed-point theorem

This result represents the extension of the Brouwer fixed-point theorem 18.19 to a infinite-dimensional Banach space.

Theorem 18.20. (Schauder 1930) *The compact operator $\Phi : \mathcal{M} \to \mathcal{M}$ has at least one fixed point when $\mathcal{M}$ is a bounded, closed convex, nonempty subset of a Banach space $\mathcal{X}$ over the field $\mathcal{F}$ (real or complex).*

Proof. **Zeidler (1995)** Let $x \in \mathcal{M}$. Replacing x with $x - x_0$, if necessary, one may assume that $0 \in \mathcal{M}$. By Theorem 18.7 on the approximation of compact operators it follows that for every $n = 1, 2, \ldots$ there exists a finite-dimensional subspace $\mathcal{X}_n$ of $\mathcal{X}$ and a continuous operator $\Phi_n : \mathcal{M} \to \mathcal{X}_n$ such that $\|\Phi_n(x) - \Phi_n(x)\| \leq n^{-1}$ for any $x \in \mathcal{M}$. Define $\mathcal{M}_n := \mathcal{M} \cap \mathcal{X}_n$. Then $\mathcal{M}_n$ is a bounded, closed, convex subset of $\mathcal{X}_n$ with $0 \in \mathcal{M}_n$ and $\Phi_n(\mathcal{M}) \subseteq \mathcal{M}$ since $\mathcal{M}$ is convex. By the Brouwer fixed-point theorem 18.19 the operator $\Phi_n : \mathcal{M}_n \to \mathcal{M}_n$ has a fixed point, say x_n, that is, for all $n = 1, 2, \ldots$ we have $\Phi_n(x_n) = x_n \in \mathcal{M}_n$. Moreover, $\|\Phi(x_n) - x_n\| \leq n^{-1}$. Since $\mathcal{M}_n \subseteq \mathcal{M}$, the sequence $\{x_n\}$ is bounded. The compactness of $\Phi : \mathcal{M} \to \mathcal{M}$ implies the existence of a sequence $\{\tilde{x}_n\}$ such that $\Phi(x_n) \to v$ when $n \to \infty$. By the previous estimate

$$\begin{aligned} \|v - x_n\| &= \big\|[v - \Phi(x_n)] + [\Phi(x_n) - x_n]\big\| \\ &\leq \big\|[v - \Phi(x_n)]\big\| + \|\Phi(x_n) - x_n\| \underset{n\to\infty}{\to} 0 \end{aligned}$$

So, $x_n \to v$. Since $\Phi(x_n) \in \mathcal{M}$ and the set $\mathcal{M}$ is closed, we get that $v \in \mathcal{M}$. And, finally, since the operator $\Phi : \mathcal{M} \to \mathcal{M}$ is continuous, it follows that $\Phi(x) = x \in \mathcal{M}$. Theorem is proven. □

Example 18.21. (Existence of solution for integral equations) *Let us solve the following integral equation*

$$\boxed{\begin{gathered} u(t) = \lambda \int_{y=a}^{b} F(t, y, u(y))\, dy \\ -\infty < a < b < \infty, \quad t \in [a, b], \quad \lambda \in \mathbb{R} \end{gathered}} \tag{18.110}$$

Define

$$Q_r := \left\{(t, y, u) \in \mathbb{R}^3 \mid t, y \in [a, b], \quad |u| \leq r\right\}$$

Proposition 18.6. Zeidler (1995) *Assume that*

(a) The function $F : Q_r \to \mathbb{R}$ is continuous;

(b) $|\lambda|\, \mu \leq r$, $\mu := \dfrac{1}{b-a} \max\limits_{(t,x,u)\in Q_r} |F(t, x, u)|$;

Setting $\mathcal{X} := C[a, b]$ and $\mathcal{M} := \{u \in \mathcal{X} \mid \|u\| \leq r\}$, it follows that the integral equation (18.110) has at least one solution $u \in \mathcal{M}$.

Proof. For all $t \in [a, b]$ define the operator

$$(Au)(t) := \lambda \int_{y=a}^{b} F(t, y, u(y))\, dy$$

Then the integral equation (18.110) corresponds to the following fixed-point problem $Au = u \in \mathcal{M}$. Notice that the operator $A : \mathcal{M} \to \mathcal{M}$ is compact and for all $u \in \mathcal{M}$

$$\|Au\| \leq |\lambda| \max_{t\in[a,b]} \left| \int_{y=a}^{b} F(t, y, u(y))\, dy \right| \leq |\lambda|\, \mu \leq r$$

Hence, $A(\mathcal{M}) \subseteq \mathcal{M}$. Thus, by the Schauder fixed-point theorem 18.20 it follows that equation (18.110) has a solution. □

18.8.4 The Leray–Schauder principle and a priori estimates

In this subsection we will again concern ourselves with the solution of the operator equation

$$\boxed{\Phi(x) = x \in \mathcal{X}} \tag{18.111}$$

using the properties of the associated parametrized equation

$$\boxed{t\Phi(x) = x \in \mathcal{X},\ t \in [0, 1)} \tag{18.112}$$

For $t = 0$ equation (18.112) has the trivial solution $x = 0$, and for $t = 1$ coincides with (18.111). Assume that the following condition holds:

(A) There is a number $r > 0$ such that if x is a solution of (18.112), then

$$\boxed{\|x\| \leq r} \tag{18.113}$$

Remark 18.10. *Here we do not assume that (18.112) has a solution and, evidently, that the assumption (A) is trivially satisfied if the set $\Phi(\mathcal{X})$ is bounded since $\|\Phi(x)\| \leq r$ for all $x \in \mathcal{X}$.*

Theorem 18.21. (Leray–Schauder 1934) *If the compact operator $\Phi : \mathcal{X} \to \mathcal{X}$ given on the Banach space $\mathcal{X}$ over the field $\mathcal{F}$ (real or complex) satisfies assumption (A), then the original equation (18.111) has a solution (nonobligatory unique).*

Proof. **Zeidler (1995)** Define the subset

$$\mathcal{M} := \{x \in \mathcal{X} \mid \|x\| \leq 2r\}$$

and the operator

$$B(x) := \begin{cases} \Phi(x) & \text{if} \quad \|\Phi(x)\| \leq 2r \\ 2r\dfrac{\Phi(x)}{\|\Phi(x)\|} & \text{if} \quad \|\Phi(x)\| > 2r \end{cases}$$

Obviously, $\|B(x)\| \leq 2r$ for all $x \in \mathcal{X}$ which implies $B(\mathcal{M}) \subseteq \mathcal{M}$. Show that $B : \mathcal{M} \to \mathcal{M}$ is a compact operator. First, notice that B is continuous because of the continuity of Φ. Then consider the sequences $\{u_n\} \in \mathcal{M}$ and $\{v_n\}$ such that (a) $\{v_n\} \in \mathcal{M}$ or (b) $\{v_n\} \notin \mathcal{M}$. In case (a) the boundedness of $\mathcal{M}$ and the compactness of Φ imply that there is a subsequence $\{v_{n_k}\}$ such that $B(v_{n_k}) = \Phi(v_{n_k}) \to z$ as $n \to \infty$. In case (b) we may choose this subsequence so that $1/\|\Phi(v_{n_k})\| \to \alpha$ and $\Phi(v_{n_k}) \to z$. Hence, $B(v_{n_k}) \to 2r\alpha z$. So, B is compact. The Schauder fixed-point theorem 18.20 being applied to the compact operator $B : \mathcal{M} \to \mathcal{M}$ provides us with a point $x \in \mathcal{M}$ such that $x = B(x)$. So, if $\|\Phi(x)\| \leq 2r$, then $B(x) = \Phi(x) = x$ and we obtain the solution of the original problem. Another case $\|\Phi(x)\| > 2r$ is impossible by assumption (A). Indeed, suppose $\Phi(x) = x$ for $\|\Phi(x)\| > 2r$. Then $x = Bx = t\Phi(x)$ with $t := 2r/\|\Phi(x)\| < 1$. This forces $\|x\| = t\|\Phi(x)\| = 2r$ which contradicts with assumption (A). Theorem is proven. □

Remark 18.11. *Theorem 18.21 turns out to be very useful for the justification of the existence of solutions for different types of partial differential equations (such as the famous Navier–Stokes equations for viscous fluids, quasi-linear elliptic etc.).*

PART III

Differential Equations and Optimization

19 Ordinary Differential Equations

Contents

19.1 Classes of ODE

In this chapter we will deal with the class of functions satisfying the following *ordinary differential equation*

$$\boxed{\begin{array}{c}\dot{x}(t) = f(t, x(t)) \quad \text{for almost all} \quad t \in [t_0, t_0 + \theta] \\ x(t_0) = x_0 \\ f : \mathbb{R} \times \mathcal{X} \to \mathcal{X}\end{array}} \tag{19.1}$$

where f is a nonlinear function and $\mathcal{X}$ is a Banach space (any concrete space of functions). *Cauchy's problem* for (19.1) consists of resolving (19.1), or, in other words, in finding a function $x(t)$ which satisfies (19.1).

For simplicity we will also use the following abbreviations:

- ODE meaning an *ordinary differential equation*,
- DRHS meaning the *discontinuous right-hand side*.

Usually the following three classes of ODE (19.1) are considered:

1. *Regular ODE*:
   $f(t, x)$ is continuous in both variables. In this case $x(t)$, satisfying (19.1), should be continuous differentiable, i.e.,

$$\boxed{x(t) \in C^1[t_0, t_0 + \theta]} \tag{19.2}$$

2. *ODE of Carathéodory's type*:
   $f(t, x)$ in (19.1) is measurable in t and continuous in x.
3. *ODE with discontinuous right-hand side*:
   $f(t, x)$ in (19.1) is continuous in t and discontinuous in x. In fact, this type of ODE equation is related to the *differential inclusion*:

$$\boxed{\dot{x}(t) \in F(t, x(t))} \tag{19.3}$$

where $F(t, x)$ is a set in $\mathbb{R} \times \mathcal{X}$. If this set for some pair (t, x) consists of one point, then $F(t, x) = f(t, x)$.

19.2 Regular ODE

19.2.1 Theorems on existence

19.2.1.1 Theorem based on the contraction principle

Theorem 19.1. (on local existence and uniqueness) *Let $f(t,x)$ be continuous in t on $[t_0, t_0+\theta]$, $(\theta \geq 0)$ and for any $t \in [t_0, t_0+\theta]$ it satisfies the, so-called,* ***local Lipschitz condition*** *in x, that is, there exist constants $c, L_f > 0$ such that*

$$\boxed{\begin{array}{c} \|f(t,x)\| \leq c \\ \|f(t,x_1) - f(t,x_2)\| \leq L_f \|x_1 - x_2\| \end{array}} \tag{19.4}$$

for all $t \in [t_0, t_0+\theta]$ and all $x, x_1, x_2 \in B_r(x_0)$ where

$$B_r(x_0) := \{x \in \mathcal{X} \mid \|x - x_0\| \leq r\}$$

Then Cauchy's problem (19.1) has a unique solution on the time-interval $[t_0, t_0+\theta_1]$, where

$$\boxed{\theta_1 < \min\left\{r/c, L_f^{-1}, \theta\right\}} \tag{19.5}$$

Proof.

1. First, show that Cauchy's problem (19.1) is equivalent to finding the continuous solution to the following integral equation

$$x(t) = x_0 + \int_{s=t_0}^{t} f(s, x(s))\, ds \tag{19.6}$$

Indeed, if $x(t)$ is a solution of (19.1), then, obviously, it is a differentiable function on $[t_0, t_0+\theta_1]$. By integration of (19.1) on $[t_0, t_0+\theta_1]$ we obtain (19.6). Inversely, suppose $x(t)$ is a continuous function satisfying (19.6). Then, by the assumption (19.4) of the theorem, it follows

$$\begin{aligned} \|f(s,x(s)) - f(s_0,x(s_0))\| &= \big\|[f(s,x(s)) - f(s,x(s_0))] \\ &+ [f(s,x(s_0)) - f(s_0,x(s_0))]\big\| \leq \|f(s,x(s)) - f(s,x(s_0))\| \\ &+ \|f(s,x(s_0)) - f(s_0,x(s_0))\| \leq L_f \|x(s) - x(s_0)\| \\ &+ \|f(s,x(s_0)) - f(s_0,x(s_0))\| \end{aligned}$$

This implies that if $s, s_0 \in [t_0, t_0+\theta_1]$ and $s \to s_0$, then the right-hand side of the last inequality tends to zero, and, hence, $f(s,x(s))$ is continuous at each point of the interval $[t_0, t_0+\theta_1]$. And, moreover, we also obtain that $x(t)$ is differentiable on this interval, satisfies (19.1) and $x(t_0) = x_0$.

2. Using this equivalence, let us introduce the Banach space $C[t_0, t_0 + \theta_1]$ of abstract continuous functions $x(t)$ with values in $\mathcal{X}$ and with the norm

$$\|x(t)\|_C := \max_{t\in[t_0,t_0+\theta_1]} \|x(t)\|_{\mathcal{X}} \tag{19.7}$$

Consider in $C[t_0, t_0 + \theta_1]$ the ball $B_r(x_0)$ and notice that the nonlinear operator $\Phi : C[t_0, t_0 + \theta_1] \to C[t_0, t_0 + \theta_1]$ defined by

$$\Phi(x) = x_0 + \int_{s=t_0}^{t} f(s, x(s))\, ds \tag{19.8}$$

transforms $B_r(x_0)$ into $B_r(x_0)$ since

$$\begin{aligned}\|\Phi(x) - x_0\|_C &= \max_{t\in[t_0,t_0+\theta_1]} \left\| \int_{s=t_0}^{t} f(s, x(s))\, ds \right\| \\ &\le \max_{t\in[t_0,t_0+\theta_1]} \int_{s=t_0}^{t} \|f(s, x(s))\|\, ds \le \theta_1 c < r\end{aligned}$$

Moreover, the operator Φ is a contraction (see Definition 14.20) on $B_r(x_0)$. Indeed, by the local Lipschitz condition (19.4), it follows that

$$\begin{aligned}\|\Phi(x_1) - \Phi(x_2)\|_C &= \max_{t\in[t_0,t_0+\theta_1]} \left\| \int_{s=t_0}^{t} [f(s, x_1(s)) - f(s, x_2(s))]\, ds \right\| \\ &\le \max_{t\in[t_0,t_0+\theta_1]} \int_{s=t_0}^{t} \|f(s, x_1(s)) - f(s, x_2(s))\|\, ds \\ &\le \theta_1 L_f \|x_1 - x_2\|_C = q\|x_1 - x_2\|_C\end{aligned}$$

where $q := \theta_1 L_f < 1$ for small enough r. Then, by Theorem (the contraction principle) 14.17, we conclude that (19.6) has a unique solution $x(t) \in C[t_0, t_0 + \theta_1]$. Theorem is proven. □

Corollary 19.1. *If in the conditions of Theorem 19.1 the Lipschitz condition (19.4) is fulfilled not locally, but* ***globally****, that is, for all $x_1, x_2 \in \mathcal{X}$ (which corresponds with the case $r = \infty$), then Cauchy's problem (19.1) has a unique solution for $[t_0, t_0 + \theta]$ for any θ big enough.*

Proof. It directly follows from Theorem 19.1 if we take $r \to \infty$. But here we prefer to present also another proof based on another type of norm different from (19.7). Again, let us use the integral equivalent form (19.6). Introduce in the Banach space $C[t_0, t_0 + \theta_1]$ the following norm equivalent to (19.7):

$$\|x(t)\|_{\max} := \max_{t\in[t_0,t_0+\theta]} \left\|e^{-L_f t} x(t)\right\|_{\mathcal{X}} \tag{19.9}$$

Then

$$\|\Phi(x_1) - \Phi(x_2)\| \le L_f \int_{s=t_0}^{t} e^{-L_f s} e^{L_f s} \|x_1(s) - x_2(s)\|_C \, ds$$

$$= L_f \int_{s=t_0}^{t} e^{L_f s} \left(e^{-L_f s} \|x_1(s) - x_2(s)\|_C\right) ds$$

$$\le L_f \int_{s=t_0}^{t} e^{L_f s} \|x_1(t) - x_2(t)\|_{\max} \, ds$$

$$= L_f \int_{s=t_0}^{t} e^{L_f s} \, ds \, \|x_1(t) - x_2(t)\|_{\max} = \left(e^{L_f t} - 1\right) \|x_1(t) - x_2(t)\|_{\max}$$

Multiplying this inequality by $e^{-L_f t}$ and taking $\max_{t \in [t_0, t_0+\theta]}$ we get

$$\|\Phi(x_1) - \Phi(x_2)\|_{\max} \le \left(1 - e^{-L_f \theta}\right) \|x_1(t) - x_2(t)\|_{\max}$$

Since $q := 1 - e^{-L_f \theta} < 1$ we conclude that Φ is a contraction. Taking θ big enough we obtain the result. Corollary is proven. □

Remark 19.1. *Sure, the global Lipschitz condition (19.4) with $r = \infty$ holds for a very narrow class of functions which is known as the class of "quasi-linear" functions, that is why Corollary 19.1 is too conservative. On the other hand, the conditions of Theorem 19.1 for finite (small enough) $r < \infty$ are not so restrictively valid for any function satisfying somewhat mild smoothness conditions.*

Remark 19.2. *The main disadvantage of Theorem 19.1 is that the solution of Cauchy's problem (19.1) exists only on the interval $[t_0, t_0 + \theta_1]$ (where θ_1 satisfies (19.5)), but not at the complete interval $[t_0, t_0 + \theta]$, which is very restrictive. For example, the Cauchy problem*

$$\dot{x}(t) = x^2(t), \quad x(0) = 1$$

has the exact solution $x(t) = \dfrac{1}{1-t}$ which exists only on $[0, 1)$ but not for all $[0, \infty)$.

The theorem presented below gives a constructive (direct) method of finding a unique solution of the problem (19.1). It has several forms. Here we present the version of this result which does not use any Lipschitz conditions: neither local, nor global.

Theorem 19.2. (Picard–Lindelöf 1890) *Consider Cauchy's problem (19.1) where the function $f(t, x)$ is continuous on*

$$\boxed{S := \left\{(t, x) \in \mathbb{R}^{1+n} \mid |t - t_0| \le \theta \le \theta, \|x - x_0\|_C \le r\right\}} \tag{19.10}$$

and the partial derivative $\dfrac{\partial}{\partial x} f : S \to \mathbb{R}^n$ *is also continuous on* S. *Define the numbers*

$$\boxed{\mathcal{M} := \max_{(t,x)\in S} \|f(t,x)\|, \quad L := \max_{(t,x)\in S} \left\| \frac{\partial}{\partial x} f(t,x) \right\|} \tag{19.11}$$

and choose the real number θ *such that*

$$\boxed{0 < \theta \leq r, \quad \theta\mathcal{M} \leq r, \quad q := \theta L < 1} \tag{19.12}$$

Then

1. *Cauchy's problem (19.1) has a unique solution on* S;
2. *the sequence* $\{x_n(t)\}$ *of functions generated iteratively by*

$$\boxed{\begin{aligned} x_{n+1}(t) &= x_0 + \int_{s=t_0}^{t} f(s, x_n(s))\, ds \\ x_0(t) &= x_0, \quad n = 0, 1, \ldots; \ t_0 - \theta \leq t \leq \theta + t_0 \end{aligned}} \tag{19.13}$$

converges to $x(t)$ *in the Banach space* $\mathcal{X}$ *with the norm (19.7) geometrically as*

$$\boxed{\|x_{n+1}(t) - x(t)\|_C \leq q^{n+1} \|x_0 - x(t)\|_C} \tag{19.14}$$

Proof. Consider the integral equation (19.6) and the integral operator Φ (19.8) given on S. So, (19.6) can be represented as

$$\Phi(x(t)) = x(t), \quad x(t) \in B_r(x_0)$$

where $\Phi : M \to \mathcal{X}$. For all $t \in [t_0, t_0 + \theta]$ we have

$$\|\Phi(x(t)) - x_0\| = \max_{t\in[t_0,t_0+\theta]} \left\| \int_{s=t_0}^{t} f(s, x_n(s))\, ds \right\|$$

$$\leq \max_{t\in[t_0,t_0+\theta]} (t - t_0) \max_{(t,x)\in S} \|f(t,x)\| \leq \theta\mathcal{M} \leq r$$

i.e., $\Phi(M) \subseteq M$. By the classical mean value theorem 16.5

$$\|f(t,x) - f(t,y)\| = \left\| \frac{\partial}{\partial x} f(t,z) \,|_{z\in[x,y]} (x - y) \right\| \leq L \|x - y\|$$

and, hence,

$$\|\Phi(x(t)) - \Phi(y(t))\| = \max_{t\in[t_0,t_0+\theta]} \left\| \int_{s=t_0}^{t} [f(s, x(s)) - f(s, y(s))]\, ds \right\|$$
$$\leq \theta L \max_{t\in[t_0,t_0+\theta]} \|x(t) - y(t)\| = q\, \|x(t) - y(t)\|_C$$

Applying now the contraction principle we obtain that (19.6) has a unique solution $x \in B_r(x_0)$. We also have

$$\|x_{n+1}(t) - x(t)\|_C = \max_{t\in[t_0,t_0+\theta]} \int_{s=t_0}^{t} [f(s, x_n(s)) - f(s, x(s))]\, ds$$
$$\leq q\, \|x_n(t) - x(t)\|_C \leq q^{n+1}\, \|x_0 - x(t)\|_C$$

Theorem is proven. □

19.2.1.2 Theorem based on the Schauder fixed-point theorem

The next theorem to be proved drops the assumption of Lipschitz continuity but, also, the assertion of uniqueness.

Theorem 19.3. (Peano 1890) *Consider Cauchy's problem (19.1) where the function $f(t, x)$ is continuous on S (19.10) where the real parameter θ is selected in such a way that*

$$\boxed{0 < \theta \leq r, \quad \theta \mathcal{M} \leq r} \tag{19.15}$$

Then Cauchy's problem (19.1) has ***at least one solution*** *on S.*

Proof.

(a) *For the Schauder fixed-point theorem use* Zeidler (1995). By the same arguments as in the proof of Theorem 19.2 it follows that the operator $\Phi : B_r(x_0) \to B_r(x_0)$ is compact (see Definition 18.14). So, by the *Schauder fixed-point theorem* 18.20 we conclude that the operator equation $\Phi(x(t)) = x(t)$, $x(t) \in B_r(x_0)$ has at least one solution. This completes the proof.

(b) *Direct proof (Hartman 2002).* Let $\delta > 0$ and $x_0(t) \in C^1[t_0 - \delta, t_0]$ satisfy on $[t_0 - \delta, t_0]$ the following conditions: $x_0(t) = x_0$, $\|x_0(t) - x_0\| \leq r$ and $\left\|x_0'(t)\right\| \leq d$. For $0 < \varepsilon \leq \delta$ define a function $x_\varepsilon(t)$ on $[t_0 - \delta, t_0 + \varepsilon]$ by putting $x_\varepsilon(t_0) = x_0$ on $[t_0 - \delta, t_0]$ and

$$x_\varepsilon(t) = x_0 + \int_{s=t_0}^{t} f(s, x_\varepsilon(s - \varepsilon))\, ds \tag{19.16}$$

on $[t_0, t_0 + \varepsilon]$. Note that $x_\varepsilon(t)$ is a C^0-function on $[t_0 - \delta, t_0 + \varepsilon]$ satisfying

$$\|x_\varepsilon(t) - x_0\| \leq r \quad \text{and} \quad \|x_\varepsilon(t) - x_\varepsilon(s)\| \leq d\,|t - s|$$

Thus, for the family of functions $\{x_{\varepsilon_n}(t)\}$, $\varepsilon_n \to 0$ whereas for $n \to \infty$ it follows that the limit $x(t) = \lim_{n\to\infty} x_{\varepsilon_n}(t)$ exists uniformly on $[t_0 - \delta, t_0 + \theta]$, which implies that

$$\left\| f\left(t, x_{\varepsilon_n}(t - \varepsilon_n)\right) - f(t, x(t)) \right\| \to 0$$

uniformly as $n \to \infty$. So, term-by-term integration of (19.16) with $\varepsilon = \varepsilon_n$ gives (19.6) and, hence, $x(t)$ is a solution of (19.1). □

The following corollary of Peano's existence theorem is often used.

Corollary 19.2. (Hartman 2002) *Let $f(t, x)$ be continuous on an open (t, x)-set of $\mathbb{E} \subseteq \mathbb{R}^{1+n}$ satisfying $\|f(t, x)\| \leq \mathcal{M}$. Let also $\mathbb{E}_0$ be a compact subset of $\mathbb{E}$. Then there exists a $\theta > 0$, depending on $\mathbb{E}$, $\mathbb{E}_0$ and $\mathcal{M}$, such that if $(t_0, x_0) \in \mathbb{E}_0$, then (19.6) has a solution on $|t - t_0| \leq \theta$.*

19.2.2 Differential inequalities, extension and uniqueness

The most important technique in ODE theory involves the "integration" of the, so-called, differential inequalities. In this subsection we present results dealing with this integration process which is extensively used throughout; there will be presented its immediate application to the extension and uniqueness problems.

19.2.2.1 Bihari and Gronwall–Bellman inequalities

Lemma 19.1.(Bihari 1956) *Let*

1. *$v(t)$ and $\xi(t)$ be nonnegative continuous functions on $[t_0, \infty)$, that is,*

$$\boxed{v(t) \geq 0, \quad \xi(t) \geq 0 \ \forall t \in [t_0, \infty), \quad v(t), \xi(t) \in C[t_0, \infty)} \tag{19.17}$$

2. *for any $t \in [t_0, \infty)$ the following inequality holds*

$$\boxed{v(t) \leq c + \int_{\tau=t_0}^{t} \xi(\tau)\, \Phi(v(\tau))\, d\tau} \tag{19.18}$$

where c is a positive constant ($c > 0$) and $\Phi(v)$ is a positive nondecreasing continuous function, that is,

$$\boxed{0 < \Phi(v) \in C[t_0, \infty) \ \forall v \in (0, \bar{v}), \quad \bar{v} \leq \infty} \tag{19.19}$$

Denote

$$\boxed{\Psi(v) := \int_{s=c}^{v} \frac{ds}{\Phi(s)} \ (0 < v < \bar{v})} \tag{19.20}$$

If in addition

$$\boxed{\int_{\tau=t_0}^{t} \xi(\tau)\, d\tau < \Psi(\bar{v}-0), \quad t \in [t_0, \infty)} \tag{19.21}$$

then for any $t \in [t_0, \infty)$

$$\boxed{v(t) \le \Psi^{-1}\left(\int_{\tau=t_0}^{t} \xi(\tau)\, d\tau\right)} \tag{19.22}$$

where $\Psi^{-1}(y)$ *is the function inverse to* $\Psi(v)$*, that is,*

$$y = \Psi(v), \quad v = \Psi^{-1}(y) \tag{19.23}$$

In particular, if $\bar{v} = \infty$ *and* $\Psi(\infty) = \infty$*, then the inequality (19.22) is fulfilled without any constraints.*

Proof. Since $\Phi(v)$ is a positive nondecreasing continuous function the inequality (19.18) implies that

$$\Phi(v(t)) \le \Phi\left(c + \int_{\tau=t_0}^{t} \xi(\tau)\, \Phi(v(\tau))\, d\tau\right)$$

and

$$\frac{\xi(t)\, \Phi(v(t))}{\Phi\left(c + \int_{\tau=t_0}^{t} \xi(\tau)\, \Phi(v(\tau))\, d\tau\right)} \le \xi(t)$$

Integrating the last inequality, we obtain

$$\int_{s=t_0}^{t} \frac{\xi(s)\, \Phi(v(s))}{\Phi\left(c + \int_{\tau=t_0}^{s} \xi(\tau)\, \Phi(v(\tau))\, d\tau\right)} ds \le \int_{s=t_0}^{t} \xi(s)\, ds \tag{19.24}$$

Denote

$$w(t) := c + \int_{\tau=t_0}^{t} \xi(\tau)\, \Phi(v(\tau))\, d\tau$$

Then evidently

$$\dot{w}(t) = \xi(t)\,\Phi(v(t))$$

Hence, in view of (19.20), the inequality (19.24) may be represented as

$$\int_{s=t_0}^{t} \frac{\dot{w}(s)}{\Phi(w(s))}\,ds = \int_{w=w(t_0)}^{w(t)} \frac{dw}{\Phi(w)} = \Psi(w(t)) - \Psi(w(t_0)) \le \int_{s=t_0}^{t} \xi(s)\,ds$$

Taking into account that $w(t_0) = c$ and $\Psi(w(t_0)) = 0$, from the last inequality it follows that

$$\Psi(w(t)) \le \int_{s=t_0}^{t} \xi(s)\,ds \tag{19.25}$$

Since

$$\Psi(v) = \frac{1}{\Phi(v)} \quad (0 < v < \bar{v})$$

the function $\Psi(v)$ has the uniquely defined continuous monotonically increasing inverse function $\Psi^{-1}(y)$ defined within the open interval $(\Psi(+0), \Psi(\bar{v} - 0))$. Hence, (19.25) directly implies

$$w(t) = c + \int_{\tau=t_0}^{t} \xi(\tau)\,\Phi(v(\tau))\,d\tau \le \Psi^{-1}\left(\int_{s=t_0}^{t} \xi(s)\,ds\right)$$

which, in view of (19.18), leads to (19.22). Indeed,

$$v(t) \le c + \int_{\tau=t_0}^{t} \xi(\tau)\,\Phi(v(\tau))\,d\tau \le \Psi^{-1}\left(\int_{s=t_0}^{t} \xi(s)\,ds\right)$$

The case $\bar{v} = \infty$ and $\Psi(\infty) = \infty$ is evident. Lemma is proven. □

Corollary 19.3. *Taking in (19.22)*

$$\boxed{\Phi(v) = v^m \ (m > 0, m \ne 1)}$$

it follows that

$$\boxed{v(t) \le \left[c^{1-m} + (1-m)\int_{\tau=t_0}^{t} \xi(\tau)\,d\tau\right]^{\frac{1}{m-1}} \quad \text{for} \quad 0 < m < 1} \tag{19.26}$$

and

$$\boxed{v(t) \leq c \left[1 - (1-m)\, c^{m-1} \int_{\tau=t_0}^{t} \xi(\tau)\, d\tau \right]^{-\frac{1}{m-1}}}$$

for

$$\boxed{m > 1 \quad \text{and} \quad \int_{\tau=t_0}^{t} \xi(\tau)\, d\tau < \frac{1}{(m-1)\, c^{m-1}}}$$

Corollary 19.4. (Gronwall 1919) *If $v(t)$ and $\xi(t)$ are nonnegative continuous functions on $[t_0, \infty)$ verifying*

$$\boxed{v(t) \leq c + \int_{\tau=t_0}^{t} \xi(\tau)\, v(\tau)\, d\tau} \tag{19.27}$$

then for any $t \in [t_0, \infty)$ the following inequality holds:

$$\boxed{v(t) \leq c \exp\left(\int_{s=t_0}^{t} \xi(s)\, ds \right)} \tag{19.28}$$

This result remains true if $c = 0$.

Proof. Taking in (19.18) and (19.20)

$$\Phi(v) = v$$

we obtain (19.193) and, hence, for the case $c > 0$

$$\Psi(v) := \int_{s=c}^{v} \frac{ds}{s} = \ln\left(\frac{v}{c}\right)$$

and

$$\Psi^{-1}(y) = c \cdot \exp(y)$$

which implies (19.28). The case $c = 0$ follows from (19.28) applying $c \to 0$. □

19.2.2.2 Differential inequalities

Here we follow Hartman (2002) completely.

Definition 19.1. *Let $f(t,x)$ be a continuous function on a plane (t,x)-set $\mathbb{E}$. By a* ***maximal solution*** *$x^0(t)$ of Cauchy's problem*

$$\boxed{\dot{x}(t) = f(t,x), \quad x(t_0) = x_0 \in \mathbb{R}} \tag{19.29}$$

is meant to be a solution of (19.29) on a maximal interval of existence such that if $x(t)$ is any solution of (19.29) then

$$\boxed{x(t) \leq x^0(t)} \tag{19.30}$$

holds (by component-wise) on the common interval of existence of $x(t)$ and $x^0(t)$. The ***minimal solution*** *is similarly defined.*

Lemma 19.2. *Let $f(t,x)$ be a continuous function on a rectangle*

$$S^+ := \left\{(t,x) \in \mathbb{R}^2 \mid t_0 \leq t \leq t_0 + \theta \leq t_0, \|x - x_0\|_C \leq r\right\} \tag{19.31}$$

and on S^+

$$\|f(t,x)\| \leq \mathcal{M} \quad \textit{and} \quad \alpha := \min\{\theta; r/\mathcal{M}\}$$

Then Cauchy's problem (19.29) has a solution on $[t_0, t_0+\alpha)$ such that every solution $x = x(t)$ of $\dot{x}(t) = f(t,x)$, $x(t_0) \leq x_0$ satisfies (19.30) on $[t_0, t_0+\alpha)$.

Proof. Let $0 < \alpha' < \alpha$. Then, by Peano's existence theorem 19.3, the Cauchy problem

$$\dot{x}(t) = f(t,x) + \frac{1}{n}, \quad x(t_0) = x_0 \tag{19.32}$$

has a solution $x = x_n(t)$ on $[t_0, t_0+\alpha']$ if n is sufficiently large. Then there exists a subsequence $\{n_k\}_{k=1,2,\ldots}$ such that the limit $x^0(t) = \lim_{k\to\infty} x_{n_k}(t)$ exists uniformly on $[t_0, t_0+\alpha']$ and $x^0(t)$ is a solution of (19.29). To prove that (19.30) holds on $[t_0, t_0+\alpha']$ it is sufficient to verify

$$x(t) \leq x_n(t) \quad \text{on} \quad [t_0, t_0+\alpha'] \tag{19.33}$$

for large enough n. If this is not true, then there exists a $t = t_1 \in (t_0, t_0+\alpha')$ such that $x(t_1) > x_n(t_1)$. Hence there exists a largest t_2 on $[t_0, t_1)$ such that $x(t_2) = x_n(t_2)$ and $x(t) > x_n(t)$. But by (19.32) $x_n'(t_2) = x'(t_2) + \frac{1}{n}e$, so that $x_n(t) > x(t)$ for $t > t_2$ near t_2. This contradiction proves (19.33). Since $\alpha' < \alpha$ is arbitrary, the lemma follows. □

Corollary 19.5. *Let $f(t,x)$ be a continuous function on an open set $\mathbb{E}$ and $(t_0, x_0) \in \mathbb{E} \subseteq \mathbb{R}^2$. Then Cauchy's problem (19.29) has a maximal and minimal solution near (t_0, x_0).*

19.2.2.3 Right derivatives

Lemma 19.3.

1. If $n = 1$ and $x \in C^1[a, b]$ then $|x(t)|$ has a right derivative

$$\boxed{D_R |x(t)| := \lim_{0<h\to 0} \frac{1}{h} [|x(t+h)| - |x(t)|]} \tag{19.34}$$

such that

$$\boxed{D_R |x(t)| = \begin{cases} x'(t) \operatorname{sign}(x(t)) & \text{if } x(t) \neq 0 \\ |x'(t)| & \text{if } x(t) = 0 \end{cases}} \tag{19.35}$$

and

$$\boxed{|D_R |x(t)|| = |x'(t)|} \tag{19.36}$$

2. If $n > 1$ and $x \in C^1[a, b]$ then $\|x(t)\|$ has a right derivative

$$\boxed{D_R \|x(t)\| := \lim_{0<h\to 0} \frac{1}{h} [\|x(t+h)\| - \|x(t)\|]} \tag{19.37}$$

such that on $t \in [a, b)$

$$\boxed{\begin{aligned} \|D_R \|x(t)\|\| &= \max_{k=1,\ldots,n} D_R |x_k(t)| \\ &\leq \|x'(t)\| := \max\{|x_1'(t)|, \ldots, |x_n'(t)|\} \end{aligned}} \tag{19.38}$$

Proof. Assertion (1) is clear when $x(t) \neq 0$ and the case $x(t) = 0$ follows from the identity

$$x(t+h) = x(t) + hx'(t) + o(h) = h[x'(t) + o(1)]$$

so that, in general, when $h \to 0$

$$|x(t+h)| = |x(t)| + h\left[|x'(t)| + o(1)\right]$$

The multidimensional case (2) follows from (1) if we take into account that

$$|x_k(t+h)| = |x_k(t)| + h\left[|x_k'(t)| + o(1)\right]$$

Taking the $\max_{k=1,\ldots,n}$ of these identities, we obtain

$$\|x(t+h)\| = \|x(t)\| + h\left[\max_{k=1,\ldots,n} |x_k'(t)| + o(1)\right]$$

whereas $h \to 0$. This proves (19.38). □

Example 19.1. *Let* $x(t) := (t - t_0)^2$. *Then* $x'(t) := 2(t - t_0)$ *is continuous and, hence,* $x(t) \in C^1$. *By Lemma 19.3 it follows that* $D_R |x(t)| = 2|t - t_0|$.

19.2.2.4 Differential inequalities

The next theorem deals with the integration of differential inequalities and is frequently used in the ODE theory.

Theorem 19.4. (Hartman 2002) *Let* $f(t, x)$ *be continuous on an open* (t, x)*-set* $\mathbb{E} \subseteq \mathbb{R}$ *and* $x^0(t)$ *be the maximal solution of (19.6). Let* $v(t)$ *be continuous on* $[t_0, t_0 + \alpha]$ *function such that*

$$\left.\begin{array}{l} v(t_0) \leq x_0, \quad (t, v) \in \mathbb{E} \\ D_R v(t) \leq f(t, v(t)) \end{array}\right\} \tag{19.39}$$

Then, on the common interval of existence of $x^0(t)$ *and* $v(t)$

$$\boxed{v(t) \leq x^0(t)} \tag{19.40}$$

Remark 19.3. *If the inequalities (19.39) are reversed with the* ***left derivative*** $D_L v(t)$ *instead of* $D_R v(t)$, *then the conclusion (19.40) must be replaced by* $v(t) \geq x_0(t)$ *where* $x_0(t)$ *is the* ***minimal solution*** *of (19.6).*

Proof. It is sufficient to show that there exists a $\delta > 0$ such that (19.40) holds for $[t_0, t_0 + \delta]$. Indeed, if this is the case and if $v(t)$ and $x^0(t)$ are defined on $[t_0, t_0 + \beta]$, then it follows that the set of t-values, where (19.40) holds, cannot have an upper bound different from β. In Lemma 19.2 let $n > 0$ be large enough and δ be chosen independent of n such that (19.32) has a solution $x = x_n(t)$ on $[t_0, t_0 + \delta]$. In view of Lemma 19.2 it is sufficient to verify that $v(t) \leq x_n(t)$ on $[t_0, t_0 + \delta]$. But the proof of this fact is absolutely identical to the proof of (19.33). Theorem is proven. □

In fact, the following several consequences of this theorem are widely used in the ODE theory.

Corollary 19.6. *If* $v(t)$ *is continuous on* $[t_0, t_f]$ *and* $D_R v(t) \leq 0$ *when* $t \in [t_0, t_f]$, *then*

$$\boxed{v(t) \leq v(t_0) \quad \textit{for any} \quad t \in [t_0, t_f]} \tag{19.41}$$

Corollary 19.7. (Lemma on differential inequalities) *Let* $f(t, x)$, $x^0(t)$ *be as in Theorem 19.4 and* $g(t, x)$ *be continuous on an open* (t, x)*-set* $\mathbb{E} \subseteq \mathbb{R}^2$ *satisfying*

$$\boxed{g(t, x) \leq f(t, x)} \tag{19.42}$$

Let also $v(t)$ *be a solution of the following ODE:*

$$\boxed{\dot{v}(t) = g(t, v(t)), \quad v(t_0) := v_0 \leq x_0} \tag{19.43}$$

on $[t_0, t_0 + \alpha]$. *Then*

$$\boxed{v(t) \leq x^0(t)} \tag{19.44}$$

holds on any common interval of existence of $v(t)$ *and* $x^0(t)$ *to the right of* t_0.

Corollary 19.8. *Let* $x^0(t)$ *be the maximal solution of*

$$\dot{x}(t) = f(t, x(t)), \quad x(t_0) := x^0 \in \mathbb{R}$$

and $x_0(t)$ *be the minimal solution of*

$$\dot{x}(t) = -f(t, x(t)), \quad x(t_0) := x_0 \geq 0$$

Let also $y = y(t)$ *be a* C^1 *vector valued function on* $[t_0, t_0 + \alpha]$ *such that*

$$\boxed{\begin{array}{c} x_0 \leq \|y(t_0)\| \leq x^0 \quad (t, y) \in \mathbb{E} \subseteq \mathbb{R}^2 \\ \dfrac{d}{dt}(\|y(t)\|) \leq f(t, \|y(t)\|) \end{array}} \tag{19.45}$$

Then the first (second) of two inequalities

$$\boxed{x_0(t) \leq \|y(t)\| \leq x^0(t)} \tag{19.46}$$

holds on any common interval of existence of $x_0(t)$ *and* $y(t)$ *(or* $x^0(t)$ *and* $y(t)$*).*

Corollary 19.9. *Let* $f(t, x)$ *be continuous and nondecreasing on* x *when* $t \in [t_0, t_0 + \alpha]$. *Let* $x^0(t)$ *be a maximal solution of (19.6) which exists on* $[t_0, t_0 + \alpha]$. *Let another continuous function* $v(t)$ *satisfy on* $[t_0, t_0 + \alpha]$ *the integral inequality*

$$\boxed{v(t) \leq v_0 + \int_{s=t_0}^{t} f(s, v(s))\, ds} \tag{19.47}$$

where $v_0 \leq x_0$. *Then*

$$\boxed{v(t) \leq x^0(t)} \tag{19.48}$$

holds on $[t_0, t_0 + \alpha]$. *This result is false if we omit:* $f(t, x)$ *is nondecreasing on* x.

Proof. Denote by $V(t)$ the right-hand side of (19.47), so that $v(t) \leq V(t)$, and, by the monotonicity property with respect to the second argument, we have

$$\dot{V}(t) = f(t, v(t)) \leq f(t, V(t))$$

By Theorem 19.4 we have $V(t) \leq x^0(t)$ on $[t_0, t_0 + \alpha]$. Thus $v(t) \leq x^0(t)$ which completes the proof. □

19.2.2.5 Existence of solutions on the complete axis $[t_0, \infty)$

Here we show that the condition $\|f(t,x)\| \leq k\|x\|$ guarantees the existence of the solutions of ODE $\dot{x}(t) = f(t, x(t))$, $x(t_0) = x_0 \in R^n$ for any $t \geq t_0$. In fact, the following more general result holds.

Theorem 19.5. (Wintner 1945) *Let for any* $t \geq t_0$ *and* $x \in \mathbb{R}^n$

$$\boxed{(x, f(t,x)) \leq \Psi\left(\|x\|^2\right)} \tag{19.49}$$

where the function Ψ *satisfies the condition*

$$\boxed{\int_{s=s_0}^{\infty} \frac{ds}{\Psi(s)} = \infty, \quad \Psi(s) > 0 \quad as \quad s \geq s_0 \geq 0} \tag{19.50}$$

Then Cauchy's problem

$$\dot{x}(t) = f(t, x(t)), \quad x(t_0) = x_0 \in R^n$$

has a solution on the complete semi-axis $[t_0, \infty)$ *for any* $x_0 \in \mathbb{R}^n$.

Proof. Notice that for the function $w(t) := \|x(t)\|^2$ in view of (19.49) we have

$$\frac{d}{dt} w(t) = 2(x(t), \dot{x}(t)) = 2(x(t), f(t, x(t)))$$

$$\leq 2\Psi\left(\|x(t)\|^2\right) = 2\Psi(w(t))$$

Then by Theorem 19.4 (see (19.44)) it follows that $w(t_0) \leq s_0$ implies $w(t) \leq s(t)$, where $s(t)$ satisfies

$$\dot{s}(t) = 2\Psi(s(t)), \quad s(t_0) = s_0 := \|x_0\|^2$$

But the solution of the last ODE is always bounded for any finite $t \geq t_0$. Indeed,

$$\int_{s=s_0}^{s(t)} \frac{ds}{\Psi(s)} = 2(t - t_0) \tag{19.51}$$

and $\Psi(s) > 0$ as $s \geq s_0$ implies that $\dot{s}(t) > 0$, and, hence, $s(t) > 0$ for all $t > t_0$. But the solution $s(t)$ can fail to exist on a bounded interval $[t_0, t_0 + a]$ only if it exists on $[t_0, t_0 + \alpha]$ with $\alpha < a$ and $s(t) \to \infty$ if $t \to t_0 + a$. But this gives the contradiction to (19.50) since the left-hand side of (19.51) tends to infinity and the right-hand side of (19.51) remains finite and equal to $2a$. □

Remark 19.4. *The admissible choices of* $\Psi(s)$ *may be, for example,* C, Cs, $Cs\ln s$,... *for large enough* s *and* C *as a positive constant.*

Remark 19.5. *Some generalizations of this theorem can be found in Hartman (2002).*

Example 19.2. *If $A(t)$ is a continuous $n \times n$ matrix and $g(t)$ is continuous on $[t_0, t_0 + a]$ vector function, then Cauchy's problem*

$$\dot{x}(t) = A(t)\,x(t) + g(t), \quad x(t_0) = x_0 \in R^n \tag{19.52}$$

has a unique solution $x(t)$ on $[t_0, t_0 + a]$. It follows from the Wintner theorem 19.5 if we take $\Psi(s) := C(1 + s)$ with $C > 0$.

19.2.2.6 The continuous dependence of the solution on a parameter and on the initial conditions

Theorem 19.6. *If the right-hand side of ODE*

$$\boxed{\dot{x}(t) = f(t, x(t), \mu), \quad x(t_0) = x_0 \in R^n} \tag{19.53}$$

*is **continuous** with respect to μ on $[\mu^-, \mu^+]$ and satisfies the condition of Theorem 19.1 with the Lipschitz constant L_f which is independent of μ, then the solution $x(t, \mu)$ of (19.53) depends **continuously** on $\mu \in [\mu^-, \mu^+] \in \mathbb{R}^m$ as well as on x_0 in some neighborhoods.*

Proof. The proof of this assertion repeats word by word the proof of Theorem 19.1. Indeed, by the same reasons as in Theorem 19.1, the solution $x(t, \mu)$ is a continuous function of both t and μ if L_f is independent of μ. As for the proof of the continuous dependence of the solution on the initial conditions, it can be transformed to the proof of the continuous dependence of the solution on the parameter. Indeed, putting

$$\tau := t - t_0, \quad z := x(t, \mu) - x_0$$

we obtain that (19.53) is converted to

$$\frac{d}{d\tau} z = f(\tau + t_0, z + x_0, \mu), \quad z(0) = 0$$

where x_0 may be considered as a new parameter so that f is continuous on x_0 by the assumption. This proves the theorem. □

19.2.3 Linear ODE

19.2.3.1 Linear vector ODE

Lemma 19.4. *The solution $x(t)$ of the linear ODE (or the corresponding Cauchy's problem)*

$$\boxed{\dot{x}(t) = A(t)\,x(t), \quad x(t_0) = x_0 \in \mathbb{R}^{n \times n}, \quad t \geq t_0} \tag{19.54}$$

where $A(t)$ is a continuous $n \times n$-matrix function, may be presented as

$$\boxed{x(t) = \Phi(t, t_0)\,x_0} \tag{19.55}$$

where the matrix $\Phi(t, t_0)$ is the, so-called, ***fundamental matrix*** *of the system (19.54) and satisfies the following matrix ODE*

$$\boxed{\frac{d}{dt}\Phi(t, t_0) = A(t)\,\Phi(t, t_0), \quad \Phi(t_0, t_0) = I} \tag{19.56}$$

and fulfills the ***group property***

$$\boxed{\Phi(t, t_0) = \Phi(t, s)\,\Phi(s, t_0)\ \forall s \in (t_0, t)} \tag{19.57}$$

Proof. Assuming (19.55), the direct differentiation of (19.55) implies

$$\dot{x}(t) = \frac{d}{dt}\Phi(t, t_0)\,x_0 = A(t)\,\Phi(t, t_0)\,x_0 = A(t)\,x(t)$$

So, (19.55) verifies (19.54). The uniqueness of such a presentation follows from Example 19.2. The property (19.57) results from the fact that

$$x(t) = \Phi(t, s)\,x_s = \Phi(t, s)\,\Phi(s, t_0)\,x(t_0) = \Phi(t, t_0)\,x(t_0)$$

Lemma is proven. □

19.2.3.2 Liouville's theorem

The next result serves as a demonstration that the transformation $\Phi(t, t_0)$ is nonsingular (or has its inverse) on any finite time interval.

Theorem 19.7. (Liouville 1836) *If $\Phi(t, t_0)$ is the solution to (19.56), then*

$$\boxed{\det\Phi(t, t_0) = \exp\left\{\int_{s=t_0}^{t} \mathrm{tr}A(s)\,ds\right\}} \tag{19.58}$$

Proof. The usual expansion for the determinant $\det\Phi(t, t_0)$ and the rule for differentiating the product of scalar functions show that

$$\frac{d}{dt}\det\Phi(t, t_0) = \sum_{j=1}^{n} \det\tilde{\Phi}_j(t, t_0)$$

where $\tilde{\Phi}_j(t, t_0)$ is the matrix obtained by replacing the jth row $\Phi_{j,1}(t, t_0), \ldots, \Phi_{j,n}(t, t_0)$ of $\Phi(t, t_0)$ by its derivatives $\dot{\Phi}_{j,1}(t, t_0), \ldots, \dot{\Phi}_{j,n}(t, t_0)$. But since

$$\dot{\Phi}_{j,k}(t, t_0) = \sum_{i=1}^{n} a_{j,i}(t)\,\Phi_{i,k}(t, t_0), \quad A(t) = \left\|a_{j,i}(t)\right\|_{j,i=1,\ldots,n}$$

it follows that

$$\det \tilde{\Phi}_j(t, t_0) = a_{j,j}(t) \det \Phi(t, t_0)$$

which gives

$$\frac{d}{dt} \det \Phi(t, t_0) = \sum_{j=1}^{n} \frac{d}{dt} \det \tilde{\Phi}_j(t, t_0)$$

$$= \sum_{j=1}^{n} a_{j,j}(t) \det \Phi(t, t_0) = \operatorname{tr}\{A(t)\} \det \Phi(t, t_0)$$

and, as a result, we obtain (19.58) which completes the proof. □

Corollary 19.10. *If for the system (19.54)*

$$\boxed{\int_{s=t_0}^{T} \operatorname{tr} A(s)\, ds > -\infty} \tag{19.59}$$

then for any $t \in [t_0, T]$

$$\boxed{\det \Phi(t, t_0) > 0} \tag{19.60}$$

Proof. It is the direct consequence of (19.58). □

Lemma 19.5. *If (19.59) is fulfilled, namely,* $\int_{s=t_0}^{T} \operatorname{tr} A(s)\, ds > -\infty$, *then the solution* $x(t)$ *on* $[0, T]$ *of the linear nonautonomous ODE*

$$\boxed{\dot{x}(t) = A(t)\, x(t) + g(t), \quad x(t_0) = x_0 \in \mathbb{R}^{n \times n}, \quad t \geq t_0} \tag{19.61}$$

where $A(t)$ *and* $f(t)$ *are assumed to be continuous matrix and vector functions, may be presented by the* ***Cauchy formula***

$$\boxed{x(t) = \Phi(t, t_0) \left[x_0 + \int_{s=t_0}^{t} \Phi^{-1}(s, t_0)\, g(s)\, ds \right]} \tag{19.62}$$

where $\Phi^{-1}(t, t_0)$ *exists for all* $t \in [t_0, T]$ *and satisfies*

$$\boxed{\frac{d}{dt} \Phi^{-1}(t, t_0) = -\Phi^{-1}(t, t_0)\, A(t), \quad \Phi^{-1}(t_0, t_0) = I} \tag{19.63}$$

Proof. By the previous corollary, $\Phi^{-1}(t, t_0)$ exists within the interval $[t_0, T]$. The direct derivation of (19.62) implies

$$\dot{x}(t) = \dot{\Phi}(t, t_0)\left[x_0 + \int_{s=t_0}^{t} \Phi^{-1}(s, t_0)\, g(s)\, ds\right] + \Phi(t, t_0)\,\Phi^{-1}(t, t_0)\, g(t)$$

$$= A(t)\Phi(t, t_0)\left[x_0 + \int_{s=t_0}^{t} \Phi^{-1}(s, t_0)\, g(s)\, ds\right] + g(t) = A(t)x(t) + g(t)$$

which coincides with (19.61). Notice that the integral in (19.62) is well defined in view of the continuity property of the participating functions to be integrated. By identities

$$\Phi(t, t_0)\,\Phi^{-1}(t, t_0) = I$$

$$\frac{d}{dt}\left[\Phi(t, t_0)\,\Phi^{-1}(t, t_0)\right] = \dot{\Phi}(t, t_0)\,\Phi^{-1}(t, t_0) + \Phi(t, t_0)\frac{d}{dt}\Phi^{-1}(t, t_0) = 0$$

it follows that

$$\frac{d}{dt}\Phi^{-1}(t, t_0) = -\Phi^{-1}(t, t_0)\left[\dot{\Phi}(t, t_0)\right]\Phi^{-1}(t, t_0)$$

$$= -\Phi^{-1}(t, t_0)\left[A(t)\,\Phi(t, t_0)\right]\Phi^{-1}(t, t_0) = -\Phi^{-1}(t, t_0)\,A(t)$$

Lemma is proven. □

Remark 19.6. *The solution (19.62) can be rewritten as*

$$\boxed{x(t) = \Phi(t, t_0)\,x_0 + \int_{s=t_0}^{t} \Phi(t, s)\, g(s)\, ds} \tag{19.64}$$

since by (19.57)

$$\boxed{\Phi(t, s) = \Phi(t, t_0)\,\Phi^{-1}(s, t_0)} \tag{19.65}$$

19.2.3.3 Bounds for norm of ODE solutions

Let $\|A\| := \sup_{\|x\|=1} \|Ax\|$ where $\|x\|$ is Euclidean or Chebishev's type.

Lemma 19.6. *Let $x(t)$ be a solution of (19.61). Then*

$$\boxed{\|x(t)\| \le (\|x(t_0)\| + \int_{s=t_0}^{t} \|g(s)\|\, ds)\exp(\int_{s=t_0}^{t} \|A(s)\|\, ds)} \tag{19.66}$$

Proof. By (19.61) it follows that

$$\|\dot{x}(t)\| \leq \|A(t)\| \|x(t)\| + \|g(t)\|$$

Let $v(t)$ be the unique solution of the following ODE:

$$\dot{v}(t) = \|A(t)\| v(t) + \|g(t)\|, \quad v(t_0) = \|x(t_0)\|$$

which solution is

$$v(t) = [v(t_0) + \int_{s=t_0}^{t} \|f(s)\| \exp(-\int_{r=t_0}^{s} \|A(r)\| dr) ds] \exp(\int_{s=t_0}^{t} \|A(s)\| ds)$$

Then, by Lemma 19.7, it follows that $\|x(t)\| \leq v(t)$ for any $t \geq t_0$ which gives (19.66). □

Corollary 19.11. *Similarly, if $w(t)$ is the solution of*

$$\dot{v}(t) = -\|A(t)\| v(t) - \|g(t)\|, \quad w(t_0) = \|x(t_0)\|$$

then $\|x(t)\| \geq v(t)$ for any $t \geq t_0$ that gives

$$\boxed{\|x(t)\| \geq (\|x(t_0)\| - \int_{s=t_0}^{t} \|g(s)\| ds) \exp(-\int_{s=t_0}^{t} \|A(s)\| ds)} \tag{19.67}$$

19.2.3.4 Stationary linear ODE

If in (19.1) $A(t) = A$ is a constant matrix, then it is easy to check that

$$\boxed{\Phi(t, t_0) := e^{A(t-t_0)} \quad \text{where} \quad e^{At} = \sum_{k=0}^{\infty} \frac{1}{k!} A^k t^k} \tag{19.68}$$

and (19.62), (19.64) become

$$\boxed{\begin{aligned} x(t) &= e^{A(t-t_0)} \left[x_0 + \int_{s=t_0}^{t} e^{-A(s-t_0)} g(s) ds \right] \\ &= e^{A(t-t_0)} x_0 + \int_{s=t_0}^{t} e^{A(t-s)} g(s) ds \end{aligned}} \tag{19.69}$$

19.2.3.5 Linear ODE with periodic matrices

In this subsection we show that the case of variable, but periodic, coefficients can be reduced to the case of constant coefficients.

Theorem 19.8. (Floquet 1883) *Let in ODE*

$$\boxed{\dot{x}(t) = A(t)\,x(t)} \tag{19.70}$$

the matrix $A(t) \in \mathbb{R}^{n\times n}$ $(-\infty < t < \infty)$ *be continuous and periodic of period* T, *that is, for any* t

$$\boxed{A(t+T) = A(t)} \tag{19.71}$$

Then the fundamental matrix $\Phi(t, t_0)$ *of (19.70) has a representation of the form*

$$\boxed{\begin{gathered}\Phi(t, t_0) = \tilde{\Phi}(t - t_0) = Z(t - t_0)\, e^{R(t-t_0)} \\ Z(\tau) = Z(\tau + T)\end{gathered}} \tag{19.72}$$

and R *is a constant* $n \times n$ *matrix.*

Proof. Since $\tilde{\Phi}(\tau)$ is a fundamental matrix of (19.70), then $\tilde{\Phi}(\tau + T)$ is fundamental too. By the group property (19.57) it follows that $\tilde{\Phi}(\tau + T) = \tilde{\Phi}(\tau)\,\tilde{\Phi}(T)$. Since $\det \tilde{\Phi}(T) \neq 0$ one can represent $\tilde{\Phi}(T)$ as $\tilde{\Phi}(T) = e^{RT}$ and hence

$$\tilde{\Phi}(\tau + T) = \tilde{\Phi}(\tau)\, e^{RT} \tag{19.73}$$

So, defining $Z(\tau) := \tilde{\Phi}(\tau)\, e^{-R\tau}$, we get

$$\begin{aligned} Z(\tau + T) &= \tilde{\Phi}(\tau + T)\, e^{-R(\tau+T)} \\ &= \left[\tilde{\Phi}(\tau + T)\, e^{-RT}\right] e^{-R\tau} \\ &= \tilde{\Phi}(\tau)\, e^{-R\tau} = Z(\tau) \end{aligned}$$

which completes the proof. □

19.2.3.6 First integrals and related adjoint linear ODE

Definition 19.2. *A function* $F = F(t, x) : \mathbb{R} \times \mathbb{C}^n \to \mathbb{C}$, *belonging to* $C^1[\mathbb{R} \times \mathbb{C}^n]$, *is called the* ***first integral*** *to ODE (19.1) if it is constant over trajectories of* $x(t)$ *generated by (19.1), that is, if for any* $t \geq t_0$ *and any* $x_0 \in \mathbb{C}^n$

$$\boxed{\begin{aligned} \frac{d}{dt}F(t, x(t)) &= \frac{\partial}{\partial t}F(t, x(t)) + \left(\frac{\partial}{\partial x}F(t, x(t)), \dot{x}(t)\right) \\ &= \frac{\partial}{\partial t}F(t, x) + \left(\frac{\partial}{\partial x}F(t, x(t)), f(t, x(t))\right) = 0 \end{aligned}} \tag{19.74}$$

In the case of linear ODE (19.54) the condition (19.74) is converted into the following:

$$\boxed{\frac{\partial}{\partial t}F(t,x)+\left(\frac{\partial}{\partial x}F(t,x(t)),A(t)\right)=0} \tag{19.75}$$

Let us try to find a first integral for (19.54) as a linear form of $x(t)$, i.e., let us try to satisfy (19.75) selecting F as

$$\boxed{F(t,x)=(z(t),\quad x(t)):=z^*(t)\,x(t)} \tag{19.76}$$

where $z^*(t)\in\mathbb{C}^n$ is from $C^1[\mathbb{C}^n]$.

The existence of the first integral for ODE (19.1) permits to decrease the order of the system to be integrated since if the equation $F(t,x(t))=c$ can be resolved with respect to one of the components, say,

$$x_\alpha(t)=\varphi(t,x_1(t),\ldots,x_{\alpha-1}(t),x_{\alpha+1}(t),\ldots,x_n(t))$$

then the order of ODE (19.1) becomes equal to $(n-1)$. If one can find all n first integrals $F_\alpha(t,x(t))=c_\alpha$ $(\alpha=1,\ldots,n)$ which are linearly independent, then the ODE system (19.1) can be considered to be solved.

Lemma 19.7. *A first integral $F(t,x)$ for (19.54) is linear on $x(t)$ as in (19.76) if and only if*

$$\boxed{\dot z(t)=-A^*(t)\,z(t),\quad z(t_0)=z_0\in\mathbb{R}^{n\times n},\quad t\ge t_0} \tag{19.77}$$

Proof.

(a) *Necessity.* If a linear $F(t,x)=(z(t),x(t))$ is a first integral, then

$$\begin{aligned}\frac{d}{dt}F(t,x(t))&=(\dot z(t),x(t))+(z(t),\dot x(t))\\&=(\dot z(t),x(t))+(z(t),A(t)x(t))\\&=(\dot z(t),x(t))+(A^*(t)z(t),x(t))\\&=(\dot z(t)+A^*(t)z(t),x(t))\end{aligned} \tag{19.78}$$

Suppose that $\dot z(t')+A^*(t')z(t')\neq 0$ for some $t'\ge t_0$. Put

$$x(t'):=\dot z(t')+A^*(t')z(t')$$

Since $x(t')=\Phi(t',t_0)x_0$ and $\Phi^{-1}(t',t_0)$ always exists, then for $x_0=\Phi^{-1}(t',t_0)\,x(t')$ we obtain

$$\begin{aligned}\frac{d}{dt}F(t',x(t'))&=(\dot z(t')+A^*(t')z(t'),x(t'))\\&=\|\dot z(t')+A^*(t')z(t')\|^2\neq 0\end{aligned}$$

which is in contradiction with the assumption that $F(t, x(t))$ is a first integral.

(b) *Sufficiency.* It directly results from (19.78).

Lemma is proven. □

Definition 19.3. *The system (19.77) is called the ODE system* ***adjoint*** *to (19.54). For the corresponding inhomogeneous system (19.61) the adjoint system is*

$$\boxed{\dot{z}(t) = -A^*(t)\, z(t) - \tilde{g}(t), \quad z(t_0) = z_0 \in \mathbb{R}^{n\times n}, \quad t \geq t_0} \tag{19.79}$$

There are several results concerning the joint behavior of (19.54) and (19.77).

Lemma 19.8. *A matrix $\Phi(t, t_0)$ is a fundamental matrix for the linear ODE (19.54) if and only if $(\Phi^*(t, t_0))^{-1} = \left(\Phi^{-1}(t, t_0)\right)^*$ is a fundamental matrix for the adjoint system (19.77).*

Proof. Since $\Phi(t, t_0)\, \Phi^{-1}(t, t_0) = I$ by differentiation it follows that

$$\frac{d}{dt}\Phi^{-1}(t, t_0) = -\Phi^{-1}(t, t_0)\frac{d}{dt}\Phi(t, t_0)\, \Phi^{-1}(t, t_0) = -\Phi^{-1}(t, t_0)\, A(t)$$

and taking the complex conjugate transpose of the last identity gives

$$\frac{d}{dt}\left(\Phi^{-1}(t, t_0)\right)^* = -A^*(t)\left(\Phi^{-1}(t, t_0)\right)^*$$

The converse is proved similarly. □

Lemma 19.9. *The direct (19.61) and the corresponding adjoint (19.79) linear systems can be presented in the Hamiltonian form, i.e.,*

$$\boxed{\dot{x}(t) = \frac{\partial}{\partial z}H(z, x), \quad \dot{z}(t) = -\frac{\partial}{\partial x}H(z, x)} \tag{19.80}$$

where

$$\boxed{H(t, z, x) := (z, f(t, x)) = (z, A(t)\, x + g(t))} \tag{19.81}$$

is called the ***Hamiltonian function*** *for the system (19.61). In the stationary homogeneous case when*

$$\boxed{\dot{x}(t) = Ax(t), \quad x(t_0) = x_0 \in \mathbb{R}^{n\times n}, \quad t \geq t_0} \tag{19.82}$$

the Hamiltonian function is a first integral for (19.82).

Proof. The representation (19.80) follows directly from (19.81). In the stationary, when $\frac{\partial}{\partial t} H(t, z, x) = 0$, we have

$$\frac{d}{dt} H(t, z, x) = \frac{\partial}{\partial t} H(t, z, x) + \left(\frac{\partial}{\partial z} H(z, x), \dot{z}\right) + \left(\frac{\partial}{\partial x} H(z, x), \dot{x}\right)$$

$$= \left(\frac{\partial}{\partial z} H(z, x), -\frac{\partial}{\partial x} H(z, x)\right) + \left(\frac{\partial}{\partial x} H(z, x), \frac{\partial}{\partial z} H(z, x)\right) = 0$$

So, $H(t, z, x)$ is a constant. □

Lemma 19.10. *If $A(t) = -A^*(t)$ is skew Hermitian, then*

$$\boxed{\|x(t)\| = \underset{t}{\text{const}}} \tag{19.83}$$

Proof. One has directly

$$\frac{d}{dt} \|x(t)\|^2 = (\dot{x}(t), x(t)) + (x(t), \dot{x}(t))$$

$$= (A(t)x(t), x(t)) + (x(t), A(t)x(t))$$

$$= (A(t)x(t), x(t)) + (A^*(t)x(t), x(t))$$

$$= \left([A(t) + A^*(t)]x(t), x(t)\right) = 0$$

which proves the result. □

Lemma 19.11. (Green's formula) *Let $A(t)$, $g(t)$ and $\tilde{g}(t)$ be continuous for $t \in [a, b]$; $x(t)$ be a solution of (19.61) and $z(t)$ be a solution of (19.79). Then for all $t \in [a, b]$*

$$\boxed{\int_{s=a}^{t} [g^{\mathsf{T}}(s) z(s) - x(s)^{\mathsf{T}} \tilde{g}(s)] ds = x^{\mathsf{T}}(t) z(t) - x^{\mathsf{T}}(a) z(a)} \tag{19.84}$$

Proof. The relation (19.84) is proved by showing that both sides have the same derivatives, since $(Ay, z) = (y, A^*z)$. □

19.2.4 Index of increment for ODE solutions

Definition 19.4. *A number τ is called a **Lyapunov order number** (or **the index of the increment**) **for a vector function** $x(t)$ defined for $t \geq t_0$, if for every $\varepsilon > 0$ there exist positive constants C_ε^0 and C_ε such that*

$$\boxed{\begin{array}{l} \|x(t)\| \leq C_\varepsilon e^{(\tau+\varepsilon)t} \quad \textit{for all large } t \\ \\ \|x(t)\| \leq C_\varepsilon^0 e^{(\tau-\varepsilon)t} \quad \textit{for some arbitrary large } t \end{array}} \tag{19.85}$$

which equivalently can be formulated as

$$\boxed{\tau = \limsup_{t\to\infty} t^{-1} \ln \|x(t)\|} \tag{19.86}$$

Lemma 19.12. *If $x(t)$ is the solution of (19.61), then it has the Lyapunov order number*

$$\boxed{\begin{aligned}\tau \leq & \limsup_{t\to\infty} t^{-1} \ln \left(\|x(t_0)\| + \int_{s=t_0}^{t} \|f(s)\| \, ds \right) \\ & + \limsup_{t\to\infty} t^{-1} \int_{s=t_0}^{t} \|A(s)\| \, ds\end{aligned}} \tag{19.87}$$

Proof. It follows directly from (19.66). □

19.2.5 Riccati differential equation

Let us introduce the symmetric $n \times n$ matrix function $P(t) = P^{\intercal}(t) \in C^1[0, T]$ which satisfies the following ODE:

$$\boxed{\left.\begin{aligned}-\dot{P}(t) &= P(t) A(t) + A(t)^{\intercal} P(t) \\ &\quad - P(t) R(t) P(t) + Q(t) \\ P(T) &= G \geq 0\end{aligned}\right\}} \tag{19.88}$$

with

$$A(t), \quad Q(t) \in \mathbb{R}^{n\times n}, \quad R(t) \in \mathbb{R}^{m\times m} \tag{19.89}$$

Definition 19.5. *We call ODE (19.88) the* ***matrix Riccati differential equation.***

Theorem 19.9. (on the structure of the solution) *Let $P(t)$ be a symmetric nonnegative solution of (19.88) defined on $[0, T]$. Then there exist two functional $n \times n$ matrices $X(t), Y(t) \in C^1[0, T]$ satisfying the following linear ODE*

$$\boxed{\begin{aligned}&\begin{pmatrix}\dot{X}(t) \\ \dot{Y}(t)\end{pmatrix} = H(t) \begin{pmatrix}X(t) \\ Y(t)\end{pmatrix} \\ &X(T) = I, \quad Y(T) = P(T) = G\end{aligned}} \tag{19.90}$$

with

$$\boxed{H(t) = \begin{bmatrix}A(t) & -R(t) \\ -Q(t) & -A^{\intercal}(t)\end{bmatrix}} \tag{19.91}$$

where $A(t)$ and $Q(t)$ are as in (19.88) and such that $P(t)$ may be uniquely represented as

$$\boxed{P(t) = Y(t)\, X^{-1}(t)} \tag{19.92}$$

for any finite $t \in [0, T]$.

Proof.

(a) Notice that the matrices $X(t)$ and $Y(t)$ exist since they are defined by the solution to the ODE (19.90).

(b) Show that they satisfy the relation (19.92). Firstly, remark that $X(T) = I$, so $\det X(T) = 1 > 0$. From (19.90) it follows that $X(t)$ is a continuous matrix function and, hence, there exists a time τ such that for all $t \in (T - \tau, T]$ $\det X(t) > 0$. As a result, $X^{-1}(t)$ exists within the small semi-open interval $(T - \tau, T]$. Then, directly using (19.90) and in view of the identities

$$X^{-1}(t)\, X(t) = I, \quad \frac{d}{dt}\left[X^{-1}(t)\right] X(t) + X^{-1}(t)\, \dot{X}(t) = 0$$

it follows that

$$\begin{aligned} \frac{d}{dt}\left[X^{-1}(t)\right] &= -X^{-1}(t)\, \dot{X}(t)\, X^{-1}(t) \\ &= -X^{-1}(t)\left[A(t)\, X(t) - R(t)\, Y(t)\right] X^{-1}(t) \\ &= -X^{-1}(t)\, A(t) + X^{-1}(t)\, R(t)\, Y(t)\, X^{-1}(t) \end{aligned} \tag{19.93}$$

and, hence, for all $t \in (T - \tau, T]$ in view of (19.88)

$$\begin{aligned} \frac{d}{dt}\left[Y(t)\, X^{-1}(t)\right] &= \dot{Y}(t)\, X^{-1}(t) + Y(t)\, \frac{d}{dt}\left[X^{-1}(t)\right] \\ &= \left[-Q(t)\, X(t) - A^{\mathsf{T}}(t)\, Y(t)\right] X^{-1}(t) \\ &\quad + Y(t)\left[-X^{-1}(t)\, A(t) + X^{-1}(t)\, R(t)\, Y(t)\, X^{-1}(t)\right] \\ &= -Q(t) - A^{\mathsf{T}}(t)\, P(t) - P(t)\, A(t) \\ &\quad + P(t)\, R(t)\, P(t) = \dot{P}(t) \end{aligned}$$

which implies $\dfrac{d}{dt}\left[Y(t)\, X^{-1}(t) - P(t)\right] = 0$, or,

$$Y(t)\, X^{-1}(t) - P(t) = \underset{t \in (T-\tau, T]}{\text{const}}$$

But for $t = T$ we have

$$\underset{t\in(T-\tau,T]}{\text{const}} = Y(T)\,X^{-1}(T) - P(T) = Y(T) - P(T) = 0$$

So, for all $t \in (T-\tau, T]$ it follows that $P(t) = Y(t)\,X^{-1}(t)$.

(c) Show that $\det X(T-\tau) > 0$. The relations (19.90) and (19.92) lead to the following presentation within $t \in t \in (T-\tau, T]$

$$\dot{X}(t) = A(t)\,X(t) - R(t)\,Y(t) = [A(t) - R(t)\,P(t)]\,X(t)$$

and, by Liouville's theorem 19.7, it follows that

$$\det X(T-\tau) = \det X(0) \exp\left\{\int_{t=0}^{T-\tau} tr\,[A(t) - R(t)\,P(t)]\,dt\right\}$$

$$1 = \det X(T) = \det X(0) \exp\left\{\int_{t=0}^{T} tr\,[A(t) - R(t)\,P(t)]\,dt\right\}$$

$$\det X(T-\tau) = \exp\left\{-\int_{t=T-\tau}^{T} tr\,[A(t) - R(t)\,P(t)]\,dt\right\} > 0$$

By continuity, again there exists a time $\tau_1 > \tau$ that $\det X(t) > 0$ for any $t \in [T-\tau, T-\tau_1]$. Repeating the same considerations we may conclude that $\det X(t) > 0$ for any $t \in [0, T]$.

(d) Show that the matrix $G(t) := Y(t)\,X^{-1}(t)$ is symmetric. One has

$$\frac{d}{dt}[Y^{\top}(t)\,X(t) - X^{\top}(t)\,Y(t)] = \dot{Y}^{\top}(t)\,X(t) + Y^{\top}(t)\,\frac{d}{dt}[X(t)]$$

$$-\frac{d}{dt}X^{\top}(t)\,Y(t) - X^{\top}(t)\,\dot{Y}(t) = [-Q(t)\,X(t) - A^{\top}(t)\,Y(t)]^{\top}\,X(t)$$

$$+Y^{\top}(t)\,[A(t)\,X(t) - R(t)\,Y(t)] - [A(t)\,X(t) - R(t)\,Y(t)]^{\top}\,Y(t)$$

$$-X^{\top}(t)\,[-Q(t)\,X(t) - A^{\top}(t)\,Y(t)] = 0$$

and $Y(T)^{\top}\,X(T) - [X(T)]^{\top}\,Y(T) = Y^{\top}(T) - Y(T) = G^{\top} - G = 0$ that implies $Y^{\top}(t)\,X(t) - X^{\top}(t)\,Y(t) = 0$ for any $t \in [0, T]$. So, $Y^{\top}(t) = X^{\top}(t)\,Y(t)\,X^{-1}(t) = X^{\top}(t)\,P(t)$ and, hence, by the transposition operation we get $Y(t) = P^{\top}(t)\,X(t)$ and $P(t) = Y(t)\,X^{-1}(t) = P^{\top}(t)$. The symmetricity of $P(t)$ is proven.

(e) The Riccati differential equation (19.88) is uniquely solvable with $P(t) = Y(t)\,X^{-1}(t) \geq 0$ on $[0, T]$ since the matrices $X(t)$ and $Y(t)$ are uniquely defined by (19.92). □

19.2.6 Linear first order partial DE

Consider the following *linear first order partial* DE

$$\boxed{\sum_{i=1}^{n} X_i(x,z)\frac{\partial z}{\partial x_i} = Z(x,z)} \tag{19.94}$$

where $x \in \mathbb{R}^n$ is a vector of n independent real variables and $z = z(x)$ is a real-valued function of the class $C^1(\mathcal{X})$, $x \in \mathcal{X} \subseteq \mathbb{R}^n$. Defining

$$X(x,z) := (X_1(x,z),\ldots,X_n(x,z))^{\mathsf{T}}, \qquad \frac{\partial z}{\partial x} := \left(\frac{\partial z}{\partial x_1},\ldots,\frac{\partial z}{\partial x_n}\right)^{\mathsf{T}}$$

equation (19.94) can be rewritten as follows

$$\boxed{\left(X(x,z),\frac{\partial z}{\partial x}\right) = Z(x,z)} \tag{19.95}$$

Any function $z = z(x) \in C^1(\mathcal{X})$ satisfying (19.95) is its solution. If so, then its full differential dz is

$$\boxed{dz = \left(\frac{\partial z}{\partial x}, dx\right) = \sum_{i=1}^{n}\frac{\partial z}{\partial x_i}dx_i} \tag{19.96}$$

Consider also the following auxiliary system of ODE:

$$\boxed{\frac{dx_1}{X_1(x,z)} = \cdots = \frac{dx_n}{X_n(x,z)} = \frac{dz}{Z(x,z)}} \tag{19.97}$$

or, equivalently,

$$\boxed{X_1^{-1}(x,z)\frac{dx_1}{dz} = \cdots = X_n^{-1}(x,z)\frac{dx_n}{dz} = Z^{-1}(x,z)} \tag{19.98}$$

or,

$$\left.\begin{aligned}\frac{dx_1}{dz} &= X_1(x,z)Z^{-1}(x,z)\\ &\cdots\\ \frac{dx_n}{dz} &= X_n(x,z)Z^{-1}(x,z)\end{aligned}\right\} \tag{19.99}$$

which is called *the system of characteristic ODE* related to (19.95). The following important result, describing the natural connection of (19.95) and (19.98), is given below.

Lemma 19.13. *If $z = z(x)$ satisfies (19.97), then it satisfies (19.95) too.*

Proof. Indeed, by (19.99) and (19.96) we have

$$dx_i = X_i(x, z)\, Z^{-1}(x, z)\, dz$$

$$dz = \sum_{i=1}^{n} \frac{\partial z}{\partial x_i} dx_i = \sum_{i=1}^{n} \frac{\partial z}{\partial x_i} X_i(x, z)\, Z^{-1}(x, z)\, dz$$

which implies (19.94). □

19.2.6.1 Cauchy's method of characteristics

The method, presented here, permits to convert the solution of a linear first order partial DE of a system of nonlinear ODE.

Suppose that we can solve the system (19.98) of ODE and its solution is

$$\boxed{x_i = x_i(z, c_i), \quad i = 1, \ldots, n} \tag{19.100}$$

where c_i are some constants.

Definition 19.6. *The solutions (19.100) are called* ***the characteristics*** *of (19.94).*

Assume that this solution can be resolved with respect to the constants c_i, namely, there exist functions

$$\boxed{\psi_i = \psi_i(x, z) = c_i \ (i = 1, \ldots, n)} \tag{19.101}$$

Since these functions are constants on the solutions of (19.98) they are the first integrals of (19.98). Evidently, any arbitrary function $\Phi : \mathbb{R}^n \to \mathbb{R}$ of constants c_i $(i = 1, \ldots, n)$ is a constant too, that is,

$$\boxed{\Phi(c_1, \ldots, c_n) = \text{const}} \tag{19.102}$$

Without the loss of a generality we can take const $= 0$, so the equation (19.102) becomes

$$\boxed{\Phi(c_1, \ldots, c_n) = 0} \tag{19.103}$$

Theorem 19.10. (Cauchy's method of characteristics) *If the first integrals (19.74) $\psi_i(x, z)$ of the system (19.98) are* ***independent****, that is,*

$$\boxed{\det \left[\frac{\partial \psi_i(x, z)}{\partial x_j} \right]_{i,j=1,\ldots,n} \neq 0} \tag{19.104}$$

then the solution $z = z(x)$ of (19.97) can be found from the algebraic equation

$$\boxed{\Phi(\psi_1(x, z), \ldots, \psi_n(x, z)) = 0} \tag{19.105}$$

where $\Phi(\psi_1, \ldots, \psi_n)$ is an arbitrary function of its arguments.

Proof. By Theorem 16.8 on an implicit function, the systems (19.74) can be uniquely resolved with respect to x if (19.104) is fulfilled. So, the obtained functions (19.100) satisfy (19.99) and, hence, by Lemma 19.13 it follows that $z = z(x)$ satisfies (19.95). □

Example 19.3. *Let us integrate the equation*

$$\sum_{i=1}^{n} x_i \frac{\partial z}{\partial x_i} = pz \quad (p \text{ is a constant}) \tag{19.106}$$

The system (19.97)

$$\frac{dx_1}{x_1} = \cdots = \frac{dx_n}{x_n} = \frac{dz}{pz}$$

has the following first integrals

$$x_i^p - z = c_i \quad (i = 1, \ldots, n)$$

So, $z = z(x)$ can be found from the algebraic equation

$$\Phi\left(x_1^p - z, \ldots, x_n^p - z\right) = 0$$

where $\Phi(\psi_1, \ldots, \psi_n)$ is an arbitrary function, for example,

$$\Phi(\psi_1, \ldots, \psi_n) := \sum_{i=1}^{n} \lambda_i \psi_i, \quad \sum_{i=1}^{n} \lambda_i \neq 0$$

which gives

$$z = \left(\sum_{i=1}^{n} \lambda_i\right)^{-1} \sum_{i=1}^{n} \lambda_i x_i^p$$

19.3 Carathéodory's type ODE

19.3.1 Main definitions

The differential equation

$$\boxed{\dot{x}(t) = f(t, x(t)), \quad t \geq t_0} \tag{19.107}$$

in the regular case (with the continuous right-hand side in both variables) is known to be equivalent to the integral equation

$$\boxed{x(t) = x(t_0) + \int_{s=t_0}^{t} f(s, x(s))\, ds} \tag{19.108}$$

Definition 19.7. *If the function $f(t, x)$ is **discontinuous** in t and **continuous** in $x \in \mathbb{R}^n$, then the functions $x(t)$, satisfying the integral equation (19.108) where the integral is understood in the Lebesgue sense, is called the **solution** of ODE (19.107).*

The material presented below follows Filippov (1988).

Let us define more exactly the conditions which the function $f(t, x)$ should satisfy.

Condition 19.1. (Carathéodory's conditions) *In the domain $\mathcal{D}$ of the (t, x)-space let the following conditions be fulfilled:*

1. *the function $f(t, x)$ be defined and continuous in x for almost all t;*
2. *the function $f(t, x)$ be measurable (see (15.97)) in t for each x;*
3.

$$\boxed{\|f(t, x)\| \leq m(t)} \tag{19.109}$$

where the function $m(t)$ is summable (integrable in the Lebesgue sense) on each finite interval (if t is unbounded in the domain $\mathcal{D}$).

Definition 19.8.

*(a) Equation (19.107), where the function $f(t, x)$ satisfies conditions 19.1, is called **Carathéodory's type ODE**.*

*(b) A function $x(t)$, defined on an open or closed interval l, is called **a solution** of Carathéodory's type ODE if*

- *it is **absolutely continuous** on each interval $[\alpha, \beta] \in l$;*
- *it satisfies **almost everywhere** this equation or, which under conditions 19.1 is the same thing, satisfies the integral equation (19.108).*

19.3.2 Existence and uniqueness theorems

Theorem 19.11. (Filippov 1988) *For $t \in [t_0, t_0 + a]$ and $x : \|x - x_0\| \leq b$ let the function $f(t, x)$ satisfy Carathéodory's conditions 19.1. Then on a closed interval $[t_0, t_0 + d]$ **there exists a solution** of Cauchy's problem*

$$\boxed{\dot{x}(t) = f(t, x(t)), \quad x(t_0) = x_0} \tag{19.110}$$

In this case one can take an arbitrary number d such that

$$\boxed{0 < d \leq a, \quad \varphi(t_0 + d) \leq b \quad \text{where} \quad \varphi(t) := \int_{s=t_0}^{t} m(s)\, ds} \tag{19.111}$$

($m(t)$ is from (19.109)).

Proof. For integer $k \geq 1$ define $h := d/k$, and on the intervals $[t_0 + ih, t_0 + (i+1)h]$ $(i = 1, 2, \ldots, k)$ construct iteratively an approximate solution $x_k(t)$ as

$$x_k(t) := x_0 + \int_{s=t_0}^{t} f(s, x_{k-1}(s))\, ds \quad (t_0 < t \leq t_0 + d) \tag{19.112}$$

(for any initial approximation $x_0(s)$, for example, $x_0(s) = \text{const}$). Remember that if $f(t, x)$ satisfies Carathéodory's conditions 19.1 and $x(t)$ is measurable on $[a, b]$, then the composite function $f(t, x(t))$ is summable (integrable in the Lebesgue sense) on $[a, b]$. In view of this and by the condition (19.111) we obtain $\|x_k(t) - x_0\| \leq b$. Moreover, for any $\alpha, \beta : t_0 \leq \alpha < \beta \leq t_0 + d$

$$\|x_k(\beta) - x_k(\alpha)\| \leq \int_{s=t_0}^{t} m(s)\, ds = \varphi(\beta) - \varphi(\alpha) \tag{19.113}$$

The function $\varphi(t)$ is continuous on the closed interval $[t_0, t_0 + d]$ and therefore uniformly continuous. Hence, for any $\varepsilon > 0$ there exists a $\delta = \delta(\varepsilon)$ such that for all $|\beta - \alpha| < \delta$ the right-hand side of (19.113) is less than ε. Therefore, the functions $x_k(t)$ are equicontinuous (see (14.18)) and uniformly bounded (see (14.17)). Let us choose (by the Arzelà's theorem 14.16) from them a uniformly convergent subsequence having a limit $x(t)$. Since

$$\|x_k(s - h) - x(s)\| \leq \|x_k(s - h) - x_k(s)\| + \|x_k(s) - x(s)\|$$

and the first term on the right-hand side is less than ε for $h = d/k < \delta$, it follows that $x_k(s - h)$ tends to $x(s)$, by the chosen subsequence. In view of continuity of $f(t, x)$ in x, and the estimate $\|f(t, x)\| \leq m(t)$ (19.109) one can pass to the limit under the integral sign in (19.112). Therefore, we conclude that the limiting function $x(t)$ satisfies equation (19.108) and, hence, it is a solution of the problem (19.110). Theorem is proven. □

Corollary 19.12. *If Carathéodory's conditions 19.1 are satisfied for $t_0 - a \leq t \leq t_0$ and $\|x - x_0\| \leq b$, then a solution exists on the closed interval $[t_0 - d, t_0]$ where d satisfies (19.111).*

Proof. The case $t \leq t_0$ is reduced to the case $t \geq t_0$ by the simple substitution of $(-t)$ for t. □

Corollary 19.13. *Let $(t_0, x_0) \in \mathcal{D} \subseteq \mathbb{R}^{1+n}$ and let there exist a summable function $l(t)$ (in fact, this is a Lipschitz constant) such that for any two points (t, x) and (t, y) of $\mathcal{D}$*

$$\boxed{\|f(t, x) - f(t, y)\| \leq l(t)\, \|x - y\|} \tag{19.114}$$

Then in the domain $\mathcal{D}$ there exists ***at most one solution*** *of the problem (19.110).*

Proof. Using (19.114) it is sufficient to check Carathéodory's conditions 19.1. □

Theorem 19.12. (on the uniqueness) *If in Corollary 19.13 instead of (19.114) there is fulfilled the inequality*

$$\boxed{(f(t,x) - f(t,y), x - y) \leq l(t)\|x - y\|^2} \tag{19.115}$$

then in the domain $\mathcal{D}$ there exists the ***unique solution*** *of the problem (19.110).*

Proof. Let $x(t)$ and $y(t)$ be two solutions of (19.110). Define for $t_0 \leq t \leq t_1$ the function $z(t) := x(t) - y(t)$ for which it follows that

$$\frac{d}{dt}\|z\|^2 = 2\left(z, \frac{d}{dt}z\right) = 2(f(t,x) - f(t,y), x - y)$$

almost everywhere. By (19.115) we obtain $\frac{d}{dt}\|z\|^2 \leq l(t)\|z\|^2$ and, hence, $\frac{d}{dt}\left(e^{-L(t)}\|z\|^2\right) \leq 0$ where $L(t) = \int_{s=t_0}^{t} l(s)\,ds$. Thus, the absolutely continuous function (i.e., it is a Lebesgue integral of some other function) $e^{-L(t)}\|z\|^2$ does not increase, and it follows from $z(t_0) = 0$ that $z(t) = 0$ for any $t \geq t_0$. So, the uniqueness is proven. □

Remark 19.7. *The uniqueness of the solution of the problem (19.110) implies that if there exists two solutions of this problem, the graphs of which lie in the domain $\mathcal{D}$, then these solutions coincide on the common part of their interval of existence.*

Remark 19.8. *Since the condition (19.114) implies the inequality (19.115) (this follows from the Cauchy–Bounyakoski–Schwarz inequality), thus the uniqueness may be considered to be proven for $t \geq t_0$ also under the condition (19.114).*

19.3.3 Variable structure and singular perturbed ODE

19.3.3.1 Variable structure ODE

In fact, if by the *structure* of ODE (19.107) $\dot{x}(t) = f(t, x(t))$ we will understand the function $f(t,x)$, then evidently any nonstationary system may be considered as a dynamic system with a *variable structure*, since for different $t_1 \neq t_2$ we will have $f(t_1, x) \neq f(t_2, x)$. From this point of view such treatment seems to be naive and has no correct mathematical sense. But if we consider the special class of ODE (19.107) given by

$$\boxed{\dot{x}(t) = f(t, x(t)) := \sum_{i=1}^{N} \chi\left(t \in [t_{i-1}, t_i)\right) f^i(x(t))} \tag{19.116}$$

where $\chi(\cdot)$ is the characteristic function of the corresponding event, namely,

$$\boxed{\chi\left(t \in [t_{i-1}, t_i)\right) := \begin{cases} 1 & \text{if} \quad t \in [t_{i-1}, t_i) \\ 0 & \text{if} \quad t \notin [t_{i-1}, t_i) \end{cases}, \quad t_{i-1} < t_i} \tag{19.117}$$

then ODE (19.116) can also be treated as ODE with "jumping" parameters (coefficients). Evidently, if $f^i(x)$ are continuous on a compact $\mathcal{D}$ and, hence, are bounded, that is,

$$\max_{i=1,\ldots,N} \max_{x \in \mathcal{D}} \left\| f^i(x) \right\| \leq M \tag{19.118}$$

then the third Carathéodory's condition (19.109) will be fulfilled on the time interval $[\alpha, \beta]$, since

$$\boxed{m(t) = M \sum_{i=1}^{N} \chi\left(t \in [t_{i-1}, t_i)\right) = MN < \infty} \tag{19.119}$$

Therefore, such ODE equation (19.116) has *at most one solution*. If, in addition, for each $i = 1, \ldots, N$ the Lipschitz condition holds, i.e.,

$$\boxed{\left(f^i(x) - f^i(y), x - y\right) \leq l_i \|x - y\|^2}$$

then, as it follows from Theorem 19.12, this equation has a *unique solution*.

19.3.3.2 Singular perturbed ODE

Consider the following ODE containing a *singular type* of perturbation:

$$\boxed{\dot{x}(t) = f(x(t)) + \sum_{i=1}^{N} \mu_i \delta(t - t_i), \quad t, t_i \geq t_0} \tag{19.120}$$

where $\delta(t - t_i)$ is the "Dirac delta-function" (15.128), μ_i is a real constant and f is a continuous function. The ODE (19.120) must be understood as the integral equation

$$\boxed{x(t) = x(t_0) + \int_{s=t_0}^{t} f(x(s))\, dt + \sum_{i=1}^{N} \mu_i \int_{s=t_0}^{t} \delta(t - t_i)\, dt} \tag{19.121}$$

The last term, by the property (15.134), can be represented as

$$\sum_{i=1}^{N} \mu_i \int_{s=t_0}^{t} d\chi(s > t_i) = \sum_{i=1}^{N} \mu_i \chi(t > t_i)$$

where $\chi\,(t \geq t_i)$ is the "Heavyside's (step) function" defined by (19.117). Let us apply the following state transformation:

$$\tilde{x}\,(t) := x\,(t) + \sum_{i=1}^{N} \mu_i \chi\,(t > t_i)$$

New variable $\tilde{x}\,(t)$ satisfies (with $\mu_0 := 0$) the following ODE:

$$\boxed{\begin{aligned} \frac{d}{dt}\tilde{x}\,(t) &= f\left(\tilde{x}\,(t) - \sum_{i=1}^{N} \mu_i \chi\,(t > t_i)\right) \\ &= \sum_{i=1}^{N} \chi\,(t > t_i)\, f\left(x\,(t) - \sum_{s=1}^{i} \mu_i\right) \\ &= \sum_{i=1}^{N} \chi\,(t > t_i)\, \tilde{f}^i\,(x\,(t)) \end{aligned}} \tag{19.122}$$

where

$$\tilde{f}^i\,(x\,(t)) := f\left(x\,(t) - \sum_{s=1}^{i} \mu_i\right)$$

Claim 19.1. *This means that* ***the perturbed ODE (19.120) are equivalent to a variable structure ODE*** *(19.116).*

19.4 ODE with DRHS

In this section we will follow Utkin (1992), Filippov (1988) and Geligle *et al.* (1978).

19.4.1 Why ODE with DRHS are important in control theory

Here we will present some motivating consideration justifying our further study of ODE with DRHS. Let us start with the simplest scalar case dealing with the following standard ODE which is *affine* (linear) on control:

$$\boxed{\dot{x}\,(t) = f\,(x\,(t)) + u(t), \quad x\,(t) = x_0 \quad \text{is given}} \tag{19.123}$$

where $x(t), u\,(t) \in \mathbb{R}$ are interpreted here as the *state* of the system (19.123) and, respectively, the *control action* applied to it at time $t \in [0, T]$. The function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is a Lipschitz function satisfying the, so-called, Lipschitz condition, that is, for any $x, x' \in \mathbb{R}$

$$\left|f\,(x) - f\,\left(x'\right)\right| \leq L\left|x - x'\right|, \quad 0 \leq L < \infty \tag{19.124}$$

Problem 19.1. *Let us try to stabilize this system at the point* $x^* = 0$ *using the, so-called, feedback control*

$$\boxed{u(t) := u(x(t))} \tag{19.125}$$

considering the following informative situations

- *the* ***complete information*** *case when the function* $f(x)$ *is exactly known;*
- *the* ***incomplete information*** *case when it is only known that the function* $f(x)$ *is bounded as*

$$|f(x)| \leq f_0 + f^+ |x|, \quad f_0 < \infty, \quad f^+ < \infty \tag{19.126}$$

(this inequality is assumed to be valid for any $x \in \mathbb{R}$*).*

There are two possibilities:

1. use any **continuous control**, namely, take $u : \mathbb{R} \longrightarrow \mathbb{R}$ as a continuous function, i.e., $u \in C$;
2. use **a discontinuous control** which will be defined below.

19.4.1.1 The complete information case

Evidently, at the stationary point $x^* = 0$ any continuous control $u(t) := u(x(t))$ should satisfy the following identity

$$f(0) + u(0) = 0 \tag{19.127}$$

For example, this property may be fulfilled if we use the control $u(x)$ containing the nonlinear compensating term

$$u_{comp}(x) := -f(x)$$

and the linear correction term

$$u_{cor}(x) := -kx, \ k > 0$$

that is, if

$$u(x) = u_{comp}(x) + u_{cor}(x) = -f(x) - kx \tag{19.128}$$

The application of this control (19.128) to the system (19.123) implies that

$$\dot{x}(t) = -kx(t)$$

and, as the result, one gets

$$x(t) = x_0 \exp(-kt) \underset{t \to 0}{\to} 0$$

So, this *continuous control (19.128) in the complete information case solves the stabilization problem (19.1).*

19.4.1.2 The incomplete information case

Several informative situations may be considered.

1. $f(x)$ ***is unknown, but a priori it is known that*** $f(0) = 0$. In this situation the Lipschitz condition (19.124) is transformed into

$$|f(x)| = |f(x) - f(0)| \leq L|x|$$

which for the Lyapunov function candidate $V(x) = x^2/2$ implies

$$\begin{aligned} \dot{V}(x(t)) &= x(t)\dot{x}(t) \\ &= x(t)[f(x(t)) + u(x(t))] \leq |x(t)||f(x(t))| \\ &\quad + x(t)u(x(t)) \leq L|x(t)|^2 + x(t)u(x(t)) \end{aligned} \tag{19.129}$$

Since $f(x)$ is unknown let us select $u(x)$ in (19.128) as

$$\begin{aligned} u(x) &= u_{cor}(x) = -kx \\ u_{comp}(x) &:= 0 \end{aligned} \tag{19.130}$$

The use of (19.130) in (19.129) leads to the following identity:

$$\dot{V}(x(t)) \leq Lx^2(t) + x(t)u(x(t)) = (L-k)x^2(t) = -2(k-L)V(x(t))$$

Selecting k big enough (this method is known as the "high-gain control") we get

$$\dot{V}(x(t)) \leq -2(k-L)V(x(t)) \leq 0$$

$$V(x(t)) \leq V(x_0)\exp\left(-2[k-L]t\right) \underset{t\to 0}{\to} 0$$

This means that in the considered informative situation *the "high-gain control" solves the stabilization problem.*

2. $f(x)$ ***is unknown and it is admissible that*** $f(0) \neq 0$. In this situation the condition (19.127) never can be fulfilled since we do not know exactly the value $f(0)$ and, hence, neither the control (19.128) nor the control (19.130) can be applied. Let us try to apply a discontinuous control, namely, let us take $u(x)$ in the form of the, so-called, *sliding-mode (or relay) control*:

$$\boxed{u(x) = -k_t \operatorname{sign}(x), \quad k_t > 0} \tag{19.131}$$

where

$$\boxed{\operatorname{sign}(x) := \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x < 0 \\ \in [-1, 1] & \text{if } x = 0 \end{cases}} \tag{19.132}$$

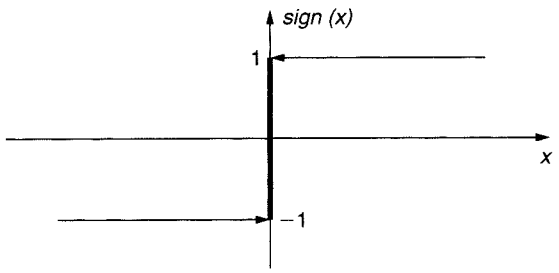

Fig. 19.1. The signum function.

(see Fig. 19.1). Starting from some $x_0 \neq 0$, analogously to (19.129) and using (19.126), we have

$$\dot{V}(x(t)) = x(t)\,\dot{x}(t) = x(t)\,[f(x(t)) + u(x(t))]$$

$$\leq |x(t)|\,|f(x(t))| + x(t)\,u(x(t)) \leq |x(t)|\left(f_0 + f^+ |x(t)|\right)$$

$$- k_t x(t)\,\text{sign}\,(x(t)) = |x(t)|\,f_0 + f^+ |x(t)|^2 - k_t |x(t)|$$

Taking

$$\boxed{\begin{array}{l} k_t = k(x(t)) := k^0 + k^1 |x(t)| \\ \qquad k^0 > f_0,\ k^1 > f^+ \end{array}} \tag{19.133}$$

we have

$$\dot{V}(x(t)) \leq -|x(t)|\left(k^0 - f_0\right) - \left(k^1 - f^+\right)|x(t)|^2$$

$$\leq -|x(t)|\left(k^0 - f_0\right) = -\sqrt{2}\left(k^0 - f_0\right)\sqrt{V(x(t))} \leq 0$$

Hence,

$$\frac{dV(x(t))}{\sqrt{V(x(t))}} \leq -\sqrt{2}\left(k^0 - f_0\right)dt$$

which leads to the following identity

$$2\left(\sqrt{V(x(t))} - \sqrt{V(x_0)}\right) \leq -\sqrt{2}\left(k^0 - f_0\right)t$$

or, equivalently,

$$\sqrt{V(x(t))} \leq \sqrt{V(x_0)} - \frac{k^0 - f_0}{\sqrt{2}}t$$

This means that the, so-called, "reaching phase", during which the system (19.123) controlled by the sliding-mode algorithm (19.131)–(19.133) reaches the origin, is equal to

$$\boxed{t^* = \frac{\sqrt{2V(x_0)}}{k^0 - f_0}} \tag{19.134}$$

Conclusion 19.1. It follows from the considerations above that **the discontinuous (in this case, sliding-mode) control** (19.131)–(19.133) **can stabilize the class of the dynamic systems** (19.123), (19.124), (19.126) **in finite time** (19.134) without the exact knowledge of its model. Besides, **the reaching phase may be done as small as you wish** by the simple selection of the gain parameter k^0 in (19.134). In other words, the discontinuous control (19.131)–(19.133) is **robust with respect to the presence of unmodeled dynamics** in (19.123) which means that it is capable of stabilizing a wide class of "black/gray-box" systems.

Remark 19.9. *Evidently, using such discontinuous control, the trajectories of the controlled system cannot stay in the stationary point $x^* = 0$ since it arrives at it in finite time but with a nonzero rate, namely, with $\dot{x}(t)$ such that*

$$\dot{x}(t) = \begin{cases} f(0) + k^0 & \text{if} \quad x(t) \longrightarrow +0 \\ f(0) - k^0 & \text{if} \quad x(t) \longrightarrow -0 \end{cases}$$

which provokes the, so-called, "chattering effect" (see Fig. 19.2). Simple engineering considerations show that some sort of smoothing (or low-pass filtering) should be applied to keep dynamics close to the stationary point $x^* = 0$.

Remark 19.10. *Notice that when $x(t) = x^* = 0$ we only know that*

$$\boxed{\dot{x}(t) \in \left[f(0) - k^0, \quad f(0) + k^0\right]} \tag{19.135}$$

*This means that we deal with a **differential inclusion** (not an equation) (19.135). So, we need to define what does it mean mathematically correctly a solution of a differential inclusion and what is it itself.*

All these questions, arising in the remarks above, will be considered below in detail and be illustrated by the corresponding examples and figures.

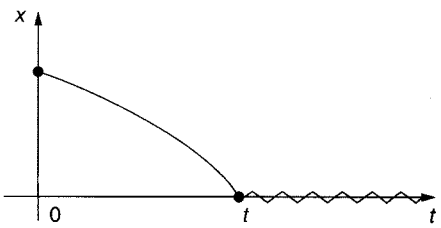

Fig. 19.2. The "chattering effect".

19.4.2 ODE with DRHS and differential inclusions

19.4.2.1 General requirements to a solution

As it is well known, a solution of the differential equation

$$\dot{x}(t) = f(t, x(t)) \tag{19.136}$$

with a continuous right-hand side is a function $x(t)$ which has a derivative and satisfies (19.136) everywhere on a given interval. This definition is not, however, valid for DE with DRHS since in some points of discontinuity the derivative of $x(t)$ does not exist. That's why the consideration of DE with DRHS requires a generalization of the concept of a solution. Anyway, such a generalized concept should necessarily meet the following requirements:

- For differential equations with a continuous right-hand side the definition of a solution must be equivalent to the usual (standard) one.
- For the equation $\dot{x}(t) = f(t)$ the solution should be the functions $x(t) = \int f(t)\,dt + c$ only.
- For any initial data $x(t_0) = x_{\text{init}}$ within a given region the solution $x(t)$ should exist (at least locally) for any $t > t_0$ and admit the possibility to be continued up to the boundary of this region or up to infinity (when $(t, x) \to \infty$).
- The limit of a uniformly convergent sequence of solutions should be a solution too.
- Under the commonly used changes of variables a solution must be transformed into a solution.

19.4.2.2 The definition of a solution

Definition 19.9. *A vector-valued function $f(t, x)$, defined by a mapping $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^p$, is said to be* ***piecewise continuous*** *in a finite domain $\mathcal{G} \subseteq \mathbb{R}^{n+1}$ if G consists of a finite number of a domain $\mathcal{G}_i$ $(i = 1, \ldots, l)$, i.e.,*

$$\mathcal{G} = \bigcup_{i=1}^{l} \mathcal{G}_i$$

such that in each of them the function $f(t, x)$ is continuous up to the boundary

$$\mathcal{M}_i := \bar{\mathcal{G}}_i \backslash \mathcal{G}_i \quad (i = 1, \ldots, l) \tag{19.137}$$

of a measure zero.

The most frequent case is the one where the set

$$\mathcal{M} = \bigcup_{i=1}^{l} \mathcal{M}_i$$

of all discontinuity points consists of a finite number of hypersurfaces

$$0 = S_k(x) \in C^1, \quad k = 1, \ldots, m$$

where $S_k(x)$ is a smooth function.

Definition 19.10. *The set* $\mathcal{M}$ *defined as*

$$\boxed{\mathcal{M} = \{x \in \mathbb{R}^n \mid S(x) = (S_1(x), \ldots, S_m(x))^{\mathsf{T}} = 0\}} \tag{19.138}$$

is called a ***manifold*** *in* $\mathbb{R}^n$. *It is referred to as* ***a smooth manifold*** *if* $S_k(x) \in C^1$, $k = 1, \ldots, m$.

Now we are ready to formulate the main definition of this section.

Definition 19.11. (A solution in Filippov's sense) *A solution* $x(t)$ *on a time interval* $[t_0, t_f]$ *of ODE* $\dot{x}(t) = f(t, x(t))$ *with DRHS* ***in Filippov's sense*** *is called a solution of the* ***differential inclusion***

$$\boxed{\dot{x}(t) \in \mathcal{F}(t, x(t))} \tag{19.139}$$

that is, an ***absolutely continuous*** *on* $[t_0, t_f]$ *function* $x(t)$ *(which can be represented as a Lebesgue integral of another function) satisfying (19.139) almost everywhere on* $[t_0, t_f]$, *where the set* $\mathcal{F}(t, x)$ *is* ***the smallest convex closed set containing all limit values of the vector-function*** $f(t, x^*)$ *for* $(t, x^*) \notin \mathcal{M}$, $x^* \to x$, $t = \text{const}$.

Remark 19.11. *The set* $\mathcal{F}(t, x)$

1. *consists of one point* $f(t, x)$ *at points of continuity of the function* $f(t, x)$;
2. *is a segment (a convex polygon, or polyhedron), which in the case when* $(t, x) \in \mathcal{M}_i$ *(19.137) has the vertices*

$$f_i(t, x) := \lim_{(t, x^*) \in \mathcal{G}_i,\ x^* \to x} f(t, x^*) \tag{19.140}$$

All points $f_i(t, x)$ *are contained in* $\mathcal{F}(t, x)$, *but it is not obligatory that all of them are vertices.*

Example 19.4. *For the scalar differential inclusion*

$$\dot{x}(t) \in -\text{sign}(x(t))$$

the set $\mathcal{F}(t, x)$ *is as follows (see Fig. 19.3):*

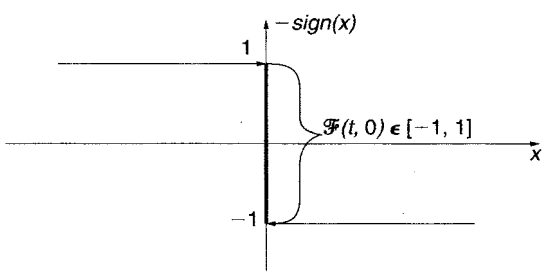

Fig. 19.3. The right-hand side of the differential inclusion $\dot{x}_t = -sign(x_t)$.

1. $\mathcal{F}(t,x) = -1 \quad if \quad x > 0;$
2. $\mathcal{F}(t,x) = 1 \quad if \quad x < 0;$
3. $\mathcal{F}(t,x) = [-1,1] \quad if \quad x = 0.$

19.4.2.3 Semi-continuous sets as functions

Definition 19.12. *A multi-valued function (or a set) $\mathcal{F} = \mathcal{F}(t,x)$ $(t \in \mathbb{R},\ x \in \mathbb{R}^n)$ is said to be*

- *a **semi-continuous** in the point (t_0, x_0) if for any $\varepsilon > 0$ there exists $\delta = \delta(t_0, x_0, \varepsilon)$ such that the inclusion*

$$\boxed{(t,x) \in \{z \mid \|z - (t_0,x_0)\| \le \delta\}} \tag{19.141}$$

implies

$$\boxed{\mathcal{F}(t,x) \in \{f \mid \|f - f(t_0,x_0)\| \le \varepsilon\}} \tag{19.142}$$

- *a **continuous** in the point (t_0, x_0) if it is semi-continuous and, additionally, for any $\varepsilon > 0$ there exists $\delta = \delta(t_0, x_0, \varepsilon)$ such that the inclusion*

$$\boxed{(t_0,x_0) \in \{z \mid \|z - (t',x')\| \le \delta\}} \tag{19.143}$$

implies

$$\boxed{\mathcal{F}(t_0,x_0) \in \{f \mid \|f - f(t',x')\| \le \varepsilon\}} \tag{19.144}$$

Example 19.5. *Consider the multi-valued functions $\mathcal{F}(t,x)$ depicted at Fig. 19.4.*

Here the functions (sets) $\mathcal{F}(t,x)$, corresponding to the plots (1)–(4), are semi-continuous.

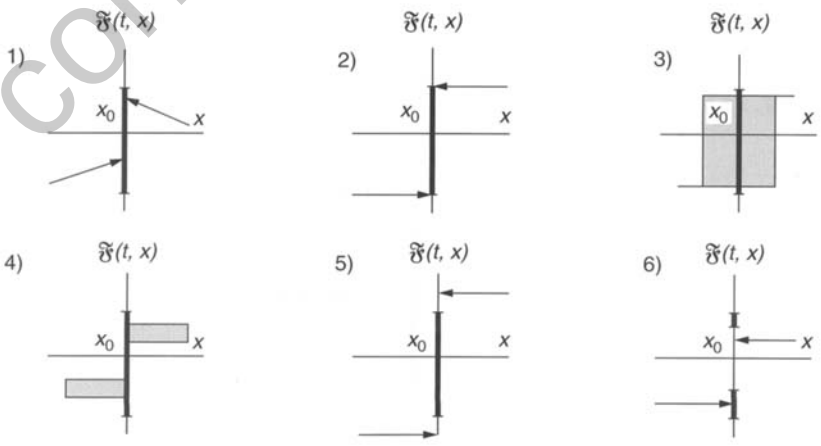

Fig. 19.4. Multi-valued functions.

19.4.2.4 Theorem on the local existence of solution

First, let us formulate some useful result which will be applied in the following considerations.

Lemma 19.14. *If $x(t)$ is absolutely continuous on the interval $t \in [\alpha, \beta]$ and within this interval $\|\dot{x}(t)\| \le c$, then*

$$\boxed{\frac{1}{\beta - \alpha}(x_\beta - x_\alpha) \subset \underset{\text{a.a. } t \in [\alpha,\beta]}{Conv} \cup \dot{x}(t)} \tag{19.145}$$

where Conv is a convex closed set containing $\cup \dot{x}(t)$ for almost all $t \in [\alpha, \beta]$.

Proof. By the definition of the Lebesgue integral

$$\frac{1}{\beta - \alpha}(x_\beta - x_\alpha) = \frac{1}{\beta - \alpha}\int_{t=\alpha}^{\beta} \dot{x}(t)\,dt = \lim_{k\to\infty} s_k$$

where

$$s_k = \sum_{i=1}^{k} \frac{\mu_i}{|\beta - \alpha|}\dot{x}(t_i), \quad \mu_i \ge 0, \ \sum_{i=1}^{k} \mu_i = |\beta - \alpha|$$

are Lebesgue sums of the integral above. But $s_k \in \underset{\text{a.a. } t \in [\alpha,\beta]}{Conv} \cup \dot{x}(t)$. Hence, the same fact is valid for the limit vectors $\lim_{k\to\infty} s_k$ which proves the lemma. □

Theorem 19.13. (on the local existence) *Suppose that*

A1. a multi-valued function (set) $\mathcal{F}(t, x)$ is a semi-continuous at each point

$$(t, x) \in D_{\gamma,\rho}(t_0, x_0) := \{(t, x) \mid \|x - x_0\| \le \gamma, |t - t_0| \le \rho\}$$

A2. the set $\mathcal{F}(t, x)$ is a convex compact and $\sup \|y\| = c$ whenever

$$y \in \mathcal{F}(t, x) \text{ and } (t, x) \in D_{\gamma,\rho}(t_0, x_0)$$

Then for any t such that $|t - t_0| \le \tau := \rho/c$ there exists an absolutely continuous function $x(t)$ (maybe not unique) such that

$$\dot{x}(t) \in \mathcal{F}(t, x), \quad x(t_0) = x_0$$

that is, the ODE $\dot{x}(t) = f(t, x(t))$ with DRHS ***has a local solution*** *in Filippov's sense (see Definition 19.11).*

Proof. Divide the interval $[t_0 - \tau, t_0 + \tau]$ into $2m$-parts $t_i^{(m)} := t_0 + j\frac{\tau}{m}$ $(i = 0, \pm 1, \ldots \pm m)$ and construct the, so-called, partially linear *Euler's curves*

$$x_m(t) := x_m\left(t_i^{(m)}\right) + \left(t - t_i^{(m)}\right) \hat{f}_i^{(m)}(t_i^{(m)}), \quad t \in \left[t_i^{(m)}, t_i^{(m+1)}\right]$$

$$x_m\left(t_0^{(m)}\right) = x_0, \quad \hat{f}_i^{(m)}\left(t_i^{(m)}\right) \in \mathcal{F}\left(t_i^{(m)}, x_m\left(t_i^{(m)}\right)\right)$$

By the assumption (A2) it follows that $x_m(t)$ is uniformly bounded and continuous on $D_{\gamma,\rho}(t_0, x_0)$. Then, by Arzelà's theorem 14.16 there exists a subsequence $\{x_{m_k}(t)\}$ which uniformly converges to some vector function $x(t)$. This limit evidently has a Lipschitz constant on $D_{\gamma,\rho}(t_0, x_0)$ and satisfies the initial condition $x(t_0) = x_0$. In view of Lemma 19.14, for any $h > 0$ we have

$$h^{-1}\left[x_{m_k}(t+h) - x_{m_k}(t)\right] \subset \underset{\text{a.a.}[t_0-\tau,t_0+\tau]}{Conv} \bigcup_{i=-m_k}^{m_k} \hat{f}_i^{(m_k)}$$

$$\subset \underset{\text{a.a. } \lambda \in \left[t_0 - \frac{\tau}{m_k}, t_0 + \frac{\tau}{m_k} + h\right]}{Conv} \bigcup_{i=-m_k}^{m_k} \hat{f}_i^{(m_k)}(\lambda) := A_k$$

Since $\mathcal{F}(t, x)$ is semi-continuous, it follows that $\sup_{x \in A_k} \inf_{y \in A} \|x - y\| \to 0$ whenever $k \to \infty$ (here $A := \underset{\text{a.a. } \lambda \in [t, t+h]}{Conv} \cup_{l=-m_l}^{m_l} f(\lambda, x_\lambda)$). The convexity of $\mathcal{F}(t, x)$ implies also that $\sup_{x \in A} \inf_{y \in \mathcal{F}(t,x)} \|x - y\| \to 0$ when $h \to 0$ which, together with previous property, proves the theorem. □

Remark 19.12. *By the same reasons as for the case of regular ODE, we may conclude that the solution of the differential inclusion (if it exists) is continuously dependent on t_0 and x_0.*

19.4.3 Sliding mode control

19.4.3.1 Sliding mode surface

Consider the special case where the function $f(t, x)$ is discontinuous on a smooth surface S given by the equation

$$\boxed{s(x) = 0, \quad s : \mathbb{R}^n \to \mathbb{R}, \quad s(\cdot) \in C^1} \tag{19.146}$$

The surface separates its neighborhood (in $\mathbb{R}^n$) into domains $\mathcal{G}^+$ and $\mathcal{G}^-$. For $t = \text{const}$ and for the point x^* approaching the point $x \in S$ from the domains $\mathcal{G}^+$ and $\mathcal{G}^-$ let us suppose that the function $f(t, x^*)$ has the following limits:

$$\begin{aligned} \lim_{(t,x^*) \in \mathcal{G}^-,\ x^* \to x} f(t, x^*) &= f^-(t, x) \\ \lim_{(t,x^*) \in \mathcal{G}^+,\ x^* \to x} f(t, x^*) &= f^+(t, x) \end{aligned} \tag{19.147}$$

Then by Filippov's definition, $\mathcal{F}(t, x)$ is a linear segment joining the endpoints of the vectors $f^-(t, x)$ and $f^+(t, x)$. Two situations are possible.

- If for $t \in (t_1, t_2)$ this segment lies on one side of the plane P tangent to the surface S at the point x, the solutions for these t pass from one side of the surface S to the other one (see Fig. 19.5 depicted at the point $x = 0$);
- If this segment intersects the plane P, the intersection point is the endpoint of the vector $f^0(t, x)$ which defines the velocity of the motion

$$\boxed{\dot{x}(t) = f^0(t, x(t))} \tag{19.148}$$

along the surface S in $\mathbb{R}^n$ (see Fig. 19.6 depicted at the point $x = 0$). Such a solution, lying on S for all $t \in (t_1, t_2)$, is often called a **sliding motion** (or **mode**). Defining the projections of the vectors $f^-(t, x)$ and $f^+(t, x)$ to the surface S ($\nabla s(x) \neq 0$) as

$$p^-(t, x) := \frac{\left(\nabla s(x), f^-(t, x)\right)}{\|\nabla s(x)\|}, \quad p^+(t, x) := \frac{\left(\nabla s(x), f^+(t, x)\right)}{\|\nabla s(x)\|}$$

one can find that when $p^-(t, x) < 0$ and $p^+(t, x) > 0$

$$f^0(t, x) = \alpha f^-(t, x) + (1 - \alpha) f^+(t, x)$$

Here α can be easily found from the equation

$$\left(\nabla s(x), f^0(t, x)\right) = 0$$

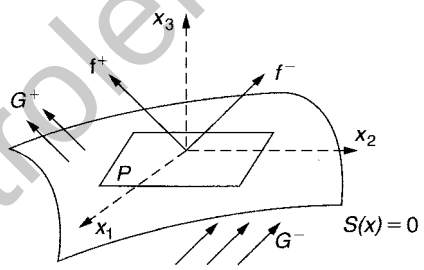

Fig. 19.5. The sliding surface and the rate vector field at the point $x = 0$.

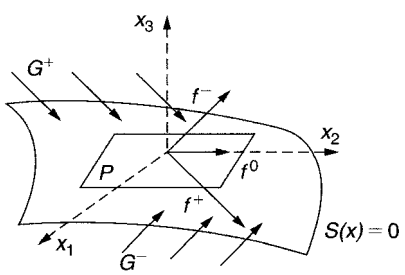

Fig. 19.6. The velocity of the motion.

or, equivalently,

$$0 = \left(\nabla s(x), \alpha f^{-}(t,x) + (1-\alpha) f^{+}(t,x)\right)$$

$$= \alpha p^{-}(t,x) + (1-\alpha) p^{+}(t,x)$$

which implies

$$\alpha = \frac{p^{+}(t,x)}{p^{+}(t,x) - p^{-}(t,x)}$$

Finally, we obtain that

$$\boxed{\begin{aligned} f^{0}(t,x) = &\frac{p^{+}(t,x)}{p^{+}(t,x) - p^{-}(t,x)} f^{-}(t,x) \\ &+ \left(1 - \frac{p^{+}(t,x)}{p^{+}(t,x) - p^{-}(t,x)}\right) f^{+}(t,x) \end{aligned}} \tag{19.149}$$

19.4.3.2 Sliding mode surface as a desired dynamic

Let us consider in this subsection several examples demonstrating that a desired dynamic behavior of a controlled system may be expressed not only in the traditional manner, using some cost (or payoff) functionals as possible performance indices, but also representing a nominal (desired) dynamic in the form of a surface (or manifold) in a space of coordinates.

First-order tracking system: consider a first-order system given by the following ODE:

$$\boxed{\dot{x}(t) = f(t, x(t)) + u(t)} \tag{19.150}$$

where $u(t)$ is a control action and $f : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is supposed to be bounded, that is,

$$|f(t, x(t))| \leq f^{+} < \infty$$

Assume that the desired dynamics (signal), which should be tracked, is given by a smooth function $r(t)$ ($|\dot{r}(t)| \leq \rho$), such that the tracking error e_t is (see Fig. 19.7)

$$e(t) := x(t) - r(t)$$

Select a desired surface s as follows

$$\boxed{s(e) = e = 0} \tag{19.151}$$

which exactly corresponds to an "ideal tracking" process. Then, designing the control $u(t)$ as

$$\boxed{u(t) := -k \ \operatorname{sign}(e(t))}$$

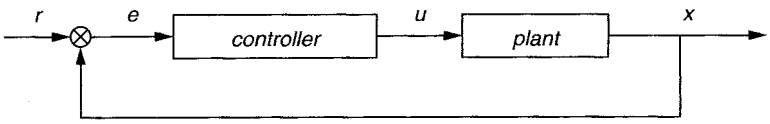

Fig. 19.7. A tracking system.

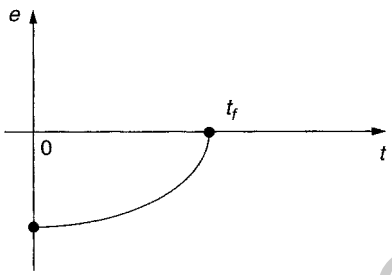

Fig. 19.8. The finite time error cancellation.

we derive that

$$\dot{e}(t) = f(t, x(t)) - \dot{r}(t) - k \operatorname{sign}(e(t))$$

and for $V(e) = e^2/2$ we have

$$\dot{V}(e(t)) = e(t)\dot{e}(t) = e(t)[f(t, x(t)) - \dot{r}(t) - k \operatorname{sign}(e(t))]$$

$$= e(t)[f(t, x(t)) - \dot{r}(t)] - k|e(t)| \leq |e(t)|[f^+ + \rho] - k|e(t)|$$

$$= |e(t)|[f^+ + \rho - k] = -\sqrt{2}[k - f^+ - \rho]\sqrt{V(e(t))}$$

and, hence,

$$\sqrt{V(e_t)} \leq \sqrt{V(e_0)} - \frac{1}{\sqrt{2}}[k - f^+ - \rho]t$$

So, taking $k > f^+ + \rho$ implies the finite time convergence of e_t (with the reaching phase $t_f = \frac{\sqrt{2V(e_0)}}{k - f^+ - \rho}$) to the surface (19.151) (see Figs. 19.8 and 19.9).

Stabilization of a second order relay system: let us consider a second order relay system given by the following ODE

$$\boxed{\begin{array}{l} \ddot{x}(t) + a_2\dot{x}(t) + a_1 x(t) = u(t) + \xi(t) \\ u(t) = -k \operatorname{sign}(\tilde{s}(t)) - \text{the relay-control} \\ \tilde{s}(t) := \dot{x}(t) + cx(t), \quad c > 0 \\ |\xi(t)| \leq \xi^+ - \text{a bounded unknown disturbance} \end{array}} \qquad (19.152)$$

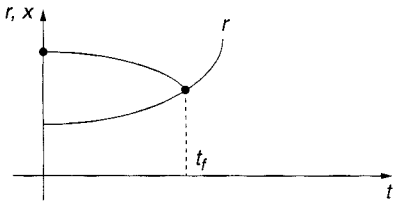

Fig. 19.9. The finite time tracking.

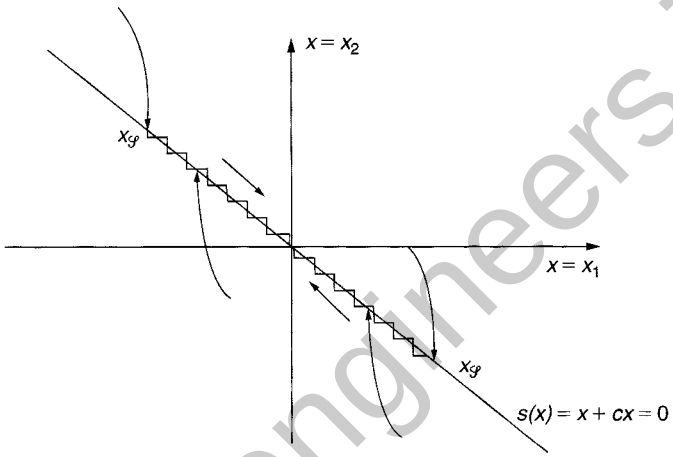

Fig. 19.10. The sliding motion on the sliding surface $s(x) = x_2 + cx_1$.

We may rewrite the dynamic ($x_1 := x$) as

$$\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= -a_1 x_1(t) - a_2 x_2(t) + u(t) + \xi(t) \\ u(t) &= -k\,\mathrm{sign}\,(x_2(t) + cx_1(t)) \end{aligned} \tag{19.153}$$

Here the *sliding surface* is

$$\boxed{s(x) = x_2 + cx_1}$$

So, the sliding motion, corresponding to the dynamics $\tilde{s}(t) := \dot{x}(t) + cx(t) = 0$, is given by (see Fig. 19.10)

$$x(t) = x_0 e^{-ct}$$

Let us introduce the following Lyapunov function candidate:

$$\boxed{V(s) = s^2/2}$$

for which the following property holds:

$$\begin{aligned}\dot{V}(s) = s\dot{s} &= s(x(t))\left[\frac{\partial s(x(t))}{\partial x_1}\dot{x}_1(t) + \frac{\partial s(x(t))}{\partial x_2}\dot{x}_2(t)\right] \\ &= s(x(t))\left[cx_2(t) - a_1x_1(t) - a_2x_2(t) + u(t) + \xi(t)\right] \\ &\le |s(x(t))|\left[|a_1|\,|x_1(t)| + (c + |a_2|)\,|x_2(t)| + \xi^+\right] - ks(x(t))\,\text{sign}(s(x(t))) \\ &= -\left[k - |a_1|\,|x_1(t)| - (c + |a_2|)\,|x_2(t)| - \xi^+\right]|s(x(t))| \le 0\end{aligned}$$

if we take

$$\boxed{k = |a_1|\,|x_1(t)| + (c + |a_2|)\,|x_2(t)| + \xi^+ + \rho, \quad \rho > 0} \tag{19.154}$$

This implies $\dot{V}(s) \le -\rho\sqrt{2V(s)}$, and, hence, the reaching time t_f (see Fig. 19.9) is

$$\boxed{t_f = \frac{\sqrt{2V(s_0)}}{\rho} = \frac{|\dot{x}_0 + cx_0|}{\rho}} \tag{19.155}$$

Sliding surface and a related LQ-problem: consider a linear multi-dimensional plant given by the following ODE

$$\boxed{\begin{array}{l}\dot{x}(t) = A(t)\,x(t) + B(t)\,u(t) + \xi(t) \\ x_0 \quad \text{is given,} \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^r \\ B^{\top}(t)\,B(t) > 0 \ \text{rank}\,[B(t)] = r \quad \text{for any} \quad t \in [t_s, t_1] \\ \xi(t) \quad \text{is known as external perturbation}\end{array}} \tag{19.156}$$

A sliding mode is said to be taking place in this system (19.156) if there exists a finite reaching time t_s, such that the solution $x(t)$ satisfies

$$\boxed{\sigma(x, t) = 0 \quad \text{for all} \quad t \ge t_s} \tag{19.157}$$

where $\sigma(x, t) : R^n \times R_+ \to R^r$ is *a sliding function* and (21.65) defines *a sliding surface* in R^{n+1}. For each $t_1 > 0$ the quality of the system (19.156) motion in the sliding surface (21.65) is characterized by *the performance index* (Utkin (1992))

$$\boxed{J_{t_s,t_1} = \frac{1}{2}\int_{t_s}^{t_1}(x(t), \quad Qx(t))\,dt, \quad Q = Q^{\top} \ge 0} \tag{19.158}$$

Below we will show that the system motion in the sliding surface (21.65) does not depend on the control function u, that's why (19.158) is a functional of x and $\sigma(x, t)$ only. Let us try to solve the following problem.

Problem formulation: for the given linear system (19.156) and $t_1 > 0$ define the optimal sliding function $\sigma = \sigma(x, t)$ (21.65) providing the optimization in the sense of (19.158) *in the sliding mode*, that is,

$$\boxed{J_{t_s,t_1} \to \inf_{\sigma \in \Xi}} \tag{19.159}$$

where Ξ is the set of the admissible smooth (differentiable on all arguments) sliding functions $\sigma = \sigma(x, , t)$. So, we wish to minimize the performance index (19.158) varying (optimizing) the sliding surface $\sigma \in \Xi$.

Introduce a new state vector z defined by

$$z = T(t)\, x \tag{19.160}$$

where the linear nonsingular transformations $T(t)$ are given by

$$T(t) := \begin{bmatrix} I_{(n-r)\times(n-r)} & -B_1(t)\,(B_2(t))^{-1} \\ 0 & (B_2(t))^{-1} \end{bmatrix} \tag{19.161}$$

Here $B_1^{(n-r)\times(n-r)}(t) \in \mathbb{R}^{r\times r}$ and $B_2(t) \in \mathbb{R}^{r\times r}$ represent the matrices $B(t)$ in the form

$$B(t) = \begin{bmatrix} B_1(t) \\ B_2(t) \end{bmatrix}, \quad \det[B_2(t)] \neq 0 \ \forall\, t \geq 0 \tag{19.162}$$

Applying (19.162) to the system (19.156), we obtain (below we will omit the time dependence)

$$\dot{z} = \begin{pmatrix} \dot{z}_1 \\ \dot{z}_2 \end{pmatrix} = \begin{pmatrix} \tilde{A}_{11}z_1 + \tilde{A}_{12}z_2 \\ \tilde{A}_{21}z_1 + \tilde{A}_{22}z_2 \end{pmatrix} + \begin{pmatrix} 0 \\ u \end{pmatrix} + \begin{pmatrix} \tilde{\xi}_1 \\ \tilde{\xi}_2 \end{pmatrix} \tag{19.163}$$

where $z_1 \in R^{n-r}$, $z_2 \in R^r$ and

$$\tilde{A} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{12} \\ \tilde{A}_{21} & \tilde{A}_{22} \end{bmatrix} = TAT^{-1} + \dot{T}T^{-1}, \quad \begin{pmatrix} \tilde{\xi}_1(t) \\ \tilde{\xi}_2(t) \end{pmatrix} = T\xi(t) \tag{19.164}$$

Using the operator T^{-1}, it follows $x = T^{-1}z$ and, hence, the performance index (19.158) in new variables z may be rewritten as

$$\boxed{\begin{aligned} J_{t_s,t_1} &= \frac{1}{2}\int_{t_s}^{t_1} (x, Qx)\, dt = \frac{1}{2}\int_{t_s}^{t_1} \left(z, \tilde{Q}^{\alpha} z\right) dt \\ &= \frac{1}{2}\int_{t_s}^{t_1} \left[\left(z_1, \tilde{Q}_{11}^{\alpha} z_1\right) + 2\left(z_1, \tilde{Q}_{12}^{\alpha} z_2\right) + \left(z_2, \tilde{Q}_{22}^{\alpha} z_2\right)\right] dt \\ \tilde{Q} &:= \left(T^{-1}\right)^{\mathsf{T}} Q T^{-1} = \begin{bmatrix} \tilde{Q}_{11} & \tilde{Q}_{12} \\ \tilde{Q}_{21} & \tilde{Q}_{22} \end{bmatrix} \end{aligned}} \tag{19.165}$$

and the sliding function $\sigma = \sigma(x, t)$ becomes

$$\sigma = \sigma\left(T^{-1}z, t\right) := \tilde{\sigma}(z, t) \tag{19.166}$$

Remark 19.13. *The matrices $\tilde{Q}_{11}, \tilde{Q}_{12}, \tilde{Q}_{21}$ and $\tilde{Q}_{22}$ are supposed to be symmetric. Otherwise, they can be symmetrized as follows:*

$$\begin{aligned} J_{t_s,t_1} &= \frac{1}{2}\int_{t_s}^{t_1}\left[(z_1, \tilde{Q}_{11}z_1) + 2(z_1, \tilde{Q}_{12}z_2) + (z_2, \tilde{Q}_{22}z_2)\right]dt \\ \tilde{Q}_{11}^{\alpha} &:= \left(\tilde{Q}_{11} + \tilde{Q}_{11}^{\mathsf{T}}\right)/2, \quad \tilde{Q}_{22} := \left(\tilde{Q}_{22} + \tilde{Q}_{22}^{\mathsf{T}}\right)/2 \\ \tilde{Q}_{12} &= \left(\tilde{Q}_{12} + \tilde{Q}_{12}^{\mathsf{T}} + \tilde{Q}_{21} + \tilde{Q}_{21}^{\mathsf{T}}\right)/2 \end{aligned} \tag{19.167}$$

Assumption (A1): we will look for the sliding function (19.166) in the form

$$\boxed{\tilde{\sigma}(z, t) := z_2 + \tilde{\sigma}_0(z_1, t)} \tag{19.168}$$

If the sliding mode exists for the system (19.163) in the sliding surface $\tilde{\sigma}(z, t) = 0$ under assumption (A1), then for all $t \geq t_s$ the corresponding sliding mode dynamics, driven by the unmatched disturbance $\tilde{\xi}_1(t)$, are given by

$$\begin{aligned} \dot{z}_1 &= \tilde{A}_{11}z_1 + \tilde{A}_{12}z_2 + \tilde{\xi}_1 \\ z_2 &= -\tilde{\sigma}_0(z_1, t) \end{aligned} \tag{19.169}$$

with the initial conditions $z_1(t_s) = (Tx(t_s))_1$. Defining z_2 as a virtual control, that is,

$$v := z_2 = -\tilde{\sigma}_0(z_1, t) \tag{19.170}$$

the system (19.169) may be rewritten as

$$\dot{z}_1 = \tilde{A}_{11}z_1 + \tilde{A}_{12}v + \tilde{\xi}_1 \tag{19.171}$$

and the performance index (19.165) becomes

$$J_{t_s,t_1} = \frac{1}{2}\int_{t_s}^{t_1}\left[\left(z_1, \tilde{Q}_{11}^{\alpha}z_1\right) + 2\left(z_1, \tilde{Q}_{12}^{\alpha}v\right) + \left(v, \tilde{Q}_{22}^{\alpha}v\right)\right]dt \tag{19.172}$$

In view of (19.171) and (19.172), the sliding surface design problem (19.159) is reduced to the following one:

$$\boxed{J_{t_s,t_1} \to \inf_{v \in R^r}} \tag{19.173}$$

But this is the standard LQ-optimal control problem. This means that the optimal control $v_t^* = v^*(z_{1,t}, t)$, optimizing the cost functional (19.172), defines the optimal sliding surface $\sigma^*(x,t)$ (see (19.171) and (19.168)) in the following manner:

$$v^*(z_{1,t}, t) = -\tilde{\sigma}_0(z_1, t)$$
$$\tilde{\sigma}(z,t) = z_2 - v^*(z_{1,t}, t) = 0$$

or, equivalently,

$$\boxed{\sigma^*(x,t) = (Tx)_2 - v^*((Tx)_1, t) = 0} \tag{19.174}$$

19.4.3.3 Equivalent control method

Equivalent control construction: here a formal procedure will be described to obtain sliding equations along the intersection of sets of discontinuity for a nonlinear system given by

$$\boxed{\begin{array}{l}\dot{x}(t) = f(t, x(t), u(t)) \\ x_0 \quad \text{is given} \\ x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^r\end{array}} \tag{19.175}$$

and the manifold $\mathcal{M}$ (19.138) defined as

$$\boxed{S(x) = (S_1(x), \ldots, S_m(x))^{\mathrm{T}} = 0} \tag{19.176}$$

representing an intersection of m submanifolds $S_i(x)$ $(i = 1, \ldots, m)$.

Definition 19.13. *Hereinafter the control $u(t)$ will be referred to (according to V. Utkin) as the **equivalent control** $u^{(eq)}(t)$ in the system (19.175) if it satisfies the equation*

$$\boxed{\begin{array}{l}\dot{S}(x(t)) = G(x(t))\,\dot{x}(t) = G(x(t))\,f(t, x(t), u(t)) = 0 \\ G(x(t)) \in \mathbb{R}^{m\times n}, \quad G(x(t)) = \dfrac{\partial}{\partial x} S(x(t))\end{array}} \tag{19.177}$$

It is quite obvious that, by virtue of the condition (19.177), a motion starting at $S(x(t_0)) = 0$ in time t_0 will proceed along the trajectories

$$\boxed{\dot{x}(t) = f\left(t, x(t), u^{(eq)}(t)\right)} \tag{19.178}$$

which lies on the manifold $S(x) = 0$.

Definition 19.14. *The above procedure is called the **equivalent control method** (Utkin 1992; Utkin* et al. *1999) and equation (19.178), obtained as a result of applying this method, will be regarded as the **sliding mode equation** describing the motion on the manifold $S(x) = 0$.*

From the geometric viewpoint, the equivalent control method implies a replacement of the undefined discontinued control on the discontinuity boundary with a continuous control which directs the velocity vector in the system state space along the discontinuity surface intersection. In other words, it realizes the velocity $f^0\left(t, x(t), u^{(eq)}(t)\right)$ (19.149) exactly corresponding to Filippov's definition of the differential inclusion in the point $x = x(t)$.

Consider now the equivalent control procedure for an important particular case of a nonlinear system which is affine on u, the right-hand side of whose differential equation is a linear function of the control, that is,

$$\boxed{\dot{x}(t) = f(t, x(t)) + B(t, x(t))\, u(t)} \tag{19.179}$$

where $f : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ and $B : \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^{n \times r}$ are all argument continuous vector and matrix, respectively, and $u(t) \in \mathbb{R}^r$ is a control action. The corresponding equivalent control should satisfy (19.177), namely,

$$\begin{aligned} \dot{S}(x(t)) &= G(x(t))\, \dot{x}(t) = G(x(t))\, f(t, x(t), u(t)) \\ &= G(x(t))\, f(t, x(t)) + G(x(t))\, B(t, x(t))\, u(t) = 0 \end{aligned} \tag{19.180}$$

Assuming that the matrix $G(x(t))\, B(t, x(t))$ is nonsingular for all $x(t)$ and t, one can find the equivalent control from (19.180) as

$$\boxed{u^{(eq)}(t) = -\left[G(x(t))\, B(t, x(t))\right]^{-1} G(x(t))\, f(t, x(t))} \tag{19.181}$$

Substitution of this control into (19.179) yields the following ODE:

$$\begin{aligned} \dot{x}(t) &= f(t, x(t)) \\ &\quad - B(t, x(t))\left[G(x(t))\, B(t, x(t))\right]^{-1} G(x(t))\, f(t, x(t)) \end{aligned} \tag{19.182}$$

which describes the sliding mode motion on the manifold $S(x) = 0$. Below the corresponding trajectories in (19.182) will be referred to as $x(t) = x^{(sl)}(t)$.

Remark 19.14. *If we deal with an uncertain dynamic model (19.175) or, particularly, with (19.179), then the equivalent control* $u^{(eq)}(t)$ ***is not physically realizable.***

Below we will show that $u^{(eq)}(t)$ may be successfully approximated (in some sense) by the output of the first-order low-pass filter with the input equal to the corresponding sliding mode control.

Sliding mode control design: let us try to stabilize the system (19.179) by applying the sliding mode approach. For the Lyapunov function $V(x) := \|S(x)\|^2/2$, considered on the trajectories of the controlled system (19.179), we have

$$\begin{aligned}\dot{V}(x(t)) &= \left(S(x(t)), \dot{S}(x(t))\right)\\ &= (S(x(t)), G(x(t))\, f(t, x(t)) + G(x(t))\, B(t, x(t))\, u(t))\\ &= (S(x(t)), G(x(t))\, f(t, x(t))) + (S(x(t)), G(x(t))\, B(t, x(t))\, u(t))\\ &\le \|S(x(t))\|\, \|G(x(t))\, f(t, x(t))\| + (S(x(t)), G(x(t))\, B(t, x(t))\, u(t))\end{aligned}$$

Taking $u(t)$ as a *sliding mode control*, i.e.,

$$\boxed{\begin{aligned}&u(t) = u^{(sl)}(t)\\ &u^{(sl)}(t) := -k_t \left[G(x(t))\, B(t, x(t))\right]^{-1} \operatorname{sign}(S(x(t)))\\ &k_t > 0, \quad \operatorname{sign}(S(x)) := (\operatorname{sign}(S_1(x)), \ldots, \operatorname{sign}(S_m(x)))^{\mathsf{T}}\end{aligned}} \tag{19.183}$$

we obtain

$$\dot{V}(x(t)) \le \|S(x)\|\, \|G(x(t))\, f(t, x(t))\| - k_t \sum_{i=1}^{m} |S_i(x(t))|$$

which, in view of the inequality, $\|S\| \ge \sum_{i=1}^{m} |S_i|$, implies

$$\dot{V}(x(t)) \le -\|S(x)\|\, (k_t - \|G(x(t))\, f(t, x(t))\|)$$

Selecting

$$\boxed{k_t = \|G(x(t))\, f(t, x(t))\| + \rho, \quad \rho > 0} \tag{19.184}$$

gives $\dot{V}(x(t)) \le -\rho\, \|S(x)\| = -\rho\sqrt{2V(x(t))}$ which provides the reaching phase in time

$$\boxed{t_f = \frac{\sqrt{2V(x_0)}}{\rho} = \frac{\|S(x_0)\|}{\rho}} \tag{19.185}$$

Remark 19.15. *If the sliding motion on the manifold $S(x) = 0$ is stable then there exists a constant $k^0 \in (0, \infty)$ such that*

$$\|G(x(t))\, f(t, x(t))\| \le k^0$$

and, hence, k_t (19.184) may be selected as a constant

$$\boxed{k_t := k = k^0 + \rho} \tag{19.186}$$

Low-pass filtering: to minimize the influence of the chattering effect arising after the reaching phase let us consider the property of the signal obtained as an output of a low-pass filter with the input equal to the sliding mode control, that is,

$$\boxed{\mu \dot{u}^{(av)}(t) + u^{(av)}(t) = u^{(sl)}(t), \quad u_0^{(av)} = 0, \quad \mu > 0} \tag{19.187}$$

where $u^{(sl)}(t)$ is given by (19.183). The next simple lemma states the relation between the, so-called, averaged control $u^{(av)}(t)$, which is the filtered output, and the input signal $u^{(sl)}(t)$.

Lemma 19.15. *If*

$$\begin{aligned} g^+ \geq \|GB(t, x(t))\| &:= \lambda_{\max}^{1/2}\left([B^\top(t, x(t))\, G^\top][GB(t, x(t))]\right) \\ &\geq \lambda_{\min}^{1/2}\left([B^\top(t, x(t))\, G^\top][GB(t, x(t))]\right) \geq \varkappa I_{r\times r}, \quad \varkappa > 0 \end{aligned} \tag{19.188}$$

then for the low-pass filter (19.187) the following properties hold:

1. The difference between the input and output signals are bounded, i.e.,

$$\boxed{\begin{gathered} u^{(av)}(t) = u^{(sl)}(t) + \zeta(t) \\ \|\zeta(t)\| \leq 2c, \quad c := \left(g^+ + \rho\right) m/\varkappa \\ \left\|\dot{u}^{(av)}(t)\right\| \leq 2c/\mu \end{gathered}} \tag{19.189}$$

2. The amplitude-frequency characteristic $A(\omega)$ of the filter is

$$\boxed{A(\omega) = \frac{1}{\sqrt{1 + (\mu\omega)^2}}, \quad \omega \in [0, \infty)} \tag{19.190}$$

whose plot is depicted at Fig. 19.11 for $\mu = 0.01$, where $y = A(\omega)$ and $x = \omega$.

Proof.

1. The solution of the ODE (19.183) and its derivative are as follows:

$$\left\|k_t\left[GB(t, x_t)\right]^{-1}\right\| \leq \left(\|Gf(t, x_t)\| + \rho\right)\left\|\left[GB(t, x_t)\right]^{-1}\right\|$$

$$= \frac{\left(\|Gf(t, x_t)\| + \rho\right)}{\lambda_{\min}^{1/2}\left([B^\top(t, x_t)\, G^\top][GB(t, x_t)]\right)} \leq \varkappa^{-1}\left(g^+ + \rho\right)$$

$$\left\|u_t^{(sl)}\right\| \leq \varkappa^{-1}\left(g^+ + \rho\right) m := c$$

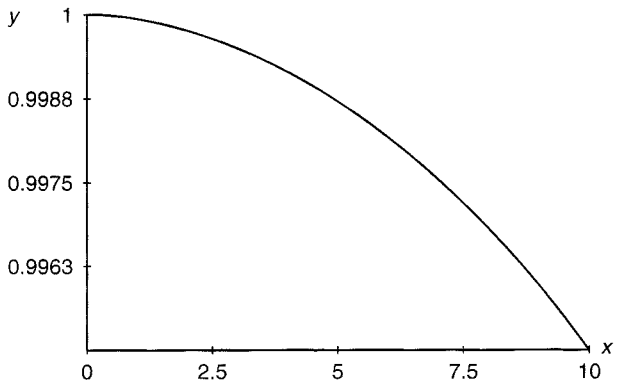

Fig. 19.11. The amplitude-phase characteristic of the low-pass filter.

and by (19.187)

$$u_t^{(av)} = \frac{1}{\mu}\int_{s=0}^{t} e^{-(t-s)/\mu} u_s^{(sl)} ds$$

$$\dot{u}_t^{(av)} = \frac{1}{\mu}\left[u_t^{(sl)} - \frac{1}{\mu}\int_{s=0}^{t} e^{-(t-s)/\mu} u_s^{(sl)} ds\right] \tag{19.191}$$

which implies

$$\mu\left\|\dot{u}_t^{(av)}\right\| \le \left\|u_t^{(sl)}\right\| + \frac{1}{\mu}\int_{s=0}^{t} e^{-(t-s)/\mu}\left\|u_s^{(sl)}\right\| ds \le c + \frac{c}{\mu}\int_{s=0}^{t} e^{-(t-s)/\mu} ds$$

$$= c + c\int_{s=0}^{t} e^{-(t-s)/\mu} d(s/\mu)$$

$$= c + c\int_{\tilde{s}=0}^{t/\mu} e^{-(t/\mu-\tilde{s})} d\tilde{s} = c + c\left(1 - e^{-t/\mu}\right) \le 2c$$

Hence, (19.189) holds.

2. Applying the Fourier transformation to (19.187) leads to the following identity:

$$\mu j\omega U^{(av)}(j\omega) + U^{(av)}(j\omega) = U^{(sl)}(j\omega)$$

or, equivalently,

$$U^{(av)}(j\omega) = \frac{1}{1+\mu j\omega} U^{(sl)}(j\omega) = \frac{1-\mu j\omega}{1+(\mu\omega)^2} U^{(sl)}(j\omega)$$

So, the amplitude-frequency characteristic

$$A(\omega) := \sqrt{[\mathrm{Re}\, U^{(av)}(j\omega)]^2 + [\mathrm{Im}\, U^{(av)}(j\omega)]^2}$$

of the filter (19.187) is as in (19.190). Lemma is proven. □

19.4.3.4 The realizable approximation of the equivalent control

By (19.191) $u_t^{(av)}$ may be represented as $u_t^{(av)} = \int_{s=0}^{t} u_s^{(sl)} d\left(e^{-(t-s)/\mu}\right)$. Consider the dynamics $x_t^{(av)}$ of the system (19.175) controlled by $u_t^{(av)}$ (19.191) at two time intervals: during the *reaching phase* and during the *sliding mode regime*.

1. **Reaching phase** ($t \in [0, t_f]$). Here the integration by part implies

$$u_t^{(av)} = \int_{s=0}^{t} u_s^{(sl)} d\left(e^{-(t-s)/\mu}\right) = u_t^{(sl)} - u_0^{(sl)} e^{-t/\mu} - \int_{s=0}^{t} \dot{u}_s^{(sl)} e^{-(t-s)/\mu} ds$$

Supposing that $u_t^{(sl)}$ (19.183) is bounded almost everywhere, i.e., $\left\|\dot{u}_t^{(sl)}\right\| \le d$. The above identity leads to the following estimation:

$$\left\|u_t^{(av)} - u_t^{(sl)}\right\| \le \left\|u_0^{(sl)}\right\| e^{-t/\mu} + d \int_{s=0}^{t} e^{-(t-s)/\mu} ds$$

$$= \left\|u_0^{(sl)}\right\| e^{-t/\mu} + \mu d \int_{s=0}^{t} e^{-(t-s)/\mu} d(s/\mu)$$

$$= \left\|u_0^{(sl)}\right\| e^{-t/\mu} + \mu d \int_{\tilde{s}=0}^{t/\mu} e^{-(t/\mu-\tilde{s})} d\tilde{s}$$

$$= \left\|u_0^{(sl)}\right\| e^{-t/\mu} + \mu d\left(1 - e^{-t/\mu}\right) = \mu d + O\left(e^{-t/\mu}\right)$$

So, $u_t^{(av)}$ may be represented as

$$\boxed{u_t^{(av)} = u_t^{(sl)} + \xi_t} \tag{19.192}$$

where ξ_t may be done as small as you wish taking μ tending to zero, since

$$\|\xi_t\| \le \mu d + O\left(e^{-t/\mu}\right)$$

As a result, the trajectories $x_t^{(sl)}$ and $x_t^{(av)}$ will differ slightly. Indeed,

$$\dot{x}_t^{(sl)} = f\left(t, x_t^{(sl)}\right) - B\left(t, x_t^{(sl)}\right) u_t^{(sl)}$$

$$\dot{x}_t^{(av)} = f\left(t, x_t^{(av)}\right) - B\left(t, x_t^{(av)}\right) u_t^{(av)}$$

Defining

$$\tilde{B} = B(t, x_t^{(av)}), \quad \tilde{G} = G(x_t^{(av)}), \quad \tilde{f} = f\left(t, x_t^{(av)}\right)$$

and omitting the arguments for simplicity, the last equation may be represented as

$$\dot{x}_t^{(sl)} = f - Bu_t^{(sl)}, \quad \dot{x}_t^{(av)} = \tilde{f} - \tilde{B}u_t^{(av)}$$

Hence by (19.192), the difference $\Delta_t := x_t^{(sl)} - x_t^{(av)}$ satisfies

$$\Delta_t = \Delta_0 - \int_{s=0}^{t} \left[\left(f - \tilde{f}\right) - Bu_s^{(sl)} + \tilde{B}u_s^{(av)}\right] ds$$

$$= \Delta_0 - \int_{s=0}^{t} \left[\left(f - \tilde{f}\right) - Bu_s^{(sl)} + \tilde{B}\left(u_s^{(sl)} + \xi_s\right)\right] ds$$

Taking into account that $\Delta_0 = 0$ (the system starts with the same initial conditions independently on an applied control) and that $f(t, x)$ and $B(x)$ are Lipschitz (with the constant L_f and L_B) on x it follows that

$$\|\Delta_t\| \le \int_{s=0}^{t} \left[\left\|f - \tilde{f}\right\| + \left\|\left(\tilde{B} - B\right)u_s^{(sl)} + \tilde{B}\xi_s\right\|\right] ds$$

$$\le \int_{s=0}^{t} \left[L_f \|\Delta_s\| + L_B \|\Delta_s\| \left\|u_s^{(sl)}\right\| + \left\|\tilde{B}\right\| \|\xi_s\|\right] ds$$

$$\le \int_{s=0}^{t} \left[\left(L_f + L_B \left\|u_s^{(sl)}\right\|\right) \|\Delta_s\| + \left\|\tilde{B}\right\| \left(\mu d + O\left(e^{-s/\mu}\right)\right)\right] ds$$

Since $O\left(e^{-t/\mu}\right) = \mu O\left(\frac{1}{\mu} e^{-t/\mu}\right) = \mu o(1) \le \mu\varepsilon$ and

$$\left\|u_s^{(sl)}\right\| \le u_+^{(sl)} < \infty, \quad \left\|\tilde{B}\right\| \le B^+ < \infty$$

we finally have

$$\|\Delta_t\| \le \int_{s=0}^{t} \left[\left(L_f + L_B u_+^{(sl)}\right) \|\Delta_s\| + B^+ \mu (d + \varepsilon)\right] ds$$

$$\le B^+ \mu (d + \varepsilon) t_f + \int_{s=0}^{t} \left(L_f + L_B u_+^{(sl)}\right) \|\Delta_s\| ds$$

Now let us apply the Gronwall lemma which says that if $v(t)$ and $\xi(t)$ are nonnegative continuous functions on $[t_0, \infty)$ verifying

$$v(t) \leq c + \int_{s=t_0}^{t} \xi(s)\, v(s)\, ds \tag{19.193}$$

then for any $t \in [t_0, \infty)$ the following inequality holds:

$$v(t) \leq c \exp\left(\int_{s=t_0}^{t} \xi(s)\, ds \right) \tag{19.194}$$

This result remains true if $c = 0$. In our case

$$v(t) = \|\Delta_t\|, \quad c = B^{+}\mu(d+\varepsilon)t_f, \quad \xi(s) = L_f + L_B u_{+}^{(sl)}$$

for any $s \in [0, t_f)$. So,

$$\boxed{\|\Delta_t\| \leq \delta := B^{+}\mu(d+\varepsilon)t_f \exp\left(\left(L_f + L_B u_{+}^{(sl)} \right) t_f \right)} \tag{19.195}$$

Claim 19.2. *For any finite reaching time t_f and any small value $\delta > 0$ there exists a small enough μ such that $\|\Delta_t\|$ is less than δ.*

2. **Sliding mode phase** ($t > t_f$). During the sliding mode phase we have

$$S\left(x_t^{(sl)}\right) = \dot{S}\left(x_t^{(sl)}\right) = G\left(f - Bu_t^{(eq)}\right) = 0 \tag{19.196}$$

if $u_t = u_t^{(eq)}$ for all $t > t_f$. Applying $u_t = u_t^{(av)}$ we cannot guarantee (19.196) already. Indeed,

$$S\left(x_t^{(av)}\right) = S\left(x_{t_f}^{(av)}\right) + \int_{s=t_f}^{t} \dot{S}\left(x_s^{(av)}\right) ds$$

and, by (19.195),

$$\left\| S\left(x_{t_f}^{(av)}\right) \right\| = \left\| S\left(x_{t_f}^{(av)}\right) - S\left(x_{t_f}^{(sl)}\right) \right\| \leq \left\| G\left(x_{t_f}^{(sl)}\right) \Delta_{t_f} \right\| \leq O(\mu)$$

Hence, in view of (19.196), $\left\| S\left(x_t^{(av)}\right) \right\| = O(\mu)$.

Claim 19.3. *During the sliding-mode phase*

$$\boxed{\left\| S\left(x_t^{(av)}\right) \right\| = O(\mu)} \tag{19.197}$$

20 Elements of Stability Theory

Contents

This chapter deals with the basic notions concerning the stability property of certain solutions or sets of solutions of the different classes of ordinary differential equations (ODE).

In the famous work of A.M. Lyapunov (Lyapunov 1892) there is given some very simple (but philosophically very profound) theorems (hereafter referred to as the *direct Lyapunov's method*) for deciding the stability or instability of an equilibrium point of an ODE. The idea of this approach consists of the generalization of the concept of "energy" and its "power" the usefulness of which lies in the fact that the decision on stability can be made by investigating the differential equation itself (in fact, its right-hand side only) but not by finding its exact solution.

The purpose of this chapter is to give an introduction to some of the fundamental ideas and problems in the field which can be successfully applied to some problems arising in *automatic control theory*.

20.1 Basic definitions

20.1.1 Origin as an equilibrium

Definition 20.1. *The vector-valued function* $y(t, y_0, t_0) \in \mathbb{R}^n$ *is said to be the* ***dynamic motion*** *satisfying*

$$\dot{y}(t) = g(t, y(t)), \quad y(t_0) = y_0$$

if

- *there exists no other function satisfying this ODE with the same initial conditions;*
- *it is differentiable in* t *for all* $t \geq t_0$.

Remark 20.1. *Evidently, by this definition,* $y(t_0, y_0, t_0) = y_0$.

Consider the, so-called, *nominal dynamic motion* $y^*(t, y_0, t_0)$ satisfying

$$\dot{y}^*(t) = g^*(t, y^*(t)), \quad y^*(t_0) = y_0^* \in \mathbb{R}^n$$

and the *dynamic motion in deviations* $x(t, x_0, t_0) \in \mathbb{R}^n$, defined by

$$\boxed{\begin{aligned} x(t, x_0, t_0) &:= y(t, y_0, t_0) - y^*(t, y_0, t_0) \\ \dot{x}(t) &= f(t, x(t)) \\ x(t_0) &= x_0 := y_0 - y_0^* \\ f(t, x) &:= g(t, x - y^*) - g^*(t, y^*) \end{aligned}} \tag{20.1}$$

Definition 20.2. *Supposing that equation (20.1) admits the dynamic motion* $x(t, 0, t_0) \equiv 0$. *We will also call it the trivial solution or* ***the equilibrium*** *which can be expressed by*

$$\boxed{f(t, 0) = 0,\ t \geq t_0} \tag{20.2}$$

Further we will assume that the solution, belonging to the initial point x_0 *in a certain neighborhood* $\|x_0\| < \delta$ *of the origin, exists for all* $t \geq t_0$ *and is uniquely determined by the initial values* x_0, t_0.[1]

In this chapter we will study different aspects of stability of the equilibrium point $x = 0$.

20.1.2 Positive definite functions

First, let us introduce the following definitions which will be intensively used hereafter.

Definition 20.3. *A real function* $V = V(t, x)$, *specified in the domain* $\|x\| \leq h$ $(x \in R^n,\ h > 0)$ *for all* $t \geq t_0$, *is called* ***positive-definite*** *if there exists a real continuous function* $W(x)$ *defined for* $\|x\| \leq h$ *such that*

1.

$$\boxed{W(0) = 0} \tag{20.3}$$

2. *for* $\|x\| > 0$

$$\boxed{W(x) > 0} \tag{20.4}$$

3. *for all* $t \geq t_0$

$$\boxed{V(t, x) \geq W(x)} \tag{20.5}$$

[1] We may assume that $f(t, x)$ is an n-dimensional vector function which is locally (uniformly on t) Lipschitz on x in a neighborhood of the point $x = 0$.

If the properties (2)–(3) are replaced by $W(x) < 0$ *and* $V(t, x) \leq W(x)$, *then the function* $V(t, x)$ *will be* ***negative-definite***.

Example 20.1.

$$V(t, x) = x_1^2 + x_2^2 + x_1 x_2 \sin t$$

is precisely such a function. Indeed,

$$\begin{aligned} V(t, x) &= x_1^2 + x_2^2 + x_1 x_2 \sin t \geq x_1^2 + x_2^2 - |x_1|\,|x_2| \\ &= (|x_1| - |x_2|)^2 + |x_1|\,|x_2| \geq |x_1|\,|x_2| := W(x) \end{aligned}$$

So, all conditions of Definition 20.3 are fulfilled for the function $W(x)$.

Definition 20.4. *Denote by* $\bar{x}(t, x_0, t_0)$ *the dynamic motion (trajectory) which satisfies (20.1) when* $x(t_0) = x_0$. *Then, if there exists the time derivative of the function* $\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0))$, *then the function* $V(t, x)$ *is said to be* ***differentiable along the integral curves (or the path)*** $\bar{x}(t, x_0, t_0)$ *of the system (20.1).*

Claim 20.1. *The* ***full time derivative*** *of the differentiable function* $\bar{V}(t)$ *is calculated as follows*

$$\boxed{\begin{aligned} \frac{d}{dt}\bar{V}(t) = \frac{d}{dt}V(t, \bar{x}(t, x_0, t_0)) &= \frac{\partial}{\partial t}V(t, \bar{x}(t, x_0, t_0)) \\ &+ \sum_{i=1}^{n} \frac{\partial V}{\partial x_i}(t, \bar{x}(t, x_0, t_0))\, f_i(t, \bar{x}(t, x_0, t_0)) \end{aligned}} \tag{20.6}$$

In short (20.6) is written as

$$\boxed{\frac{d}{dt}V(t, \bar{x}) = \frac{\partial}{\partial t}V(t, \bar{x}) + \sum_{i=1}^{n} \frac{\partial V}{\partial x_i}(t, \bar{x})\, f_i(t, \bar{x})} \tag{20.7}$$

20.2 Lyapunov stability

20.2.1 Main definitions and examples

Definition 20.5. *The equilibrium zero-point (or zero-state)* $x = 0$ *of the system given by ODE (20.1) is said to be*

1. ***Lyapunov stable***, *or* ***locally stable***, *if for any* $\varepsilon > 0$ *there exist* $t_0' \geq t_0 \geq 0$ *and* $\delta = \delta(t_0', \varepsilon) > 0$ *such that for all* $t \geq t_0$ *we have* $\|\bar{x}(t, x_0, t_0')\| < \varepsilon$ *whenever* $x(t_0') = x_0$ *and* $\|x_0\| < \delta$;
2. ***uniformly Lyapunov stable***, *or* ***uniformly locally stable,*** *if it is Lyapunov stable for any* $t_0' \geq t_0$, *that is,* δ *is independent on* t_0';

3. ***asymptotically locally stable*** *if it is locally stable and, additionally,* $\bar{x}\left(t, x_0, t_0'\right) \to 0$ *as* $t \to \infty$*;*
4. ***asymptotically uniformly locally stable*** *if it is uniformly locally stable and, additionally,* $\bar{x}\left(t, x_0, t_0'\right) \to 0$ *as* $t \to \infty$.
5. ***exponentially locally stable*** *if*

- *it is asymptotically uniformly locally stable, and,*
- *additionally, there exists two positive constants* α *and* β *such that*

$$\boxed{\|x(t_0)\| \leq \alpha \|x(t_0)\| e^{-\beta(t-t_0)}} \tag{20.8}$$

Definition 20.6. *The equilibrium zero-point (or zero-state)* $x = 0$ *of the system given by ODE (20.1) is said to be* ***unstable*** *if at least one of two requirements holds:*

- *either the solution* $x(t)$ *of (20.1) is* ***noncontinuable*** *in* t *from* $t = t_0$ *up to* ∞ *in any neighborhood of the zero-state* $x = 0$*; or*
- *when for any* $\delta > 0$ *and any* $t' \geq t_0$ *there exists* $\varepsilon = \varepsilon(\delta, t')$ *and* $t'' \geq t'$ *such that* $\|x(t'')\| > \varepsilon$ *in spite of the fact that* $\|x(t')\| < \delta$.

The illustrations of Lyapunov and asymptotic types of stability are given by Figs. 20.1 and 20.2.

Example 20.2. (The linear oscillator) *Consider the model of the linear oscillator given by*

$$\boxed{\begin{array}{c} \ddot{x}(t) + d\dot{x}(t) + \omega^2 x(t) = 0 \\ t \geq t_0 := 0,\ \omega > 0,\ x(0) = x_0,\ \dot{x}(0) = \dot{x}_0 \quad \textit{are given} \end{array}} \tag{20.9}$$

(a) ***The no-friction case*** $d = 0$: $x_1(t) := x(t)$, $x_2(t) := \dot{x}(t)$

$$\dot{x}_1(t) = x_2(t),\ \dot{x}_2(t) = -\omega^2 x_1(t)$$

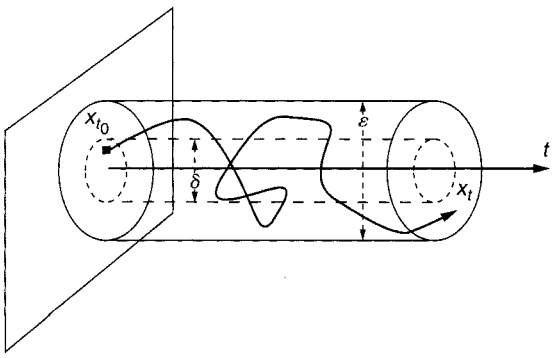

Fig. 20.1. Lyapunov's stability illustration.

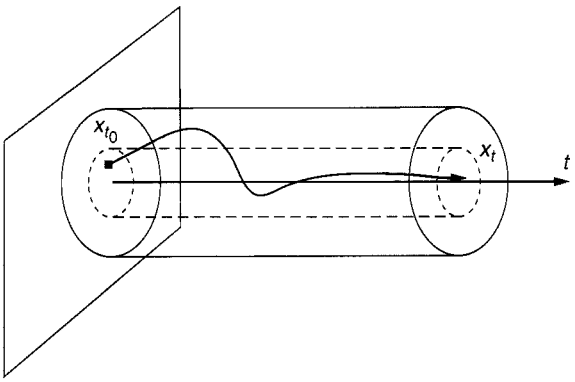

Fig. 20.2. Asymptotic local stability illustration.

and

$$x_1(t) = x(t) = x_0 \cos \omega t + \frac{\dot{x}_0}{\omega} \sin \omega t$$
$$x_2(t) = \dot{x}(t) = -x_0 \omega \sin \omega t + \frac{\dot{x}_0}{\omega} \cos \omega t$$

So, for $|x_0| \leq \delta$ *and* $|\dot{x}_0| \leq \delta$ *we have*

$$|x(t)| \leq |x_0| + \left|\frac{\dot{x}_0}{\omega}\right| \leq \delta\left(1 + \omega^{-1}\right) \leq \varepsilon$$

$$|\dot{x}(t)| = |x_0 \omega| + |\dot{x}_0| \leq \delta(\omega + 1) \leq \varepsilon$$

$$\textit{if} \quad \delta := \varepsilon / \max\left\{1 + \omega^{-1}; 1 + \omega\right\}$$

This means that the state $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ *is* ***uniformly locally stable.***

(*b*) ***The friction case*** $d > 0$: $x_1(t) := x(t)$, $x_2(t) := \dot{x}(t)$

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = -\omega^2 x_1(t) - d x_2(t)$$

and

$$x_1(t) = x(t) = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$
$$x_2(t) = \dot{x}(t) = c_1 \lambda_1 e^{\lambda_1 t} + c_2 \lambda_2 e^{\lambda_2 t}$$
$$\lambda_{1,2} = -\frac{d}{2} \pm \sqrt{\left(\frac{d}{2}\right)^2 - \omega^2}$$

In any case, Re $\lambda_i < 0$ $(i = 1, 2)$ *which implies*

$$|x_1(t)| = |x(t)| \le |c_1| e^{Re\,\lambda_1 t} + |c_2| e^{Re\,\lambda_2 t} \to 0 \text{ as } t \to \infty$$
$$|x_2(t)| = |\dot{x}(t)| \le |\lambda_1| |c_1| e^{Re\,\lambda_1 t} + |\lambda_2| |c_2| e^{Re\,\lambda_2 t}$$
$$\le \max_{i=1,2} (|\lambda_i| |c_i|) e^{\max\limits_{i=1,2}(Re\,\lambda_i)t} \to 0 \text{ as } t \to \infty$$

Therefore the stationary state $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$ *is* ***exponentially locally stable*** *with* $\alpha = \max\limits_{i=1,2} (|\lambda_i| |c_i|)$ *and* $\beta = -\max\limits_{i=1,2} (Re\,\lambda_i)$.[2]

20.2.2 *Criteria of stability: nonconstructive theory*

The theory, designed by A.M. Lyapunov (Doctor thesis 1892, the first translations from Russian are in Lyapunov 1907), says that **if for a system (20.1) there exists a Lyapunov (positive-definite "energetic") function, then the zero-state is Lyapunov stable**. So, this theory deals with the, so-called, sufficient conditions of stability. But there exists another concept (see Zubov 1962, 1964) which states that **if the zero-state of (20.1) is locally stable, then obligatory there exists a corresponding Lyapunov function**. This means exactly that the existence of a Laypunov function is also a necessary condition of stability.

In this subsection we will present the "joint result" (due to Zubov (1962)) on the necessary and sufficient conditions of local stability or, in other words, the *criterion of local, asymptotic and exponential stability* for nonlinear systems governed by (20.1).

20.2.2.1 *Criterion of Lyapunov (local) stability*

Theorem 20.1. (The criterion of stability (Zubov 1964)) *The zero-state of the system (20.1) is* ***Lyapunov (or, locally) stable*** *if and only if there exists a function* $V(t, x)$*, called the* ***Lyapunov function****, satisfying the following conditions:*

1. $V(t, x)$ *is* ***defined*** *for* $\|x\| \le h$ *and* $t \ge t_0$*;*
2. $V(t, 0) = 0$ *for all* $t \ge t_0$ *and is* ***continuous*** *in* x *for all* $t \ge t_0$ *in the point* $x = 0$*;*
3. $V(t, x)$ *is* ***positive-definite****, that is, there exists a function* $W(x)$ *such that*

$$V(t, x) \ge W(x) \quad \text{for all} \quad t \ge t_0$$
$$W(0) = 0, \; W(x) > 0 \quad \text{for} \quad \|x\| > 0$$

4. *the function* $\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0))$ ***does not increase***[3] *in* $t \ge t_0$ *for all* x_0 *satisfying* $\|x_0\| \le h$.

---

[2] In fact, here it is proven more accurately: this state is globally exponentially stable (the exact definition in subsection 20.2.3), since the property $\|\bar{x}(t, x_0, t_0')\| \le \alpha \exp(-\beta t) \underset{t \to \infty}{\to} 0$ is true for any x_0 and any $t_0' \ge 0$.

[3] Notice that here it is not required for $\bar{V}_t$ to be t-differentiable.

Proof.

(a) *Sufficiency.* Suppose that there exists a function $V(t, x)$ satisfying all conditions (1)–(4) of Theorem 20.1. Take $\varepsilon < h$ and consider the sphere $\|x\| = \varepsilon$. By condition (3)

$$\inf_{x:\|x\|=\varepsilon} W(x) := \lambda > 0$$

By the continuity of $V(t, x)$ (the properties (1)–(2)) it follows that there is a number $\delta = \delta(t_0, \varepsilon) > 0$ such that $V(t_0, x) < \lambda$ as $\|x\| < \delta$. Take any point x_0 satisfying $\|x_0\| < \delta$. Then $V(t_0, x_0) < \lambda$ and by property (4) the function $V(t, \bar{x}(t, x_0, t_0))$ is not increasing for $t \geq t_0$ which implies

$$V(t, \bar{x}(t, x_0, t_0)) \leq V(t_0, x_0) < \lambda \tag{20.10}$$

Hence, $\|x(t, x_0, t_0)\| < \varepsilon$ for all $t \geq t_0$, otherwise there exists an instant time $\tilde{t} > t_0$ such that $\left\|x\left(\tilde{t}, x_0, t_0\right)\right\| = \varepsilon$ and, therefore,

$$V\left(\tilde{t}, \bar{x}\left(\tilde{t}, x_0, t_0\right)\right) \geq W\left(\bar{x}\left(\tilde{t}, x_0, t_0\right)\right) \geq \lambda$$

which contradicts (20.10). The sufficiency is proven.

(b) *Necessity.* Let the stationary zero-point be Lyapunov stable. Consider the solution $\bar{x}(t, x_0, t_0)$ of (20.1) for $\|x_0\| \leq h$. Define the function $V(t, x)$ as follows

$$V(t, x) := \sup_{s \geq t} \|\bar{x}(s, x, t)\| \tag{20.11}$$

where $\bar{x}(s, x, t)$ is the solution (20.1) started at the point x at time t. So, the condition (1) of Theorem 20.1 holds. Since $x = 0$ is a stationary point (an equilibrium), then $V(t, 0) = 0$ which follows from (20.11). Additionally, this function is continuous at the point $x = 0$ for any $t \geq t_0$. Indeed, for $\epsilon > 0$, by the stability property, there exists $\delta = \delta(t_0, \epsilon)$ such that $\|\bar{x}(t, x_0, t_0)\| < \epsilon$ whenever $\|x_0\| < \delta$. By (20.11), $V(t_0, x_0) < \epsilon$ which proves the fulfillment of condition (2) of Theorem 20.1. One can see also that when $\|x_0\| > 0$

$$V(t_0, x_0) \geq \|\bar{x}(t_0, x_0, t_0)\| = \|x_0\| := W(x_0) > 0$$

So, $V(t, x)$ is positive-definite and condition (3) is also fulfilled. To demonstrate the validity of condition (4) it is sufficient to establish that

$$V\left(t', \bar{x}\left(t', x_0, t_0\right)\right) \leq V\left(t'', \bar{x}\left(t'', x_0, t_0\right)\right) \quad \text{if} \quad t'' \geq t' \geq t_0$$

To do this, it is sufficient to notice that by formula (20.11)

$$\begin{aligned} V(t', \bar{x}(t', x_0, t_0)) &= \sup_{s \geq t'} \|\bar{x}(s, \bar{x}(t', x_0, t_0), t')\| \\ &\geq \sup_{s \geq t''} \|\bar{x}(s, \bar{x}(t', x_0, t_0), t')\| = \sup_{s \geq t''} \|\bar{x}(s, \bar{x}(t'', x_0, t_0), t'')\| \\ &= V(t'', \bar{x}(t'', x_0, t_0)) \end{aligned}$$

which means that the function does not increase along the solutions $x(t, x_0, t_0)$. This completes the proof of the necessity of the conditions of Theorem 20.1. □

Remark 20.2. *Condition (4) of Theorem 20.1 seems to be restrictive since the integral curves* $\bar{x}(t, x_0, t_0)$ *are not given analytically and, therefore, are unknown if we do not know the exact analytical solution of (20.1). However, this problem can be slightly simplified if we remember the following fact:*

"By one of the Lebesgue theorems (see Corollary 15.5), the derivative of a monotone function exists almost everywhere."

Therefore, by the condition (4) the function $\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0))$ *is monotone on any integral curve* $\bar{x}(t, x_0, t_0)$ *and, hence, there exists the derivative* $\frac{d}{dt}\bar{V}(t)$. *Admitting also the existence of the partial derivatives* $\frac{\partial V(t,x)}{\partial t}$ *and* $\frac{\partial V(t,x)}{\partial x_i}$ *for all* $t \geq t_0$ *and all* x *in a neighborhood of the origin, condition (4) can be verified by checking the inequality*

$$\boxed{\frac{d}{dt}\bar{V}(t) = \frac{\partial V(t,x)}{\partial t} + \sum_{i=1}^{n} \frac{\partial V(t,x)}{\partial x_i} f_i(t,x) \leq 0} \tag{20.12}$$

Sure, the derivatives $\frac{\partial V(t,x)}{\partial t}$ *and* $\frac{\partial V(t,x)}{\partial x_i}$ *of the function* $V(t,x)$, *as it is defined in (20.11), cannot be calculated analytically. So, this means that Theorem 20.1 makes only a "philosophical sense", but not a practical one: it says that* ***any system with the stable zero-state has a Lyapunov function****.*

Corollary 20.1. (Lyapunov 1892)[4] *If the function* $V(t,x)$ *is positive-definite and continuous in* x *at* $x = 0$ *uniformly in* t *for all* $t \geq t_0$ *and*

$$\boxed{\frac{d}{dt}V(t, \bar{x}(t, x_0, t_0)) \leq 0} \tag{20.13}$$

then the stationary point $x = 0$ *of the system (20.1) is* ***uniformly local stable****.*

Proof. In the proof of Theorem 20.1 a number $\delta = \delta(t_0, \epsilon)$ is selected from the condition (20.10) such that $V(t_0, x_0) < \lambda$ for $\|x_0\| < \delta$. Since $V(t,x)$ is continuous in x at the point $x = 0$ uniformly in t for all $t \geq t_0$, then there exists a number $\delta = \delta(\varepsilon)$ such that $V(t_0, x_0) < \lambda$ for all $\|x_0\| < \delta(\varepsilon)$ at all t_0 which proves the corollary. □

20.2.2.2 Criterion of asymptotic stability

Theorem 20.2. (The criterion of AS (Zubov 1964)) *The state* $x = 0$ *of the system (20.1) is* ***asymptotically stable*** *if and only if all assumptions of Theorem 20.1 hold and in the condition (4) the function* $\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0))$ *decreases monotonically up to zero, that is,*

$$\boxed{\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0)) \downarrow 0 \text{ as } t \to \infty} \tag{20.14}$$

[4] This result is referred to as *the second Lyapunov's theorem (method).*

Proof.

(a) *Necessity.* If $x = 0$ is asymptotically stable, then $\|\bar{x}(s, x, t)\| \to 0$ as $t \to \infty$, and, therefore, by the construction (20.11), it follows that

$$\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0)) = \sup_{s \geq t} \|\bar{x}(s, x, t)\| \underset{t \to \infty}{\to} 0$$

Monotonicity results from the inequality

$$\bar{V}(t) := V(t, \bar{x}(t, x_0, t_0)) = \sup_{s \geq t} \|\bar{x}(s, x, t)\|$$
$$\geq \sup_{s \geq t' > t} \|\bar{x}(s, x, t)\| = V(t', \bar{x}(t', x_0, t_0)) = \bar{V}(t')$$

(b) *Sufficiency.* If $\sup_{s \geq t} \|\bar{x}(s, x, t)\| \underset{t \to \infty}{\to} 0$, then it follows that

$$\bar{x}(t, x_0, t_0) \underset{t \to \infty}{\to} 0$$

Theorem is proven. □

20.2.2.3 Criterion of exponential stability

Theorem 20.3. (on exponential stability) *For any solution $\bar{x}(t, x_0, t_0)$ of (20.1) to be* ***exponentially stable*** *(see Definition 20.5) it is necessary and sufficient that there exist two positive-definite functions $V(t, x)$ and $W(t, x)$ such that*

1. for any x and any $t \geq t_0$ there exists $\bar{\beta} > 0$ for which

$$\boxed{W(t, x) \geq \bar{\beta} V(t, x)} \tag{20.15}$$

2. the functions $V(t, x)$ and $W(t, x)$ are related by

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = -W(t, \bar{x}(t, x_0, t_0))} \tag{20.16}$$

Proof.

(a) *Necessity.* Let the solution $\bar{x}(t, x_0, t_0)$ of (20.1) be exponentially stable with some $\alpha > 0$ and $\beta > 0$. Define $W(t, x)$ and $V(t_0, x_0)$

$$W(t, x) := \|x\|^2, \quad V(t, x) := \int_{s=t}^{\infty} W(s, \bar{x}(s, x, t))\, ds \tag{20.17}$$

The relation (20.16) is evident. Show that $V(t, x)$, as it is defined above, satisfies the condition (20.15). By (20.8) we have

$$V(t, x) = \int_{s=t}^{\infty} \|\bar{x}(s, x, t)\|^2\, ds$$
$$\leq \int_{s=t}^{\infty} \alpha^2 \|x\|^2 \exp(-2\beta(s - t))\, ds \leq \frac{\alpha^2}{2\beta} \|x\|^2 = \frac{\alpha^2}{2\beta} W(t, x)$$

This means that $V(t, x)$ satisfies (20.15) with $\bar{\beta} := \dfrac{2\beta}{\alpha^2}$.

(b) *Sufficiency*. Show that (20.16) and (20.15) imply (20.8). Considering $V(t, \bar{x}(t, x_0, t_0)) \neq 0$ (if not, we already have the stability since $\bar{x}(s, x_0, t_0) = 0$ for all $s \geq t$) and integrating (20.16) lead to

$$\int_{s=t_0}^{t} \frac{dV}{V} = -\int_{s=t_0}^{t} \frac{W(s, \bar{x}(s, x_0, t_0))}{V(s, \bar{x}(s, x_0, t_0))} ds$$

and, hence, by (20.15)

$$V(t, \bar{x}(t, x_0, t_0)) = V(t_0, x_0) \exp\left(-\int_{s=t_0}^{t} \frac{W(s, \bar{x}(s, x_0, t_0))}{V(s, \bar{x}(s, x_0, t_0))} ds\right)$$
$$V(t_0, x_0) \exp\left(-\int_{s=t_0}^{t} \beta \, ds\right) = V(t_0, x_0) \exp\left(-\bar{\beta}[t - t_0]\right) \tag{20.18}$$

which corresponds to the exponential stability (20.8) with $\alpha = V(t_0, x_0)$ and $\beta = \bar{\beta}$. Theorem is proven. □

20.2.2.4 Criterion of instability

Theorem 20.4. (Criterion of instability (Zubov 1964)) *For the state $x = 0$ of the system (20.1) to be unstable, it is* ***necessary and sufficient*** *that there exist two scalar continuous functions $V(t, x)$ and $W(t, x) \geq 0$, defined in (t, x)-domain Ω (which includes the point $x = 0$), such that*

1. *$V(t, x)$ is bounded in Ω;*
2. *for any $t \geq t_0$ and $\delta > 0$ there exists $x(t') : \|x(t')\| < \delta$, $t' \geq t$ such that in this point the inequality $V(t', x(t')) > 0$ holds;*
3. *there exists the time derivative*

$$\boxed{\begin{array}{l} \dfrac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) \\ = \lambda V(t, \bar{x}(t, x_0, t_0)) + W(t, \bar{x}(t, x_0, t_0)), \ \lambda > 0 \end{array}} \tag{20.19}$$

Proof.

(a) *Necessity*. Suppose there is instability. This means exactly that there exists $\epsilon > 0$ such that for any $t \geq t_0$ and any $\delta > 0$ it is possible to find $x(t') : \|x(t')\| < \delta$ such that the inequality $\left\|\bar{x}\left(\tilde{t}, x(t'), t'\right)\right\| < \epsilon$ fails to hold for all $\tilde{t} \geq t' \geq t$. Take any point $(x(t'), t')$ satisfying the inequalities $\|x(t')\| < \delta$, $t' \geq t_0$ and $0 < \delta < \epsilon$. Then two cases may occur: either (1) $\left\|\bar{x}\left(\tilde{t}, x(t'), t'\right)\right\| \leq \epsilon$ for all $\tilde{t} \geq t'$, or (2) there exists an instant $\bar{t} = \bar{t}(x(t'), t')$ when $\left\|\bar{x}\left(\bar{t}, x(t'), t'\right)\right\| = \epsilon$ and $\left\|\bar{x}\left(\tilde{t}, x(t'), t'\right)\right\| < \epsilon$ for all $t' \leq \tilde{t} < \bar{t}$. Let $V\left(\tilde{t}, x(t')\right) = 0$ in the first case and $V\left(\tilde{t}, x(t')\right) = e^{-(\tilde{t}-\bar{t})} = e^{\bar{t}-\tilde{t}}$ in the second one. Thus the function $V(t, x)$ is defined at any point of the set

$\{(t, x) \mid \|x(t')\| < \epsilon,\ t \geq t_0\}$ and is bounded therein since $V(t, x) < 1$. So, condition (1) holds. The second type points exist in a neighborhood of the point $x = 0$ and, hence, by the construction, condition (2) holds too. Show that condition (3) is also satisfied. Indeed, in the first case, when $V(t', x(t')) = 0$ for all $\tilde{t} \geq t'$, we have $\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = V(t, \bar{x}(t, x_0, t_0))$ along any such motion. In the second case, when $V(\tilde{t}, x(t')) = e^{\tilde{t}-\tilde{t}}$, we also have $\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = V(t, \bar{x}(t, x_0, t_0))$ and, hence, condition (3) holds with $\lambda = 1$ and $W = 0$.

(b) *Sufficiency.* Suppose that all the conditions of the theorem hold. Show that $x = 0$ is unstable. Assume conversely that $x = 0$ is stable and, hence, $V(t, x)$ is bounded (uniformly on t_0) in Ω. By property (2) there exists a point $(t', x(t')) \in \Omega$ such that $V(t', x(t')) > 0$. But, by property (3), integrating ODE (20.19) with the initial condition $V(t', x(t'))$ implies

$$V\left(t, \bar{x}\left(t, x\left(t'\right), t'\right)\right) \geq V\left(t', x\left(t'\right)\right) e^{\lambda(t-t')} \quad \text{for} \quad t \geq t'$$

which contradicts the boundedness of $V(t, x)$ on Ω. So, the point $x = 0$ is unstable. Theorem is proven. □

Example 20.3. (The "mathematical point" in a potential field) *Consider a mathematical point with mass m which can move in the (x, y)-plane over the potential convex curve $\Pi = \Pi(x)$ which corresponds to its vertical position, i.e., $y = \Pi(x)$. Then its velocity $v(t)$, the kinetic T and the potential V energies are as follows*

$$\begin{aligned} v^2(t) &:= \dot{x}^2(t) + \dot{y}^2(t) \\ T &= \frac{m}{2} v^2(t) = \frac{m}{2} \dot{x}^2(t) \left(1 + [\Pi'(x(t))]^2\right) \\ V &= mgy(t) = mg\Pi(x(t)) \end{aligned}$$

So, the Lagrange dynamic equation (see, for example, Gantmacher (1990))

$$\frac{d}{dt} \frac{\partial}{\partial \dot{x}} L - \frac{\partial}{\partial x} L = 0, \quad L := T - V$$

for this case becomes

$$\ddot{x}(t) \left(1 + [\Pi'(x(t))]^2\right) + \Pi'(x(t)) \left[\dot{x}^2(t) \Pi''(x(t)) + g\right] = 0$$

and, hence, for $x_1(t) := x(t)$, $x_2(t) := \dot{x}(t)$ we have

$$\left.\begin{aligned} \dot{x}_1(t) &= x_2(t) \\ \dot{x}_2(t) &= \phi(x_1(t), x_2(t)) := -\Pi'(x_1(t)) \frac{\left[x_2^2(t) \Pi''(x_1(t)) + g\right]}{\left(1 + [\Pi'(x_1(t))]^2\right)} \end{aligned}\right\}$$

In view of convexity $\Pi''(x_1(t)) \geq 0$. This results in the conclusion that the set of all possible stationary points consists of all points $\begin{pmatrix} x_1 : \Pi'(x_1) = 0 \\ x_2 = \dot{x}_1 = 0 \end{pmatrix}$. If the function

*$\Pi = \Pi(x)$ is **strictly convex** (see Definition 21.2) with the minimum in $x = 0$, then the zero-state is (globally) **uniformly asymptotically stable** (see Fig. 20.3a). If it is **only convex** such that the minimum is attained in a neighborhood of $x = 0$ (see Fig. 20.3b), then the zero-state is **unstable**.*

Remark 20.3. *All criteria presented above are nonconstructive in the sense that each of them demands the exact knowledge of the solution $\bar{x}(t, x_0, t_0)$ of (20.1).*

The next subsection deals only with the **sufficient conditions of global asymptotic stability** which are based on some properties of the right-hand side of ODE (20.1) which makes them **constructive** and **easily verified**.

20.2.3 Sufficient conditions of asymptotic stability: constructive theory

This constructive theory of *stability*, more precisely, *asymptotic stability*, exists due to fundamental investigations of Lyapunov (1892), Barbashin (1951), Krasovskii (1952), Antosiewicz (1958), Letov (1962), Rumyantzev (1963), Chetaev (1965), Zubov (1964), Halanay (1966) and others.

20.2.3.1 Sufficient conditions for asymptotic stability: General result

Theorem 20.5. (on asymptotic local stability (Zubov 1964)) *Assume that there exists a positive-definite function $V(t, x)$ which is continuous in the point $x = 0$ uniformly on t for all $t \geq t_0$ and satisfying the following ODE*

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = -W(t, \bar{x}(t, x_0, t_0))} \tag{20.20}$$

on the trajectories of the system (20.1) where $W(t, x)$ is a positive-definite function (see Definition 20.3). Then the stationary point $x = 0$ of the system (20.1) is asymptotically locally stable uniformly on t_0.

Proof. Suppose that such function $V(t, x)$ exists. Show that $\bar{x}(t, x_0, t_0) \to 0$ as $t \to \infty$ whenever $\|x_0\|$ is small enough, that is, show that for any $\varepsilon > 0$ there exists $T = T(\varepsilon)$ such that $\|\bar{x}(t, x_0, t_0)\| < \varepsilon$ for all $t > T$. Notice that by the uniform continuity

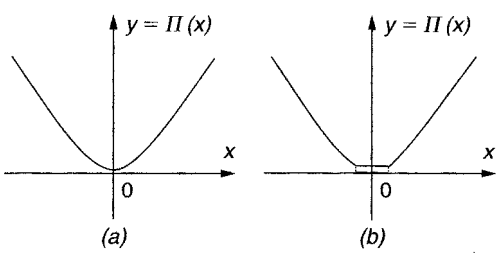

Fig. 20.3. Potential surfaces.

$V(t, x)$ and by Corollary 20.1 all trajectories $\bar{x}(t, x_0, t_0)$ of (20.1) remain within the region where $\|\bar{x}(t, x_0, t_0)\| < \varepsilon$ if $\|x_0\| < \delta$. So, the function $W(t, \bar{x}(t, x_0, t_0))$ remains bounded too. Suppose that $\|\bar{x}(t, x_0, t_0)\|$ does not converge to zero. By monotonicity of $V(t, \bar{x}(t, x_0, t_0))$, this means that there exists $\epsilon > 0$ and a moment $T = T(\epsilon)$ such that for all $t \geq T(\epsilon)$ we have $\|\bar{x}(t, x_0, t_0)\| > \epsilon$. Since $W(t, x)$ is a positive-definite function, it follows that

$$W(t, \bar{x}(t, x_0, t_0)) > \alpha > 0$$

for all $t \geq T(\epsilon)$ and, hence, by (20.20) we have

$$\begin{aligned} V(t, \bar{x}(t, x_0, t_0)) &= V(t_0, \bar{x}(t_0, x_0, t_0)) - \int_{s=t_0}^{t} W(s, \bar{x}(s, x_0, t_0))\, ds \\ &\leq V(t_0, \bar{x}(t_0, x_0, t_0)) - \alpha t \to -\infty \end{aligned}$$

which contradicts the condition that $V(t, x)$ is a positive-definite function. The fact that this result is uniform on t_0 follows from Corollary 20.1. Theorem is proven. □

20.2.3.2 Asymptotic stability for stationary system

Consider the stationary (autonomous) ODE

$$\boxed{\dot{x}(t) = f(x(t)), \quad f(0) = 0, \quad t \geq t_0} \tag{20.21}$$

Notice that the right-hand side of (20.21) can be represented as follows:

$$\boxed{\begin{aligned} f(x) &= Ax + h(x) \\ h(x) &:= f(x) - Ax \end{aligned}} \tag{20.22}$$

Below we will give **three very important results** concerning the asymptotic stability property of the zero-state $x = 0$ for the **stationary systems** governed by (20.21).

Theorem 20.6. (Lyapunov 1892)[5] *If*

1. the matrix $A \in \mathbb{R}^{n \times n}$ in (20.22) is stable (Hurwitz), i.e., for all $i = 1, \ldots, n$

$$\boxed{\operatorname{Re} \lambda_i(A) < 0} \tag{20.23}$$

2. and

$$\boxed{\frac{h(x)}{\|x\|} \to 0 \quad \textit{whenever} \quad \|x\| \to 0} \tag{20.24}$$

*then the stationary point $x = 0$ is **exponentially locally stable**.*

[5] This result is referred to as *the first Lyapunov's theorem* (method).

Proof. By the Lyapunov Lemma 9.1 for any $0 < Q = Q^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ and a given stable A there exists $0 < P = P^{\mathsf{T}} \in \mathbb{R}^{n\times n}$ such that

$$AP + PA^{\mathsf{T}} = -Q$$

Then for the Lyapunov function $V(x) := x^{\mathsf{T}} Px$ we have

$$\begin{aligned}
&\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = 2x^{\mathsf{T}}(t) P\dot{x}(t) \\
&\quad = 2x^{\mathsf{T}}(t) P[Ax(t) + h(x(t))] = 2x^{\mathsf{T}}(t) PAx(t) \\
&\qquad + 2x^{\mathsf{T}}(t) Ph(x(t)) = x^{\mathsf{T}}(t)[PA + A^{\mathsf{T}}P]x(t) \\
&\qquad + 2x(t)^{\mathsf{T}} Ph(x(t)) = -x^{\mathsf{T}}(t) Qx(t) + 2x^{\mathsf{T}}(t) Ph(x(t)) \\
&\qquad \le -x^{\mathsf{T}}(t) Qx(t) + 2\|x(t)\| \|P\| = -x^{\mathsf{T}}(t) Qx(t) \\
&\qquad + 2\|x(t)\|^2 \|P\| \|h(x(t))\| / \|x(t)\| \le -x^{\mathsf{T}}(t) Qx(t) \\
&\qquad + x^{\mathsf{T}}(t) Qx(t) \left(2\|P\| \lambda_{\max}\left(Q^{-1}\right) \|h(x(t))\| / \|x(t)\|\right)
\end{aligned}$$

By assumption (2) of this theorem, for $\varepsilon < \left[2\|P\| \lambda_{\max}\left(Q^{-1}\right)\right]^{-1}$ always exists δ such that if $\|x(t_0)\| \le \delta$, then $\|h(x(t_0))\| / \|x(t_0)\| \le \varepsilon$. Taking the corresponding $x(t_0)$ from the last differential inequality we find that

$$\begin{aligned}
&\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) \\
&\quad \le -x^{\mathsf{T}}(t) Qx(t) \left[1 - 2\|P\| \lambda_{\max}\left(Q^{-1}\right) \varepsilon\right] = -\alpha V(t, \bar{x}(t, x_0, t_0)) \\
&\quad \alpha := 1 - 2\|P\| \lambda_{\max}\left(Q^{-1}\right) \varepsilon > 0
\end{aligned}$$

and, hence,

$$V(t, \bar{x}(t, x_0, t_0)) \le V(t_0, \bar{x}(t_0, x_0, t_0)) e^{-\alpha t} \to 0 \text{ as } t \to \infty$$

Theorem is proven. □

Remark 20.4. *If the function $f(x)$ is differentiable at $x = 0$, then the matrix A in the Lyapunov theorem 20.6 is*

$$\boxed{A = \frac{\partial}{\partial x} f(x)\,|_{x=0}} \tag{20.25}$$

that is, A is ***the linear approximation*** *of the nonlinear vector function $f(x)$ in the origin.*

Example 20.4. *Consider the second-order ODE*

$$\ddot{x}(t) + \beta\dot{x}(t) - k\left(\frac{a}{c - x(t)} - x(t)\right) = 0, \qquad a, c, \beta, k > 0$$

which can be rewritten as the following extended first-order ODE

$$\left.\begin{aligned} &x_1(t) := x(t),\ \ x_2(t) := \dot{x}(t) \\ &\dot{x}_1(t) = f_1(x_1(t), x_2(t)) := x_2(t) \\ &\dot{x}_2(t) = f_2(x_1(t), x_2(t)) := k\left(\frac{a}{c - x_1(t)} - x_1(t)\right) - \beta x_2(t) \end{aligned}\right\}$$

The stationary points

$$\begin{aligned} x_1^* &= \frac{1}{2}\left(c + \sqrt{c^2 - 4a}\right), \quad x_2^* = 0 \\ x_1^{**} &= \frac{1}{2}\left(c - \sqrt{c^2 - 4a}\right), \quad x_2^{**} = 0 \end{aligned}$$

exist if $c^2 \geq 4a$. *Using the Taylor expansion in a small neighborhood of the stationary points, the function* $f(x_1, x_2)$ *can be represented as*

$$f(x_1, x_2) = \underbrace{f_2\left(x_1^{*,**}, x_2^{*,**}\right)}_{0} + A^{(*,**)}x + o(\|x\|) = A^{(*,**)}x + o(\|x\|)$$

where

$$A^{(*)} = \frac{\partial}{\partial x} f(x)\,|_{x=x^*} = \begin{bmatrix} 0 & 1 \\ b^* & -\beta \end{bmatrix}, \quad b^* := -k\left(1 + \frac{a}{\left(c - x_1^*\right)^2}\right)$$

$$A^{(**)} = \frac{\partial}{\partial x} f(x)\,|_{x=x^{**}} = \begin{bmatrix} 0 & 1 \\ b^{**} & -\beta \end{bmatrix}, \quad b^{**} := -k\left(1 + \frac{a}{\left(c - x_2^*\right)^2}\right)$$

In both cases the eigenvalues of A *satisfy the quadratic equation*

$$\lambda^2 + \beta\lambda - b = 0$$

and are equal

$$\lambda_{1,2} = \frac{-\beta \pm \sqrt{\beta^2 + 4b}}{2}, \quad b = b^{*,**} = -\frac{kc^2}{2a}\left(1 \pm \sqrt{1 - \frac{4a}{c^2}}\right) \leq 0$$

Hence, the points $x^{*,**}$ *are exponentially locally stable by the Lyapunov theorem 20.6 when* $c^2 > 4a$, *since Re* $\lambda_{1,2} < 0$.

Example 20.5. (Lefschetz 1965) *For which* a, b *and* $f(0)$ *the following system*

$$x^{(3)}(t) + f(\dot{x}(t))\ddot{x}(t) + a\dot{x}(t) + bx(t) = 0 \tag{20.26}$$

is asymptotically locally stable? Here the function $f(z)$ is assumed to be differentiable in the point $z = 0$. Let us apply the Lyapunov theorem 20.6 and represent (20.26) in the matrix form:

$$x_1(t) := x(t), \quad x_2(t) := \dot{x}(t), \quad x_3(t) := \ddot{x}(t)$$

$$\bar{x}(t) := (x_1(t), x_2(t), x_3(t))^{\mathsf{T}}$$

$$\begin{pmatrix} \dot{x}_1(t) \\ \dot{x}_2(t) \\ \dot{x}_3(t) \end{pmatrix} = A \begin{pmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{pmatrix} + h(\bar{x}(t))$$

where

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -b & -a & -f(0) \end{bmatrix}, \quad h(\bar{x}) = \begin{pmatrix} 0 \\ 0 \\ [-f'(0)\, x_2 + o(\|x_2\|)]\, x_3 \end{pmatrix}$$

Firstly, notice that

$$h(\bar{x}) / \|\bar{x}\| \to 0 \text{ as } \|\bar{x}\| \to 0$$

So, to answer the initial question we should try to find the conditions when the matrix A is stable. The corresponding characteristic polynomial is

$$p_A(\lambda) = \lambda^2 (f(0) + \lambda) + b + a\lambda$$

and it is Hurwitz if and only if (see Criterion 9.3)

$$f(0) > 0, \quad a > 0, \quad b > 0, \quad af(0) > b$$

which gives the relation among the parameters guaranteeing the exponential local stability.

20.3 Asymptotic global stability

20.3.1 Definition of asymptotic global stability

Definition 20.7. *The equilibrium zero-point (or zero-state) $x = 0$ of the system given by ODE (20.1) is said to be* ***asymptotically globally stable***[6] *if $\bar{x}(t, x_0, t_0) \to 0$ when $t \to \infty$ for any initial state $x_0 = x(t_0)$ of a bounded norm.*

[6] Sometimes such asymptotically globally stable systems are called, by R. Kalman, **mono-stable** (or having **the dichotomy property**) systems.

20.3.2 Asymptotic global stability for stationary systems

Theorem 20.7. (Barbashin & Krasovskii 1952) *To guarantee the* ***asymptotic global stability*** *of the unique stationary point $x = 0$ of (20.21) with the continuous right-hand side it is sufficient to show the existence of a differentiable function $V = V(x)$ such that*
(a) $V(0) = 0$ and for any $x \neq 0$

$$\boxed{V(x) > 0} \tag{20.27}$$

(b) for $\|x\| \to \infty$

$$\boxed{V(x) \to \infty} \tag{20.28}$$

(c) for any $t_0,\ x_0 \in \mathbb{R}$ and any $\bar{x}(t, x_0, t_0) \neq 0$

$$\boxed{\frac{d}{dt} V(\bar{x}(t, x_0, t_0)) < 0} \tag{20.29}$$

Proof. By assumption (c) $V(\bar{x}(t, x_0, t_0))$ monotonically decreases and is bounded from below. Therefore, by the Weierstrass theorem 14.9 $V(\bar{x}(t, x_0, t_0)) \downarrow V^*$ monotonically. By property (b) $\bar{x}(t, x_0, t_0)$ remains to be bounded. Suppose that $V^* > 0$. Hence, by property (a) there exist $\varepsilon > 0$ and $T = T(\varepsilon) \geq t_0$ such that $\inf\limits_{t \geq T(\varepsilon)} \|\bar{x}(t, x_0, t_0)\| > \varepsilon$. Therefore, by (20.29) we get

$$\sup_{t \geq T(\varepsilon)} \frac{d}{dt} V(\bar{x}(t, x_0, t_0)) < -\varepsilon',\ \varepsilon' > 0$$

which implies the inequality

$$0 < V^* < V(\bar{x}(t, x_0, t_0)) = V(\bar{x}(T, x(T), T)) + \int_{s=T}^{t} \frac{d}{dt} V(\bar{x}(s, x_0, t_0))\, ds \leq V(\bar{x}(T, x(T), T)) - \varepsilon'(t - T)$$

making the right-hand side negative for large enough t. This leads to the contradiction. So, $V^* = 0$. Theorem is proven. □

Theorem 20.8. (Krasovskii 1952) *Assume that*
(a) the function $f(x)$ in (20.21) is differentiable everywhere in x and the stationary point $x = 0$ (where $f(x) = 0$) is unique;
(b) there exists a positive-definite matrix $B = B^{\mathsf{T}} > 0$ such that the functional matrix $M(x)$ defined by

$$\boxed{M(x) := \frac{\partial}{\partial x} f(x)^{\mathsf{T}} B + B \frac{\partial}{\partial x} f(x)} \tag{20.30}$$

is strictly negative on x, that is, for all $x \in \mathbb{R}^n$

$$\boxed{\lambda_{\max}(M(x)) \leq -c,\ c > 0} \tag{20.31}$$

Then the point $x = 0$ is ***asymptotically globally stable***.

Proof. Take in Theorem 20.7

$$V(x) := \frac{1}{2} f(x)^{\top} B f(x) \geq 0$$

Notice that $V(x)$ is nonnegative in $\mathbb{R}^n$. Then

$$\begin{aligned}\frac{d}{dt} V(\bar{x}(t, x_0, t_0)) &= f^{\top}(\bar{x}(t, x_0, t_0)) M(\bar{x}(t, x_0, t_0)) f(\bar{x}(t, x_0, t_0)) \\ &\leq \lambda_{\max}(M(x)) \| f(\bar{x}(t, x_0, t_0)) \|^2 \\ &\leq -c \| f(\bar{x}(t, x_0, t_0)) \|^2 < 0\end{aligned}$$

if $\bar{x}(t, x_0, t_0) \neq 0$. Hence, by Theorem 20.7 we have the asymptotic global stability that proves the desired result. □

Example 20.6. *Consider the following ODE:*

$$\left.\begin{aligned}\dot{x}_{1,t} &= -a x_{1,t} + b x_{2,t}, \quad a > 1/4, \quad -a < b \\ \dot{x}_{2,t} &= \sin x_{1,t} - x_{2,t}\end{aligned}\right\} \tag{20.32}$$

Let us demonstrate how Theorem 20.8 works. The stationary points here satisfy the relation

$$-a x_1^* + b x_2^* = 0, \quad \sin x_1^* - x_2^* = 0$$

and are equal to $x_1^* = x_2^* = 0$. It is the unique equilibrium point since the derivative of the function $\phi(x_1) := -a x_1 + b \sin x_1$, which zeros we are interested in, at the point $x_1 = 0$ is

$$\phi'(x_1) := -a + b \cos x_1 < -a - b < 0$$

and remains negative for all x_1 (see the corresponding graphics of $\phi(x_1)$). Then we have

$$\frac{\partial}{\partial x} f(x) = \begin{bmatrix} -a & b \\ \cos x_1 & -1 \end{bmatrix}$$

Taking in Theorem 20.8 $B = I$ we get

$$M(x) = \begin{bmatrix} -2a & b + \cos x_1 \\ b + \cos x_1 & -2 \end{bmatrix}$$

This matrix will be strictly negative. Indeed, by the Sylvester criteria 7.20 we have

$$-2a < 0$$
$$4a - (b + \cos x_1)^2 = 4a - b^2 - 2b\cos x_1 - \cos^2 x_1$$
$$\leq 4a - b^2 - 2b\cos x_1 \leq -4a - b^2 + 2|b| < 0 \text{ as } a > 1/4$$

So, the state $x_1^* = x_2^* = 0$ is asymptotically globally stable.

20.3.3 Asymptotic global stability for nonstationary system

Let us consider again the general ODE (20.1). The results below deal with the asymptotic stability of the origin on the trajectories of this system.

We need a definition to use throughout this subsection.

Definition 20.8. *The class $\mathcal{K}$ of functions $f : \mathbb{R} \to \mathbb{R}$ is said to be* ***Hahn's class*** *if it contains all nonnegative functions satisfying the following conditions:*

1. *$f \in C(-\infty, \infty)$, i.e., f is continuous in $\mathbb{R}$;*
2. *f is strictly monotone, i.e., for any $x \in \mathbb{R}$ and any $\varepsilon > 0$*

$$\boxed{f(x+\varepsilon) > f(x)} \tag{20.33}$$

3.

$$\boxed{f(0) = 0} \tag{20.34}$$

Theorem 20.9. (Antosiewicz 1958) *If*

1. *the stationary point $x^* = 0$ of (20.1) is uniformly (on t_0) locally stable;*
2. *there exists a function $V(t, x)$ which is continuously differentiable in both variables and, additionally,*
   (a) for any $x \in \mathbb{R}^n$ and any $t \geq t_0$

$$\boxed{V(t,x) \geq a(\|x\|)} \tag{20.35}$$

   (b) for any $t \geq t_0$

$$\boxed{V(t,0) = 0} \tag{20.36}$$

   (c) for any $x \in \mathbb{R}^n$ and any $t \geq t_0$

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) \leq -b(\|\bar{x}(t, x_0, t_0)\|)} \tag{20.37}$$

where the functions $a(\cdot), b(\cdot)$ are from Hahn's class $\mathcal{K}$, then the stationary point $x^* = 0$ of (20.1) is ***asymptotically globally stable****.*

Proof. By the conditions of the theorem it follows that any trajectory $\bar{x}(t, x_0, t_0)$ obligatory will arrive at some small neighborhood Ω containing the point $x = 0$ and will never leave it. Indeed, by condition (2c) $\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) < 0$ whenever $\|x\| > 0$ and, hence, decreases reaching Ω in a finite time t' uniformly on t_0. By condition (1) after this moment it will always belong to Ω. Suppose that $\|\bar{x}(t, x_0, t_0)\|$ does not converge to zero. By monotonicity of $V(t, \bar{x}(t, x_0, t_0))$ (see the condition (20.37)), this means that there exists $\epsilon > 0$ and a moment $T = T(\epsilon)$ such that for all $t \geq T(\epsilon)$ we have $\|\bar{x}(t, x_0, t_0)\| > \epsilon$. Since $b(\|x\|)$ belongs to Hahn's class $\mathcal{K}$, it follows that

$$b(\|\bar{x}(t, x_0, t_0)\|) > \alpha > 0$$

and, hence, by (20.37)

$$V(t, \bar{x}(t, x_0, t_0)) \leq V(t_0, \bar{x}(t_0, x_0, t_0)) - \int_{s=t'}^{t} b\left(\left\|\bar{x}\left(s, x\left(t'\right), t'\right)\right\|\right) ds \leq V(t_0, x_0) - \alpha t \to -\infty$$

which contradicts the assumption (20.35) that $V(t, x) \geq a(\|x\|)$. So, $\|\bar{x}(t, x_0, t_0)\| \to 0$ as $t \to \infty$. Theorem is proven. □

Corollary 20.2. (Halanay 1966) *The result of Theorem 20.9 remains true if instead of assumption (2c) there is*

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) \leq -cV(t, \bar{x}(t, x_0, t_0)),\ c(\cdot) \in \mathcal{K}} \tag{20.38}$$

Proof. It is sufficient to take in Theorem 20.9

$$b(\|x\|) := c(a(\|x\|))$$

since $c(V(t, x)) \geq c(a(\|x\|))$. □

Theorem 20.10. (Chetaev 1965) *Let there exist the function $k(\cdot) \in C$, $a(\cdot) \in \mathcal{K}$ and $V(t, x) \in C^1$ such that*

1. for any $x \in \mathbb{R}^n$ and any $t \geq t_0$

$$\boxed{V(t, x) \geq k(t) a(\|x\|)} \tag{20.39}$$

2. for any $t \geq t_0$

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) \leq 0} \tag{20.40}$$

3. $k(\cdot) \in \mathcal{K}$, *i.e., for any* $t \geq t_0$ $k(t) \geq 0$ *and* $k(t) \to \infty$ *as* $t \to \infty$.
Then the stationary point $x^* = 0$ *of (20.1) is* ***asymptotically globally stable***.

Proof. Evidently, by condition (2), $V(t, \bar{x}(t, x_0, t_0))$ is a nondecreasing function of t and, hence, in view of condition (1),

$$\infty > \limsup_{t\to\infty} V(t, \bar{x}(t, x_0, t_0)) \geq \liminf_{t\to\infty} V(t, \bar{x}(t, x_0, t_0))$$
$$\geq \liminf_{t\to\infty} k(t)\, a(\|\bar{x}(t, x_0, t_0)\|) \geq 0$$

But, by condition (3), $k(t) \to \infty$ and, hence, $a(\|\bar{x}(t, x_0, t_0)\|) \to 0$ as $t \to \infty$. And since $a(\cdot) \in \mathcal{K}$ it follows that $\|\bar{x}(t, x_0, t_0)\| \to 0$ as $t \to \infty$. Theorem is proven. □

20.4 Stability of linear systems

20.4.1 Asymptotic and exponential stability of linear time-varying systems

Consider the linear time-varying system given by the following ODE:

$$\boxed{\dot{x}(t) = A(t)\, x(t),\ x(t_0) = x_0,\ t \geq t_0} \tag{20.41}$$

Its solution can be presented as

$$\boxed{x(t) = \Phi(t, t_0)\, x_0} \tag{20.42}$$

where $\Phi(t, t_0)$ is the corresponding fundamental matrix defined by (19.56). This presentation, evidently, implies the following proposition.

Proposition 20.1. *The system (20.41) is*
(a) locally ***stable*** *if and only if*

$$\boxed{c := \sup_{t\geq t_0} \|\Phi(t, t_0)\| < \infty} \tag{20.43}$$

(b) ***asymptotically globally stable*** *if and only if*

$$\boxed{\|\Phi(t, t_0)\| \to 0 \quad \textit{whereas} \quad t \to \infty} \tag{20.44}$$

Let $A(t)$ satisfy the inequality

$$\int_{s=t}^{\infty} \operatorname{tr} A(s)\, ds > -\infty \tag{20.45}$$

which by the Liouville's theorem 19.7 implies that $\det \Phi(t, t_0) \neq 0$ for all $t \geq t_0$, and as a result there exists a constant $a > 0$ such that

$$\int_{s=t}^{\infty} \Phi^{\mathsf{T}}(s, t)\, \Phi(s, t)\, ds \geq a \tag{20.46}$$

Below we present the criterion of exponential stability expressed in terms of *Lyapunov's approach (see Theorem 20.3)*. It is interesting to note that *for the class of linear systems this approach turns out to be constructive.*

Theorem 20.11. (on exponential stability (Zubov 1964)) *For the solution $\bar{x}(t, x_0, t_0)$ of (20.41) satisfying (20.45) to be* ***exponentially globally stable*** *(see Definition 20.5) it is necessary and sufficient to show the existence of two quadratic forms*

$$\boxed{V(t, x) = x^{\mathsf{T}} P(t) x \text{ and } W(t, x) = x^{\mathsf{T}} Q(t) x} \tag{20.47}$$

such that

1. *both quadratic forms are positive definite and increase no quicker than the quadratic function, i.e.,*

$$\boxed{\begin{array}{c} a_1 \|x\|^2 \leq V(t, x) \leq a_2 \|x\|^2 \\ b_2 \|x\|^2 \geq W(t, x) \geq b_1 \|x\|^2 \\ a_1, a_2, b_1, b_2 \quad \text{are positive constants} \end{array}} \tag{20.48}$$

2. *$V(t, x)$ and $W(t, x)$ are related as*

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = -W(t, \bar{x}(t, x_0, t_0))} \tag{20.49}$$

Proof.

(a) *Necessity.* Let $\|\bar{x}(t, x_0, t_0)\| \leq \alpha \exp(-\beta(t - t_0))$. Define

$$\begin{aligned} W(t, x) &:= \|x\|^2 \\ V(t, x) &:= \int_{t=t}^{\infty} W(s, \bar{x}(s, x, t))\, ds \end{aligned} \tag{20.50}$$

Evidently, in this case $b_1 = b_2 = 1$. Show that $V(t, x)$ as it is defined above satisfies the conditions (20.48) and (20.49). By (20.42) and in view of (20.46) we have

$$\begin{aligned} V(t, x) &= \int_{s=t}^{\infty} \bar{x}^{\mathsf{T}}(s, x, t)\, \bar{x}(s, x, t)\, ds \\ &= x^{\mathsf{T}} \left[\int_{s=t}^{\infty} \Phi^{\mathsf{T}}(s, t)\, \Phi(s, t)\, ds \right] x \geq a \|x\|^2 \end{aligned}$$

By the inequality (20.8) one has

$$a \|x\|^2 \leq V(t,x) \leq \int_{s=t}^{\infty} \alpha^2 \|x\|^2 e^{-2\beta(s-t)} ds = \frac{1}{2\beta}\alpha^2 \|x\|^2$$

This means that $V(t,x)$ satisfies (20.48) with $a_1 = a$ and $a_2 = \frac{1}{2\beta}\alpha^2$. Obviously, $V(t,x)$, as it is defined in (20.50), satisfies (20.49).

(b) *Sufficiency*. Show that (20.49) together with (20.47) and (20.48) imply (20.8). Integrating (20.49) for $V \neq 0$ leads to

$$\int_{s=t_0}^{t} \frac{dV}{V} = -\int_{s=t_0}^{t} \frac{W(s, \bar{x}(s, x_0, t_0))}{V(s, \bar{x}(s, x_0, t_0))} ds$$

and, hence,

$$V(t, \bar{x}(t, x_0, t_0)) = V(t_0, x_0) \exp\left(-\int_{s=t_0}^{t} \frac{W(s, \bar{x}(s, x_0, t_0))}{V(s, \bar{x}(s, x_0, t_0))} ds\right) \tag{20.51}$$

Notice that

$$\frac{b_1}{a_2} \leq \frac{\bar{x}^\top Q(s)\bar{x}}{\bar{x}^\top P(s)\bar{x}} = \frac{W(s, \bar{x}(s, x_0, t_0))}{V(s, \bar{x}(s, x_0, t_0))}$$

Application of these estimates in (20.51) gives

$$a_1 \|\bar{x}(t, x_0, t_0)\|^2 \leq V(t, \bar{x}(t, x_0, t_0)) \leq V(t_0, x_0) \exp\left(-\frac{b_1}{a_2}[t - t_0]\right)$$

Using the estimates (20.15) for $V(t_0, x_0)$ implies

$$a_1 \|\bar{x}(t, x_0, t_0)\|^2 \leq V(t, \bar{x}(t, x_0, t_0)) \leq a_2 \|x_0\|^2 \exp\left(-\frac{b_1}{a_2}[t - t_0]\right)$$

and, therefore, we obtain the exponential global stability (20.8) for the zero-state $x = 0$ with

$$\alpha := \sqrt{\frac{a_2}{a_1}}, \quad \beta := \frac{b_1}{2a_2}$$

Theorem is proven. □

Remark 20.5. *Since*

$$\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = \frac{d\bar{x}^\top}{dt} P(t)\bar{x} + \bar{x}^\top \frac{d}{dt} P(t)\bar{x} + \bar{x}^\top P(t) \frac{d\bar{x}}{dt}$$

$$= \bar{x}^\top \left[A^\top(t) P(t) + \frac{d}{dt} P(t) + P(t) A(t)\right] \bar{x}$$

and

$$W(t, \bar{x}(t, x_0, t_0)) = \bar{x}^\top Q(t)\bar{x}$$

in view of (20.49), we have

$$\bar{x}^\top \left[A^\top(t) P(t) + \frac{d}{dt} P(t) + P(t) A(t) \right] \bar{x} = -\bar{x}^\top Q(t)\bar{x}$$

which is true on any trajectory $\bar{x} = \bar{x}(t, x_0, t_0)$. *This transforms the nonconstructive form of Theorem 20.3 into the constructive one as in Theorem 20.11.*

Corollary 20.3. *For any solution* $\bar{x}(t, x_0, t_0)$ *of (20.41) satisfying (20.45) to be* ***exponentially stable*** *(see Definition 20.5) it is* ***necessary and sufficient*** *that there exist two symmetric positive definite matrices* $P(t)$ *and* $Q(t)$ *such that*

$$\boxed{\frac{d}{dt} P(t) + A^\top(t) P(t) + P(t) A(t) + Q(t) = 0} \tag{20.52}$$

Equation (20.52) is known as the ***differential Lyapunov equation.*** *If* $A(t) = A$ *is a constant matrix then we may take* $P(t) = P$, $Q(t) = Q$ *and (20.52) is converted into the* ***algebraic Lyapunov equation***

$$\boxed{A^\top P + PA + Q = 0} \tag{20.53}$$

Its property is seen in Lemma 9.1.

20.4.2 Stability of linear system with periodic coefficients

Consider again the linear system (20.41) where the matrix $A(t)$ is periodic with the period T, i.e., for any $t \geq t_0$

$$\boxed{A(t) = A(t + T)} \tag{20.54}$$

Proposition 20.2. *For equation (20.41) the fundamental matrix is*

$$\Phi(t, t_0) = \tilde{\Phi}(t - t_0) = Z(t - t_0)\, e^{R(t - t_0)}$$

$$Z(\tau) = Z(\tau + T)$$

and, hence, all stability properties depend on the properties of the matrix R. But according to (19.73) for $\tau = 0$ *it follows that*

$$e^{RT} = \tilde{\Phi}^{-1}(0)\, \tilde{\Phi}(T)$$

So, we can derive the following results.

Corollary 20.4.
1. the system (20.41) is exponentially stable if and only if

$$Re\ \lambda_i\,(R) < 0 \text{ for all } i = 1, \ldots, n$$

2. the system (20.41) is Lyapunov stable if and only if

$$Re\ \lambda_i\,(R) \leq 0 \text{ for all } i = 1, \ldots, n$$

and the multiplicity μ_i of the eigenvalues $Re\ \lambda_i\,(B) = 0$ does not exceed 1, i.e.,

$$\mu_i = 1$$

20.4.3 BIBO stability of linear time-varying systems

Consider again a linear nonstationary system governed by the following ODE

$$\boxed{\dot{x}\,(t) = A\,(t)\,x\,(t) + \omega\,(t), \quad x\,(t_0) = x_0 \in \mathbb{R}^n, \quad t \geq t_0} \tag{20.55}$$

where ω_t is an exogenous input from L_∞ (see 22.154), i.e.,

$$\boxed{\|\omega\|_{L_\infty} := \underset{t \geq t_0}{ess\ \sup} \|\omega\,(t)\| < \infty} \tag{20.56}$$

Below we present the criterion explaining when $\|x\|_{L_\infty}$ is also bounded for any x_0. Such systems are called *BIBO* (*bounded input–bounded output*) *stable*.

Theorem 20.12. (Criterion of BIBO stability) *For the system (20.55) $\|x\|_{L_\infty} < \infty$ whereas $\|\omega\|_{L_\infty} < \infty$ if and only if the corresponding fundamental matrix $\Phi\,(t, t_0)$ satisfies the following conditions:*
1.

$$\boxed{c := \sup_{t \geq t_0} \|\Phi\,(t, t_0)\| < \infty}$$

2.

$$\boxed{\int_{t=t_0}^{\infty} \|\Phi\,(t, t_0)\|\,dt < \infty} \tag{20.57}$$

Proof. By (19.64) we have

$$x\,(t) = \Phi\,(t, t_0)\,x_0 + \int_{s=t_0}^{t} \Phi\,(t, s)\,\omega\,(s)\,ds$$

Sufficiency follows directly from this formula since

$$\|x(t)\| \leq \|\Phi(t, t_0)\, x_0\| + \left\| \int_{s=t_0}^{t} \Phi(t, s)\, \omega(s)\, ds \right\|$$

$$\leq c\, \|x_0\| + \|\omega\|_{L_\infty} \int_{s=t_0}^{t} \|\Phi(t, s)\|\, ds < \infty$$

Let us prove *necessity*. Suppose that $\|x\|_{L_\infty} < \infty$. Taking $\omega(t) \equiv 0$ we have $x(t) = \Phi(t, t_0)\, x_0$ that proves the necessity of which condition (1). Take now $x_0 = 0$ and suppose that condition (2) is violated, that is, there exists at least one element (i_0, j_0) of the matrix $\Phi(t, t_0)$ such that

$$\int_{s=t_0}^{t} \left|\Phi_{i_0 j_0}(t, s)\right| ds \to \infty \text{ as } t \to \infty$$

Hence,

$$\|x(t)\| = \left\| \int_{s=t_0}^{t} \Phi(t, s)\, \omega(s)\, ds \right\|$$

$$= \sqrt{\sum_{i=1}^{n} \left| \int_{s=t_0}^{t} \sum_{j=1}^{n} \Phi_{ij}(t, s)\, \omega_j(s)\, ds \right|^2} \geq \sqrt{\left| \sum_{j=1}^{n} \int_{s=t_0}^{t} \Phi_{i_0 j}(t, s)\, \omega_j(s)\, ds \right|^2}$$

Taking then

$$\omega_j(s) := \begin{cases} \operatorname{sign} \Phi_{i_0 j_0}(t, s) & \text{if} \quad j = j_0 \\ 0 & \text{if} \quad j \neq j_0 \end{cases}$$

from the last inequality we obtain

$$\|x(t)\| \geq \int_{s=t_0}^{t} \left|\Phi_{i_0 j_0}(t, s)\right| ds \to \infty \text{ as } t \to \infty$$

But this contradicts the assumption that $\|x\|_{L_\infty} < \infty$. Theorem is proven. □

Example 20.7. *Consider the system given by*

$$\dot{x}(t) + [a + \sin(\omega_0 t)]\, x(t) = \omega(t), \quad x(t_0) = x_0 \in \mathbb{R}, \quad a > 0$$

The corresponding transition matrix (in this case it is a scalar function) is as follows

$$\Phi(t,s) = \exp\left\{-\int_{\tau=s}^{t}[a+\sin(\omega_0\tau)]\,d\tau\right\}$$
$$= \exp\left\{-a(t-s)+\frac{1}{\omega_0}[\cos(\omega_0 t)-\cos(\omega_0 s)]\right\}$$

So, it is bounded and, hence, the first condition of the theorem is fulfilled. Let us check the second one:

$$\int_{t=t_0}^{\infty}|\Phi(t,t_0)|\,dt = \lim_{t\to\infty}\int_{s=t_0}^{t}\exp\left\{a(s-t)+\frac{[\cos(\omega_0 t)-\cos(\omega_0 s)]}{\omega_0}\right\}ds$$
$$< \exp\left\{\frac{2}{\omega_0}\right\}\limsup_{t\to\infty}\int_{s=t_0}^{t}\exp\{-a(t-s)\}\,ds = \frac{1}{a}\exp\left\{\frac{2}{\omega_0}\right\} < \infty$$

This means that the second condition (20.57) of the theorem is also valid for any $a > 0$, and, hence, this system is BIBO stable for any $a > 0$.

20.5 Absolute stability

20.5.1 Linear systems with nonlinear feedbacks

Consider the dynamic system given by the following ODE:

$$\boxed{\begin{array}{l}\dot{x}(t) = Ax(t) + bu(t), \quad t \geq t_0\\ u(t) = \varphi(y(t)), \quad y(t) = c^{\mathsf{T}}x(t)\\ A \in \mathbb{R}^{n\times n};\ b, \quad c \in \mathbb{R}^n; \quad u(t), \quad y(t) \in \mathbb{R}\end{array}} \tag{20.58}$$

It can be interpreted (see Fig. 20.4) as a linear system given by the transfer function

$$\boxed{H(s) = c^{\mathsf{T}}(sI - A)^{-1}b} \tag{20.59}$$

with a nonlinear feedback

$$\boxed{u = \varphi(y)} \tag{20.60}$$

We will consider the class $\mathcal{F}$ of continuous functions $\varphi(y)$ (nonlinear feedbacks) satisfying

$$\boxed{0 \leq \frac{\varphi(y)}{y} \leq k \text{ for } y \neq 0,\ \varphi(0) = 0} \tag{20.61}$$

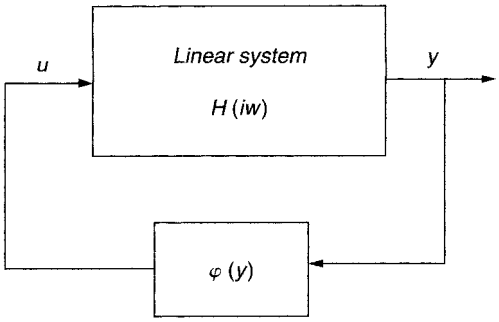

Fig. 20.4. A linear system with a nonliner feedback.

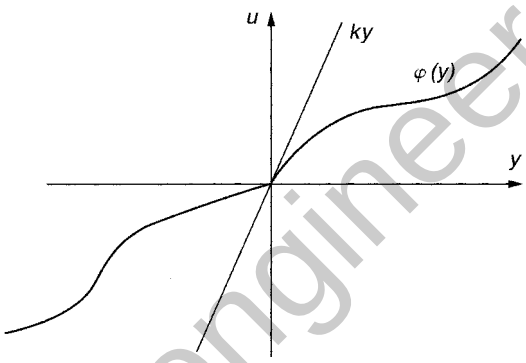

Fig. 20.5. A function $u = \varphi(y)$ from the class $\mathcal{F}$.

Definition 20.9. *The nonlinear system (20.58) is said to be* ***absolutely stable in the class*** $\mathcal{F}$ *if the solution* $x(t) \equiv 0$ *(or zero-state) is asymptotically globally stable (see Definition 20.7) for any nonlinear feedback (20.60) satisfying (20.61).*

In this section we are interested in finding the conditions guaranteeing the absolute stability of the system (20.58) in the class $\mathcal{F}$.

20.5.2 Aizerman and Kalman conjectures

Proposition 20.3. (Conjecture of M.A. Aizeman, 1949) *Let the system (20.58) be stable for any*

$$\varphi(y) = \alpha y, \alpha \in [0, k]$$

It seems to be true that this system remains stable for any feedback $\varphi(y)$ *satisfying (20.61), namely, for any* $\varphi(y)$ *such that*

$$0 \leq \frac{\varphi(y)}{y} \leq k \quad for \quad y \neq 0, \quad \varphi(0) = 0$$

Proposition 20.4. (Conjecture of R. Kalman, 1957) *Let the system (20.58) be stable for any*

$$\varphi(y) = \alpha y, \alpha \in [0, k]$$

It seems to be natural to admit that this system remains stable for any feedback $\varphi(y)$ satisfying the conditions:

$$\varphi(y) \text{ is differentiable}, \quad \varphi(0) = 0$$
$$0 \leq \varphi'(y) \leq k$$

Claim 20.2. *Conjectures of Both M.A. Aizerman and R. Kalman are not valid.*

Proof. See a number of counterexamples in Pliss (1964). □

Counterexample (Pliss 1964) Let

$$H(s) = \frac{s^2}{\left[(s+0.5)^2 + (0.9)^2\right]\left[(s+0.5)^2 + (1.1)^2\right]}$$

The closed-loop system is stable for any $u = ky$ with

$$k \in [-0,7124, \infty)$$

It follows for example from the Routh–Hurwitz criterion (see Theorem 9.2), applied to the closed-loop system. But for

$$\varphi(y) = \begin{cases} y^3 & \text{for} \quad |y| \leq \sqrt{|k|} \\ ky & \text{for} \quad |y| > \sqrt{|k|} \end{cases}$$

in this system auto-oscillations arise, and, hence, there is no asymptotic stability.
These conjectures were proposed before the keystone result of V.M. Popov who found the exact conditions of absolute stability of the linear system (20.58) with any feedback satisfying (20.61).

20.5.3 Analysis of absolute stability

To guarantee the absolute stability of the system (20.58)–(20.61), according to the Barbashin–Krasovskii theorem 20.7, it is sufficient that there exists the Lyapunov function $V(x)$ such that

(a) $V(0) = 0$ and

$$V(x) > 0 \quad \text{for any} \quad x \neq 0$$

(b)

$$V(x) \to \infty \quad \text{whereas} \quad \|x\| \to \infty$$

(c) for any $t \geq t_0$

$$\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) < 0$$

In Lurie & Postnikov (1944) it was suggested that the function $V(x)$ as a *quadratic form plus the integral of the nonlinear feedback* be found, that is,

$$\boxed{V(x) = x^{\top} P x + q \int_{y=0}^{y=c^{\top} x} \varphi(y)\, dy} \tag{20.62}$$

where $P = P^{\top}$ is a real matrix and q is a real number.

Remark 20.6. *Notice that if*

$$\boxed{P = P^{\top} > 0 \quad \text{and} \quad q \geq 0} \tag{20.63}$$

the function $V(x)$ (20.62) satisfies conditions (a) and (b) given above. Indeed, by the condition (20.61) we have $y\varphi(y) \geq 0$ for any $y \in \mathbb{R}$, and, therefore, $\int_{y=0}^{y} \varphi(y)\, dy \geq 0$. Below we shall see that q is also admitted to be negative.

Calculating the time derivative of (20.62) on the trajectories of (20.58) we obtain

$$\begin{gathered} \frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = 2x^{\top}(t) P(Ax(t) + bu(t)) + qu(t)\dot{y}(t) \\ \dot{y}(t) = c^{\top}(Ax(t) + bu(t)) \end{gathered} \tag{20.64}$$

The right-hand side is a quadratic form of variables x and u, namely,

$$Q_0(x, u) := 2x^{\top} P(Ax + bu) + quc^{\top}(Ax + bu) \tag{20.65}$$

So, (20.64) is

$$\boxed{\frac{d}{dt} V(t, \bar{x}(t, x_0, t_0)) = Q_0(x(t), u(t))} \tag{20.66}$$

Therefore, to fulfill condition (c), given above, one must fulfill the condition

$$Q_0(x, u) < 0 \quad \text{for all real} \quad x \in \mathbb{R}^n \quad \text{and all} \quad u \in \mathbb{R}$$

which, by (7.6), is equivalent to fulfilling of the following inequality:

$$\boxed{\begin{gathered} Q_0(z, u) < 0 \\ \text{for all complex} \quad z \in \mathbb{C}^n \quad \text{and all complex} \quad u \in \mathbb{C} \end{gathered}} \tag{20.67}$$

Next, return to the constraints (20.61), and notice that they can be rewritten as

$$0 \le u/y \le k,\ 0 \le u^2 \le kuy,\ uy \ge k^{-1}u^2$$

or, equivalently, as

$$\boxed{Q_1(x,u) := u\left(k^{-1}u - y\right) \le 0,\ y = c^{\top}x} \tag{20.68}$$

Notice that $Q_1(x,u)$ is also a quadratic form of x and u, and in fact defines the constraint for these variables.

Theorem 20.13. (Gelig *et al.* 1978) *Suppose that*

- *the matrix A in (20.58) has no pure imaginary eigenvalues;*
- *the nonlinear feedback $\varphi(y)$ is from the class $\mathcal{F}$ (20.61).*

To guarantee the existence of the function $V(x)$ of the form (20.62) for which

$$\frac{d}{dt}V\left(t, \bar{x}\left(t, x_0, t_0\right)\right) < 0$$

*in any points $x = x(t) \in \mathbb{R}^n$ and any $u = u(t) \in \mathbb{R}$ satisfying the constraints (20.61), it is **necessary and sufficient** that for all $\omega \in [-\infty, \infty]$ the following "**frequency inequality**" (it is known as **Popov's inequality**) would be fulfilled:*

$$\boxed{k^{-1} - \operatorname{Re}\left[(1 + \tilde{q}i\omega)\,H(i\omega)\right] > 0} \tag{20.69}$$

where the complex function $H(s)$ is the transfer function (20.59) of the linear subsystem and $\tilde{q}$ is a real number.

Remark 20.7. *In fact, the frequency inequality (20.69) is the necessary and sufficient condition for fulfilling only condition (c) of the Barbashin–Krasovskii theorem 20.7 which, together with conditions (a) and (b), is sufficient for asymptotic global stability of the zero-state of the system given by (20.58).*

Proof.

(a) *Sufficiency.* Evidently, inside the constraint (20.68) there exists a point $(\mathring{x}, \mathring{u})$ such that $Q_1(\mathring{x}, \mathring{u}) < 0$. Hence, S-procedure (see subsection 12.3.2) may be applied in its version given in Corollary 12.2, that is, for some $\tau \ge 0$ and any (x,u) $(\|x\| + |u| \ne 0)$ define the quadratic form

$$Q_\tau(x,u) := Q_0(x,u) - \tau Q_1(x,u) \tag{20.70}$$

Obviously,

$$\frac{d}{dt}V(t,x) = Q_\tau(x,u) + \tau Q_1(x,u)$$

and, hence, the condition

$$Q_\tau(x,u) < 0 \quad (\|x\| + |u| \neq 0) \tag{20.71}$$

is sufficient to guarantee that $\dfrac{d}{dt}V(t,x) < 0$. Expanding the quadratic form $Q_\tau(x,u)$ up to its Hermitian form (20.67) (see by (7.6)) we get that the condition (20.71) is equivalent to the following

$$\begin{aligned} Q_\tau(z,u) &= Q_0(z,u) - \tau Q_1(z,u) \\ &= 2\mathrm{Re}\, z^* P(Az+bu) + q\,\mathrm{Re}\left[u^* c^\intercal (Az+bu)\right] \\ &\quad -\tau \mathrm{Re}\left[u^*\left(k^{-1}u - y\right)\right] < 0 \end{aligned}$$

$$(\|z\| + |u| \neq 0)$$

Let now z and u be connected in such a way that (after the application of the Laplace transformation (17.73))

$$i\omega z = Az + bu$$

where ω is a real number for which $\det[A - i\omega I] \neq 0$. Then

$$\mathrm{Re} z^* P(Az+bu) = \mathrm{Re} i\omega (z^* P z) = 0$$

and

$$Q_\tau(z,u) = q\,\mathrm{Re}\left[i\omega u^* y\right] - \tau \mathrm{Re}\, u^*\left(k^{-1}u - y\right)$$

By (20.59), we also have that

$$y = H(i\omega)u$$

So, finally, we obtain

$$Q_\tau(z,u) = Q_\tau(z,u) = \mathrm{Re}\left[qi\omega H(i\omega) + \tau H(i\omega) - \tau k^{-1}\right]|u|^2 < 0$$

whereas $(|u| \neq 0)$. Since for $\omega = 0$ it follows that $\tau > 0$, dividing by τ and denoting $\tilde{q} = q/\tau$ we get (20.69).

(b) *Necessity.* It follows directly from the properties of S-procedure (see Theorem 12.3), if we take into account that it gives necessary and sufficient conditions of the "equivalency" of the sets defined by the inequalities $Q_0(x,u) < 0$ under the constraints $Q_1(x,u) < 0$ and $Q_\tau(x,u) < 0$ $(\|x\| + |u| \neq 0)$. Theorem is proven. □

Remark 20.8. *Theorem 20.13 still* ***does not guarantee*** *the global asymptotic stability for each admissible nonlinear feedback, since we have still not proved the validity of properties (a) and (b) of the Barbashin–Krasovskii theorem 20.7 for the Lyapunov function (20.62).*

20.5.4 Popov's sufficient conditions

The next theorem gives such additional conditions.

Theorem 20.14. (Sufficient conditions (Popov 1961)) *Let*

(a) in (20.58) the matrix A is stable (Hurwitz);
(b) for some $\tilde{q}$ (not obligatory nonnegative) and for all $\omega \in [-\infty, \infty]$ Popov's frequency condition (20.69) holds.
Then the system (20.58), (20.61) is ***absolutely stable*** *in class $\mathcal{F}$.*

Proof. To complete the proof we need to prove the validity of conditions (a) and (b). Let (20.69) hold. For $u = 0$ the inequality (20.71) implies

$$Q_\tau(x, u) = 2x^\top PAx = x^\top (PA + A^\top P)\, x < 0$$

or, equivalently,

$$PA + A^\top P < 0$$

So, by the Lyapunov Lemma 9.1, it follows that $P > 0$.

(a) For $q = \tilde{q}r \geq 0$, by the condition (20.61) we have $y\varphi(y) \geq 0$ for any $y \in \mathbb{R}$, and, therefore, $\int_{y=0}^{y} \varphi(y)\, dy \geq 0$. Hence we derive that $V(0) = 0$, $V(x) > 0$ for any $x \neq 0$ and $V(x) \to \infty$ whereas $\|x\| \to \infty$.

(b) Let $q = \tilde{q}r < 0$. Taking $u = \mu y$ $(0 \leq \mu \leq k)$ we get $\dfrac{d}{dt}V(t, x) < 0$ for the system (20.58), (20.68) with $u = \mu y$. Then, the matrix $A_\mu := A + \mu bc^\top$ of the corresponding closed-loop system has no eigenvalues on the imaginary axis since A_μ is Hurwitz for any $\mu : 0 \leq \mu \leq k$. For such a system the direct substitution shows that the Lyapunov function (20.62) is

$$V_\mu(x) = x^\top \left(P + \frac{q\mu}{2} cc^\top\right) x$$

and, since $\dfrac{d}{dt}V(t, x) < 0$, it follows (by the same Lyapunov Lemma 9.1) that

$$P + \frac{q\mu}{2} cc^\top > 0$$

which gives $k \neq \infty$. For $\mu = k$ we get

$$V_k(x) = x^\top \left(P + \frac{qk}{2} cc^\top\right) x > 0 \quad \text{for} \quad x \neq 0$$

Let now $\varphi \in \mathcal{F}$. In this case the form $V(x)$ (20.62) can be represented as

$$V(x) = x^{\mathsf{T}} P x + q \int_{y=0}^{y=c^{\mathsf{T}}x} \varphi(y)\, dy$$
$$= x^{\mathsf{T}} \left(P + \frac{qk}{2} cc^{\mathsf{T}} \right) x - q \int_{y=0}^{y=c^{\mathsf{T}}x} [ky - \varphi(y)]\, dy$$

So, $V(0) = 0$, $V(x) > 0$ for any $x \neq 0$ and $V(x) \to \infty$ whereas $\|x\| \to \infty$ completes the proof. □

Remark 20.9. *As it is mentioned in Gelig, Leonov & Yakubovich (1978), Popov's frequency condition (20.69) guarantees the absolute stability for the system (20.58), (20.61) with a much wider class of nonlinear feedback, namely, for all nolinearities that satisfy the condition*

$$\boxed{\int_{t=t_0}^{t_n} \left[u(t) \left(y(t) - k^{-1} u(t) \right) + q u(t) \dot{y}(t) \right] dt \geq -\gamma > -\infty} \tag{20.72}$$

for some $\{t_n\}, t_n \to \infty$. *It may include also unstable linear systems (where A is not obligatory stable).*

Remark 20.10. *For* $\tilde{q} = 0$ *in (20.69) the corresponding Popov's frequency condition is called "the circle criterion" which is valid also for a much wider class of nolinearities including multi-valued functions such as hysteresis elements.*

20.5.5 Geometric interpretation of Popov's conditions

Let us represent the transfer function $H(i\omega)$ as

$$\boxed{\begin{array}{l} H(i\omega) = U(\omega) + iV(\omega) \\ U(\omega) := \operatorname{Re} H(i\omega),\ V(\omega) := \operatorname{Im} H(i\omega) \end{array}} \tag{20.73}$$

Then Popov's frequency condition (20.69) can be represented as follows:

$$\boxed{\tilde{q}\omega V(\omega) > U(\omega) - k^{-1}} \tag{20.74}$$

Definition 20.10. *The line*

$$\boxed{\tilde{q}\omega V(\omega) = U(\omega) - k^{-1}} \tag{20.75}$$

in the plane $(U(\omega), \omega V(\omega))$ *is called* ***Popov's line****, with* $\tan\psi = \tilde{q}$ *(see Fig. 20.6).*

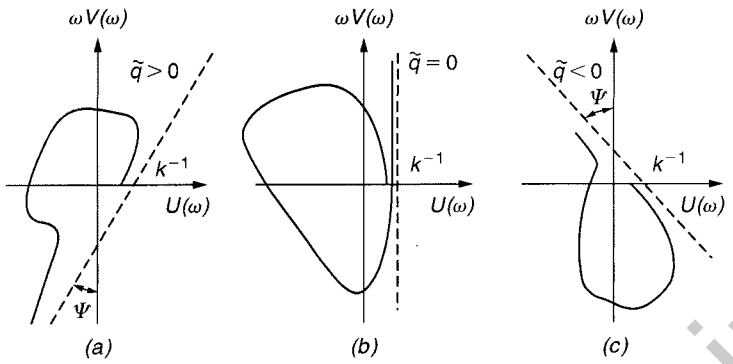

Fig. 20.6. The geometric interpretation of Popov's "criterion".

Claim 20.3. *Popov's frequency condition (20.69) is fulfilled if there exists a real number* $\tilde{q}$ *such that the "modified godograph"* $(U(\omega), \omega V(\omega))$ *of the transfer function* $H(i\omega)$ ***lies "above"*** *Popov's line (see versions (a)–(c) of Fig. 20.6) for all* $\omega \in [0, \infty]$. *If such a line does not exist (it cannot be drawn), then such a system cannot be referred to as absolutely stable.*

Example 20.8.

$$H(i\omega) = \frac{1 - i\omega}{2 + i\omega} = \frac{2 - \omega^2}{4 + \omega^2} - i\frac{3\omega}{4 + \omega^2}$$

$$U(\omega) = \frac{2 - \omega^2}{4 + \omega^2}, \quad \omega V(\omega) = -\frac{3\omega^2}{4 + \omega^2}$$

The corresponding godograph is depicted at Fig. 20.7. The Popov's line can be drawn with $\tilde{q} < 0$ *crossing the point* $(k^{-1}, 0)$ *for any* $k > 0.5$.

20.5.6 Yakubovich–Kalman lemma

As mentioned above, Popov's frequency condition (20.69) can be generalized for a significantly wide class of dynamic systems. All of them are based on the verification of negativity (positivity) of some Hermitian (quadratic) forms obtained as a time-derivative of a Lyapunov function (usually of the form (20.62)) on the trajectories of a linear system resembling (20.58) and (20.61). This verification can be done using the, so-called, "Yakubovich–Kalman lemma" known also as the "frequency theorem". Its simplified version oriented to the systems governed by (20.58) and (20.61) is given below.

Lemma 20.1. (Yakubovich 1973) *Let the pair* (A, b) *in (20.58) be controllable (see Criterion 3 in Theorem (9.8)) and*

$$\boxed{\Phi(s) = (sI - A)^{-1}} \tag{20.76}$$

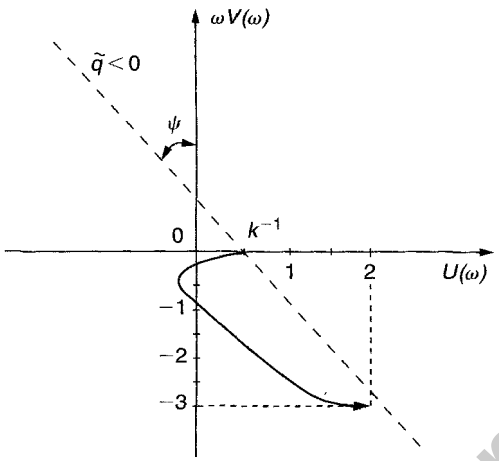

Fig. 20.7. Analysis of the admissible zone for the nonlinar feedback.

Let

$$G(x,u) := x^*Gx + 2\mathrm{Re}\,(x^*gu) - |\gamma|^2 u^*u$$

be a Hermitian form of (x,u) *where* $x \in \mathbb{C}^n$ *and* $u \in \mathbb{C}$.

1. *If*

$$\boxed{G[\Phi(i\omega)bu, u] \leq 0} \tag{20.77}$$

for all $u \in \mathbb{C}$ *and all* $\omega \in [-\infty, \infty]$, *then there exist a Hermitian matrix* $P = P^* \in \mathbb{C}^{n\times n}$, *a vector* $h \in \mathbb{C}^n$ *and a constant* $\gamma \in \mathbb{C}$ *such that the following identity holds*

$$\boxed{2\mathrm{Re}\, x^*P(Ax + bu) + G(x,u) = -|h^*x - \gamma u|^2} \tag{20.78}$$

where P, h *and* γ *satisfy the equations (the "resolving Lourie's equations")*

$$\boxed{\left.\begin{array}{l} PA + A^\top P + hh^* + G = 0 \\ Pb - h\gamma + g = 0 \end{array}\right\}} \tag{20.79}$$

(If A, b *and* G *are real, then* $P = P^\top$, h *and* γ *can be also real.)*

2. *If*

$$\boxed{G[\Phi(i\omega)bu, u] < 0} \tag{20.80}$$

for all $\omega \in [-\infty, \infty]$ *and all* $u \neq 0$, *then there exists a Hermitian matrix* $P = P^* \in \mathbb{C}^{n\times n}$ *such that*

$$\boxed{2\mathrm{Re}\, x^*P(Ax + bu) + G(x,u) < 0} \tag{20.81}$$

whenever $\|x\| + |u| \neq 0$.

3. *If for all* $u \in \mathbb{C}$

$$\boxed{G(0, u) \equiv -|\gamma|^2 |u|^2, \ \gamma \neq 0} \tag{20.82}$$

then

$$\boxed{\begin{gathered}\det\left(sI - \left(A - b\bar{k}^*\right)\right) = |\gamma|^{-1} g_h(s) \\ \bar{k} := |\gamma|^{-1} h\end{gathered}} \tag{20.83}$$

where the polynomial $g_h(s)$ *can be selected as stable (Hurwitz) by the corresponding selection of* h.

Proof. To prove (1) it is sufficient to compare the vector and matrix parameters in both parts of equation (20.78) which leads to (20.79). Indeed, the right-hand side of (20.78) is

$$\begin{aligned}-|h^*x - \gamma u|^2 &= x^*hh^*x - |\gamma|^2 u^*u + 2x^*(h\gamma)u \\ &= -x^*(hh^*)x - |\gamma|^2 u^*u + 2\mathrm{Re}\, x^*(h\gamma)u\end{aligned} \tag{20.84}$$

In turn, using the identity

$$2\mathrm{Re}\, x^*PAx = x^*(PA + A^\top P)x$$

the left-hand side of (20.78) can be represented as

$$\begin{aligned}&2\mathrm{Re}\, x^*P(Ax + bu) + G(x, u) = 2\mathrm{Re}\, x^*PAx \\ &\quad + 2\mathrm{Re}\, x^*Pbu + x^*Gx + 2\mathrm{Re}\,(x^*gu) - |\gamma|^2 u^*u \\ &\quad = x^*(PA + A^\top P + G)x + 2\mathrm{Re}\left(x^*[g + Pb]u\right) - |\gamma|^2 u^*u\end{aligned} \tag{20.85}$$

Comparing the coefficients in (20.84) and (20.85) we get

$$\begin{aligned}PA + A^\top P + G &= -hh^* \\ g + Pb &= h\gamma\end{aligned}$$

which coincides with (20.79).

To prove (2) let us rewrite (20.81) as follows

$$G[\Phi(i\omega)bu, u] = -\Pi(i\omega)|u|^2$$

where $\Pi(i\omega)$ is continuous and satisfies

$$\Pi(i\omega) \geq \Pi_0 > 0$$

Introduce also the regularized Hermitian form $G_\varepsilon(x, u)$ defined by

$$G_\varepsilon(x, u) = G(x, u) + \varepsilon\left(\|x\|^2 + |u|^2\right), \quad \varepsilon > 0$$

Then, evidently,

$$G_\varepsilon(x,u) \le -\Pi_0 |u|^2 + \varepsilon(1+c)\left(\|x\|^2 + |u|^2\right) \le 0$$

where the constant c satisfies the inequality $\|\Phi(i\omega)b\|^2 \le c$. But, by property (1) it follows that there exists $P = P^*$ such that

$$2\mathrm{Re}\, x^* P(Ax + bu) + G_\varepsilon(x,u) = -|h^*x - \gamma u|^2 \le 0$$

which proves (20.81). Property (3) follows directly from the assumptions of the theorem. □

Remark 20.11. *This lemma is also valid when $b = B$ is a matrix. This generalization can be found in Vidyasagar (1993).*

Corollary 20.5. *The matrix algebraic Riccati equation*

$$\boxed{\begin{array}{c} PA + A^\top P - K^\top P K + Q = 0 \\ RK = B^\top P \end{array}} \tag{20.86}$$

has a unique positive definite solution $P = P^\top > 0$ and the corresponding K such that the matrix $[A - BK]$ is stable, if $R > 0$, the pair (A, B) is controllable and one of two conditions is fulfilled:

(a)

$$\boxed{Q > 0}$$

(b)

$$\boxed{\begin{array}{c} Q = C^\top C \\ \text{the pair } (C, A) \text{ is observable} \end{array}}$$

Proof. The existence of P and K satisfying (20.86) is equivalent to fulfilling the identity

$$\begin{aligned} &2\mathrm{Re}\left[x^*(-P)(Ax + By)\right] - (x^*Qx + y^*Ry) \\ &\quad = -(y + Kx)^* R (y + Kx) \end{aligned}$$

valid for all complex x and y. By the Yakubovich–Kalman Lemma 20.1 to have this identity it is sufficient that

$$G\left[\Phi(i\omega)by, y\right] = -(x^*Qx + y^*Ry) \le 0$$

Since $R > 0$ this property holds. Moreover, the strict condition (20.80) also holds and

$$G(0, y) = -y^*Ry < 0$$

if $y \neq 0$. Hence, from Lemma 20.1 it follows also that there exists a unique solution P, K for which $\bar{A} := [A - BK]$ is stable. Since (20.86) is equivalent to

$$P\bar{A} + \bar{A}^{\top}P = -K^{\top}RK - Q = \bar{Q}, \quad K = R^{-1}B^{\top}P$$

then for $Q > 0$ we have $\bar{Q} > 0$ and, by the Lyapunov Lemma 9.1, there exists $P > 0$ resolving the last matrix equation. If $Q = C^{\top}C$, the existence of the positive definite solution also follows from Lemma 9.1 (statement (2)). □

Remark 20.12. *It seems to be useful to compare the statement of Corollary 20.5 with Theorem 10.7 which gives the same conditions for the existence of a strictly positive solution of the matrix Riccati equation making the closed-loop system stable.*

21 Finite-Dimensional Optimization

Contents

This chapter deals with the simplest problems of optimization in finite-dimensional spaces starting with unconstrained optimization of smooth convex functions and proceeds to investigate the influence of different complicating factors such as nonsmoothness, singularity of a minimum point and constraints of equality and inequality types. Each class of problems is analyzed in a similar way: first, the necessary conditions of extremality are derived, then sufficient conditions of an extremum are proved, followed by the results concerning existence, uniqueness and the stability of a solution. Finally some numerical methods (with their analysis) are presented. The selected method of the presentation follows (Polyak 1987).

21.1 Some properties of smooth functions

21.1.1 Differentiability remainder

Definition 21.1. *The function* $f : \mathbb{R}^n \to \mathbb{R}$ *is said to be*

1. ***differentiable*** *at a point* $x \in \mathbb{R}^n$ *if there exists a vector* $a \in \mathbb{R}^n$ *such that for all* $y \in \mathbb{R}^n$

$$\boxed{\begin{gathered} f(x+y) = f(x) + (a, y) + o(\|y\|) \\ a = \nabla f(x) = \left[\frac{\partial}{\partial x_j} f(x)\right]_{i=1,\ldots,n} \end{gathered}} \tag{21.1}$$

(the vector a *is usually called the* ***gradient*** *of* $f(x)$ *at the point* $x \in \mathbb{R}^n$*);*

2. ***twice differentiable*** *at a point* $x \in \mathbb{R}^n$ *if there exists a symmetric matrix* $H \in \mathbb{R}^{n\times n}$ *such that for all* $y \in \mathbb{R}^n$

$$\boxed{\begin{gathered} f(x+y) = f(x) + (\nabla f(x), y) + (Hy, y) + o\left(\|y\|^2\right) \\ H = \nabla^2 f(x) = \left[\frac{\partial^2}{\partial x_i \partial x_j} f(x)\right]_{i,j=1,\ldots,n} \end{gathered}} \tag{21.2}$$

(the matrix H *is called the* ***matrix of second derivatives of Hessian*** *of* $f(x)$ *at the point* $x \in \mathbb{R}^n$*).*

Lemma 21.1. (on a finite increment)

1. If $f(x)$ is differentiable on $[x, x+y]$, then

$$\boxed{f(x+y) = f(x) + (\nabla f(x), y) + \int_{\tau=0}^{1} (\nabla f(x+\tau y) - \nabla f(x), y)\, d\tau} \tag{21.3}$$

2. If $f(x)$ is twice differentiable on $[x, x+y]$, then

$$\boxed{f(x+y) = f(x) + (\nabla f(x), y) + \frac{1}{2}\left(\nabla^2 f(x)\, y, y\right) + \int_{t=0}^{1}\int_{\tau=0}^{t} \left(\left[\nabla^2 f(x+\tau y) - \nabla^2 f(x)\right] y, y\right) d\tau\, dt} \tag{21.4}$$

Proof. For any $x, y \in \mathbb{R}^n$ define the function

$$\phi(\tau) := f(x+\tau y) \tag{21.5}$$

which is, obviously, differentiable (twice differentiable) if $f(x)$ is differentiable (twice differentiable). The identity (21.3) follows from the Newton–Leibniz formula

$$\phi(1) = \phi(0) + \int_{\tau=0}^{1} \phi'(\tau)\, d\tau \tag{21.6}$$

and (21.4) results from the Taylor formula

$$\phi(1) = \phi(0) + \phi'(0) + \int_{t=0}^{1}\int_{\tau=0}^{t} \phi''(\tau)\, d\tau\, dt$$

Lemma is proven. □

Corollary 21.1.

(a) If $\nabla f(x)$ satisfies the Lipschitz condition on $[x, x+y]$, that is,

$$\|\nabla f(u) - \nabla f(v)\| \le L_f \|u - v\|, u, v \in [x, x+y] \tag{21.7}$$

then for all $x, y \in \mathbb{R}^n$

$$|f(x+y) - f(x) - (\nabla f(x), y)| \le \frac{L_f}{2}\|y\|^2 \tag{21.8}$$

(b) If for all $x, y \in \mathbb{R}^n$

$$\left\|\nabla^2 f(x+\tau y)\right\| \leq L_{\nabla^2}, \quad \tau \in [0,1] \tag{21.9}$$

then for all $x, y \in \mathbb{R}^n$

$$|f(x+y) - f(x) - (\nabla f(x), y)| \leq \frac{L_{\nabla^2}}{2}\|y\|^2 \tag{21.10}$$

(c) If for all $x, y \in \mathbb{R}^n$

$$\left\|\nabla^2 f(x+y) - \nabla^2 f(x)\right\| \leq L_{\nabla^2}\|y\| \tag{21.11}$$

then for all $x, y \in \mathbb{R}^n$

$$\left|f(x+y) - f(x) - (\nabla f(x), y) - \frac{1}{2}\left(\nabla^2 f(x)\, y, y\right)\right| \leq \frac{L_{\nabla^2}}{6}\|y\|^3 \tag{21.12}$$

Proof. The inequality (21.8) follows directly from (21.3), (21.7) and (21.11) if we take into account that

$$\left|\int_{\tau=0}^{1} (\nabla f(x+\tau y) - \nabla f(x), y)\, d\tau\right| \leq \int_{\tau=0}^{1} |(\nabla f(x+\tau y) - \nabla f(x), y)|\, d\tau$$
$$\leq \int_{\tau=0}^{1} \|\nabla f(x+\tau y) - \nabla f(x)\|\,\|y\|\, d\tau \leq \int_{\tau=0}^{1} L_f \tau \|y\|^2\, d\tau \leq \frac{L_f}{2}\|y\|^2$$

The inequalities (21.10) and (21.12) result from (21.4) and (21.9) since

$$\left|\int_{t=0}^{1}\int_{\tau=0}^{t} \left(\nabla^2 f(x+\tau y)\, y, y\right) d\tau\, dt\right| \leq \int_{t=0}^{1}\int_{\tau=0}^{t} \left|\left(\nabla^2 f(x+\tau y)\, y, y\right)\right| d\tau\, dt$$
$$\leq \int_{t=0}^{1}\int_{\tau=0}^{t} \left\|\nabla^2 f(x+\tau y)\right\|\|y\|^2\, d\tau\, dt \leq \int_{t=0}^{1}\int_{\tau=0}^{t} L_{\nabla^2}\|y\|^2\, d\tau\, dt \leq \frac{L_{\nabla^2}}{2}\|y\|^2$$

and

$$\left| \int_{t=0}^{1} \int_{\tau=0}^{t} \left(\left[\nabla^2 f(x+\tau y) - \nabla^2 f(x) \right] y, y \right) d\tau\, dt \right|$$

$$\leq \int_{t=0}^{1} \int_{\tau=0}^{t} \left| \left(\left[\nabla^2 f(x+\tau y) - \nabla^2 f(x) \right] y, y \right) \right| d\tau\, dt$$

$$\leq \int_{t=0}^{1} \int_{\tau=0}^{t} \left\| \nabla^2 f(x+\tau y) - \nabla^2 f(x) \right\| \|y\|^2 d\tau\, dt$$

$$\leq \int_{t=0}^{1} \int_{\tau=0}^{t} L_{\nabla^2} \|y\|^3 \tau d\tau\, dt \leq \frac{L_{\nabla^2}}{6} \|y\|^3$$

which proves the corollary. □

Exercise 21.1. *It is easy to check that*

1.

$$\boxed{\nabla \left\| [Ax-b]_+ \right\|^2 = 2A^{\mathsf{T}} [Ax-b]_+} \tag{21.13}$$

where

$$[z]_+ := \left([z_1]_+, \ldots, [z_n]_+ \right)$$
$$[z_i]_+ := \begin{cases} z_i & if \quad z_i \geq 0 \\ 0 & if \quad z_i < 0 \end{cases} \tag{21.14}$$

2. If $x \neq 0$

$$\boxed{\begin{aligned} \nabla \|x\| &= \frac{x}{\|x\|} \\ \nabla^2 \|x\| &= \frac{1}{\|x\|} I - \frac{x x^{\mathsf{T}}}{\|x\|^3} \end{aligned}} \tag{21.15}$$

3.

$$\boxed{\nabla^2 (c, x)^2 = 2cc^{\mathsf{T}}} \tag{21.16}$$

The following lemma will be useful in the considerations below.

Lemma 21.2. (Polyak 1987) *Let*

(a) $f(x)$ be differentiable on $\mathbb{R}^n$;
(b) $\nabla f(x)$ satisfy the Lipschitz condition (21.7) with the constant L_f;

(c) $f(x)$ *be bounded from below, i.e.,* $f(x) \geq f^* > -\infty$ *for all* $x \in \mathbb{R}^n$.

Then

$$\boxed{\|\nabla f(x)\|^2 \leq 2L_f (f(x) - f^*)} \tag{21.17}$$

Proof. Putting in (21.8) $y := -L_f^{-1} \nabla f(x)$ we obtain

$$\begin{aligned} f^* \leq f(x+y) &= f(x - L_f^{-1} \nabla f(x)) \\ &\leq f(x) - \left(\nabla f(x), L_f^{-1} \nabla f(x)\right) + \frac{L_f}{2} \left\|L^{-1} \nabla f(x)\right\|^2 \\ &= f(x) - \frac{1}{2L_f} \left\|L_f^{-1} \nabla f(x)\right\|^2 \end{aligned}$$

which implies (21.17). Lemma is proven. □

21.1.2 Convex functions

21.1.2.1 Main definition

Definition 21.2. *A scalar valued function* $f(x)$ *defined on* $\mathbb{R}^n$ *is said to be*

1. ***convex*** *(see Fig. 21.1) if for any* $x, y \in \mathbb{R}^n$ *and any* $\alpha \in [0, 1]$

$$\boxed{f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y)} \tag{21.18}$$

2. ***strictly convex*** *if for any* $x \neq y \in \mathbb{R}^n$ *and any* $\alpha \in [0, 1]$

$$\boxed{f(\alpha x + (1-\alpha) y) < \alpha f(x) + (1-\alpha) f(y)} \tag{21.19}$$

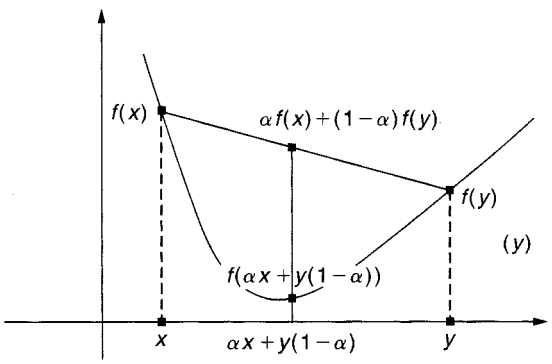

Fig. 21.1. A convex function.

3. ***strongly convex with the constant*** $l > 0$ *if for any* $x, y \in \mathbb{R}^n$ *and any* $\alpha \in [0, 1]$

$$\boxed{\begin{array}{c} f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y) \\ - l\alpha (1-\alpha) \|x - y\|^2 \end{array}} \tag{21.20}$$

4. ***concave*** *(strictly, strongly with the given constant) if the function* $[-f(x)]$ *is convex (strictly, strongly with the same constant).*

21.1.2.2 Some properties of convex (not obligatory differentiable) functions

Claim 21.1. *From Definition 21.2 it follows directly that*

(a) the affine function $f(x) = (a, x) + b$ *is both convex and concave;*
(b) if the functions $f_i(x)$ *are convex (concave) then the functions*

$$f(x) = \sum_{i=1}^{k} \gamma_i f_i(x), \quad \gamma_i \geq 0 \ (i = 1, \ldots, k)$$

and

$$f(x) = \max_{i=1,\ldots,k} f_i(x)$$

are convex (concave) too.

Claim 21.2. *If* $f(x)$ *is convex on* $\mathbb{R}^n$, *then for any* $x^{(1)}, \ldots, x^{(k)} \in \mathbb{R}^n$ *and any* $\alpha_1, \ldots, \alpha_k$ *such that* $\alpha_i \geq 0$ $(i = 1, \ldots, k)$, $\sum_{i=1}^{k} \alpha_i = 1$ *the following inequality holds*

$$f\left(\sum_{i=1}^{k} \alpha_i x^{(i)}\right) \leq \sum_{i=1}^{k} \alpha_i f\left(x^{(i)}\right) \tag{21.21}$$

Proof. It follows directly from the Jensen inequality (16.152). □

Lemma 21.3. *Any convex function is continuous.*

Proof. Let us prove this result by contradiction. Suppose that there exists a point x where a convex function $f(x)$ is discontinuous. This means that in any δ-neighborhood of x, containing x as an internal point, there are always two points x' and x'' such that $|f(x') - f(\alpha x'' + (1-\alpha) x')| > \varepsilon > 0$ for any $0 < \alpha^- < \alpha < \alpha^+ < 1$. Assuming that $f(x') \geq f(x'')$, we have $f(x') + \varepsilon > f(x'')$. But, by the convexity property, we also have

$$f\left(\alpha x'' + (1-\alpha) x'\right) \leq \alpha f\left(x''\right) + (1-\alpha) f\left(x'\right)$$

for any $\alpha \in [0, 1]$, or equivalently,

$$f(x') + \varepsilon > f(x'') \geq f(x') + \frac{1}{\alpha}\left[f(\alpha x'' + (1-\alpha) x') - f(x')\right] \geq f(x') + \frac{\varepsilon}{\alpha}$$

which, for α satisfying $0 < \alpha^- < \alpha < \alpha^+ < 1$, implies contradiction. □

Corollary 21.2. *If $f(x)$ is convex, then the set*

$$Q_\alpha := \{x \in R^n \mid f(x) \leq \alpha\} \tag{21.22}$$

is convex and closed.

Lemma 21.4. *Any convex function $f(x)$ at any arbitrary point has a* ***one side derivative*** *in any direction y and this derivative is uniformly bounded with respect to this direction.*

Proof. One has

$$\begin{aligned} f'_+(x, y) &:= \lim_{\alpha \to +0} \frac{f(x + \alpha y) - f(x)}{\alpha} \\ &= \lim_{\alpha \to +0} \frac{f((1-\alpha)x + \alpha(x+y)) - f(x)}{\alpha} \\ &\leq \lim_{\alpha \to +0} \frac{\alpha f(x+y) + (1-\alpha) f(x) - f(x)}{\alpha} \\ &\leq f(x+y) - f(x) \leq \max_{z:\|z\|=\|y\|} [f(x+z) - f(x)] \end{aligned} \tag{21.23}$$

which completes the proof. □

Corollary 21.3. (Rademacher theorem) *Any convex function is differentiable almost everywhere (excepting a set of measure zero).*

21.1.2.3 Some properties of convex differentiable functions

Lemma 21.5. *If a function $f(x)$ on $\mathbb{R}^n$ is differentiable, then*

1. *its* ***convexity*** *is equivalent to the inequality*

$$\boxed{f(x+z) \geq f(x) + (\nabla f(x), z)} \tag{21.24}$$

   valid for all $x, y \in \mathbb{R}^n$;
2. *its* ***strict convexity*** *is equivalent to the inequality*

$$\boxed{f(x+z) > f(x) + (\nabla f(x), z)} \tag{21.25}$$

   valid for all $x \in \mathbb{R}^n$ and all $z \neq 0$ ($z \in \mathbb{R}^n$);
3. *its* ***strong convexity*** *is equivalent to the inequality*

$$\boxed{f(x+z) \geq f(x) + (\nabla f(x), z) + \frac{l}{2}\|z\|^2} \tag{21.26}$$

   valid for all $x, z \in \mathbb{R}^n$.

Proof.

(a) *Necessity*. Suppose that (21.18), (21.19) and (21.20) hold. Show that (21.24), (21.25) and (21.26) result from them. Taking there $1-\alpha=\delta\to 0$ and substituting $y=z+x$ we obtain the result. Indeed, from (21.18) we have

$$\frac{1}{\delta}\left[f\left(x+\delta\left(y-x\right)\right)-f\left(x\right)\right]\leq f\left(y\right)-f\left(x\right)$$

and, when $\delta\to 0$ in the left-hand side of this inequality, it follows that

$$\left(\nabla f\left(x\right),\left(y-x\right)\right)\leq f\left(y\right)-f\left(x\right)$$

which leads to (21.24) if we take $y=z+x$. The inequalities (21.19) and (21.20) are derived analogously.

(b) *Sufficiency*. Suppose that (21.24) holds. Define the function

$$g_{\alpha}\left(x,y\right):=\alpha f\left(x\right)+\left(1-\alpha\right)f\left(y\right)-f\left(\alpha x+\left(1-\alpha\right)y\right)$$

To prove (21.18) we need to prove that for all $x,z\in\mathbb{R}^n$ and all $\alpha\in[0,1]$

$$g_{\alpha}\left(x,y\right)\geq 0 \tag{21.27}$$

First, notice that all $x,z\in\mathbb{R}^n$

$$g_{\alpha=0}\left(x,y\right)=g_{\alpha=1}\left(x,y\right)=0 \tag{21.28}$$

All stationary points $\alpha^*\in[0,1]$ (if they exist) of the function $g_{\alpha}\left(x,y\right)$ satisfy the identity

$$g'_{\alpha=\alpha^*}\left(x,y\right)=f\left(x\right)-f\left(y\right)-\left(\nabla f\left(\alpha^*x+\left(1-\alpha^*\right)y\right),x-y\right)=0$$

or, equivalently,

$$f\left(x\right)-f\left(y\right)=\left(\nabla f\left(\alpha^*x+\left(1-\alpha^*\right)y\right),x-y\right)$$

For any stationary point α^*, after the application of the inequality (21.24) and in view of the last identity, we have

$$\begin{aligned}g_{\alpha=\alpha^*}\left(x,y\right)&:=\alpha^*f\left(x\right)+\left(1-\alpha^*\right)f\left(y\right)-f\left(\alpha^*x+\left(1-\alpha^*\right)y\right)\\&=\alpha^*\left[f\left(x\right)-f\left(y\right)\right]+f\left(y\right)-f\left(\alpha^*x+\left(1-\alpha^*\right)y\right)\\&\geq\alpha^*\left[f\left(x\right)-f\left(y\right)\right]-\left(\nabla f\left(\alpha^*x+\left(1-\alpha^*\right)y\right),\alpha^*x-y\right)=0\end{aligned} \tag{21.29}$$

which, together with (21.28), implies (21.27). Indeed, if we assume that there exists a point α' such that $g_{\alpha=\alpha'}\left(x,y\right)<0$, then, by the continuity and taking into account (21.28), it follows that there should be a minimum point α^* where also $g_{\alpha=\alpha^*}\left(x,y\right)<0$ which contradicts with (21.29). So, for all $\alpha\in[0,1]$ it follows that $g_{\alpha}\left(x,y\right)\geq 0$. The validity of (21.19) and (21.20) may be proven by the same manner. Lemma is proven. □

Corollary 21.4. *If a convex function $f(x)$ is differentiable on $\mathbb{R}^n$ then for any $x, y \in \mathbb{R}^n$*

$$\boxed{(\nabla f(x), (y - x)) \leq f(y) - f(x)} \tag{21.30}$$

and

$$\boxed{(\nabla f(x) - \nabla f(y), (x - y)) \geq 0} \tag{21.31}$$

which means that ***the gradient of a convex function is a monotone operator.***

Proof. The inequality (21.30) is proven just above. Changing y to x and x to y we also have

$$(\nabla f(y), (x - y)) \leq f(x) - f(y)$$

Adding this inequality with (21.30) we obtain (21.31). □

Corollary 21.5. *If a function $f(x)$ is twice differentiable on $\mathbb{R}^n$, then*

1. *its* ***convexity*** *is equivalent to the matrix inequality*

$$\boxed{\nabla^2 f(x) \geq 0} \tag{21.32}$$

   valid for all $x \in \mathbb{R}^n$;
2. *the matrix inequality*

$$\boxed{\nabla^2 f(x) > 0} \tag{21.33}$$

   valid for all $x \in \mathbb{R}^n$ implies its ***strict convexity****;*
3. *its* ***strong convexity*** *is equivalent to the matrix inequality*

$$\boxed{lI \leq \nabla^2 f(x) \leq L_f I} \tag{21.34}$$

   valid for all $x \in \mathbb{R}^n$;
4. *if x^* is an optimal (minimal) point of strongly convex function $f(x)$ (with a constant $l > 0$), then (taking $x = x^*$, $y = x$ and $\nabla f(x^*) = 0$) the inequalities above lead to the following ones:*

$$\boxed{f(x) \geq f(x^*) + \frac{l}{2}\|x - x^*\|^2} \tag{21.35}$$

$$\boxed{(\nabla f(x), (x - x^*)) \geq l\|x - x^*\|^2} \tag{21.36}$$

$$\boxed{\|\nabla f(x)\| \geq l\|x - x^*\|} \tag{21.37}$$

Example 21.1.

1. *The function* $f(x) = x^2$ *is strongly convex (and, hence, convex and strictly convex) with* $l = 2$.
2. *The functions* $f(x) = x^4$ *and* $f(x) = e^x$ *are strictly convex (and, hence, convex, but not strongly convex).*

The next two lemmas will be used hereinafter.

Lemma 21.6. (Polyak 1987) *Let*

(a) $f(x)$ *be convex and twice differentiable on* $\mathbb{R}^n$;
(b) $\nabla f(x)$ *satisfies the Lipschitz condition (21.7) with the constant* L_f.

Then for all $x, y \in \mathbb{R}^n$

$$\boxed{(\nabla f(x) - \nabla f(y), x - y) \geq L_f^{-1} \|\nabla f(x) - \nabla f(y)\|^2} \tag{21.38}$$

Proof. By (21.6) we have

$$\begin{aligned}\nabla f(y) &= \nabla f(x) + \int_{\tau=0}^{1} \frac{d}{d\tau} \nabla f(x + \tau(y - x))\, d\tau \\ &= \nabla f(x) + \int_{\tau=0}^{1} \nabla^2 f(x + \tau(y - x))(y - x)\, d\tau = A(y - x)\end{aligned}$$

where, by the strict convexity condition and the property (21.32),

$$A = A^{\mathsf{T}} := \int_{\tau=0}^{1} \nabla^2 f(x + \tau(y - x))\, d\tau \geq 0$$

which, implies (in view of the inequalities $\|A\| I \geq A$ and $\|A\| \leq L_f$)

$$\begin{aligned}(\nabla f(y) - \nabla f(x), (y - x)) &= (y - x)^{\mathsf{T}} A (y - x) \\ &= \frac{1}{\|A\|} (y - x)^{\mathsf{T}} A^{1/2} (\|A\| I) A^{1/2} (y - x) \\ &\geq \frac{1}{\|A\|} (y - x)^{\mathsf{T}} A^{1/2} A A^{1/2} (y - x) \\ &= \frac{1}{\|A\|} (y - x)^{\mathsf{T}} A^{\mathsf{T}} A (y - x) = \frac{1}{\|A\|} \|A(y - x)\|^2 \\ &\geq \frac{1}{L_f} \|A(y - x)\|^2 = \frac{1}{L_f} \|\nabla f(y) - \nabla f(x)\|^2\end{aligned}$$

Lemma is proven. □

Lemma 21.7. *If*

1. *the function $f(x)$ is differentiable strongly convex on $\mathbb{R}^n$ (see (21.20)) with the constant $l > 0$,*
2. *x^* is its minimum point, then for all $x \in \mathbb{R}^n$*

$$\boxed{\|\nabla f(x)\|^2 \geq 2l\,[f(x) - f(x^*)]} \tag{21.39}$$

Proof. The inequality (21.26) can be rewritten as follows

$$\left\|\sqrt{\frac{1}{2l}}\nabla f(x)\right\|^2 \geq f(x) - f(x+z) + \left\|\sqrt{\frac{1}{2l}}\nabla f(x) + \sqrt{\frac{l}{2}}z\right\|^2$$

which, for $z := x^* - x$, leads to the following inequality

$$\frac{1}{2l}\|\nabla f(x)\|^2 \geq f(x) - f(x^*) + \left\|\sqrt{\frac{1}{2l}}\nabla f(x) + \sqrt{\frac{l}{2}}z\right\|^2 \geq f(x) - f(x^*)$$

which completes the proof. □

21.2 Unconstrained optimization

21.2.1 Extremum conditions

Definition 21.3.

- *The point x^* is called a **local minimum** of $f(x)$ on $\mathbb{R}^n$ if there exists $\delta > 0$ such that $f(x) \geq f(x^*)$ for all x satisfying $\|x - x^*\| \leq \delta$.*
- *The point x^* is called a **global minimum** (simply minimum) of the function $f(x)$ on $\mathbb{R}^n$ if $f(x) \geq f(x^*)$ for all $x \in \mathbb{R}^n$.*

21.2.1.1 Necessary conditions

Theorem 21.1. (on necessary conditions) *Let x^* be a **local** minimum of $f(x)$ on $\mathbb{R}^n$.*

1. *The first-order necessary condition (Fermat). If $f(x)$ is differentiable at x^*, then*

$$\boxed{\nabla f(x^*) = 0} \tag{21.40}$$

2. *The second-order necessary condition. If $f(x)$ is twice differentiable at x^*, then*

$$\boxed{\nabla^2 f(x^*) \geq 0} \tag{21.41}$$

Proof. To prove (1) suppose that $\nabla f(x^*) \neq 0$. Then $\tau > 0$ we have

$$\begin{aligned} f(x^* - \tau\nabla f(x^*)) &= f(x^*) + (\nabla f(x^*), -\tau\nabla f(x^*)) + o(\tau) \\ &= f(x^*) - \tau\|\nabla f(x^*)\|^2 + o(\tau) > f(x^*) \end{aligned}$$

for small enough τ which contradicts the fact that x^* is a minimum. So, (21.40) holds. To prove (2) let us use (21.2) which for any $y \in \mathbb{R}^n$ and a small positive $\tau > 0$ gives

$$\begin{aligned} f(x^*) &\le f(x^* + \tau y) \\ &= f(x^*) + (\nabla f(x^*), \tau y) + \left(\nabla^2 f(x^*)\, \tau y, \tau y\right) + o\left(\tau^2\right) \\ &= f(x^*) + \tau^2 \left(\nabla^2 f(x^*)\, y, y\right) + o\left(\tau^2\right) \end{aligned}$$

or, equivalently,

$$0 \le \tau^2 \left(\nabla^2 f(x^*)\, y, y\right) + o\left(\tau^2\right)$$

Dividing by τ^2 and tending τ to zero implies $\left(\nabla^2 f(x^*)\, y, y\right)$ which is equivalent to (21.41). Theorem is proven. □

21.2.1.2 Sufficient conditions

Theorem 21.2.

1. *The first-order sufficient condition. Let $f(x)$ be a convex on $\mathbb{R}^n$ function differentiable at a point x^* such that the first-order necessary condition (21.40) holds, that is, $\nabla f(x^*) = 0$. Then x^* is a global minimum point of $f(x)$ on $\mathbb{R}^n$.*
2. *The second-order sufficient condition. Let $f(x)$ be twice differentiable at a point x^* and*

$$\boxed{\nabla f(x^*) = 0, \quad \nabla^2 f(x^*) > 0} \tag{21.42}$$

Then x^* is a local minimum point.

Proof. The first-order sufficient condition follows directly from (21.24) since for any $z \in \mathbb{R}^n$

$$f(x^* + z) \ge f(x^*) + (\nabla f(x^*), z) = f(x^*)$$

The second-order sufficient condition follows from the Taylor formula (21.2) since

$$\begin{aligned} f(x^* + \tau z) = f(x^*) &+ (\nabla f(x^*), \tau z) \\ &+ \tau^2 \left(\nabla^2 f(x^*)\, y, y\right) + o\left(\tau^2 \|y\|^2\right) \ge f(x^*) \\ &+ \tau^2 \lambda_{\min} \left(\nabla^2 f(x^*)\right) \|y\|^2 + o\left(\tau^2 \|y\|^2\right) \ge f(x^*) \end{aligned}$$

for small enough $\tau > 0$ which proves the result. □

21.2.2 Existence, uniqueness and stability of a minimum

21.2.2.1 Existence of a minimum

Theorem 21.3. (Weierstrass) *If*
(a) $f(x)$ is continuous on $\mathbb{R}^n$,

(b) the set

$$\boxed{Q_\alpha := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\}} \tag{21.43}$$

is nonempty and bounded for some $\alpha \in \mathbb{R}$.
Then there exists a global minimum of $f(x)$ *on* $\mathbb{R}^n$.

Proof. For some vector-sequence $\{x^{(k)}\}$ we have

$$f\left(x^{(k)}\right) \to \inf_{x \in \mathbb{R}^n} f(x) < \alpha \quad \text{as} \quad k \to \infty$$

Then $x^{(k)} \in Q_\alpha$ for large enough k. But the set Q_α is a compact and, hence, the sequence $\{x^{(k)}\}$ has a limit $x^* \in Q_\alpha$. But from the continuity of $f(x)$ it follows that $f(x^*) = \inf_{x \in \mathbb{R}^n} f(x)$ which proves the theorem. □

21.2.2.2 Uniqueness of a minimum

Definition 21.4.

1. *A minimum point is called* ***locally unique*** *if there are no other minimum points in some neighborhood of this point.*
2. x^* *is said to be a* ***nonsingular minimum point*** *if the second-order sufficient conditions (21.42) hold, that is, if* $\nabla f(x^*) = 0$, $\nabla^2 f(x^*) > 0$.

Theorem 21.4. *A nonsingular minimum point is locally unique.*

Proof. Suppose that x^* is a nonsingular minimum point, but there exists another minimum point $x^{**} \neq x^*$ in any small neighborhood of x^*, that is, $f(x^*) = f(x^{**})$ when $\|x^{**} - x^*\| < \delta$ for any small enough δ. Then we have

$$\begin{aligned} f(x^{**}) &= f(x^* + (x^{**} - x^*)) = f(x^*) + (\nabla f(x^*), x^{**} - x^*) \\ &\quad + \left(\nabla^2 f(x^*)(x^{**} - x^*), (x^{**} - x^*)\right) + o\left(\|x^{**} - x^*\|^2\right) \\ &= f(x^*) + \left(\nabla^2 f(x^*)(x^{**} - x^*), (x^{**} - x^*)\right) + o\left(\|x^{**} - x^*\|^2\right) > f(x^*) \end{aligned}$$

since

$$\left(\nabla^2 f(x^*)(x^{**} - x^*), (x^{**} - x^*)\right) > o\left(\|x^{**} - x^*\|^2\right)$$

So, we have obtained the contradiction that x^* is a minimum point. Theorem is proven. □

Proposition 21.1. *A minimum point of a strictly convex function is globally unique.*

Proof. It follows directly from the definition (21.19). Indeed, putting in (21.19) $y := x^*$ we get (for $\alpha > 0$)

$$\begin{aligned} 0 &< \frac{1}{\alpha}\left[f(y + \alpha(x - y)) - f(y)\right] \\ &= \frac{1}{\alpha}\left[f(x^* + \alpha(x - x^*)) - f(x^*)\right] < f(x) - f(x^*) \end{aligned}$$

□

21.2.2.3 Stability of a minimum

Definition 21.5. *A local minimum point x^* of $f(x)$ is called*

(a) ***locally stable*** *if every local minimizing sequence converges to x^*, that is, there exists $\delta > 0$ such that*

$$f\left(x^{(k)}\right) \underset{k\to\infty}{\to} f(x^*), \left\|x^{(k)} - x^*\right\| \leq \delta$$

implies

$$x^{(k)} \underset{k\to\infty}{\to} x^*$$

(b) globally stable if any minimizing sequence converges to x^*.

Theorem 21.5. (Polyak 1987) *A local minimum point x^* of a continuous function $f(x)$ is locally stable if it is locally unique.*

Proof. Let $\left\{x^{(k)}\right\}$ be a local minimizing sequence. By the compactness of a unit sphere in $\mathbb{R}^n$, from any sequence there can be extracted a convergent subsequence, namely, there exists $\left\{x^{(k_i)}\right\}$ such that $x^{(k_i)} \to \bar{x}$. But from the definition of a local minimizing sequence one gets $\|\bar{x} - x^*\| \leq \delta$. By the continuity property, we have

$$f(\bar{x}) = \lim_{i\to\infty} f\left(x^{(k_i)}\right) = f(x^*)$$

which implies $\bar{x} = x^*$ since x^* is a local minimum point. The same is true for any other convergent subsequence, so, $x^{(k)} \to x^*$, and therefore x^* is locally stable. □

The next result turns out to be often useful in different applications.

Lemma 21.8. (on regularized (perturbed) functions) *The stability property implies that a minimum point of a nonperturbed functions is closed to a minimum point of a perturbed function, namely, if x^* is a* ***nonsingular minimum point*** *of $f(x)$ and $g(x)$ is* ***continuously differentiable*** *in a neighborhood of x^*, then, for sufficiently small $\varepsilon > 0$, the function $F_\varepsilon(x) := f(x) + \varepsilon g(x)$ has a local minimum point x_ε^* in a neighborhood of x^*, and*

$$\boxed{x_\varepsilon^* = x^* - \varepsilon\left[\nabla^2 f(x^*)\right]^{-1} \nabla g(x^*) + o(\varepsilon)} \tag{21.44}$$

Proof. By Definition 21.4 it follows that x_ε^* satisfies

$$\nabla F_\varepsilon\left(x_\varepsilon^*\right) = \nabla f\left(x_\varepsilon^*\right) + \varepsilon \nabla g\left(x_\varepsilon^*\right) = 0$$

and, hence, we have

$$0 = \nabla F_\varepsilon\left(x_\varepsilon^*\right) = \nabla f\left(x_\varepsilon^*\right) + \varepsilon \nabla g\left(x_\varepsilon^*\right)$$
$$= \nabla f\left(x^* + \left(x_\varepsilon^* - x^*\right)\right) + \varepsilon \nabla g\left(x_\varepsilon^*\right) = \nabla^2 f\left(x^*\right)\left(x_\varepsilon^* - x^*\right)$$
$$+ \varepsilon\left[\nabla g\left(x^*\right) + \left(\nabla g\left(x_\varepsilon^*\right) - \nabla g\left(x^*\right)\right)\right] + o\left(\left\|x_\varepsilon^* - x^*\right\|^2\right)$$

By the continuity property of $\nabla g(x)$ at the point $x = x^*$ it follows that for any $\tilde{\varepsilon} > 0$ there exists $\tilde{\delta} > 0$ such that $\|\tilde{x} - x^*\| \leq \tilde{\delta}$ implies $\|\nabla g(\tilde{x}) - \nabla g(x^*)\| < \tilde{\varepsilon}$. Taking here $\tilde{x} := x_\varepsilon^*$, $\tilde{\delta} := k\varepsilon$ (where, maybe, $k < 1$) and $\tilde{\varepsilon} := \varepsilon$, from the last identity it follows that

$$0 = \nabla^2 f\left(x^*\right)\left(x_\varepsilon^* - x^*\right) + \varepsilon \nabla g\left(x^*\right) + o(\varepsilon)$$

which implies (21.44). Lemma is proven. □

Remark 21.1. *When $g(x) \geq 0$ for all $x \in \mathbb{R}^n$ then $g(x)$ is usually called a* ***regularizing term*** *and the function $F_\varepsilon(x)$ is called the* ***regularized function***.

21.2.3 Some numerical procedure of optimization

Let us consider the following numerical procedure for finding a minimum point $x^* \in \mathbb{R}^n$ of the function $f(x)$ on $\mathbb{R}^n$ using only the value of its gradient $\nabla f(x_n)$ in a current point x_n:

$$\boxed{\begin{array}{c} x_{n+1} = x_n - \gamma_n H_{n+1} \nabla f\left(x_n\right) \\ x_0 = \mathring{x}, \quad n = 0, 1, 2, \ldots \\ 0 < \gamma_n \in \mathbb{R}, \quad 0 < H_n = H_n^{\mathsf{T}} \in \mathbb{R}^{n \times n} \end{array}} \tag{21.45}$$

21.2.3.1 Strong (argument) convergence

Theorem 21.6. (on strong (argument) convergence)
Assume that

1. *x^* is an optimal point of strongly convex differentiable function $f(x)$ with a constant $l > 0$ (such point always exists by Proposition (21.1));*
2. *$\nabla f(x)$ satisfies the Lipschitz condition (21.7) with the constant L_f;*
3. *for any $n = 0, 1, \ldots$*

$$\alpha_n := \lambda_{\max}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) - 2\gamma_n l \lambda_{\min}\left(H_n\right)$$
$$+ \gamma_n^2 L_f^2 \lambda_{\max}\left(H_{n+1}\right) \lambda_{\max}\left(H_n\right) \leq q < 1 \tag{21.46}$$

Then for the sequence $\{x_n\}$ generated by (21.45) with any initial conditions $\mathring{x}$ we have the following exponential convergence:

$$W_n := \left\|x_n - x^*\right\|_{H_n^{-1}}^2 = O\left(q^n\right) = O\left(\left[e^{\ln q}\right]^n\right) \to 0 \tag{21.47}$$

whereas $n \to \infty$.

Proof. We have the following recursion:

$$\begin{aligned} W_{n+1} &:= \|x_{n+1} - x^*\|^2_{H_{n+1}^{-1}} \\ &= \|(x_n - x^*) - \gamma_n H_{n+1} \nabla f(x_n)\|^2_{H_{n+1}^{-1}} = \|x_n - x^*\|^2_{H_{n+1}^{-1}} \\ &\quad - 2\left(\gamma_n H_{n+1}^{-1} H_{n+1} \nabla f(x_n), x_n - x^*\right) + \|\gamma_n H_{n+1} \nabla f(x_n)\|^2_{H_{n+1}^{-1}} \\ &= \|x_n - x^*\|^2_{H_n^{-1/2}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) H_1^{-1/2}} \\ &\quad - 2\gamma_n (\nabla f(x_n), x_n - x^*) + \gamma_n^2 \|\nabla f(x_n)\|^2_{H_{n+1}} \end{aligned} \tag{21.48}$$

Recall that

$$H_{n+1}^{-1} = H_n^{-1/2}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) H_n^{-1/2} \le \lambda_{\max}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) H_n^{-1}$$

By the condition (21.36),

$$(\nabla f(x_n), x_n - x^*) \ge l \|x_n - x^*\|^2$$

and by (21.7)

$$\|\nabla f(x_n) - \nabla f(x^*)\| = \|\nabla f(x_n)\| \le L_f \|x_n - x^*\|$$

We also have

$$\|x_n - x^*\|^2 = \left((x_n - x^*), H_n^{-1/2} H_n H_n^{-1/2} (x_n - x^*)\right)$$

$$\lambda_{\min}(H_n) W_n \le \|x_n - x^*\|^2 \le \lambda_{\max}(H_n) W_n$$

Substitution of these inequalities into (21.48) implies

$$\begin{aligned} W_{n+1} &\le \lambda_{\max}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) \|x_n - x^*\|^2_{H_n^{-1}} \\ &\quad - 2\gamma_n l \|x_n - x^*\|^2 + \gamma_n^2 L_f^2 \lambda_{\max}(H_{n+1}) \|x_n - x^*\|^2 \\ &\le \lambda_{\max}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) W_n - 2\gamma_n l \lambda_{\min}(H_n) W_n \\ &\quad + \gamma_n^2 L_f^2 \lambda_{\max}(H_{n+1}) \lambda_{\max}(H_n) W_n \\ &= \alpha_n W_n \le q W_n \le \cdots \le q^n W_0 \to 0 \end{aligned}$$

Theorem is proven. □

Corollary 21.6. (on the gradient method convergence) *If in (21.45) we take*

$$\gamma_n := \gamma, \quad H_n := I \tag{21.49}$$

we get the ***gradient method***

$$x_{n+1} = x_n - \gamma \nabla f(x_n), \quad x_0 = \mathring{x}, \quad n = 0, 1, 2, \ldots$$
$$0 < \gamma < 2l/L_f^2 \tag{21.50}$$

which converges exponentially as

$$W_n := \|x_n - x^*\|^2 = O(q^n)$$

with

$$q = 1 - \gamma\left(2l - \gamma L_f^2\right)$$

Corollary 21.7. (on the modified Newton's method) *If in (21.45) we take*

$$\gamma_n := \gamma, \quad H_n := \left[\nabla^2 f(x_{n-1})\right]^{-1} \tag{21.51}$$

we get the ***modified Newton's method***

$$x_{n+1} = x_n - \gamma\left[\nabla^2 f(x_n)\right]^{-1} \nabla f(x_n)$$
$$x_0 = \mathring{x}, \quad n = 0, 1, 2, \ldots \tag{21.52}$$
$$\gamma = l^3/L_f^3$$

which converges exponentially as

$$W_n := \|x_n - x^*\|_{H_n^{-1}}^2 = O(q^n)$$

with

$$\alpha_n := \lambda_{\max}\left(H_n^{1/2} H_{n+1}^{-1} H_n^{1/2}\right) - 2\gamma l \lambda_{\min}(H_n)$$
$$+ \gamma^2 L_f^2 \lambda_{\max}(H_{n+1})\lambda_{\max}(H_n)$$
$$\leq L_f/l - 2\gamma l/L_f + \gamma^2 L_f^2/l^2 = L_f/l - l^4/L_f^4 := q < 1$$

within the class of strongly convex functions satisfying

$$0.755 \leq l/L_f < 1$$

Proof. It follows from the estimates (in view of (21.34)) that

$$\lambda_{\max}(H_{n+1}) = \lambda_{\max}\left(\left[\nabla^2 f(x_n)\right]^{-1}\right) = \frac{1}{\lambda_{\min}\left(\left[\nabla^2 f(x_n)\right]\right)} \leq l^{-1}$$
$$\lambda_{\min}(H_n) = \frac{1}{\lambda_{\max}\left(\left[\nabla^2 f(x_n)\right]\right)} \geq 1/L_f \tag{21.53}$$

and

$$\lambda_{\max}\left(H_n^{1/2}H_{n+1}^{-1}H_n^{1/2}\right) \le \lambda_{\max}\left(H_{n+1}^{-1}\right)\lambda_{\max}\left(H_n\right)$$
$$= \lambda_{\max}\left(\nabla^2 f\left(x_n\right)\right)\lambda_{\max}\left(\left[\nabla^2 f\left(x_n\right)\right]^{-1}\right) \le L_f \frac{1}{\lambda_{\min}\left(\nabla^2 f\left(x_n\right)\right)} \le L_f/l$$

if we use them in (21.46). Moreover, defining $x := l/L_f < 1$, we get

$$q = L_f/l - l^4/L_f^4 = x^{-1} - x^4 < 1$$

if $x \ge 0.755$. □

Below we show that there exist other modifications of Newton's method working within much wider classes of functions.

21.2.3.2 Weak (functional) convergence

Theorem 21.7. (on weak (functional) convergence) *Assume that*

1. *the function $f(x)$ is differentiable and is bounded from below, i.e.,*

$$f(x) \ge f^* > -\infty \tag{21.54}$$

2. *$\nabla f(x)$ satisfies the Lipschitz condition (21.7) with the constant L_f;*
3. *for all $n = 0, 1, 2, \ldots$*

$$\lambda_{\max}\left(H_{n+1}\right) \le \lambda^+$$
$$\gamma_n \le \bar{\gamma} := \frac{2\left(1-\varkappa\right)}{L_f\lambda^+}, \quad \varkappa \in (0,1) \tag{21.55}$$
$$\gamma_n\lambda_{\min}\left(H_{n+1}\right) \ge c_- > 0$$

Then for the sequence $\{x_n\}$ generated by (21.45) with any initial conditions $\hat{x}$ we have the following property:

$$f\left(x_{n+1}\right) \le f\left(x_n\right)$$
$$\|\nabla f\left(x_n\right)\| \to 0 \quad as \quad n \to \infty \tag{21.56}$$

whereas $n \to \infty$.

Proof. By (21.3) and (21.53) we have

$$f\left(x_{n+1}\right) = f\left(x_n\right) - \gamma_n \|\nabla f\left(x_n\right)\|^2_{H_{n+1}}$$
$$- \gamma_n \int_{\tau=0}^{1} \left(\nabla f\left(x_n - \tau\gamma_n H_{n+1}\nabla f\left(x_n\right)\right) - \nabla f\left(x_n\right), H_{n+1}\nabla f\left(x_n\right)\right) d\tau$$
$$\le f\left(x_n\right) - \gamma_n \|\nabla f\left(x_n\right)\|^2_{H_{n+1}} + \gamma_n^2 \frac{L_f}{2} \|H_{n+1}\nabla f\left(x_n\right)\|^2$$
$$\le f\left(x_n\right) - \gamma_n\left(1 - \gamma_n\frac{L_f}{2}\lambda_{\max}\left(H_{n+1}\right)\right)\|\nabla f\left(x_n\right)\|^2_{H_{n+1}} \le f\left(x_n\right)$$
$$- \gamma_n\varkappa\|\nabla f\left(x_n\right)\|^2_{H_{n+1}} \le f\left(x_n\right) - \varkappa\gamma_n\lambda_{\min}\left(H_{n+1}\right)\|\nabla f\left(x_n\right)\|^2$$

or, equivalently,

$$c_- \|\nabla f(x_n)\|^2 \le \gamma_n \lambda_{\min}(H_{n+1}) \|\nabla f(x_n)\|^2 \le \varkappa^{-1}[f(x_n) - f(x_{n+1})] \tag{21.57}$$

We also have

$$f(x_{n+1}) \le f(x_n)$$

which, by the Weierstrass theorem 14.9 in view of the boundedness from below, implies the existence of the limit

$$\lim_{n\to\infty} f(x_n) = \bar{f} > -\infty$$

Summing the inequalities (21.57) on $n = 0, 1, .., T$ and taking $T \to \infty$, we get

$$\sum_{n=0}^{\infty} \|\nabla f(x_n)\|^2 \le (\varkappa c_-)^{-1}[f(x_0) - \bar{f}] < \infty$$

which implies $\|\nabla f(x_n)\| \to 0$ as $n \to \infty$. Theorem is proven. □

Corollary 21.8. (on the gradient method) *If in (21.45) we take*

$$\gamma_n := \gamma, \quad H_n := I$$

we get the ***gradient method***

$$x_{n+1} = x_n - \gamma \nabla f(x_n), \quad x_0 = \mathring{x}, \quad n = 0, 1, 2, \ldots$$

$$0 < \gamma_n \le \bar{\gamma} := \frac{2(1-\varkappa)}{L_f}$$

for which

$$\lim_{n\to\infty} f(x_n) = \bar{f} > -\infty, \quad f(x_{n+1}) \le f(x_n)$$
$$\|\nabla f(x_n)\| \to 0 \text{ as } n \to \infty \tag{21.58}$$

for any initial conditions $\mathring{x} \in \mathbb{R}^n$.

Corollary 21.9. (on the modified Newton's method) *If in (21.45) we take*

$$\gamma_n := \gamma, \quad H_n := [\nabla^2 f(x_{n-1})]^{-1}$$

and if we get the ***modified Newton's method***

$$x_{n+1} = x_n - \gamma \left[\nabla^2 f(x_n)\right]^{-1} \nabla f(x_n)$$
$$x_0 = \mathring{x}, \quad n = 0, 1, 2, \ldots$$
$$0 < \gamma \leq \bar{\gamma} := \frac{2(1-\varkappa)}{L_f \lambda^+}$$

then again (21.58) is valid.

Remark 21.2. (One more modification) *Let the function $f(x)$ be twice differentiable, its second derivative satisfies the Lipschitz condition with the constant L_{∇^2} and be strongly convex (with the constant $l > 0$) on $\mathbb{R}^n$. Taking in (21.45)*

$$\gamma_n := \begin{cases} \gamma \leq \bar{\gamma} = \dfrac{2(1-\varkappa)\, l}{L_{\nabla^2}}, \quad \varkappa \in (0,1) & \text{if} \quad \|\nabla f(x_n)\| \geq \dfrac{2l^2}{L_{\nabla^2}} \\ 1 & \text{if} \quad \|\nabla f(x_n)\| < \dfrac{2l^2}{L_{\nabla^2}} \end{cases}$$
$$H_n := \left[\nabla^2 f(x_{n-1})\right]^{-1}$$

we get the ***modified Newton's method with a switched step-parameter*** *for which, at the end of optimization, we obtain the, so-called,* ***"quadratic" exponential convergence***

$$\|x_n - x^*\| \leq \frac{2l}{L_f} q^{2^n}, \quad q = \frac{L_f}{2l^2} \|\nabla f(x_{n^*})\| \tag{21.59}$$
$$x_{n^*} := \inf_{n=0,1,2,\ldots} \left\{ x_n \mid \frac{L_{\nabla^2}}{2l^2} \|\nabla f(x_n)\| < 1 \right\}$$

starting from x_{n^*} which is sufficiently close to the minimum point x^* such that

$$\frac{L_{\nabla^2}}{2l^2} \|\nabla f(x_{n^*})\| < 1$$

Proof. First, notice that in this case $\lambda_{\max}(H_{n+1}) \leq \lambda^+ = l^{-1}$. Evidently, by Theorem 21.7 we have $\|\nabla f(x_n)\| \to 0$ as $n \to \infty$, and, hence, there exists the number n^* for which $\frac{L_{\nabla^2}}{2l^2} \|\nabla f(x_{n^*})\| < 1$. Consider $n \geq n^*$. Taking $x = x_n$ and $y := -\gamma \left[\nabla^2 f(x_n)\right]^{-1} \nabla f(x_n)$ in the inequality $\left\|\nabla f(x+y) - \nabla f(x) - \nabla^2 f(x)\, y\right\| \leq \frac{L_{\nabla^2}}{2} \|y\|^2$ we get $x_{n+1} = x + y$ and

$$\left\|\nabla f(x_{n+1}) - \nabla f(x_n) - \nabla^2 f(x_n) \left(-\left[\nabla^2 f(x_n)\right]^{-1} \nabla f(x_n)\right)\right\|$$
$$= \|\nabla f(x_{n+1})\| \leq \frac{L_{\nabla^2}}{2} \left\|\left[\nabla^2 f(x_n)\right]^{-1}\right\|^2 \|\nabla f(x_n)\|^2 \leq \frac{L_{\nabla^2}}{2l^2} \|\nabla f(x_n)\|^2$$

Denoting $z_n := \|\nabla f(x_n)\|$ we obtain $z_{n+1} \leq \left(L_{\nabla^2}/2l^2\right) z_n^2$. Integrating this inequality we get $z_n \leq \frac{2l^2}{L_{\nabla^2}} q^{2^n}$. Applying then the inequality (21.37) we get (21.59). Corollary is proven. □

Many other methods of unconstrained optimization can be found in (Polyak 1987) and the references within.

21.3 Constrained optimization

21.3.1 Elements of convex analysis

21.3.1.1 Convex sets

Definition 21.6.

1. *A set $Q \subseteq \mathbb{R}^n$ is **convex** if it contains any segment with the endpoints lying in Q, i.e., $\lambda x + (1-\lambda)\, y \in Q$ for any $\lambda \in [0, 1]$ whenever $x, y \in Q$.*
2. *The **convex hull** $\operatorname{conv} Q$ of a set $Q \subseteq \mathbb{R}^n$ is an intersection of all convex sets containing Q, or, equivalently, Carathéodory's lemma, see Rockafellar (1970),*

$$\operatorname{conv} Q = \left\{ x = \sum_{i=1}^{n+1} \lambda_i x_i \mid x_i \in Q,\ \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\} \tag{21.60}$$

Claim 21.3.

- *It is easy to check that*
- *if Q is bounded and closed then*

$$\operatorname{conv} Q = Q$$

- *if Q is convex then the sets*

$$\alpha Q := \{x = \alpha x' \in \mathbb{R}^n \mid \alpha \in \mathbb{R},\quad x' \in Q\}$$
$$AQ := \left\{x = Ax' \in \mathbb{R}^n \mid A \in \mathbb{R}^{n \times m},\quad x' \in Q\right\}$$

  are convex too;
- *if Q_1 and Q_2 are convex then $Q_1 \cap Q_2$ is convex.*

Claim 21.4. *For a convex function $f(x)$ the set*

$$Q_\alpha := \{x \in \mathbb{R}^n \mid f(x) \leq \alpha\} \tag{21.61}$$

is convex.

Proof. Let $x', x'' \in Q_\alpha$. Then, by the convexity property (21.18),

$$\begin{aligned} f(\lambda x' + (1-\lambda)\, x'') &\leq \lambda f(x') + (1-\lambda)\, f(x'') \\ &\leq \lambda\alpha + (1-\lambda)\,\alpha = \alpha \end{aligned}$$

which means that $z := \lambda x' + (1-\lambda)\, x'' \in Q_\alpha$. □

Definition 21.7. *A function $f(x)$ is called **quasi-convex** if the sets Q_α (21.61) are convex for any α.*

Remark 21.3. *If the sets Q_α are convex, then $f(x)$ is not obligatory convex, for example, $f(x) = e^{-x^2}$.*

21.3.1.2 Projections and their properties

Definition 21.8. *The projection of the point $x \in \mathbb{R}^n$ onto the set $Q \subseteq \mathbb{R}^n$ is a point $\pi_Q\{x\} \in Q$ such that*

$$\boxed{\pi_Q\{x\} = \arg\min_{y \in Q} \|x - y\|} \tag{21.62}$$

Proposition 21.2. *The following assertions seem to be evident:*

(a) If $x \in Q$ then

$$\pi_Q\{x\} = x$$

(b)

$$\pi_Q\{x\} = \arg\min_{y \in Q} \|x - y\|^2$$

*(c) If Q is closed convex then $\pi_Q\{x\}$ is **unique**, since $\phi(y) := \|x - y\|^2$ is a strictly convex function and, hence, has a unique minimum point.*

Lemma 21.9. *If Q is closed convex then*

1. for all $x \in \mathbb{R}^n$ and all $y \in Q$

$$\boxed{\left(x - \pi_Q\{x\}, \quad y - \pi_Q\{x\}\right) \leq 0} \tag{21.63}$$

2. for all $x, y \in \mathbb{R}^n$

$$\boxed{\left\|\pi_Q\{x\} - \pi_Q\{y\}\right\| \leq \|x - y\|} \tag{21.64}$$

Proof.

1. Since, by Definition 21.8 and in view of the closeness and the convexity of Q, for any $y \in Q$ we have

$$\left\|x - \pi_Q\{x\}\right\|^2 \leq \|x - y\|^2$$

and

$$\left\|x - \pi_Q\{x\}\right\|^2 = \left\|(x - y) + \left(y - \pi_Q\{x\}\right)\right\|^2 = \|x - y\|^2 + 2\left(x - \pi_Q\{x\}, y - \pi_Q\{x\}\right) + \left\|y - \pi_Q\{x\}\right\|^2 \leq \|x - y\|^2$$

or, equivalently,

$$2\left(x - \pi_Q\{x\}, y - \pi_Q\{x\}\right) \leq -\left\|y - \pi_Q\{x\}\right\|^2 \leq 0$$

2. By Definition 21.8 we have

$$\|\pi_Q\{x\} - \pi_Q\{y\}\|^2 \le \|x - \pi_Q\{y\}\|^2$$
$$\|\pi_Q\{y\} - \pi_Q\{x\}\|^2 \le \|y - \pi_Q\{x\}\|^2$$

Summing both inequalities with the weights $\alpha \in [0, 1]$ and $(1-\alpha)$ one gets:

$$\begin{aligned}
\|\pi_Q\{x\} - \pi_Q\{y\}\|^2 &\le \alpha\|x - \pi_Q\{y\}\|^2 + (1-\alpha)\|y - \pi_Q\{x\}\|^2 \\
&= \alpha\|(x-y) + (y - \pi_Q\{y\})\|^2 + (1-\alpha)\|(y-x) + (x - \pi_Q\{x\})\|^2 \\
&= \|x-y\|^2 + \alpha\|y - \pi_Q\{y\}\|^2 + (1-\alpha)\|x - \pi_Q\{x\}\|^2 \\
&\quad + 2\alpha(x-y, y - \pi_Q\{y\}) + 2(1-\alpha)(y-x, x - \pi_Q\{x\}) \\
&= \|x-y\|^2 + \alpha\|y - \pi_Q\{y\}\|^2 + (1-\alpha)\|x - \pi_Q\{x\}\|^2 \\
&\quad + 2\alpha([x - \pi_Q\{y\}] - [y - \pi_Q\{y\}], y - \pi_Q\{y\}) \\
&\quad + 2(1-\alpha)([y - \pi_Q\{x\}] - [x - \pi_Q\{x\}], x - \pi_Q\{x\}) = \|x-y\|^2 \\
&\quad + 2(\alpha[x - \pi_Q\{x\}], [y - \pi_Q\{y\}]) + 2(x - \pi_Q\{x\}, (1-\alpha)[y - \pi_Q\{y\}]) \\
&\quad - \alpha\|y - \pi_Q\{y\}\|^2 - (1-\alpha)\|x - \pi_Q\{x\}\|^2 \le \|x-y\|^2 \\
&\quad + 2\sqrt{\alpha}\|x - \pi_Q\{x\}\|\,\|y - \pi_Q\{y\}\| \\
&\quad + 2\sqrt{1-\alpha}\|x - \pi_Q\{x\}\|\,\|y - \pi_Q\{y\}\| \\
&\quad - \alpha\|y - \pi_Q\{y\}\|^2 - (1-\alpha)\|x - \pi_Q\{x\}\|^2 \le \|x-y\|^2 \\
&\quad - \left(\sqrt{1-\alpha}\|x - \pi_Q\{x\}\|^2 - \sqrt{\alpha}\|x - \pi_Q\{x\}\|\right)^2 \le \|x-y\|^2
\end{aligned}$$

Lemma is proven. □

21.3.1.3 Separation theorems

Here we will formulate and prove the theorem, named *the separation theorem, see Fig. 21.2 for a finite dimensional space (in infinite dimensional spaces this result is known as the Hahn–Banach theorem)*, which plays a key role in constrained optimization.

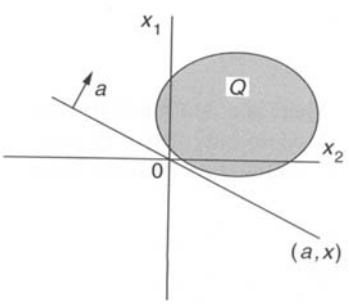

Fig. 21.2. Illustration of the separation theorem.

Theorem 21.8. (Alexeev *et al.* 1979)
Let $Q \subseteq R^n$ be a convex subspace (or a set) of $\mathbb{R}^n$ which does not contain the point 0, that is, $0 \notin Q$. Then there exists a vector $a = (a_1, \ldots, a_n)^\top \in \mathbb{R}^n$ such that for any $x = (x_1, \ldots, x_n)^\top \in Q$ the following inequality holds:

$$\boxed{(a, x) = \sum_{i=1}^{n} a_i x_i \geq 0} \tag{21.65}$$

*In other words, the plane $\sum_{i=1}^{n} \alpha_i x_i = 0$ **separates** the space $\mathbb{R}^n$ in two subspaces, one of which contains the set Q completely.*

Proof. Let lin Q be a minimal linear subspace of $\mathbb{R}^n$ containing Q. Only two cases are possible:

$$\text{lin } Q \neq \mathbb{R}^n \text{ or lin } Q = \mathbb{R}^n$$

1. If lin $Q \neq \mathbb{R}^n$, then lin Q is a proper subspace in $\mathbb{R}^n$ and, therefore, there exists a hyperplane $\sum_{i=1}^{n} \alpha_i x_i = 0$ containing Q as well as the point 0. This plane may be selected as the one we are interested in.
2. If lin $Q = \mathbb{R}^n$, then from vectors belonging to Q we may select n-linearly independent ones forming a basis in $\mathbb{R}^n$. Denote them by

$$e^1, \ldots, e^n \left(e^i \in Q,\ i = 1, \ldots, n\right)$$

Consider then the two convex sets (more exactly cones): a nonnegative "orthant" $\mathcal{K}_1$ and a "convex cone" $\mathcal{K}_2$ defined by

$$\begin{aligned} \mathcal{K}_1 &:= \left\{ x \in \mathbb{R}^n : x = \sum_{i=1}^{n} \beta_i e^i, \quad \beta_i < 0 \right\} \\ \mathcal{K}_2 &:= \left\{ \begin{array}{l} x \in \mathbb{R}^n : x = \displaystyle\sum_{i=1}^{s} a_i \bar{e}^i, \quad a_i \geq 0, \quad \bar{e}^i \in Q \\ i = 1, \ldots, s \quad (s \in \mathbb{N} \text{ is any natural number}) \end{array} \right\} \end{aligned} \tag{21.66}$$

These two cones are not crossed, that is, they do not contain a common point. Indeed, suppose that there exists a vector

$$\bar{x} = -\sum_{i=1}^{n} \bar{\beta}_i e^i, \quad \bar{\beta}_i > 0$$

which also belongs to $\mathcal{K}_2$. Then one is obliged to find $s \in N$, $\bar{a}_i \geq 0$ and $\bar{e}^i$ such that $\bar{x} = \sum_{i=1}^{s} \bar{a}_i \bar{e}^i$. But this is possible only if $0 \in Q$, since, in this case, the point 0 might be represented as a convex combination of some points from Q, i.e.,

$$0 = \frac{\displaystyle\sum_{i=1}^{s} \bar{a}_i \bar{e}^i - \bar{x}}{\displaystyle\sum_{i=1}^{s} \bar{a}_i + \sum_{i=1}^{n} \bar{\beta}_i} = \frac{\displaystyle\sum_{i=1}^{s} \bar{a}_i \bar{e}^i + \sum_{i=1}^{n} \bar{\beta}_i e^i}{\displaystyle\sum_{i=1}^{s} \bar{a}_i + \sum_{i=1}^{n} \bar{\beta}_i} \tag{21.67}$$

$$= \sum_{i=1}^{s} \frac{(\bar{a}_i + \bar{\beta}_i)}{\sum_{j=1}^{s} (\bar{a}_i + \bar{\beta}_j)} e^i$$

But this contradicts with the assumption that $0 \notin Q$. So,

$$\mathcal{K}_1 \cap \mathcal{K}_2 = \varnothing \tag{21.68}$$

3. Since $\mathcal{K}_1$ is an open set, then any point $x \in \mathcal{K}_1$ cannot belong to conv $\mathcal{K}_2$ in the same time. Note that conv $\mathcal{K}_2$ is a closed and convex set. Let us consider any point $x^0 \in \mathcal{K}_1$, for example, $x^0 = -\sum_{i=1}^{n} \bar{e}^i$ and try to find the point $y^0 \in$ conv $\mathcal{K}_2$ closer to x^0. Such point obligatory exists, namely, it is the point minimizes the continuous function $f(y) := \|x - y\|$ within all y belonging to the compact

$$\text{conv } \mathcal{K}_2 \cap \left\{x \in \mathcal{K}_1 : \left\|x - \left\|x^0\right\|\right\| \leq \varepsilon - \text{small enough}\right\}$$

4. Then let us construct the hyperplane H orthogonal to the $(x^0 - y^0)$ and show that this is the plane that we are interested in, that is, show that $0 \in H$ and Q belongs to a half closed subspace separated by this surface, namely,

$$\text{int } H \cap \text{conv } \mathcal{K}_2 = \varnothing$$

and, since, $Q \subseteq \text{conv } \mathcal{K}_2$, then

$$Q \subseteq \overline{\left(\mathbb{R}^n \setminus \text{int } H\right)}$$

By contradiction, let us suppose that there exists a point $\tilde{y} \in \left(\text{int } H \cap \bar{\mathcal{K}}_2\right)$. Then the angle $\angle x^0 y^0 \tilde{y}$ is less than $\pi/2$, and, besides, since conv $\mathcal{K}_2$ is convex, it follows that $\left[y^0, \tilde{y}\right] \in$ conv $\mathcal{K}_2$. Let us take the point $\tilde{y}' \in \left(y^0, \tilde{y}\right)$ such that $\left(x^0, \tilde{y}'\right) \perp \left(y^0, \tilde{y}\right)$ and show that $\tilde{y}'$ is not a point from conv $\mathcal{K}_2$ close to x^0. Indeed, the points y^0, $\tilde{y}$ and $\tilde{y}'$ belong to the same line and $\tilde{y}' \in$ int H. But, if $\tilde{y}' \in \left[y^0, \tilde{y}\right]$ and $\tilde{y}' \in$ conv $\mathcal{K}_2$, then obligatory $\left\|x^0 - \tilde{y}'\right\| < \left\|x^0 - y^0\right\|$ (a shortest distance is less than any other one). At the same time, $\tilde{y}' \in \left(y^0, \tilde{y}\right)$, so $\left\|x^0 - \tilde{y}\right\| < \left\|x^0 - y^0\right\|$. Also we have $0 \in H$, since if not, the line $[0, \infty)$, crossing y^0 and belonging to $\bar{\mathcal{K}}_2$, should obligatory have the common points with conv $\mathcal{K}_2$. Theorem is proven. □

Definition 21.9. *The convex sets Q_1 and Q_2 in $\mathbb{R}^n$ are said to be disjoint (or **separable**) if there exists a number α and a vector $a \in \mathbb{R}^n$ $(a \neq 0)$ such that $(a, x) \geq \alpha$ for all $x \in Q_1$ and $(a, x) \leq \alpha$ for all $x \in Q_2$. These sets are called strictly disjoint (or **strictly separable**) if $(a, x) \geq \alpha_1$ for all $x \in Q_1$ and $(a, x) \leq \alpha_2$ for all $x \in Q_2$ and $\alpha_2 < \alpha_1$.*

Lemma 21.10. (Polyak 1987) *If Q_1, Q_2 be convex disjoint (separable) sets in $\mathbb{R}^n$ and, additionally, Q_2 be closed and bounded. Then Q_1 and Q_2 are **strictly separable**.*

Proof. First, notice that the function

$$\varphi_1(x) := \left\|x - \pi_{Q_1}(x)\right\|$$

is convex. So, by the properties of Q_2, this function attains its minimum on Q_2. Denote

$$a_2 := \arg\min_{x \in Q_2} \varphi_1(x), \quad a_1 := \pi_{Q_2}(a_2)$$

Then $a_2 \neq a_1$ and

$$\|a_1 - a_2\| = \min\{\|x - y\| : x \in Q_1,\ y \in Q_2\}$$
$$a_2 = \pi_{Q_2}(a_1)$$

Hence, by the disjointedness definition,

$$(a_1 - a_2, x) \geq (a_1 - a_2, a_1) := \alpha_1 \quad \text{for all} \quad x \in Q_1$$
$$(a_1 - a_2, x) \geq (a_1 - a_2, a_2) := \alpha_2 \quad \text{for all} \quad x \in Q_2$$

which implies

$$\alpha_1 - \alpha_2 = \|a_1 - a_2\|^2 > 0$$

Lemma is proven. □

21.3.1.4 Subgradient

Definition 21.10. *A vector $a \in \mathbb{R}^n$ for which*

$$\boxed{f(x + y) \geq f(x) + (a, y)} \tag{21.69}$$

for all $y \in \mathbb{R}^n$ is called the ***subgradient*** *of the convex function $f(x)$ at the point $x \in \mathbb{R}^n$ and is denoted by $a = \partial f(x)$. Sure, in the nonsmooth points there exist a set of subgradients denoted by $\mathcal{D}f(x)$ (see Fig. 21.3).*

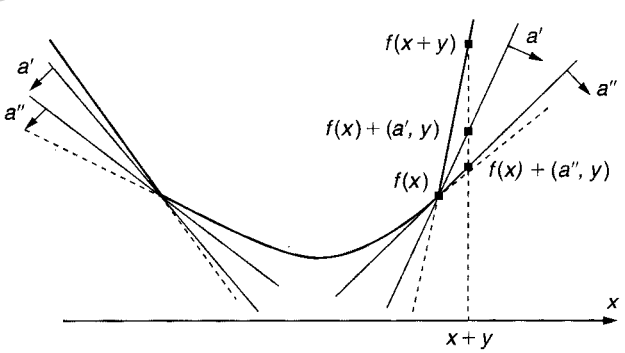

Fig. 21.3. Subgradients.

The following properties of the subgradients of convex functions seem to be evident.

Claim 21.5.

1. *Analogously to (21.31) the subgradient of a convex function is a monotone operator, i.e., for any* $x, y \in R^n$ *and any* $\partial f(x) \in \mathcal{D}f(x)$, $\partial f(y) \in \mathcal{D}f(y)$

$$(\partial f(x) - \partial f(y), x - y) \geq 0 \tag{21.70}$$

2. $f'_+(x, y)$, *as it is defined in (21.23), for any* $x, y \in R^n$ *can be calculated as follows*

$$f'_+(x, y) = \max_{a \in \mathcal{D}f(x)} (a, y) \tag{21.71}$$

3.

$$\partial \left(\sum_{i=1}^{m} \gamma_i f_i(x) \right) = \sum_{i=1}^{m} \gamma_i \partial f_i(x), \quad \gamma_i \geq 0 \tag{21.72}$$

4. *For the convex functions* $f_1(x)$ *and* $f_2(x)$ *we have*

$$\mathcal{D} \max \{f_1(x); f_2(x)\} = \operatorname{Conv} [\mathcal{D}f_1(x) \cup \mathcal{D}f_2(x)] \tag{21.73}$$

5. *For any matrix* $A \in R^{n \times n}$ *and any* $x \in R^n$

$$\partial f(Ax) = A^{\mathsf{T}} \partial f(x) \tag{21.74}$$

Exercise 21.2. The following relations seem to be useful:

1.

$$\partial \|x\| = \begin{cases} \dfrac{x}{\|x\|} & \text{if} \quad x \neq 0 \\ a \quad \text{with} \quad \|a\| \leq 1 & \text{if} \quad x = 0 \end{cases} \tag{21.75}$$

2.

$$\partial \sum_{i=1}^{m} \left| (a^i, x) - b_i \right| = \sum_{i=1}^{m} a^i \operatorname{sign} \left| (a^i, x) - b_i \right| \tag{21.76}$$

Lemma 21.11. *The set* $\mathcal{D}f(x)$ *at any point* $x \in \mathbb{R}^n$ *is nonempty, convex, closed and bounded.*

Proof. Consider in $\mathbb{R}^{n+1}$ the set $Q_\alpha := \{x, \alpha : f(x) \leq \alpha\}$ (which is called the epigraph of $f(x)$). Obviously, this set is convex, and, by Lemma 21.3, it has an interior point, since the points $\{x, f(x)\}$ form its boundary. By the convexity of Q_α, there exists a supporting hyperplane for Q_α at the point x, given by $\{a, -1\}$ for some a. Thus, a is a subgradient of $f(x)$ at x. The convexity, closedness and boundedness follow from Lemma 21.4. □

21.3.2 Optimization on convex sets

Here we will be interested in the following optimization problem

$$\boxed{\min_{x \in Q} f(x)} \tag{21.77}$$

where Q is a convex (not obligatory bounded) set and $f(x)$ is assumed to be smooth (differentiable) on Q if any special assumptions are accepted.

Definition 21.11. *We say that the point $x^* \in Q$ is*

*(a) a **local minimum** of $f(x)$ on Q if $f(x^*) \leq f(x)$ for all $x \in Q$ and such that $\|x - x^*\| \leq \delta$, $\delta > 0$;*
*(b) a **global minimum** of $f(x)$ on Q if $f(x^*) \leq f(x)$ for all $x \in Q$.*

21.3.2.1 Necessary first-order minimum condition

Theorem 21.9. (The necessary condition) *Let*

1. *$f(x)$ be differentiable at the **global minimum point** x^*;*
2. *the set Q be a convex set.*

Then for all $x \in Q$

$$\boxed{(\nabla f(x^*), x - x^*) \geq 0} \tag{21.78}$$

Proof. We will prove this theorem by contradiction. Suppose that $(\nabla f(x^*), \mathring{x} - x^*) < 0$ for some $\mathring{x} \in Q$. Then, by the convexity of Q, the point $x_\alpha := x^* + \alpha(\mathring{x} - x^*) \in Q$ for all $\alpha \in [0, 1]$, and, hence, for small enough α

$$f(x_\alpha) = f(x^*) + \alpha(\nabla f(x^*), \mathring{x} - x^*) + o(\alpha) < f(x^*)$$

which contradicts the assumption that x^* is a minimum point. Theorem is proven. □

21.3.2.2 Sufficient first-order minimum condition

Theorem 21.10. (The sufficient condition of optimality) *Let*

1. *$f(x)$ be differentiable at the point $x^* \in Q$;*
2. *the set Q be a convex set;*
3. *for all $x \in Q$ the following inequality holds*

$$\boxed{(\nabla f(x^*), x - x^*) \geq \rho \|x - x^*\|, > 0} \tag{21.79}$$

*Then the point x^* is **a local minimum point** on Q.*

Proof. Take $\varepsilon \geq \varepsilon_1 > 0$, so that

$$|f(x) - f(x^*) - (\nabla f(x^*), x - x^*)| \leq \frac{\alpha}{2} \|x - x^*\|$$

for all $x \in Q$ such that $\|x - x^*\| \le \varepsilon_1$. Then, by (21.79),

$$f(x) \ge f(x^*) + (\nabla f(x^*), x - x^*) - \frac{\alpha}{2}\|x - x^*\| \ge f(x^*) + \frac{\alpha}{2}\|x - x^*\|$$

which means the local optimality of x^*. Theorem is proven. □

Remark 21.4. *Notice that x^* in (21.79) cannot be an interior point of Q, and, therefore, under the conditions of Theorem 21.10 the minimum is attained at a boundary point of Q.*

21.3.2.3 Criterion of optimality for convex (not obligatory differentiable) functions
Theorem 21.11. (The criterion of optimality) *Let*

1. *$f(x)$ be convex on $\mathbb{R}^n$;*
2. *$Q \subseteq \mathbb{R}^n$ be a convex set.*

Then the point $x^* \in Q$ is a global minimum on Q if and only if

$$\boxed{(\partial f(x^*), x - x^*) \ge 0} \tag{21.80}$$

for some subgradient $\partial f(x^*) \in \mathcal{D}f(x^*)$ and all $x \in Q$.

Proof.

(a) *Necessity*. Suppose that there is no such subgradient. Then the sets $\mathcal{D}f(x^*)$ and $S := \{y \in \mathbb{R}^n : (y, x - x^*) \ge 0,\ x \in Q\}$ do not intersect. Notice that S is convex and closed. By Lemma 21.11, the set $\mathcal{D}f(x^*)$ is convex, closed and bounded. So, in the separation lemma 21.10, there exists $c \in \mathbb{R}^n$ such that $(c, a) \le -\alpha < 0$ for all $a \in \mathcal{D}f(x^*)$ and $(c, y) > 0$ for all $y \in S$. Denote by Γ the closure of the cone generated by all feasible directions, i.e.,

$$\Gamma := \left\{x \in \mathbb{R}^n : x = \lim_{k\to\infty} \lambda_k (x^k - x^*), \quad \lambda_k > 0, \quad x^k \in Q\right\}$$

If $c \ne \Gamma$, then again there exists a vector b such that $(b, x) \ge 0$ for all $x \in \Gamma$ and $(c, b) < 0$. But $b \in S$ and, therefore, the inequality $(c, b) < 0$ contradicts the condition $(c, y) > 0$ for all $y \in S$. So, $c = \Gamma$, and, hence, one can find the sequences $\lambda_k > 0$ and $x^k \in Q$ such that $\lambda_k (x^k - x^*) \to c$. Taking k large enough such that

$$\left\|\lambda_k (x^k - x^*) - c\right\| \le \alpha/(2a^+), \quad a^+ := \max_{a \in \mathcal{D}f(x^*)} \|a\|, \quad \alpha > 0$$

we obtain

$$\begin{aligned} f'_+\left(x, \lambda_k (x^k - x^*)\right) &:= \lim_{\alpha\to+0} \frac{f\left(x + \alpha\lambda_k (x^k - x^*)\right) - f(x)}{\alpha} \\ &= \max_{a \in \mathcal{D}f(x^*)} \left(a, \lambda_k (x^k - x^*)\right) \\ &= \max_{a \in \mathcal{D}f(x^*)} (a, c) + \max_{a \in \mathcal{D}f(x^*)} \left(a, \lambda_k (x^k - x^*) - c\right) \\ &= -\alpha + \frac{1}{2}\alpha = -\frac{1}{2}\alpha \end{aligned}$$

and, hence, $f'_+\left(x, \lambda_k\left(x^k - x^*\right)\right) < 0$. Therefore, for sufficiently small $\gamma > 0$, it follows that

$$f\left(x^* + \gamma\left(x^k - x^*\right)\right) < f\left(x^*\right)$$

which contradicts the assumption that x^* is a minimum point. The necessity is proven.

(b) *Sufficiency.* Let (21.80) hold for all $x \in Q$ and some subgradient $\partial f\left(x^*\right)$. Then

$$f(x) \geq f\left(x^*\right) + \left(\partial f\left(x^*\right), x - x^*\right) \geq f\left(x^*\right)$$

i.e., x^* is a global minimum point on Q. Lemma is proven. □

Remark 21.5. *In the case of no constraints, the criterion (21.80) becomes*

$$\boxed{0 \in \mathcal{D} f\left(x^*\right)} \tag{21.81}$$

21.3.3 Mathematical programing and Lagrange principle

21.3.3.1 Nonlinear programing problem

The general problem of nonlinear programing is formulated as follows:

$$\boxed{\begin{array}{c} \min\limits_{x \in \mathbb{R}^n} f(x) \\ \\ \begin{array}{ll} g_i(x) \leq 0, & i = 1, \ldots, r \\ h_j(x) = 0, & j = 1, \ldots, r' \end{array} \end{array}} \tag{21.82}$$

Notice that any equality constraint $h_i(x) = 0$ can be represented as two inequality-type constraints:

$$\{x \in \mathbb{R}^n \mid h_i(x) = 0\} = \{x \in \mathbb{R}^n \mid h_i(x) \leq 0,\ -h_i(x) \leq 0\}$$

So, the general nonlinear programing problem (21.82) can be represented as

$$\boxed{\begin{array}{c} \min\limits_{x \in \mathbb{R}^n} f(x) \\ \\ g_i(x) \leq 0, \quad i = 1, \ldots, m \end{array}} \tag{21.83}$$

where $m = r + 2r'$ with

$$g_i(x) = \begin{cases} h_j(x) & \text{if} \quad i = r + j \\ -h_j(x) & \text{if} \quad i = r + r' + j \end{cases} \quad \left(j = 1, \ldots, r'\right)$$

21.3.3.2 Lagrange principle

The theorem below follows the scheme of presentation given in (Alexeev *et al.* 1979) where the same result is formulated in a more general case in Banach (not obligatory in finite dimensional) space. *It permits to represent the given constraint optimization problem as another (but already) unconstrained optimization problem.*

Theorem 21.12. (Lagrange principle) *Consider the general nonlinear programing problem (21.83) where the functions* $f(x)$ *and* $g_i(x)$ $(i = 1, \ldots, m)$ *are assumed to be differentiable but not convex.*

A. *(**Necessary conditions, Karush–John**). If* x^* *is a local minimum point, then there exist nonnegative constants* $\mu^* \geq 0$ *and* $v_i^* \geq 0$ $(i = 1, \ldots, m)$ *such that the following two conditions hold:*

1. *"local minimality condition to Lagrange function"*

$$\boxed{L(x, \mu, v) := \mu f(x) + \sum_{i=1}^{m} v_i g_i(x)} \tag{21.84}$$

namely,

$$L(x^*, \mu^*, v^*) \leq L(x, \mu^*, v^*) \tag{21.85}$$

or, equivalently,

$$\mu^* \nabla f(x^*) + \sum_{i=1}^{m} v_i^* \nabla g_i(x^*) = 0 \tag{21.86}$$

2. *"complementary slackness":*

$$v_i^* g_i(x^*) = 0 \quad (i = 1, \ldots, m) \tag{21.87}$$

B. *(**Sufficient conditions**). If* $\mu^* > 0$ *(**the regular case**), or equivalently, if the vectors* $\nabla g_i(x^*)$ $(i = 1, \ldots, m)$, *corresponding to the **active indices** for which* $v_i^* > 0$, *are linearly independent, then conditions (1)–(2) above turn out to be **sufficient** to guarantee that* x^* *is a local minimum point;*

C. *To guarantee the existence of* $\mu^* > 0$ *it is **sufficient** that the, so-called, **Slater's condition** holds, namely, that in a neighborhood* $\Omega(x^*)$ *of* x^* *there exists* $\bar{x}$ *such that*

$$g_i(\bar{x}) < 0 \quad (i = 1, \ldots, m) \tag{21.88}$$

Proof. First, define the set

$$\begin{aligned} C := \{\eta \in \mathbb{R}^{m+1} \mid \exists\, x \in \Omega(x^*) : \\ f(x) - f(x^*) < \eta_0, \quad g_i(x) \leq \eta_i \ (i = 1, \ldots, m) \end{aligned} \tag{21.89}$$

A. *The set C is nonempty and convex.* Indeed, the vector η with positive components belongs to C since in (21.89) we may take $x = x^*$. So, C is nonempty. Let us show that it is convex. To do that we need to prove the existence of a vector $x^\alpha \in \Omega(x^*)$ ($\Omega(x^*)$ can always be selected as a convex set) such that for any $\eta^\alpha := \eta + \alpha(\eta' - \eta)$, $\alpha \in [0, 1]$ we have

$$f(x^\alpha) - f(x^*) < \eta_0^\alpha, \quad g_i(x^\alpha) \le \eta_i^\alpha \quad (i = 1, \ldots, m)$$

if for some $x, x' \in \Omega(x^*)$

$$\begin{aligned} &f(x) - f(x^*) < \eta_0, \quad g_i(x) \le \eta_i \ \ (i = 1, \ldots, m) \\ &f(x') - f(x^*) < \eta_0', \quad g_i(x') \le \eta_i' \quad (i = 1, \ldots, m) \end{aligned}$$

Denote $x^\alpha := x + \alpha(x' - x)$, $\alpha \in [0, 1]$ which, by the convexity of $\Omega(x^*)$, also belongs to $\Omega(x^*)$. Since the functions $f(x)$ and $g_i(x)$ $(i = 1, \ldots, m)$ are differentiable in $\Omega(x^*)$, it follows that

$$\begin{aligned} &f(x) - f(x^*) = (\nabla f(x^*), x - x^*) + o(\|x - x^*\|) < \eta_0 \\ &f(x') - f(x^*) = (\nabla f(x^*), x' - x^*) + o(\|x' - x^*\|) < \eta_0 \\ &o(\|x^\alpha - x^*\|) = o(\|x' - x^*\|) + o(\|x - x^*\|) \end{aligned}$$

and, hence,

$$\begin{aligned} f(x^\alpha) - f(x^*) &= (\nabla f(x^*), x^\alpha - x^*) + o(\|x^\alpha - x^*\|) \\ &= \alpha[(\nabla f(x^*), x' - x^*) + o(\|x' - x^*\|)] \\ &\quad + (1 - \alpha)[(\nabla f(x^*), x - x^*) + o(\|x - x^*\|)] \\ &< \alpha\eta_0' + (1 - \alpha)\eta_0 = \eta_0^\alpha \end{aligned}$$

Analogously, $g_i(x^\alpha) \le \eta_i^\alpha$, which implies $\eta^\alpha \in C$. So, C is nonempty and convex.

B. *The point 0 does not belong to C.* Indeed, if so, in view of the definition (21.89), there exists a point $\bar{x} \in X_0$ satisfying

$$\begin{aligned} &f(\bar{x}) - f(x^*) < 0 \\ &g_i(\bar{x}) \le 0 \ \ (i = 1, \ldots, m) \end{aligned} \tag{21.90}$$

which is in contradiction to the fact that x^* is a local solution of the problem. So, $0 \notin C$. Based on this fact and taking into account the convexity property of C, we may apply the separation principle (see Theorem 21.8): there exist constants $(\mu^*, \nu_1^*, \ldots, \nu_m^*)$ such that for all $\eta \in C$

$$\mu^*\eta_0 + \sum_{i=1}^{m} \nu_i^*\eta_i \ge 0 \tag{21.91}$$

C. *Multipliers μ^* and ν_i^* $(i = 1, \ldots, m)$ in (21.91) are nonnegative.* In (A) we have already mentioned that any vector $\eta \in \mathbb{R}^{L+1}$ with positive components belongs to

C, and, particularly, the vector $\left(\underbrace{\varepsilon, \ldots, \varepsilon, 1}_{l_0}, \varepsilon, \ldots, \varepsilon\right)$ $(\varepsilon > 0)$. Substitution of this vector into (21.91) leads to the following inequalities

$$\left.\begin{aligned} \nu_{l_0}^* &\geq -\mu^* \varepsilon - \varepsilon \sum_{i=l_0}^{m} \nu_l^* \quad \text{if} \quad 1 \leq l_0 \leq m \\ \mu^* &\geq -\varepsilon \sum_{i=1}^{m} \nu_l^* \quad \text{if} \quad l_0 = 0 \end{aligned}\right\} \tag{21.92}$$

Tending ε to zero in (21.92) implies the nonnegativity property for the multipliers μ^* and ν_i^* $(i = 1, \ldots, m)$.

D. *Multipliers* ν_i^* $(i = 1, \ldots, m)$ *satisfy the complementary slackness condition* (21.87). Indeed, if $g_{l_0}(x^*) = 0$, then the identity $\nu_{l_0}^* g_{l_0}(x^*) = 0$ is trivial. Suppose that $g_{l_0}(x^*) < 0$. Then the point

$$\left(\underbrace{\delta, 0, \ldots, 0, g_{l_0}(x^*)}_{l_0}, 0, \ldots, 0\right) \quad (\delta > 0) \tag{21.93}$$

belongs to the set C. To check this it is sufficient to take $x = x^*$ in (21.89). Substitution of this point into (21.91) implies

$$\nu_{l_0}^* g_{l_0}(x^*) \geq -\mu^* \delta \tag{21.94}$$

Tending δ to zero we obtain that $\nu_{l_0}^* g_{l_0}(x^*) \geq 0$, and since $g_{l_0}(x^*) < 0$, it follows that $\nu_{l_0}^* \leq 0$. But in (C) it has been proven that $\nu_{l_0}^* \geq 0$. So, $\nu_{l_0}^* = 0$, and, hence, $\nu_{l_0}^* g_{l_0}(x^*) = 0$.

E. *Minimality condition to Lagrange function*. As it follows from (21.89), for $x \in \Omega(x^*)$ the point

$$(f(x) - f(x^*) + \delta, g_1(x), \ldots, g_m(x))$$

belongs to C for any $\delta > 0$. Substitution of this point into (21.91), in view of (D), yields

$$\begin{aligned} L(x, \mu^*, \nu^*) &:= \mu^* f(x) + \sum_{i=1}^{m} \nu_i^* g_i(x) \\ &= \left(\mu^* [f(x) - f(x^*) + \delta] + \sum_{i=1}^{m} \nu_i^* g_i(x)\right) \\ &\quad + \mu^* f(x^*) - \mu^* \delta \geq \mu^* f(x^*) - \mu^* \delta \\ &= \mu^* f(x^*) + \sum_{i=1}^{m} \nu_i^* g_i(x^*) - \mu^* \delta = L(x^*, \mu^*, \nu^*) - \mu^* \delta \end{aligned} \tag{21.95}$$

Taking $\delta \to 0$ we obtain (21.85).

F. *If* $\mu^* > 0$ *(the regular case), then the conditions (A1) and (A2) are sufficient for the optimality*. Indeed, in this case it is clear that we may take $\mu^* = 1$, and, hence, for any x satisfying $g_i(x) \leq 0$ $(i = 1, \ldots, m)$

$$f(x) \geq f(x) + \sum_{i=1}^{m} v_i^* g_i(x) = L(x, 1, v^*)$$
$$\geq L(x^*, 1, v^*) = f(x^*) + \sum_{i=1}^{m} v_i^* g_i(x^*) = f(x^*)$$

This means that x^* is the optimal solution. Notice also that, by (21.86), if $\mu^* = 0$ it follows that

$$\sum_{i=1}^{m} v_i^* \nabla g_i(x^*) = \sum_{i:v_i^*>0}^{m} v_i^* \nabla g_i(x^*) = 0$$

which means linear dependence of the vectors $\nabla g_i(x^*)$ corresponding to the active constraints.

G. *Slater's condition of the regularity*. Suppose that Slater's condition is fulfilled, but $\mu^* = 0$. We directly obtain the contradiction. Indeed, since not all v_i^* are equal to zero simultaneously, it follows that

$$L(\bar{x}, 0, v^*) = \sum_{l=1}^{L} v_l^* g_l(\bar{x}) < 0 = L(x^*, 0, v^*)$$

which is in contradiction with (E). Theorem is proven. □

21.3.3.3 Convex programing

Theorem 21.13. (Kuhn & Tucker 1951) *Suppose that*

1. *all functions* $f(x)$ *and* $g_i(x)$ $(i = 1, \ldots, m)$ *in the general nonlinear programing problem (21.83) are* ***differentiable***[1] *and* ***convex*** *in* $\mathbb{R}^n$;
2. ***Slater's condition*** *("the existence of an internal point") holds, i.e., there exists* $\bar{x} \in \mathbb{R}^n$ *such that*

$$g_i(\bar{x}) < 0 \quad (i = 1, \ldots, m) \tag{21.96}$$

Then, for a point $x^* \in \mathbb{R}^n$ *to be a global solution of (21.83) it is necessary and sufficient to show the existence of nonnegative constants* $v_i^* \geq 0$ $(i = 1, \ldots, m)$ *such that the, so-called,* ***saddle-point property*** *for the Lagrange function (21.84) holds for any* $x \in \mathbb{R}^n$ *and any* $v_i \geq 0$ $(i = 1, \ldots, m)$

$$L(x, 1, v^*) \geq L(x^*, 1, v^*) \geq L(x^*, 1, v) \tag{21.97}$$

[1] Here we present the version of the theorem dealing with differential functions. In fact the same result remains valid without the assumption on differentiability (see (Polyak 1987)).

or, in another form,

$$\begin{aligned}\min_{x\in\mathbb{R}^n} L(x,1,\nu^*) &= \min_{x\in\mathbb{R}^n}\max_{\nu\geq 0} L(x,1,\nu)\\ &= L(x^*,1,\nu^*)\\ &= \max_{\nu\geq 0}\min_{x\in\mathbb{R}^n} L(x,1,\nu) = \max_{\nu\geq 0} L(x^*,1,\nu)\end{aligned} \tag{21.98}$$

Proof.

(a) *Necessity.* By Slater's condition (see item C in Theorem 21.12) we deal with the regular case, and, therefore, we may take $\mu^* = 1$. If x^* is a global minimum of (21.83), then, in view of the convexity condition, it is a local minimum also. Hence, by Theorem 21.12, we have $L(x^*, 1, \nu^*) \leq L(x, 1, \nu^*)$. On the other hand,

$$\begin{aligned}L(x^*,1,\nu^*) &= f(x^*) + \sum_{i=1}^{m} \nu_i^* g_i(x^*) = f(x^*)\\ &\geq f(x^*) + \sum_{i=1}^{m} \nu_i g_i(x^*) = L(x^*(T),1,\nu)\end{aligned} \tag{21.99}$$

which proves the necessity.

(b) *Sufficiency.* Suppose that (21.97) holds. Then

$$\begin{aligned}L(x^*,1,\nu^*) &= f(x^*) + \sum_{i=1}^{m} \nu_i^* g_i(x^*)\\ &\geq L(x^*,1,\nu) = f(x^*) + \sum_{i=1}^{m} \nu_i g_i(x^*)\end{aligned}$$

which implies

$$\sum_{i=1}^{m} \nu_i^* g_i(x^*) \geq \sum_{i=1}^{m} \nu_i g_i(x^*)$$

for all $\nu_i \geq 0$ $(i = 1, \ldots, m)$. This is possible if and only if (this can be proven by the contradiction)

$$g_i(x^*) \leq 0, \quad \nu_i^* g_i(x^*) = 0 \quad (i = 1, \ldots, m)$$

So, we have

$$\begin{aligned}L(x^*,1,\nu^*) &= f(x^*) + \sum_{i=1}^{m} \nu_i^* g_i(x^*) = f(x^*)\\ &\leq L(x,1,\nu^*) = f(x) + \sum_{i=1}^{m} \nu_i^* g_i(x) \leq f(x) + \sum_{i=1}^{m} \nu_i^* g_i(x^*) = f(x)\end{aligned}$$

which means that x^* is a solution of (21.83). Theorem is proven. □

Remark 21.6. *The construction of the Lagrange function in the form (21.84)*

$$L(x, \mu, \nu) = \mu f(x) + \sum_{i=1}^{m} \nu_i g_i(x)$$

with $\mu \geq 0$ is very essential. Indeed, the usage of this form only as $L(x, 1, \nu)$, when the regularity conditions are not valid, may provoke a serious error in the optimization process. The following counterexample demonstrates this effect. Consider the simple constrained optimization problem formulated as

$$\left.\begin{aligned} h_0(x) &:= x_1 \to \min_{x \in \mathbb{R}^2} \\ g(x) &:= x_1^2 + x_2^2 \leq 0 \end{aligned}\right\} \tag{21.100}$$

This problem evidently has the unique solution $x_1 = x_2 = 0$. But the direct usage of the Lagrange principle with $\mu = 1$ leads to the following contradiction:

$$\left.\begin{aligned} &L(x, 1, \nu^*) = x_1 + \nu^* \left(x_1^2 + x_2^2\right) \to \min_{x \in \mathbb{R}^2} \\ &\frac{\partial}{\partial x_1} L(x^*, 1, \nu^*) = 1 + 2\nu^* x_1^* = 0 \\ &\frac{\partial}{\partial x_2} L(x^*, 1, \nu^*) = 2\nu^* x_2^* = 0 \\ &\nu^* \neq 0, \quad x_2^* = 0, \quad x_1^* = -\frac{1}{2\nu^*} \neq 0 \end{aligned}\right\} \tag{21.101}$$

Notice that for this example Slater's condition (21.96) is not valid.

21.3.4 Method of subgradient projection to simplest convex sets

Let us consider the constrained optimization problem (21.77)

$$\min_{x \in Q} f(x)$$

where the function $f(x)$ is supposed to be convex, and the set Q is convex and having a simple structure such that the projection operation $\pi_Q\{x\}$ (21.62) can be easily realized. Consider also the following iterative procedure:

$$\boxed{x_{n+1} = \pi_Q\{x_n - \gamma_n \partial f(x_n)\}} \tag{21.102}$$

where $\partial f(x_n)$ is any subgradient form $\mathcal{D} f(x_n)$, and $\gamma_n \geq 0$ is "the step of the procedure". Denote by $x^* \in Q$ the solution of the constrained optimization problem (21.77) with the convex f.

Theorem 21.14. (on strong convergence) *Suppose that*

1. $f(x)$ has on Q a unique global minimum point x^*, that is,

$$f(x) > f(x^*) \text{ for all } x \in Q, \ x \neq x^* \tag{21.103}$$

2. *for any* $x \in Q$ *and any* $\partial f(x) \in \mathcal{D} f(x)$

$$\|\partial f(x)\|^2 \leq c_0 + c_1 \|x - x^*\|^2 \tag{21.104}$$

3. *the step size* γ_n *satisfies the conditions*

$$\gamma_n \geq 0, \quad \sum_{n=0}^{\infty} \gamma_n = \infty, \quad \sum_{n=0}^{\infty} \gamma_n^2 < \infty \tag{21.105}$$

Then for any initial value $x_0 \in Q$ *the vector sequence* $\{x_n\}$*, generated by the procedure (21.102), converges to* x^* *whereas* $n \to \infty$.

Proof. By the projection operator property (21.64), and in view of the inequality (21.70)

$$(\partial f(x) - \partial f(y), x - y) \geq 0$$

it follows that

$$\begin{aligned}
\|x_{n+1} - x^*\|^2 &= \left\|\pi_Q \{x_n - \gamma_n \partial f(x_n)\} - x^*\right\|^2 \\
&\leq \|x_n - \gamma_n \partial f(x_n) - x^*\|^2 = \|x_n - x^*\|^2 - 2\gamma_n (\partial f(x_n), x_n - x^*) \\
&\quad + \gamma_n^2 \|\partial f(x_n)\|^2 = \|x_n - x^*\|^2 - 2\gamma_n (\partial f(x_n) - \partial f(x^*), x_n - x^*) \\
&\quad - 2\gamma_n (\partial f(x^*), x_n - x^*) + \gamma_n^2 \left(c_0 + c_1 \|x - x^*\|^2\right) \\
&\leq \left(1 + c_1 \gamma_n^2\right) \|x_n - x^*\|^2 - 2\gamma_n (\partial f(x^*), x_n - x^*) + \gamma_n^2 c_0
\end{aligned}$$

So, defining $v_n := \|x_n - x^*\|^2$, we have

$$v_{n+1} \leq \left(1 + c_1 \gamma_n^2\right) v_n + \gamma_n^2 c_0 - 2\gamma_n (\partial f(x^*), x_n - x^*) \tag{21.106}$$

By (21.80) it follows that $(\partial f(x^*), x_n - x^*) \geq 0$, and hence, (21.106) implies

$$v_{n+1} \leq \left(1 + c_1 \gamma_n^2\right) v_n + \gamma_n^2 c_0 \tag{21.107}$$

Let us consider the sequence (*Gladishev's transformation*) $\{w_n\}$ defined by

$$w_n := v_n \prod_{i=n}^{\infty} \left(1 + c_1 \gamma_i^2\right) + c_0 \sum_{i=n}^{\infty} \gamma_i^2 \prod_{s=n+1}^{\infty} \left(1 + c_1 \gamma_s^2\right) \tag{21.108}$$

The variable w_n is correctly defined since, in view of the inequality $1 + x \leq e^x$, it follows that

$$\prod_{i=n}^{\infty} \left(1 + c_1 \gamma_i^2\right) \leq \exp\left(c_1 \sum_{i=n}^{\infty} \gamma_i^2\right) < \infty$$

For this variable, using (21.107), we have

$$\begin{aligned} w_{n+1} &= v_{n+1} \prod_{i=n+1}^{\infty} \left(1 + c_1 \gamma_i^2\right) + c_0 \sum_{i=n+1}^{\infty} \gamma_i^2 \prod_{s=n+2}^{\infty} \left(1 + c_1 \gamma_s^2\right) \\ &\leq v_n \prod_{i=n}^{\infty} \left(1 + c_1 \gamma_i^2\right) + \gamma_n^2 c_0 \prod_{i=n+1}^{\infty} \left(1 + c_1 \gamma_i^2\right) + c_0 \sum_{i=n+1}^{\infty} \gamma_i^2 \prod_{s=n+2}^{\infty} \left(1 + c_1 \gamma_s^2\right) \\ &= v_n \prod_{i=n}^{\infty} \left(1 + c_1 \gamma_i^2\right) + c_0 \sum_{i=n}^{\infty} \gamma_i^2 \prod_{s=n+1}^{\infty} \left(1 + c_1 \gamma_s^2\right) = w_n \end{aligned}$$

So, $0 \leq w_{n+1} \leq w_n$, and, hence, by the Weierstrass theorem, this sequence converges, i.e.,

$$w_n \to w^* \quad \text{as} \quad n \to \infty$$

In view of (21.108), it follows that v_n also converges (in fact, to the same limit point), that is

$$v_n \to v^* \quad \text{as} \quad n \to \infty$$

Returning to (21.106), after summing these inequalities, we obtain

$$v_{n+1} \leq v_0 + c_1 \sum_{i=0}^{n} \gamma_i^2 v_i + c_0 \sum_{i=0}^{n} \gamma_i^2 - 2 \sum_{i=0}^{n} \gamma_i \left(\partial f\left(x^*\right), x_i - x^*\right)$$

or, equivalently,

$$\begin{aligned} 2 \sum_{i=0}^{n} \gamma_i \left(\partial f\left(x^*\right), x_i - x^*\right) &\leq v_0 + c_1 \sum_{i=0}^{n} \gamma_i^2 v_i + c_0 \sum_{i=0}^{n} \gamma_i^2 - v_{n+1} \\ &\leq v_0 + c_1 \sum_{i=0}^{n} \gamma_i^2 v_i + c_0 \sum_{i=0}^{n} \gamma_i^2 \leq \text{const} < \infty \end{aligned}$$

Taking $n \to \infty$, we get

$$\sum_{i=0}^{\infty} \gamma_n \left(\partial f\left(x^*\right), x_n - x^*\right) < \infty$$

Since, by the assumption of this theorem, $\sum_{i=0}^{\infty} \gamma_n = \infty$, we may conclude that there exists a subsequence $\{n_k\}$ such that $\left(\partial f\left(x^*\right), x_{n_k} - x^*\right) \to 0$ whereas $k \to \infty$. But, by the uniqueness of the global minimum (see the condition (21.103)), we derive that $x_{n_k} \to x^*$ as $k \to \infty$, or equivalently, $v_{n_k} \to 0$. But $\{v_n\}$ converges, and, therefore, all subsequences have the same limit v^*, which implies $v^* = 0$. Theorem is proven. □

21.3.5 Arrow–Hurwicz–Uzawa method with regularization

Consider again the general *non-linear convex programing problem* in the form (21.83), i.e.,

$$\begin{aligned} &\min_{x \in Q \subset \mathbb{R}^n} f(x) \\ &g_i(x) \leq 0, \quad i = 1, \ldots, m \end{aligned} \tag{21.109}$$

when all functions are assumed to be convex (not obligatory strictly convex), the set Q is a convex compact, and Slater's condition (21.96) is fulfilled. Additionally, we will assume that

$$Q \bigcap \left(\bigcap_{i=1,\ldots,m} \{x \in \mathbb{R}^n : g_i(x) \leq 0\} \right) \neq \varnothing$$

Associate with this problem the following regularized Lagrange function

$$\boxed{\begin{aligned} L_\delta(x, 1, \nu) &:= L(x, 1, \nu) + \frac{\delta}{2}\left(\|x\|^2 - \|\nu\|^2\right), \quad \delta \geq 0 \\ L(x, 1, \nu) &= f(x) + \sum_{i=1}^{m} \nu_i g_i(x) \end{aligned}} \tag{21.110}$$

First, notice that the function $L_\delta(x, 1, \nu)$ for $\delta > 0$ is strictly convex on x for any fixed ν, and it is strictly concave on ν for any fixed x, and, hence, it has the unique saddle point $(x^*(\delta), \nu^*(\delta))$ for which the following inequalities hold: for any ν with nonnegative components and any $x \in \mathbb{R}^n$

$$L_\delta(x, 1, \nu^*(\delta)) \geq L_\delta(x^*(\delta), 1, \nu^*(\delta)) \geq L_\delta(x^*(\delta), 1, \nu) \tag{21.111}$$

As for the function $L(x, 1, \nu)$, it may have several (not obligatory unique) saddle points (x^*, ν^*). The next proposition describes the dependence of the saddle point $(x^*(\delta), \nu^*(\delta))$ of the regularized Lagrange function (21.110) on the regularizing parameter δ and analyses its asymptotic behavior when $\delta \to 0$.

Proposition 21.3.

1. *For any $x \in \mathbb{R}^n$ and any ν with nonnegative components the following inequality holds:*

$$\begin{aligned} &\left(x - x^*(\delta), \frac{\partial}{\partial x} L_\delta(x, 1, \nu)\right) - \left(\nu - \nu^*(\delta), \frac{\partial}{\partial \nu} L_\delta(x, 1, \nu)\right) \\ &\geq \frac{\delta}{2}\left(\|x - x^*(\delta)\|^2 + \|\nu - \nu^*(\delta)\|^2\right) \end{aligned} \tag{21.112}$$

2. *For any* $\delta, \delta' > 0$ *there exists* $0 < c < \infty$ *such that the following "Lipschitz-type" continuity property holds:*

$$\left\|x^*(\delta) - x^*\left(\delta'\right)\right\| + \left\|\nu^*(\delta) - \nu^*\left(\delta'\right)\right\| \leq c\left|\delta - \delta'\right| \tag{21.113}$$

3. *When* $0 < \delta_n \to 0$

$$\begin{aligned} &\left(x^*(\delta_n), \nu^*(\delta_n)\right) \to \left(x^{**}, \nu^{**}\right) \text{ as } n \to \infty \\ &\left(x^{**}, \nu^{**}\right) = \underset{(x^*, \nu^*)}{\arg\min}\left(\|x^*\|^2 + \|\nu^*\|^2\right) \end{aligned} \tag{21.114}$$

Proof.

1. In view of (21.30) for any $x, y \in \mathbb{R}^n$ we have

$$\begin{aligned} &(\nabla f(x), (y - x)) \leq f(y) - f(x) \\ &(\nabla f(x), (x - y)) \geq f(x) - f(y) \end{aligned}$$

So, since $L(x, 1, \nu)$ is convex on x for any fixed ν, and it is linear on ν, in view of (21.30) it follows that

$$\begin{aligned} &\left(x - x^*(\delta), \frac{\partial}{\partial x} L_\delta(x, 1, \nu)\right) \\ &\quad = \left(x - x^*(\delta), \frac{\partial}{\partial x} L(x, 1, \nu)\right) + \delta\left(x - x^*(\delta), x\right) \\ &\quad \geq L(x, 1, \nu) - L\left(x^*(\delta), 1, \nu\right) + \delta\left(x - x^*(\delta), x\right) \end{aligned} \tag{21.115}$$

and

$$\begin{aligned} &\left(\nu - \nu^*(\delta), \frac{\partial}{\partial \nu} L_\delta(x, 1, \nu)\right) \\ &\quad = \left(\nu - \nu^*(\delta), \frac{\partial}{\partial \nu} L(x, 1, \nu)\right) - \delta\left(\nu - \nu^*(\delta), \nu\right) \\ &\quad = \sum_{i=1}^{m}\left(\nu_i - \nu_i^*(\delta)\right) g_i(x) - \delta\left(\nu - \nu^*(\delta), \nu\right) \end{aligned} \tag{21.116}$$

which leads to the following relation:

$$\begin{aligned} &\left(x - x^*(\delta), \frac{\partial}{\partial x} L_\delta(x, 1, \nu)\right) - \left(\nu - \nu^*(\delta), \frac{\partial}{\partial \nu} L_\delta(x, 1, \nu)\right) \\ &\quad \geq L(x, 1, \nu) - L\left(x^*(\delta), 1, \nu\right) + \delta\left(x - x^*(\delta), x\right) \\ &\qquad - \sum_{i=1}^{m}\left(\nu_i - \nu_i^*(\delta)\right) g_i(x) + \delta\left(\nu - \nu^*(\delta), \nu\right) \\ &\quad = \left[f(x) + \sum_{i=1}^{m} \nu_i^*(\delta) g_i(x) + \frac{\delta}{2}\|x\|^2 - \frac{\delta}{2}\|\nu^*\|^2\right] \end{aligned}$$

$$
\begin{aligned}
&- \left[L(x^*(\delta), 1, \nu) + \frac{\delta}{2} \|x^*(\delta)\|^2 - \frac{\delta}{2} \|\nu\|^2 \right] \\
&+ \left[\frac{\delta}{2} \|x\|^2 - \delta(x^*(\delta), x) + \frac{\delta}{2} \|x^*(\delta)\|^2 \right] \\
&+ \left[\frac{\delta}{2} \|\nu\|^2 - \delta(\nu^*(\delta), \nu) + \frac{\delta}{2} \|\nu^*\|^2 \right] = L_\delta(x, 1, \nu^*(\delta)) \\
&- L_\delta(x^*(\delta), 1, \nu) + \frac{\delta}{2} \left(\|x - x^*(\delta)\|^2 + \|\nu - \nu^*(\delta)\|^2 \right) \\
&\geq \frac{\delta}{2} \left(\|x - x^*(\delta)\|^2 + \|\nu - \nu^*(\delta)\|^2 \right)
\end{aligned} \tag{21.117}
$$

which proves (21.112).

2. The inequality (21.113) results from the linear dependence of the regularized Lagrange function on δ.
3. Taking in (21.117) $x = x^*$, $\nu = \nu^*$ (one of the saddle points of the nonregularized Lagrange function) and applying the inequalities

$$
\begin{aligned}
&L(x^*, 1, \nu^*) - L(x^*(\delta), 1, \nu^*) \leq 0 \\
&\sum_{i=1}^{m} \left(\nu_i^* - \nu_i^*(\delta) \right) g_i(x^*) = - \sum_{i=1}^{m} \nu_i^*(\delta) g_i(x^*) \geq 0
\end{aligned}
$$

leads to the following relation:

$$
\begin{aligned}
0 &\leq L(x^*, 1, \nu^*) - L(x^*(\delta), 1, \nu^*) + \delta(x^* - x^*(\delta), x^*) \\
&\quad - \sum_{i=1}^{m} \left(\nu_i^* - \nu_i^*(\delta) \right) g_i(x^*) + \delta(\nu^* - \nu^*(\delta), \nu^*) \\
&\leq \delta(x^* - x^*(\delta), x^*) + \delta(\nu^* - \nu^*(\delta), \nu^*)
\end{aligned}
$$

Dividing both sides by $\delta > 0$ we obtain

$$
0 \leq (x^* - x^*(\delta), x^*) + (\nu^* - \nu^*(\delta), \nu^*) \tag{21.118}
$$

which is valid for any saddle point (x^*, ν^*) of the nonregularized Lagrange function. We also have

$$
\begin{aligned}
L_\delta(x^*(\delta), 1, \nu^*) &= L(x^*(\delta), 1, \nu^*) + \frac{\delta}{2} \|x^*(\delta)\|^2 - \frac{\delta}{2} \|\nu^*\|^2 \\
&\leq L(x^*, 1, \nu^*(\delta)) + \frac{\delta}{2} \|x^*\|^2 - \frac{\delta}{2} \|\nu^*(\delta)\|^2 \\
&= L_\delta(x^*, 1, \nu^*(\delta))
\end{aligned}
$$

Dividing by $\delta > 0$ implies

$$
\begin{aligned}
\|x^*(\delta)\|^2 + \|\nu^*(\delta)\|^2 \leq \frac{2}{\delta} &\left[L(x^*(\delta), 1, \nu^*) - L(x^*, 1, \nu^*(\delta)) \right] \\
&+ \|x^*\|^2 + \|\nu^*\|^2 < \infty
\end{aligned}
$$

This means that the left-hand side is uniformly bounded on δ, and, hence, if $\delta \to 0$, there exists a subsequence δ_k $(k \to \infty)$ on which there exist the limits

$$x^*(\delta_k) \to \tilde{x}^*, \quad \nu^*(\delta_k) \to \tilde{\nu}^* \quad \text{as} \quad k \to \infty$$

Suppose that there exist two limit points for two different convergent subsequences, i.e., there exist the limits

$$x^*(\delta_{k'}) \to \bar{x}^*, \quad \nu^*(\delta_{k'}) \to \bar{\nu}^* \quad \text{as} \quad k' \to \infty$$

Then for $\delta = \delta_k \to 0$ and $\delta = \delta_{k'} \to 0$ in (21.118) we have

$$0 \le (x^* - \tilde{x}^*, x^*) + (\nu^* - \tilde{\nu}^*, \nu^*)$$
$$0 \le (x^* - \bar{x}^*, x^*) + (\nu^* - \bar{\nu}^*, \nu^*)$$

From these inequalities it follows that points $(\tilde{x}^*, \tilde{\nu}^*)$ and $(\bar{x}^*, \bar{\nu}^*)$ correspond to the minimum point of the function

$$s(x^*, \nu^*) := \frac{1}{2}\left(\|x^*\|^2 + \|\nu^*\|^2\right)$$

defined for all possible saddle points of the nonregularized Lagrange function. But the function $s(x^*, \nu^*)$ is strictly convex, and, hence, its minimum is unique which gives $\tilde{x}^* = \bar{x}^*$, $\tilde{\nu}^* = \bar{\nu}^*$. Proposition is proven. □

Consider the following numerical procedure[2]

$$\boxed{\begin{aligned} x_{n+1} &= \pi_Q\left\{x_n - \gamma_n \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, \nu_n)\right\} \\ \nu_{n+1} &= \left[\nu_n + \gamma_n \frac{\partial}{\partial \nu} L_{\delta_n}(x_n, 1, \nu_n)\right]_+ \end{aligned}} \tag{21.119}$$

where the operator $[\cdot]_+$ acts from $\mathbb{R}^n$ into $\mathbb{R}^n$ as follows:

$$[x]_+ = \left([x_1]_+, \ldots, [x_n]_+\right)$$
$$[x_i]_+ := \begin{cases} x_i & \text{if} \quad x_i \ge 0 \\ 0 & \text{if} \quad x_i < 0 \end{cases} \tag{21.120}$$

Theorem 21.15. *Assume that*

1. *$f(x)$ and $g_i(x)$ $(i = 1, \ldots, m)$ are convex and differentiable in $\mathbb{R}^n$;*
2. *the estimates (x_n, ν_n) are generated by the procedure (21.119);*

[2] In Arrow *et al.* (1958) this procedure is considered with $\delta = 0$, that is why the corresponding convergence analysis looks incomplete.

3. *the step size* γ_n *and the regularizing parameter* δ_n *satisfy the following conditions:*

$$0 < \gamma_n \to 0,\ 0 < \delta_n \to 0 \quad as \quad n \to \infty$$

$$\sum_{n=0}^{\infty} \gamma_n \delta_n = \infty$$

$$\frac{\gamma_n}{\delta_n} \to \lambda \quad which\ is\ small\ enough,\ \frac{|\delta_{n+1} - \delta_n|}{\gamma_n \delta_n} \to 0 \quad as \quad n \to \infty$$

Then for any initial value $x_0 \in Q$ *and any* v_0 *with nonnegative components the vector sequences* $\{x_n\}$, $\{v_n\}$, *generated by the procedure (21.119), converge as* $n \to \infty$ *to* x^{**}, v^{**} *defined by (21.114).*

Proof. Using the property (21.64) of the projection operator we get the following recursion for the variable $w_n := \|x_n - x^*(\delta_n)\|^2 + \|v_n - v^*(\delta_n)\|^2$:

$$\begin{aligned}
w_{n+1} &\le \left\| x_n - \gamma_n \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) - x^*(\delta_{n+1}) \right\|^2 \\
&\quad + \left\| v_n + \gamma_n \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) - v^*(\delta_{n+1}) \right\|^2 \\
&= \left\| [x_n - x^*(\delta_n)] - \gamma_n \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) + [x^*(\delta_n) - x^*(\delta_{n+1})] \right\|^2 \\
&\quad + \left\| [v_n - v^*(\delta_n)] + \gamma_n \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) + [v^*(\delta_n) - v^*(\delta_{n+1})] \right\|^2 \\
&= w_n + \|x^*(\delta_n) - x^*(\delta_{n+1})\|^2 + \|v^*(\delta_n) - v^*(\delta_{n+1})\|^2 \\
&\quad - 2\gamma_n \left[\left(x_n - x^*(\delta_n), \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) \right) - \left(v_n - v^*(\delta_n), \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) \right) \right] \\
&\quad + 2 \left(x^*(\delta_n) - x^*(\delta_{n+1}), [x_n - x^*(\delta_n)] - \gamma_n \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) \right) \\
&\quad + 2 \left(v^*(\delta_n) - v^*(\delta_{n+1}), [v_n - v^*(\delta_n)] + \gamma_n \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) \right) \\
&\quad + \gamma_n^2 \left(\left\| \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) \right\|^2 + \left\| \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) \right\|^2 \right)
\end{aligned}$$

Taking into account that

$$\left\| \frac{\partial}{\partial x} L_{\delta_n}(x_n, 1, v_n) \right\|^2 \le c_0 + c_1 \|v_n\|^2 \le C_0 + C_1 w_n$$

$$\left\| \frac{\partial}{\partial v} L_{\delta_n}(x_n, 1, v_n) \right\|^2 \le c_2$$

and applying the inequalities (21.112) and (21.113) the last recursion can be estimated as follows:

$$\begin{aligned} w_{n+1} &\le w_n + c^2\,|\delta_{n+1}-\delta_n|^2 - \gamma_n\delta_n w_n \\ &\quad + 2\,\|x^*(\delta_n)-x^*(\delta_{n+1})\|\left(\|x_n-x^*(\delta_n)\|+\gamma_n\left(\sqrt{C_0}+\sqrt{C_1}\sqrt{w_n}\right)\right) \\ &\quad + 2\,\|v^*(\delta_n)-v^*(\delta_{n+1})\|\left(\|v_n-v^*(\delta_n)\|+\gamma_n\sqrt{c_2}\right) \\ &\quad + \gamma_n^2\left(C_0+C_1w_n+c_2\right) \le w_n\left(1-\gamma_n\delta_n+C_1\gamma_n^2\right)+c^2\,|\delta_{n+1}-\delta_n|^2 \\ &\quad + C_2\,|\delta_{n+1}-\delta_n|\sqrt{w_n}+C_3\,|\delta_{n+1}-\delta_n|+\gamma_n^2\left(C_0+c_2\right) \\ &\le w_n\left(1-\gamma_n\delta_n\left[1-C_1\gamma_n/\delta_n\right]\right)+C_2\,|\delta_{n+1}-\delta_n|\sqrt{w_n} \\ &\quad + C_4\,|\delta_{n+1}-\delta_n|+\gamma_n^2C_5 \end{aligned} \tag{21.121}$$

Here c_i $(i=1,2)$ and C_i $(i=0,\ldots,5)$ are positive constants. By Lemma 16.17 applied for the case $r=1/2$ and for

$$\alpha_n := \gamma_n\delta_n - C_1\gamma_n^2 = \gamma_n\delta_n\left[1-C_1\gamma_n/\delta_n\right] = \gamma_n\delta_n\left[1+o(1)\right]$$
$$\beta_n := C_4\,|\delta_{n+1}-\delta_n| + \gamma_n^2C_5,\ \ \bar{\delta}_n := C_2\,|\delta_{n+1}-\delta_n|$$

in view of the conditions of this theorem, we have

$$c \text{ is any large real number}, \quad b=0, \quad d=0$$

and, hence, $u(c)=0$ which proves the theorem. □

Corollary 21.10. *Within the class of numerical sequences*

$$\boxed{\begin{aligned} \gamma_n &:= \frac{\gamma_0}{(n+n_0)^\gamma}, \quad \gamma_0, n_0, \gamma > 0 \\ \delta_n &:= \frac{\delta_0}{(n+n_0)^\delta}, \quad \delta_0, \quad \delta > 0 \end{aligned}} \tag{21.122}$$

the conditions of Theorem 21.15 are satisfied if

$$\boxed{\gamma+\delta \le 1, \quad \gamma \ge \delta, \quad \gamma < 1} \tag{21.123}$$

Proof. It follows from the estimates that

$$\gamma_n\delta_n = O\left(\frac{1}{n^{\gamma+\delta}}\right)$$

$$\begin{aligned} |\delta_{n+1}-\delta_n| &= O\left(\frac{1}{n^\delta}-\frac{1}{(n+1)^\delta}\right) \\ &= O\left(\frac{1}{(n+1)^\delta}\left[\left(1+\frac{1}{n}\right)^\delta-1\right]\right) \end{aligned}$$

$$= O\left(\frac{1}{(n+1)^{\delta}}\left[\left(\frac{1}{n}\right)^{\delta} + o(1)\right]\right) = O\left(\frac{1}{n^{\delta+1}}\right)$$

and

$$\frac{|\delta_{n+1} - \delta_n|}{\gamma_n \delta_n} = O\left(\frac{1}{n^{1-\gamma}}\right)$$

which implies (21.123). Corollary is proven. □

Corollary 21.11. *Within the class (21.122) of the parameters of the procedure (21.119) the rate $\varkappa$ of convergence*

$$\|x_n - x^{**}\|^2 + \|v_n - v^{**}\|^2 = O\left(\frac{1}{n^{\varkappa}}\right)$$

is equal to

$$\boxed{\varkappa = \min\{\gamma - \delta;\ 1 - \gamma;\ \delta\}} \tag{21.124}$$

The maximal rate $\varkappa^*$ of convergence is attained for

$$\boxed{\gamma = \gamma^* = 2/3, \quad \delta = \delta^* = 1/3} \tag{21.125}$$

Proof. By Lemma 16.16 and in view of (21.121) it follows that for $\varkappa_0$ characterizing the rate of convergence

$$w_n := \|x_n - x^*(\delta_n)\|^2 + \|v_n - v^*(\delta_n)\|^2 = O\left(\frac{1}{n^{\varkappa_0}}\right)$$

we have $\varkappa_0 = \min\{\gamma - \delta;\ 1 - \gamma\}$. But, by (21.113), it follows that

$$\|x_n - x^{**}\|^2 + \|v_n - v^{**}\|^2 = w_n + O(\delta_n)$$

$$= O\left(\frac{1}{n^{\varkappa_0}}\right) + O\left(\frac{1}{n^{\delta}}\right) = O\left(\frac{1}{n^{\min\{\varkappa_0;\delta\}}}\right)$$

which implies (21.124). The maximal value $\varkappa^*$ of $\varkappa$ is attained when $\gamma - \delta = 1 - \gamma = \delta$, i.e., when (21.125) holds which completes the proof. □

Remark 21.7. *Many other numerical methods, solving the general nonlinear programing problem (21.83), are discussed in (Polyak 1987).*

22 Variational Calculus and Optimal Control

Contents

Since the fabric of the universe is most perfect, and is the work of a most wise Creator, nothing whatsoever takes place in the universe in which some form of maximum and minimum does not appear.

Leonard Euler, 1744.

22.1 Basic lemmas of variation calculus

The following lemmas represent the basic instrument for proving the main results of *variation calculus theory and optimal control theory* (see, for example, Gel'fand & Fomin (1961), Ivanov & Faldin (1981) and Troutman (1996)).

22.1.1 Du Bois–Reymond lemma

First, let us prove the following simple auxiliary result.

Lemma 22.1. *If* $0 \leq p \in C[a,b]$ *and* $\int_{x=a}^{b} p(x)\,dx = 0$, *then* $p(x) = 0$ *for all* $x \in [a,b]$.

Proof. Since $p(x) \geq 0$, for any $x \in [a,b]$ we have

$$0 \leq P(x) := \int_{t=a}^{x} p(t)\,dt \leq \int_{t=a}^{b} p(t)\,dt = 0$$

So, $P(x) \equiv 0$ on $[a,b]$, and, hence, $P'(x) \equiv 0$ too. This exactly means $p(x) = P'(x) \equiv 0$ which proves the statement of this lemma. □

Lemma 22.2. (Du Bois–Reymond) *If* $h \in C[a,b]$ *is a continuous on* $[a,b]$ *scalar function of the scalar argument and*

$$\int_{x=a}^{b} h(x)\, v'(x)\, dx = 0 \tag{22.1}$$

for all $v \in D_1 := \{v \in C^1[a,b] : v(a) = v(b) = 0\}$. *Then*

$$h(x) = c = \text{const } on \ [a,b]$$

Proof. For a constant c, the function $v(x) := \int_{t=a}^{x} [h(t) - c]\, dt$ is in $C^1[a,b]$ (it has a continuous derivative) and $v'(x) = h(x) - c$ so that $v(a) = 0$. It will be in D_1 if, additionally, it satisfies the condition $v(b) = 0$, that is, if $v(b) := \int_{t=a}^{b} [h(t) - c]\, dt = 0$, or $c = (b-a)^{-1} \int_{t=a}^{b} h(t)\, dt$. Thus, for these c and $v(x)$, in view of (22.1), we have

$$0 \le \int_{x=a}^{b} [h(x) - c]^2\, dx = \int_{x=a}^{b} [h(x) - c]\, v'(x)\, dx$$
$$= \int_{x=a}^{b} h(x)\, v'(x)\, dx - cv(x)\,|_{x=a}^{x=b} = 0$$

and, by Lemma 22.1, it follows that $[h(x) - c]^2 \equiv 0$ which completes the proof. □

The next lemma generalizes Lemma 22.2.

Lemma 22.3. *If* $h \in C[a,b]$ *and for some* $m = 1, 2, \ldots$

$$\int_{x=a}^{b} h(x)\, v^{(m)}(x)\, dx = 0$$

for all $v \in D_m$ *where*

$$D_m := \{v \in C^m[a,b] : v^{(k)}(a) = v^{(k)}(b) = 0,\ k = 0, 1, \ldots, m-1\}$$

Then on $[a,b]$ *the function* $h(x)$ *is a polynomial of a degree* $l < m$, *that is,*

$$h(x) = c_0 + c_1 x + \ldots + c_l x^l$$

Proof. By a translation, we may assume that $a = 0$. The function

$$H(x) := \int_{t_1=0}^{x} \left(\cdots \left(\int_{t_{m-1}=0}^{t_{m-2}} \left(\int_{t=0}^{t_{m-1}} h(t)\, dt \right) dt_{m-1} \right) \cdots \right) dt_1$$

is in $C^m[0, b]$ with the derivative $H^{(m)}(x) = h(x)$, and, fulfilling the identities

$$H(0) = H'(0) = \ldots = H^{(m-1)}(0) = 0$$

Then, if q is a polynomial of a degree $l < m$, then $P(x) := x^m q(x)$ vanishes at $x = 0$ together with $P^{(i)}(x)$ for $i < m$, while $p(x) := P^{(m)}(x)$ is another polynomial of the degree $l < m$. Define $v(x) := H(x) - P(x)$, so that $v^{(m)}(x) = h(x) - p(x)$. Show next that with the proper choice of $q(x)$ we can make $v^{(k)}(b) = 0$ $(k = 0, 1, \ldots, m-1)$. Supposing that this choice has been made, the resulting $v \in D_m$, and, moreover,

$$\int_{x=0}^{b} p(x) v^{(m)}(x)\, dx = -\int_{x=0}^{b} p'(x) v^{(m-1)}(x)\, dx$$
$$= \cdots = (-1)^m \int_{x=0}^{b} p^{(m)}(x) v(x)\, dx = 0$$

since the boundary term vanishes. By the assumptions of this lemma it follows that

$$0 \le \int_{x=0}^{b} [h(x) - p(x)]^2\, dx = \int_{x=0}^{b} [h(x) - p(x)] v^{(m)}(x)\, dx$$
$$= \int_{x=0}^{b} h(x) v^{(m)}(x)\, dx = 0$$

So, by Lemma 22.1, we get $h(x) = p(x)$ on $[0, b]$. Lemma is proven. □

Lemma 22.4. *If* $g, h \in C[a, b]$ *and*

$$\int_{x=a}^{b} [g(x) v(x) + h(x) v'(x)]\, dx = 0$$

for all $v \in D_1 := \{v \in C^1[a, b] : v(a) = v(b) = 0\}$. *Then* $h \in C^1[a, b]$ *and*

$$h'(x) = g(x) \quad \textit{for all} \quad x \in [a, b]$$

Proof. Denote $G(x) := \int_{t=a}^{x} g(t)\, dt$ for $x \in [a, b]$. Then $G \in C^1[a, b]$ and $G'(x) = g(x)$. The integration by part implies

$$0 = \int_{x=a}^{b} [g(x) v(x) + h(x) v'(x)]\, dx = \int_{x=a}^{b} [h(x) - G(x)] v'(x)\, dx$$
$$+ G'(x) v(x)\,|_{x=a}^{x=b} = \int_{x=a}^{b} [h(x) - G(x)] v'(x)\, dx$$

By Lemma 22.2, it follows that $h(x) - G(x) \equiv c = \text{const}$, so that

$$h(x) = G(x) + c, \quad h'(x) = G'(x) = g(x)$$

which proves the lemma. □

Corollary 22.1. *If* $g \in C[a, b]$ *and* $\int_{x=a}^{b} g(x)\, v(x)\, dx = 0$ *for all* $v \in D_1$, *then* $g(x) \equiv 0$ *on* $[a, b]$.

Proof. It is sufficient to put $h(x) \equiv 0$ in Lemma 22.4. □

This corollary also admits the following generalization.

22.1.2 Lagrange lemma

Lemma 22.5. (Lagrange) *If* $g \in C[a, b]$ *and for some* $m = 0, 1, \ldots$

$$\int_{x=a}^{b} g(x)\, v(x)\, dx = 0$$

for all $v \in D_m$ *where*

$$D_m := \left\{ v \in C^m[a, b] : v^{(k)}(a) = v^{(k)}(b) = 0,\ k = 0, 1, \ldots, m-1 \right\}$$

then $g(x) \equiv 0$ *on* $[a, b]$.

Proof. Suppose, by contradiction, that $g(c) > 0$ for some $c \in (a, b)$. Then by continuity, there exists an interval $[\alpha, \beta] \subseteq (a, b)$ which contains c and such that

$$|g(x) - g(c)| \leq g(c)/2$$

or, equivalently, $g(x) \geq g(c)/2 > 0$. On the other hand, the function

$$v(x) := \begin{cases} (x-a)(\beta - x)^{m+1} & \text{for} \quad x \in [\alpha, \beta] \\ 0 & \text{for} \quad x \notin [\alpha, \beta] \end{cases}$$

is in $C^m[a, b]$ and nonnegative. This implies that the product $g(x)\, v(x)$ is also continuous, nonnegative, and not identically zero. Thus, $\int_{x=a}^{b} g(x)\, v(x)\, dx > 0$, which contradicts the hypotheses of this lemma. Lemma is proven. □

The vector-valued version of Lemma 22.4 is also admitted.

Lemma 22.6. *If* $d = 1, 2, 3, \ldots$ *and* $G, H \in \left(C[a, b]\right)^n$ $(G(x), H(x) \in \mathbb{R}^n)$ *so that*

$$\int_{x=a}^{b} \left[(G(x), H(x)) + \left(H(x), V'(x)\right) \right] dx = 0$$

for any $V \in D_1 := \left\{v \in \left(C^1[a,b]\right)^n : V(a) = V(b) = \bar{0}\right\}$ *(here* $\bar{0}$ *is the zero-vector), then* $H \in \left(C^1[a,b]\right)^n$ *and* $H'(x) = G(x)$ *on* $[a,b]$.

Proof. It follows directly from Lemma 22.4 after its application to each individual component taking $V(x) = \left(\underbrace{0,\ldots,v(x)}_{i},0,\ldots,0\right)^{\mathrm{T}}$ for $i = 1,\ldots,n$. □

Corollary 22.2. *If* $H \in \left(C[a,b]\right)^n$ *and* $\int_{x=a}^{b}(G(x),H(x))\,dx = 0$ *for all* $V \in D_1$, *then* $H(x) = \text{const} \in \mathbb{R}^n$ *on* $[a,b]$.

Proof. It is sufficient to take $G(x) \equiv 0$ in Lemma 22.6. □

22.1.3 Lemma on quadratic functionals

Lemma 22.7. (Gel'fand & Fomin 1961) *If* $q, p \in C[a,b]$ *and*

$$I(y) = \int_{x=a}^{b}\left[q(x)\,y^2(x) + p(x)\left(y'(x)\right)^2\right]dx \geq 0 \tag{22.2}$$

for any function $y \in D_1 := \left\{y \in C^1[a,b] : y(a) = y(b) = 0\right\}$, *then* $p(x) \geq 0$ *on* $[a,b]$.

Proof. By the contradiction method, suppose that there exists $x_0 \in (a,b)$ such that $p(x_0) < 0$. Select the function

$$y(x) = y_\sigma(x) := \begin{cases} \sqrt{\sigma}\left(1 + \dfrac{x - x_0}{\sigma}\right) & \text{if} \quad x_0 - \sigma \leq x \leq x_0 \\ \sqrt{\sigma}\left(1 - \dfrac{x - x_0}{\sigma}\right) & \text{if} \quad x_0 \leq x \leq x_0 + \sigma \\ 0 & \text{if} \quad x \notin [x_0 - \sigma, x_0 + \sigma] \end{cases}$$

(here $\sigma > 0$ is small enough). Then $\left(y'_\sigma(x)\right)^2 = \sigma^{-1}$ and, by the mean-value theorem,

$$\left|\int_{x=a}^{b} q(x)\,y_\sigma^2(x)\,dx\right| = \left|\int_{x=x_0-\sigma}^{x_0+\sigma} q(x)\,y_\sigma^2(x)\,dx\right| \leq |q(x_0)|\,2\sigma^2$$

and

$$\int_{x=a}^{b} p(x)\left(y'_\sigma(x)\right)^2 dx = \int_{x=x_0-\sigma}^{x_0+\sigma} p(x)\left(y'_\sigma(x)\right)^2 dx = 2p(\bar{x})$$

where $\bar{x} \in [x_0 - \sigma, x_0 + \sigma]$. For small enough σ, so that $p(\bar{x}) < 0$, we have

$$I(y) \leq \left| \int_{x=a}^{b} q(x) y_\sigma^2(x) dx \right| + \int_{x=a}^{b} p(x) \left(y_\sigma'(x)\right)^2 dx$$
$$\leq |q(x_0)| 2\sigma^2 + 2p(\bar{x}) < 0$$

which contradicts the hypotheses of this lemma. So, $p(x_0) \geq 0$ for any internal point of the interval $[a, b]$. As for the boundary points a, b, the values $p(a), p(b)$ also cannot be negative since, if so, by the continuity property, $p(x)$ should be negative in a small internal neighborhood which is impossible. Lemma is proven. □

22.2 Functionals and their variations

Here we will briefly remember the main definitions of the first and second variation of functionals in some functional Banach space $\mathcal{B}$.

Definition 22.1. *The functional $J(y)$ defined in a Banach space $\mathcal{B}$ with a norm $\|\cdot\|_{\mathcal{B}}$ is said to be*

1. ***strongly differentiable** (in the **Fréchet sense**) at the "point" $y_0 \in \mathcal{B}$, if there exists a linear (with respect to variation $h \in \mathcal{B}$) functional $\varphi_1(y_0, h)$ such that for any $h \in \mathcal{B}$*

$$\begin{aligned} \Delta J(y_0, h) &:= J(y_0 + h) - J(y_0) \\ &= \varphi_1(y_0, h) + \alpha(y_0, h) \|h\|_{\mathcal{B}} \end{aligned} \tag{22.3}$$

where $\alpha(y_0, h) \to 0$ as $\|h\|_{\mathcal{B}} \to 0$;

2 ***twice strongly differentiable** (in the **Fréchet sense**) at the "point" $y_0 \in \mathcal{B}$, if $\Delta J(y_0, h)$ (22.3) can be represented as*

$$\Delta J(y_0, h) = \varphi_1(y_0, h) + \frac{1}{2}\varphi_2(y_0, h) + \alpha(y_0, h) \|h\|_{\mathcal{B}}^2 \tag{22.4}$$

*where $\varphi_1(y_0, h)$ is a **linear** (with respect to the variation $h \in \mathcal{B}$) functional, $\varphi_2(y_0, h)$ is a **quadratic functional** with respect to the variation $h \in \mathcal{B}$, i.e., for any $\lambda_1, \lambda_2 \in \mathbb{R}$ and any $h, h_1, h_2, \tilde{h}, \tilde{h}_1, \tilde{h}_2 \in \mathcal{B}$*

$$\begin{aligned} &\varphi_2(y_0, h) := \tilde{\varphi}_2\left(y_0, h, \tilde{h}\right)|_{h=\tilde{h}\in\mathcal{B}} \\ &\tilde{\varphi}_2\left(y_0, \lambda_1 h_1 + \lambda_2 h_2, \tilde{h}\right) = \lambda_1 \tilde{\varphi}_2\left(y_0, h_1, \tilde{h}\right) + \lambda_2 \tilde{\varphi}_2\left(y_0, h_2, \tilde{h}\right) \\ &\tilde{\varphi}_2\left(y_0, h, \lambda_1 \tilde{h}_1 + \lambda_2 \tilde{h}_2\right) = \lambda_1 \tilde{\varphi}_2\left(y_0, h, \tilde{h}_1\right) + \lambda_2 \tilde{\varphi}_2\left(y_0, h, \tilde{h}_2\right) \end{aligned}$$

and, again, $\alpha(y_0, h) \to 0$ as $\|h\|_{\mathcal{B}} \to 0$.

Definition 22.2. *The functionals $\varphi_1(y_0, h)$ and $\varphi_2(y_0, h)$, defined above, are called the* ***first and second strong differentials*** *of $J(y)$ and are denoted (according to (18.93)) by*

$$d\Phi(x_0 \mid h) = \langle h, \Phi'(x_0)\rangle := \varphi_1(y_0, h)$$
$$d^2\Phi(x_0 \mid h) := \varphi_2(y_0, h) \tag{22.5}$$

There exist other differentials, namely, weak differentials (in the Gâteaux sense). For details concerning these functionals and their relation with strong ones see section 18.7.2. Below we shall use only Fréchet differentials.

22.3 Extremum conditions

22.3.1 Extremal curves

The, so-called, *variation principle* (see Theorem 18.18) will be actively used in this section for the solution of various problems of variation calculus theory.

Definition 22.3.

1. *A functional $J(y)$, defined in a Banach space $\mathcal{B}$ with a norm $\|\cdot\|_{\mathcal{B}}$, has a* ***local extremum in a region*** *$\mathcal{G}$ (defining some additional constraints to the class of admissible functions) at the curve (function) $y_0 \in \mathcal{B} \cap \mathcal{G}$, if there exists a neighborhood*

$$\Omega_\delta := \{y \in \mathcal{B} \cap \mathcal{G} \mid \|y - y_0\|_{\mathcal{B}} < \delta\}$$

such that for all $y \in \Omega_\delta$ one has $J(y) \geq J(y_0)$. The function y_0 is said to be an ***extremal curve.***
2. *If $J(y) \geq J(y_0)$ for all $y \in \mathcal{B} \cap \mathcal{G}$, then the extremal curve y_0 is said to be a* ***global extremum*** *of the functional $J(y)$ on $\mathcal{B} \cap \mathcal{G}$.*

22.3.2 Necessary conditions

Reformulate here Theorem 18.18 for the case of the Fréchet differential existence.

Theorem 22.1. (on the necessary conditions) *Let the curve $y_0 \in \operatorname{int}(\mathcal{B} \cap \mathcal{G})$ be a local extremal (minimal) curve of the functional $J(y)$ which is assumed to be strongly (Fréchet) differentiable at the "point" y_0.*

1. ***(The first-type necessary conditions)*** *Then for any admissible $h \in \mathcal{B} \cap \mathcal{G}$*

$$\boxed{\varphi_1(y_0, h) \equiv 0} \tag{22.6}$$

2. ***(The second-type necessary conditions)*** *If, additionally, $J(y)$ is twice strongly (Fréchet) differentiable at the "point" y_0, then for any admissible $h \in \mathcal{B} \cap \mathcal{G}$*

$$\boxed{\varphi_2(y_0, h) \geq 0} \tag{22.7}$$

Proof. Let $J(y) \geq J(y_0)$ for all y within some Ω_δ.

1. Then, by Definition 22.3 and in view of the property (22.3), for any $(y_0 + h) \in \Omega_\delta$ we have

$$\Delta J(y_0, h) := J(y_0 + h) - J(y_0) = \varphi_1(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}} \geq 0$$

If $\varphi_1(y_0, h_0) < 0$ for some admissible h_0, then for small enough α it follows that

$$\Delta J(y_0, h) = \varphi_1(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}} < 0$$

which contradicts the optimality of y_0 in Ω_δ. Suppose that $\varphi_1(y_0, h) > 0$ for some admissible h_0. Since the strong (Fréchet) differential is linear on h and $y_0 \in \text{int}(\mathcal{B} \cap \mathcal{G})$, it follows that $(y_0 - h_0) \in \Omega_\delta$ and, therefore,

$$\varphi_1(y_0, -h_0) = -\varphi_1(y_0, h_0) < 0$$

and, as the result, again

$$\Delta J(y_0, h) = -\varphi_1(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}} < 0$$

for small enough α, which contradicts the optimality of y_0 in Ω_δ.

2. According to the first-type necessary condition (22.6) in view of the optimality of the curve $y_0 \in \text{int}(\mathcal{B} \cap \mathcal{G})$ we have

$$\Delta J(y_0, h) = \frac{1}{2}\varphi_2(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}}^2 \geq 0$$

Again, if for some admissible h_0 $((y_0 - h_0) \in \Omega_\delta)$ we suppose $\varphi_2(y_0, h_0) < 0$, then for small enough $\varepsilon > 0$ we have

$$\begin{aligned}\Delta J(y_0, \varepsilon h_0) &= \frac{1}{2}\varphi_2(y_0, \varepsilon h_0) + \alpha(y_0, \varepsilon h_0)\|\varepsilon h_0\|_{\mathcal{B}}^2 \\ &= \varepsilon^2\left[\frac{1}{2}\varphi_2(y_0, h_0) + \alpha(y_0, \varepsilon h_0)\|h_0\|_{\mathcal{B}}^2\right] < 0\end{aligned}$$

since $\alpha(y_0, \varepsilon h_0) \to 0$ whereas $\varepsilon \to 0$ for any h_0. But the last inequality contradicts the condition of optimality of the curve y_0. □

22.3.3 *Sufficient conditions*

Theorem 22.2. (on the sufficient conditions) *Let*

1. *the functional $J(y)$ be twice strongly (Fréchet) differentiable in $\mathcal{B} \cap \mathcal{G}$;*
2. *for some $y_0 \in \mathcal{B} \cap \mathcal{G}$ and any admissible h : $(y_0 - h) \in \mathcal{B} \cap \mathcal{G}$, $h \in \Omega_\delta$*

$$\boxed{\varphi_1(y_0, h) \equiv 0}$$

and

$$\boxed{\varphi_2(y_0, h) \geq k \|h\|_{\mathcal{B}}^2, \quad k > 0}$$

Then y_0 is the unique local minimal curve of the functional $J(y)$ on $\mathcal{B} \cap \mathcal{G}$.

Proof.

1. By (22.4) and in view of the condition of this theorem, for any admissible h

$$\Delta J(y_0, h) = \frac{1}{2}\varphi_2(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}}^2$$

For any small enough $\varepsilon > 0$ it follows that

$$\begin{aligned} \Delta J(y_0, \varepsilon h) &= \frac{1}{2}\varphi_2(y_0, \varepsilon h) + \alpha(y_0, \varepsilon h)\|\varepsilon h\|_{\mathcal{B}}^2 \\ &= \varepsilon^2\left[\frac{1}{2}\varphi_2(y_0, h) + \alpha(y_0, \varepsilon h)\|h\|_{\mathcal{B}}^2\right] \\ &\geq \varepsilon^2\left[\frac{1}{2}k\|h\|_{\mathcal{B}}^2 + \alpha(y_0, \varepsilon h)\|h\|_{\mathcal{B}}^2\right] \\ &= \varepsilon^2\|h\|_{\mathcal{B}}^2\left[\frac{1}{2}k + \alpha(y_0, \varepsilon h)\right] \geq 0 \end{aligned} \tag{22.8}$$

since $\alpha(y_0, \varepsilon h) \to 0$ whereas $\varepsilon \to 0$ for any admissible h. This means that y_0 is a local extremal (minimum) curve.

2. Suppose that in Ω_δ there exists two extremal curves y_0 and y_0' such that $J(y_0) = J(y_0') \leq J(y)$. Then, taking $h := y_0' - y_0$ in (22.8), we get

$$\begin{aligned} \Delta J(y_0, h) = 0 &= \frac{1}{2}\varphi_2(y_0, h) + \alpha(y_0, h)\|h\|_{\mathcal{B}}^2 \\ &\geq \|h\|_{\mathcal{B}}^2\left[\frac{1}{2}k + \alpha(y_0, h)\right] \end{aligned}$$

For small enough δ and any $h \in \Omega_\delta$, we have $\frac{1}{2}k + \alpha(y_0, h) > 0$ which together with the previous inequality leads to the following conclusions: $0 = \|h\|_{\mathcal{B}}^2$, or, equivalently, $y_0' = y_0$. Theorem is proven. □

22.4 Optimization of integral functionals

In this section we will consider the following *main problem* of variation calculus:

$$\boxed{\int_{x=a}^{b} F(x, y, y')\,dx \to \min_{y \in C^1[a,b]}} \tag{22.9}$$

where the function $F : \mathbb{R}^3 \to \mathbb{R}$ is assumed to be *twice differentiable* in all arguments.

22.4.1 Curves with fixed boundary points

22.4.1.1 Scalar case

Here, we will additionally suppose that we are looking for the extremum (minimum) of the integral function in (22.9) over all continuously differentiable curves $y \in C^1[a, b]$ satisfying the following boundary conditions:

$$\boxed{y(a) = \alpha, \quad y(b) = \beta} \tag{22.10}$$

The necessary conditions for this problem are given below.

1. *The first-type necessary condition* (22.6):

$$\begin{aligned} 0 = \varphi_1(y, h) &= \int_{x=a}^{b} \left(\frac{\partial}{\partial y} F h + \frac{\partial}{\partial y'} F h' \right) dx \\ &= \int_{x=a}^{b} \left(\frac{\partial}{\partial y} F - \frac{d}{dx} \frac{\partial}{\partial y'} F \right) h \, dx \end{aligned} \tag{22.11}$$

   where the variation curves $h \in C^1[a, b]$ and satisfies the boundary conditions $h(a) = h(b)$.

2. *The second-type necessary condition* (22.7):

$$\begin{aligned} 0 &\le \varphi_2(y, h) \\ &= \int_{x=a}^{b} \left[\frac{\partial^2}{\partial y^2} F h^2 + 2 \frac{\partial^2}{\partial y \partial y'} F h h' + \frac{\partial^2}{\partial y' \partial y'} F \left(h'\right)^2 \right] dx \\ &= \int_{x=a}^{b} \left[\left(\frac{\partial^2}{\partial y^2} F - \frac{d}{dx} \frac{\partial^2}{\partial y \partial y'} F \right) h^2 + \frac{\partial^2}{\partial y' \partial y'} F \left(h'\right)^2 \right] dx \end{aligned} \tag{22.12}$$

   where $y(x)$ satisfies (22.11) and the variation curves $h \in C^1[a, b]$ and fulfills the boundary conditions $h(a) = h(b)$.

In (22.11) and (22.12) the following relations obtained by the integration by parts are used:

$$\begin{aligned} \int_{x=a}^{b} \frac{\partial}{\partial y'} F h' \, dx &= \left(\frac{\partial}{\partial y'} F \right) h \, |_{x=a}^{x=b} - \int_{x=a}^{b} \left(\frac{d}{dx} \frac{\partial}{\partial y'} F \right) h \, dx \\ &= \int_{x=a}^{b} \left(-\frac{d}{dx} \frac{\partial}{\partial y'} F \right) h \, dx \end{aligned}$$

and

$$\int_{x=a}^{b} 2\frac{\partial^2}{\partial y\,\partial y'}F hh'dx = \left(\frac{\partial^2}{\partial y\,\partial y'}F\right)h^2\Big|_{x=a}^{x=b} - \int_{x=a}^{b}\left(\frac{d}{dx}\frac{\partial^2}{\partial y\,\partial y'}F\right)h^2 dx$$
$$= \int_{x=a}^{b}\left(-\frac{d}{dx}\frac{\partial^2}{\partial y\,\partial y'}F\right)h^2\,dx$$

Theorem 22.3. (Euler–Lagrange) *The **first-type necessary condition** for a curve $y \in C^1[a,b]$, satisfying (22.10), to be an extremal curve is:*

$$\boxed{\frac{\partial}{\partial y}F(x,y,y') - \frac{d}{dx}\frac{\partial}{\partial y'}F(x,y,y') = 0} \tag{22.13}$$

Proof. It follows directly from (22.11) if we apply Lemma 22.4. □

The condition (22.13) is referred to as the *Euler–Lagrange condition.*

Theorem 22.4. (Legendre) *The **second-type necessary condition** for a curve $y \in C^1[a,b]$, satisfying (22.10) and (22.13), to be an extremal (minimizing) curve is:*

$$\boxed{\frac{\partial^2}{\partial y'\,\partial y'}F(x,y,y') \geq 0} \tag{22.14}$$

Proof. It follows directly from (22.12) if we apply Lemma 22.7. □

The condition (22.14) is often referred to as *the Legendre condition.*

The *sufficient conditions*, guaranteeing that a curve $y(x)$ is minimizing for the functional $J(y)$, may be formulated in the following manner.

Theorem 22.5. (Jacobi) *If for some curve $y \in C^1[a,b]$, verifying (22.10), the following conditions are fulfilled:*

1. *it satisfies the **Euler–Lagrange necessary condition** (22.13);*
2. *it satisfies the **strong Legendre condition***

$$\boxed{\frac{\partial^2}{\partial y'\,\partial y'}F(x,y,y') \geq k > 0} \tag{22.15}$$

3. *there exists a function $u \in C^1[a,b]$, which is not equal to zero on $[a,b]$, and satisfying the next ODE (the **Jacobi ODE**):*

$$\boxed{\begin{aligned} &Qu - \frac{d}{dx}(Pu') = 0 \\ &Q := \frac{\partial^2}{\partial y^2}F - \frac{d}{dx}\frac{\partial^2}{\partial y\,\partial y'}F, \quad P := \frac{\partial^2}{\partial y'\partial y'}F \end{aligned}} \tag{22.16}$$

then this curve provides the local minimum to the functional $J(y)$.

Proof. It follows directly from Theorem 22.2 if we define $\|h\|_{\mathcal{B}}^2$ for $\mathcal{B} = C^1[a,b]$ (see (18.6)) as

$$\|h\|_{C^1[a,b]} := \max_{x\in[a,b]} |h(t)| + \max_{x\in[a,b]} |h'(x)|$$

Indeed, by the assumptions of this theorem, we have

$$\begin{aligned}
\varphi_2(y,h) &= \int_{x=a}^{b} \left[Qh^2 + P\left(h'\right)^2\right] dx = \int_{x=a}^{b} \left[(Qu)\frac{h^2}{u} + P\left(h'\right)^2\right] dx \\
&= \int_{x=a}^{b} \left[\frac{d}{dx}\left(Pu'\right)\frac{h^2}{u} + P\left(h'\right)^2\right] dx = \left(Pu'\frac{h^2}{u}\right)\Big|_{x=a}^{x=b} \\
&\quad + \int_{x=a}^{b} \left[-Pu'\frac{d}{dx}\left(\frac{h^2}{u}\right) + P\left(h'\right)^2\right] dx = \int_{x=a}^{b} P\left[\left(h'\right)^2 - u'\frac{d}{dx}\left(\frac{h^2}{u}\right)\right] dx \\
&= \int_{x=a}^{b} P\left[\left(h'\right)^2 - 2h'\left(\frac{hu'}{u}\right) + \frac{h^2}{u^2}\left(u'\right)^2\right] dx = \int_{x=a}^{b} P\left(h' - \frac{hu'}{u}\right)^2 dx \\
&\geq k\int_{x=a}^{b} \left(h' - \frac{hu'}{u}\right)^2 dx = k\|h\|_{C^1[a,b]}^2 \int_{x=a}^{b} \left(\tilde{h}' - \frac{\tilde{h}u'}{u}\right)^2 dx
\end{aligned}$$

where $\tilde{h} := h/\|h\|_{C^1[a,b]}$ satisfies

$$\left\|\tilde{h}\right\|_{C^1[a,b]} \leq 1, \quad \left\|\tilde{h}'\right\|_{C^1[a,b]} \leq 1$$

Notice also that

$$\varkappa := \inf_{\tilde{h}:\|\tilde{h}\|_{C^1[a,b]}\leq 1,\ \|\tilde{h}'\|_{C^1[a,b]}\leq 1,\ \tilde{h}(a)=\tilde{h}(b)=0} \int_{x=a}^{b} \left(\tilde{h}' - \frac{\tilde{h}u'}{u}\right)^2 dx > 0$$

since, if not, one has $\tilde{h}' - \dfrac{\tilde{h}u'}{u} = 0$, and, as a result, $\tilde{h} = u$. But, by the assumption of this theorem, $u(a) \neq 0$ and $u(b) \neq 0$, which contradicts with $\tilde{h}(a) = \tilde{h}(b) = 0$. So, finally, it follows that

$$\varphi_2(y,h) \geq k\|h\|_{C^1[a,b]}^2 \int_{x=a}^{b} \left(\tilde{h}' - \frac{\tilde{h}u'}{u}\right)^2 dx \geq k\varkappa\|h\|_{C^1[a,b]}^2$$

Theorem is proven. □

Example 22.1. ("Brachistochrone" problem) *The problem (formulated by Johann Bernoulli, 1696) consists of finding the curve AB (see Fig. 22.1) such that, during the sliding over of this curve in a gravity field with the initial velocity equal to zero, the*

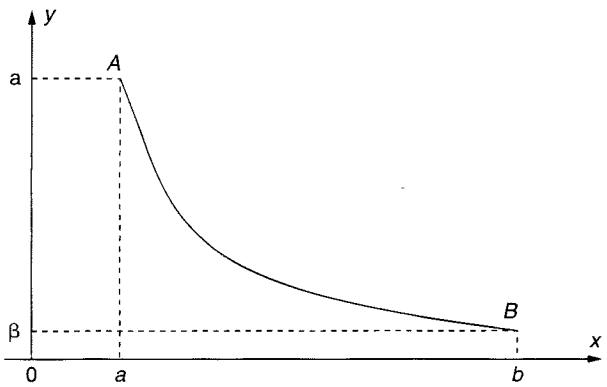

Fig. 22.1. Illustration of the Brachistochrone problem.

material point of the mass m can realize the sliding from the initial point A to the final point B in the shortest time.

So, we have to minimize $J(y) = T(y)$ *where* $y(x)$ *is the altitude of the point at the curve at the point* $x \in [a, b]$ *which satisfies the relation*

$$mgy + \frac{mv^2}{2} = E = \text{const}$$

Since, by the initial condition, when $y = y(a) = \alpha$ *we have* $v = 0$ *which implies* $E = mg\alpha$. *That is why*

$$v = \sqrt{2\left(\frac{E}{m} - gy\right)} = \sqrt{2g(\alpha - y)}$$

$$v = \frac{ds}{dt} = \sqrt{1 + (y')^2}\,\frac{dx}{dt}$$

which leads to the following formula for the functional $T(y)$:

$$T(y) = \int_{x=a}^{b} \frac{dx}{v} = \frac{1}{\sqrt{2g}} \int_{x=a}^{b} \sqrt{\frac{1 + (y')^2}{(\alpha - y)}}\,dx$$

Since the function $F(y, y') = \sqrt{\dfrac{1 + (y')^2}{(\alpha - y)}}$ *does not depend on* x, *the first-type Euler–Lagrange condition (22.13) for this functional has the* ***first integral***

$$F(y, y') - y' \frac{\partial}{\partial y'} F(y, y') = \frac{1}{c} \tag{22.17}$$

Indeed,

$$0 = \frac{\partial}{\partial y} F - \frac{d}{dx}\frac{\partial}{\partial y'} F = \frac{\partial}{\partial y} F - \left(\frac{\partial^2}{\partial y\,\partial y'} F\right) y' - \left(\frac{\partial^2}{\partial y'\,\partial y'} F\right) y''$$

Multiplying both sides by y' we get

$$0 = \frac{d}{dx}\left(F - y'\frac{\partial}{\partial y'}F\right)$$

which implies (22.17). In our case it is equivalent to the following ODE:

$$\frac{1}{c} = \sqrt{\frac{1+(y')^2}{(\alpha - y)}} - (y')^2\frac{\left(1+(y')^2\right)^{-1/2}}{\sqrt{(\alpha - y)}}$$
$$= \frac{\sqrt{1+(y')^2}\left(1 - \dfrac{(y')^2}{1+(y')^2}\right)}{\sqrt{(\alpha - y)}} = \frac{1}{\sqrt{(\alpha - y)}\sqrt{1+(y')^2}}$$

Squaring gives

$$c^2 = (\alpha - y)\left[1 + (y')^2\right]$$

or,

$$1 = y'\sqrt{\frac{(\alpha - y)}{c^2 - (\alpha - y)}} \tag{22.18}$$

With the introduction of the dependent variable $\theta = \theta(x)$ such that

$$(\alpha - y) = c^2\sin^2\frac{\theta}{2} = \frac{c^2}{2}(1 - \cos\theta), \quad \theta \in [0, 2\pi]$$

then

$$c^2 - (\alpha - y) = c^2\cos^2\frac{\theta}{2}, \quad y' = c^2\theta'\sin\frac{\theta}{2}\cos\frac{\theta}{2}$$

By substitution of these expressions into (22.18) yields

$$1 = c^2\theta'\sin^2\frac{\theta}{2} \quad or \quad 1 = \frac{c^2}{2}\theta'(1 - \cos\theta)$$

Integrating gives $\dfrac{c^2}{2}[\theta - \sin\theta] = x - c_1$. *Denoting* $c_2 := \dfrac{c^2}{2}$, *we get the* ***parametric (Brachistochrone) curve*** *(of the cycloid type)*

$$\boxed{\begin{aligned} x &= c_2[\theta - \sin\theta] + c_1, \quad c_2 > 0 \\ y &= \alpha - c_2(1 - \cos\theta), \quad \theta \in [0, 2\pi] \end{aligned}} \tag{22.19}$$

The constants c_1 and c_2 can be found from the boundary conditions

$$\theta = \theta_A = 0 : A = c_1, \quad y(A) = \alpha$$
$$\theta = \theta_B : B = A + c_2[\theta_B - \sin\theta_B], \quad y(B) = \beta = \alpha - c_2(1 - \cos\theta_B)$$

Example 22.2. *Let us try to answer the following question: which stable linear system of the first order, given by*

$$\dot{x}(t) = ax(t), \quad x(0) = x_0, \quad x(\infty) = 0 \tag{22.20}$$

provides the minimum of the functional

$$J(x) := \int_{t=0}^{\infty} \left[\alpha_0 x^2(t) + \alpha_1 \dot{x}^2(t)\right] dt \tag{22.21}$$

where $\alpha_0, \alpha_1 > 0$.

1. ***Variation calculus application.*** *The first-type necessary condition of the optimality for this functional is*

$$\alpha_0 x(t) - \alpha_1 \ddot{x}(t) = 0$$

since $F = \alpha_0 x^2 + \alpha_1 \dot{x}^2$. *Its general solution is*

$$x(t) = c_1 e^{-kt} + c_2 e^{kt}, \quad k := \sqrt{\alpha_0/\alpha_1}$$

Taking into account the boundary conditions we get: $c_2 = 0$, $c_1 = x_0$ *which gives* $x(t) = x_0 e^{-kt}$, *or, equivalently,* $\dot{x}(t) = -kx(t)$. *So, the optimal* $a = a^*$ *in (22.20) is* $a = a^* = -\sqrt{\alpha_0/\alpha_1}$. *Let us show that the obtained curve is minimizing. To do this we need to check conditions (2) and (3) of the Jacobi theorem 22.5. First, notice that* $\dfrac{\partial^2}{\partial \dot{x}\, \partial \dot{x}} F = 2\alpha_1 > 0$. *So, condition (2) is fulfilled. The Jacobi equation (22.16) is as follows*

$$\alpha_0 u - \alpha_1 u'' = 0$$

$$Q := \frac{\partial^2}{\partial y^2} F - \frac{d}{dx} \frac{\partial^2}{\partial y\, \partial y'} F = 2\alpha_0, \quad P := \frac{\partial^2}{\partial y'\, \partial y'} F = 2\alpha_1$$

It has a nontrivial solution $u(t) = u_0 e^{-kt} > 0$ *or* $u_0 > 0$. *This means that condition (3) is also valid. So, the curve* $\dot{x}(t) = -kx(t)$, $x(0) = x_0$ *is minimizing.*

2. ***Direct method.***[1] *Assuming that the minimal value of the cost functional (22.21) is finite, i.e.,* $x(t) \to 0$ *with* $t \to \infty$, *we can represent it in the following equivalent form*

$$J(x) := \int_{t=0}^{\infty} \left[\alpha_0 x^2(t) + 2\sqrt{\alpha_0 \alpha_1}\, x(t)\, \dot{x}(t) + \alpha_1 \dot{x}^2(t)\right] dt$$

$$- 2\sqrt{\alpha_0 \alpha_1} \int_{t=0}^{\infty} x(t)\, \dot{x}(t)\, dt = \int_{t=0}^{\infty} \left[\sqrt{\alpha_0}\, x(t) + \sqrt{\alpha_1}\, \dot{x}(t)\right]^2 dt$$

$$+ \sqrt{\alpha_0 \alpha_1}\, x^2(0) \geq \sqrt{\alpha_0 \alpha_1}\, x^2(0)$$

[1] The author has been informed of this elegant and simple solution by Prof. V. Utkin.

The last inequality becomes the equality when

$$\sqrt{\alpha_0}x\,(t) + \sqrt{\alpha_1}\dot{x}\,(t) = 0$$

or, equivalently, when

$$\dot{x}\,(t) = -kx\,(t), \quad k = \sqrt{\alpha_0/\alpha_1}$$

which coincides with the previous result.

22.4.1.2 Vector case

Consider the following optimization problem

$$\boxed{\int_{x=a}^{b} F\left(x, y_1, \ldots y_n, y_1', \ldots, y_n'\right) dx \to \min_{y_i \in C^1[a,b]}} \tag{22.22}$$

where the function $F : \mathbb{R}^{2n+1} \to \mathbb{R}$ is assumed to be *twice differentiable* in all arguments, and the functions y_i satisfy the following boundary conditions

$$\boxed{y_i\,(a) = \alpha_i, \quad y_i\,(b) = \beta_i \;\; (i = 1, \ldots, n)} \tag{22.23}$$

Theorem 22.6. (The Euler–Lagrange vector form) *The **first-type necessary condition** for curve $y_i \in C^1\,[a, b]$, satisfying (22.23), to be extremal curves is:*

$$\boxed{\begin{array}{l} \dfrac{\partial}{\partial y_i} F\,(x, \mathbf{y}, \mathbf{y}') - \dfrac{d}{dx}\dfrac{\partial}{\partial y_i'} F\,(x, \mathbf{y}, \mathbf{y}') = 0 \\ \mathbf{y} := (y_1, \ldots y_n)^{\mathsf{T}}, \quad \mathbf{y}' := \left(y_1', \ldots y_n'\right)^{\mathsf{T}} \end{array}} \tag{22.24}$$

Proof. It follows directly from the identity

$$\begin{aligned} 0 = \varphi_1\,(y, h) &= \int_{x=a}^{b} \sum_{i=1}^{n} \left(\frac{\partial}{\partial y_i} F h_i + \frac{\partial}{\partial y_i'} F h_i' \right) dx \\ &= \int_{x=a}^{b} \sum_{i=1}^{n} \left(\frac{\partial}{\partial y_i} F_i - \frac{d}{dx}\frac{\partial}{\partial y_i'} F \right) h_i\, dx \end{aligned}$$

if we take into account the independence of the variation functions h_i $(i = 1, \ldots, n)$, and apply Lemma 22.4. □

Theorem 22.7. (The Legendre vector form) *The **second-type necessary condition** for curve $y_i \in C^1\,[a, b]$, satisfying (22.23) and (22.24), to be extremal curves is:*

$$\boxed{\nabla^2_{\mathbf{y}',\mathbf{y}'} F\,(x, \mathbf{y}, \mathbf{y}') \geq 0} \tag{22.25}$$

where the matrix in (22.25) is defined by

$$\nabla^2_{\mathbf{y}',\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right) := \left[\frac{\partial^2}{\partial y'_i \, \partial y'_j} F\left(x, \mathbf{y}, \mathbf{y}'\right)\right]_{i,j=1,\ldots,n}$$

Proof. It follows directly from (22.7) and the relation

$$\begin{aligned} 0 \le \varphi_2(y, h) = \int_{x=a}^{b} & \left[\left(\nabla^2_{\mathbf{y},\mathbf{y}} F\left(x, \mathbf{y}, \mathbf{y}'\right)\mathbf{h}, \mathbf{h}\right)\right. \\ & \left. + 2\left(\nabla^2_{\mathbf{y},\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)\mathbf{h}, \mathbf{h}'\right) + \left(\nabla^2_{\mathbf{y}',\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)\mathbf{h}', \mathbf{h}'\right)\right] dx \\ = \int_{x=a}^{b} & \left[\left(\left[\nabla^2_{\mathbf{y},\mathbf{y}} F\left(x, \mathbf{y}, \mathbf{y}'\right) - \frac{d}{dx}\nabla^2_{\mathbf{y},\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)\right]\mathbf{h}, \mathbf{h}\right)\right. \\ & \left. + \left(\nabla^2_{\mathbf{y}',\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)\mathbf{h}', \mathbf{h}'\right)\right] dx \end{aligned}$$

if we apply the vector version of Lemma 22.7. □

The next theorem gives the *sufficient conditions* for the vector function $\mathbf{y}$ to be a minimizer in the problem (22.22).

Theorem 22.8. (The Jacobi vector form) *If for some vector function* $\mathbf{y} \in \left(C^1[a, b]\right)^n$, *verifying (22.23), the following conditions are fulfilled:*

1. *it satisfies the* ***Euler–Lagrange necessary condition*** *(22.24);*
2. *it satisfies the* ***strong Legendre vector condition***

$$\boxed{\nabla^2_{\mathbf{y}',\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right) \ge k I_{n\times n}, \quad k > 0} \tag{22.26}$$

3. *there exists vector functions* $\mathbf{u}^i \in \left(C^1[a, b]\right)^n \ (i = 1, \ldots, n)$, *such that*

$$\boxed{\begin{array}{l} \det\left[\mathbf{u}^1(x) \cdots \mathbf{u}^n(x)\right] \ne 0 \quad \textit{for all} \quad x \in [a, b] \\ \mathbf{u}^i(a) = 0, \ \dfrac{d}{dx}\mathbf{u}^i(a) = \mathbf{e}^i := \left(\underbrace{0, \ldots, 1}_{i}, 0, \ldots, 0\right)^{\mathsf{T}} \end{array}} \tag{22.27}$$

and satisfying the next ODE (the ***Jacobi vector form ODE****):*

$$\boxed{\mathbf{Q}\mathbf{u} - \frac{d}{dx}\left(\mathbf{P}\mathbf{u}'\right) = 0} \tag{22.28}$$

with

$$\mathbf{Q} := \nabla^2_{\mathbf{y},\mathbf{y}} F\left(x, \mathbf{y}, \mathbf{y}'\right) - \frac{d}{dx}\nabla^2_{\mathbf{y},\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)$$

$$\mathbf{P} := \nabla^2_{\mathbf{y}',\mathbf{y}'} F\left(x, \mathbf{y}, \mathbf{y}'\right)$$

then this vector function provides the local minimum to the functional $J(y)$.

Proof. It follows directly from the vector form of Theorem 22.2 if we define $\|\mathbf{h}\|^2_{\mathcal{B}}$ for $\mathcal{B} = \left(C^1[a,b]\right)^n$ (see (18.6)) as

$$\|\mathbf{h}\|^2_{\left(C^1[a,b]\right)^n} := \sum_{i=1}^{n} \left(\max_{x\in[a,b]} \left|h^i(t)\right| + \max_{x\in[a,b]} \left|\frac{d}{dx}h^i(x)\right| \right)$$

and, practically, repeat the proof of Theorem 22.5. □

22.4.1.3 Integral functional depending on derivatives of an order more than one

Let us consider the first-type necessary condition of the optimality for the integral functions

$$\boxed{J(y) = \int_{x=a}^{b} F(x, y, y^{(1)}, \ldots, y^{(n)})\, dx} \tag{22.29}$$

within the functions $y \in C^n[a,b]$, satisfying the boundary conditions

$$\boxed{y^{(i)}(a) = \alpha_i, \quad y^{(i)}(b) = \beta_i \ (i = 1, \ldots, n)} \tag{22.30}$$

First, notice that

$$\varphi_1(y,h) = \int_{x=a}^{b} \left[\sum_{i=0}^{n} \frac{\partial F}{\partial y^{(i)}} h^{(i)}\right] dx, \quad y^{(0)} := y$$

where the variation functions $h^{(i)}$ satisfy the conditions

$$h^{(i)}(a) = h^{(i)}(b) = 0 \ (i = 1, \ldots, n)$$

Integrating each integral term $\int_{x=a}^{b} \frac{\partial F}{\partial y^{(i)}} h^{(i)}\, dx$ of this equality i-times we derive

$$\varphi_1(y,h) = \int_{x=a}^{b} \left[\frac{\partial F}{\partial y} + \sum_{i=1}^{n} (-1)^i \frac{d}{dx^i} \frac{\partial F}{\partial y^{(i)}}\right] h\, dx \tag{22.31}$$

Theorem 22.9. (Euler–Poisson) *Let the curve $y \in C^n[a,b]$, satisfying (22.30), be an extremal (minimum or maximum) curve for the functional (22.29). Then it should satisfy the following ODE:*

$$\boxed{\frac{\partial F}{\partial y} + \sum_{i=1}^{n} (-1)^i \frac{d}{dx^i} \frac{\partial F}{\partial y^{(i)}} = 0} \tag{22.32}$$

Proof. It follows directly from (22.31) applying the Lagrange lemma 22.5. □

22.4.2 Curves with non-fixed boundary points

Problem formulation: among all smooth curves $y = y(x)$ with the boundary points (x_0, y_0) and (x_1, y_1), belonging to the given curves ("sliding surfaces") $y = \varphi(x)$ and $y = \psi(x)$, find one which minimizes the functional

$$\boxed{J(y) = \int\limits_{x=x_0:y(x_0)=\varphi(x_0)}^{x_1:y(x_1)=\psi(x_1)} F(x, y(x), y'(x))\, dx} \tag{22.33}$$

Realizing the function variation $(y + h)$ such that

$$\begin{aligned} &h(x_i) \simeq \delta y(x_i) - y'(x_i)\,\delta x_i, \quad i = 0, 1 \\ &\delta y(x_0) \simeq \varphi'(x_0)\,\delta x_0, \quad \delta y(x_1) \simeq \psi'(x_1)\,\delta x_1 \end{aligned} \tag{22.34}$$

and integrating by parts, we derive

$$\begin{aligned} \varphi_1(y_0, h) &= \int\limits_{x=x_0}^{x_1} \left(\frac{\partial F}{\partial y} h + \frac{\partial F}{\partial y'} h'\right) dx + F(x, y, y')\,\delta x \,|_{x=x_0}^{x=x_1} \\ &= \int\limits_{x=x_0}^{x_1} \left(\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'}\right) h\, dx + \frac{\partial F}{\partial y'} h \,|_{x=x_0}^{x=x_1} + F(x, y, y')\,\delta x \,|_{x=x_0}^{x=x_1} \end{aligned} \tag{22.35}$$

Remark 22.1. *Since the problem in 22.4.2 contains as a partial case the problem (22.9) with fixed boundary points, then any solution of the problem in 22.4.2 should satisfy the Euler–Lagrange condition (22.13), which simplifies (22.35) up to*

$$\varphi_1(y_0, h) = \frac{\partial F}{\partial y'} h \,|_{x=x_0}^{x=x_1} + F(x, y, y')\,\delta x \,|_{x=x_0}^{x=x_1} \tag{22.36}$$

Applications (22.34) to (22.36) imply

$$\begin{aligned} \varphi_1(y_0, h) &= \frac{\partial F}{\partial y'}\delta y \,|_{x=x_0}^{x=x_1} + \left(F - \frac{\partial F}{\partial y'} y'\right) \delta x \,|_{x=x_0}^{x=x_1} \\ &= \left(\frac{\partial F}{\partial y'}\psi' + F - \frac{\partial F}{\partial y'} y'\right) |_{x=x_1}\, \delta x_1 \\ &\quad + \left(\frac{\partial F}{\partial y'}\varphi' + F - \frac{\partial F}{\partial y'} y'\right) |_{x=x_0}\, \delta x_0 \end{aligned} \tag{22.37}$$

Theorem 22.10. *If some curve $y = y(x) \in C^1$ with the boundary points (x_0, y_0) and (x_1, y_1), belonging to the given curves ("sliding surfaces") $y = \varphi(x)$ and $y = \psi(x)$ provides an extremum to the functional (22.33), then*

1. *it satisfies the* ***Euler–Lagrange condition*** *(22.13), i.e.,*

$$\frac{\partial F}{\partial y} - \frac{d}{dx}\frac{\partial F}{\partial y'} = 0$$

2. *it should, additionally, satisfy the, so-called,* ***transversality conditions***

$$\boxed{\begin{aligned}\left[F + \frac{\partial F}{\partial y'}(\psi' - y')\right]|_{x=x_1} &= 0\\ \left[F + \frac{\partial F}{\partial y'}(\psi' - y')\right]|_{x=x_0} &= 0\end{aligned}} \tag{22.38}$$

Proof. The relations (22.38) result from (22.37) if we take into account that the variations δx_0 and δx_1 are independent. □

Example 22.3.

$$J(y) = \int_{x_0=0:y(x_0)=0}^{x_1:y(x_1)=2-x} \frac{\sqrt{1+(y'(x))^2}}{x}dx$$

The Euler–Lagrange condition (22.13) is $-\frac{d}{dx}\frac{y'}{x\sqrt{1+(y')^2}} = 0$, *which leads to the following relation*

$$\frac{y'}{x\sqrt{1+(y')^2}} = c_1, \quad y' = \pm\frac{xc_1}{\sqrt{1-x^2c_1^2}}$$
$$y = \mp\frac{1}{c_1}\sqrt{1-x^2c_1^2} + c_2, \quad (y-c_2)^2 = 1/c_1^2 - x^2$$

The boundary condition $y(0) = 0$ *gives* $c_1^2 = 1/c_2^2$. *So, finally, the Euler–Lagrange condition (22.13) is* $(y-c_2)^2 + x^2 = c_2^2$. *The transversality conditions (22.38) imply* $0 = y' = \frac{x}{y-c_2} = 1$, *which, together with* $y = 2 - x$, *gives* $c_2 = 2$. *Finally, the extremal curve is as follows:*

$$(y-2)^2 + x^2 = 4$$

22.4.3 Curves with a nonsmoothness point

If an extremal curve has a nonsmooth point $x^* \in [a,b]$ in the problem (22.9) with a fixed boundary point, that is,

$$y(x) \quad \text{is continuous in} \quad x^*$$

$$y'(x)|_{x\to x^*-0} \neq y'(x)|_{x\to x^*+0}$$

then

$$\varphi_1(y_0,h) = \varphi_{1,1}(y_0,h) + \varphi_{1,2}(y_0,h) = 0$$

where (we can consider at each semi-interval $[a, x^*]$ and $[x^*, b]$ the boundary point x^* as a non-fixed one)

$$\varphi_{1,1}(y_0, h) = \frac{\partial F}{\partial y'}\delta y\,|_{x=x^*-0} + \left(F - \frac{\partial F}{\partial y'}y'\right)\delta x\,|_{x=x^*-0}$$

$$\varphi_{1,2}(y_0, h) = -\frac{\partial F}{\partial y'}\delta y\,|_{x=x^*+0} - \left(F - \frac{\partial F}{\partial y'}y'\right)\delta x\,|_{x=x^*+0}$$

which gives

$$\begin{aligned} 0 &= \varphi_1(y_0, h) \\ &= \left(\frac{\partial F}{\partial y'}\,|_{x=x^*-0} - \frac{\partial F}{\partial y'}\,|_{x=x^*+0}\right)\delta y \\ &\quad + \left[\left(F - \frac{\partial F}{\partial y'}y'\right)|_{x=x^*-0} - \left(F - \frac{\partial F}{\partial y'}y'\right)|_{x=x^*+0}\right]\delta x \end{aligned}$$

Since here δy and δx are admitted to be arbitrary, we obtain the, so-called, *Weierstrass–Erdmann conditions*:

$$\boxed{\begin{aligned} &\frac{\partial F}{\partial y'}\,|_{x=x^*-0} = \frac{\partial F}{\partial y'}\,|_{x=x^*+0} \\ &\left(F - \frac{\partial F}{\partial y'}y'\right)|_{x=x^*-0} = \left(F - \frac{\partial F}{\partial y'}y'\right)|_{x=x^*+0} \end{aligned}} \tag{22.39}$$

Example 22.4. *Consider the functional*

$$J(y) = \int_{x=a}^{b} (y')^2 (1 - y')^2\, dx$$

The boundary points are assumed to be fixed. The Euler–Lagrange condition (22.13) gives $y = c_1 x + c_2$. *The Weierstrass–Erdmann conditions (22.39) are*

$$2y'(1 - y')(1 - 2y')\,|_{x=x^*-0} = 2y'(1 - y')(1 - 2y')\,|_{x=x^*+0}$$

$$-(y')^2(1 - y')(1 - 3y')\,|_{x=x^*-0} = -(y')^2(1 - y')(1 - 3y')\,|_{x=x^*+0}$$

which are fulfilled for extremal curves such that

$$y'\,|_{x=x^*-0} = 0, \quad y'\,|_{x=x^*+0} = 1 \quad \textit{or} \quad y'\,|_{x=x^*-0} = 1, \quad y'\,|_{x=x^*+0} = 0$$

22.5 Optimal control problem

22.5.1 Controlled plant, cost functionals and terminal set

Consider the controlled plant given by the following system of ordinary differential equations (ODE)

$$\left.\begin{array}{l} \dot{x}(t) = f(x(t), u(t), t), \quad \text{a.e.} \quad t \in [0, T] \\ x(0) = x_0 \end{array}\right\} \tag{22.40}$$

where $x = \left(x^1, \ldots, x^n\right)^T \in \mathbb{R}^n$ is its state vector, $u = \left(u^1, \ldots, u^r\right)^T \in \mathbb{R}^r$ is the control that may run over a given control region $U \subset \mathbb{R}^r$ with *the cost functional*

$$J(u(\cdot)) := h_0(x(T)) + \int_{t=0}^{T} h(x(t), u(t), t)\, dt \tag{22.41}$$

containing the integral term as well as the terminal one and with *the terminal set* $\mathcal{M} \subseteq \mathbb{R}^n$ given by the inequalities

$$\mathcal{M} = \{x \in \mathbb{R}^n : g_l(x) \leq 0 \quad (l = 1, \ldots, L)\} \tag{22.42}$$

The time process or *horizon* T is supposed to be fixed or nonfixed and may be finite or infinite.

Definition 22.4.

(a) The function (22.41) is said to be given in ***Bolza form****.*
(b) If in (22.41) $h_0(x) = 0$ *we obtain the cost functional in* ***Lagrange form****, that is,*

$$J(u(\cdot)) = \int_{t=0}^{T} h(x(t), u(t), t)\, dt \tag{22.43}$$

(c) If in (22.41) $h(x, u, t) = 0$ *we obtain the cost functional in* ***Mayer form****, that is,*

$$J(u(\cdot)) = h_0(x(T)) \tag{22.44}$$

Usually the following assumptions are assumed to be in force:

(A1) (U, d) is a separable metric space (with the metric d) and $T > 0$.
(A2) The maps

$$\left.\begin{array}{l} f : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R}^n \\ h : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R} \\ h_0 : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R} \\ g_l : \mathbb{R}^n \to \mathbb{R}^n \quad (l = 1, \ldots, L) \end{array}\right\} \tag{22.45}$$

are measurable and there exist a constant L and a continuity modulus $\bar{\omega} : [0, \infty) \to [0, \infty)$ such that for $\varphi = f(x, u, t)$, $h(x, u, t)$, $h_0(x, u, t)$, $g_l(x)$ $(l = 1, \ldots, L)$ the following inequalities hold:

$$\left.\begin{aligned} \left\|\varphi(x, u, t) - \varphi(\hat{x}, \hat{u}, t)\right\| &\le L\left\|x - \hat{x}\right\| + \bar{\omega}\left(d(u, \hat{u})\right) \\ \forall t \in [0, T], \quad x, \hat{x} \in \mathbb{R}^n, \quad u, \hat{u} \in U& \\ \|\varphi(0, u, t)\| \le L \quad \forall u, t \in U \times [0, T]& \end{aligned}\right\} \tag{22.46}$$

(A3) The maps f, h, h_0 and g_l $(l = 1, \ldots, L)$ are from C^1 in x and there exists a continuity modulus $\bar{\omega} : [0, \infty) \to [0, \infty)$ such that for $\varphi = f(x, u, t)$, $h(x, u, t)$, $h_0(x, u, t)$, $g_l(x)$ $(l = 1, \ldots, L)$ the following inequalities hold:

$$\left.\begin{aligned} &\left\|\frac{\partial}{\partial x}\varphi(x, u, t) - \frac{\partial}{\partial x}\varphi(\hat{x}, \hat{u}, t)\right\| \\ &\quad \le \bar{\omega}\left(\left\|x - \hat{x}\right\| + d(u, \hat{u})\right) \\ &\forall t \in [0, T], \quad x, \hat{x} \in \mathbb{R}^n, \quad u, \hat{u} \in U \end{aligned}\right\} \tag{22.47}$$

22.5.2 Feasible and admissible control

Definition 22.5. *A function $u(t)$, $t_0 \le t \le T$, is said to be*

*(a) a **feasible control** if it is measurable and $u(t) \in U$ for all $t \in [0, T]$. Denote the set of all feasible controls by*

$$\mathcal{U}[0, T] := \left\{u(\cdot) : [0, T] \to U \mid u(t) \quad \text{is measurable}\right\} \tag{22.48}$$

*(b) an **admissible** or **realizing the terminal condition** (22.42), if the corresponding trajectory $x(t)$ satisfies the terminal condition, that is, satisfies the inclusion $x(T) \in \mathcal{M}$. Denote the set of all admissible controls by*

$$\mathcal{U}_{admis}[0, T] := \left\{u(\cdot) : u(\cdot) \in \mathcal{U}[0, T], \quad x(T) \in \mathcal{M}\right\} \tag{22.49}$$

In view of Theorem 19.1 on the existence and the uniqueness of an ODE solution, it follows that under assumptions (A1)–(A2) for any $u(t) \in \mathcal{U}[0, T]$ equation (1.4) admits a unique solution $x(\cdot) := x(\cdot, u(\cdot))$ and the functional (22.41) is well defined.

22.5.3 Problem setting in the general Bolza form

Based on the definitions given above, the optimal control problem (OCP) can be formulated as follows.

Problem 22.1. (OCP in Bolza form)

$$\textit{Minimize (22.41)} \quad \textit{over} \quad \mathcal{U}_{admis}[0, T] \tag{22.50}$$

Problem 22.2. (OCP with a fixed terminal term) *If in the problem (22.50)*

$$\mathcal{M} = \left\{x \in \mathbb{R}^n : g_1(x) = x - x_f \leq 0, \quad g_2(x) = -\left(x - x_f\right) \leq 0\right\} \tag{22.51}$$

then it is called the ***optimal control problem with fixed terminal term*** x_f.

Definition 22.6. *Any control* $u^*(\cdot) \in \mathcal{U}_{admis}[0, T]$ *satisfying*

$$J(u^*(\cdot)) = \inf_{u(\cdot) \in \mathcal{U}_{admis}[0,T]} J(u(\cdot)) \tag{22.52}$$

is called an ***optimal control****. The corresponding state trajectories* $x^*(\cdot) := x^*(\cdot, u^*(\cdot))$ *are called an* ***optimal state trajectory****, and* $(x^*(\cdot), u^*(\cdot))$ *is called an* ***optimal pair****.*

22.5.4 Mayer form representation

Summary 22.1. *Introduce* $(n+1)$*-dimensional space* $\mathbb{R}^{n+1}$ *of the variables* $x = (x_1, \ldots, x_n, x_{n+1})^\top$ *where the first n coordinates satisfy (22.40) and the component* x_{n+1} *is given by*

$$x_{n+1}(t) := \int_{\tau=0}^{t} h(x(\tau), u(\tau), \tau)\, d\tau \tag{22.53}$$

or, in the differential form,

$$\dot{x}_{n+1}(t) = h(x(t), u(t), t) \tag{22.54}$$

with the initial condition for the last component given by

$$x_{n+1}(0) = 0 \tag{22.55}$$

As a result, the initial optimization problem in the Bolza form (22.50) can be reformulated in the space $\mathbb{R}^{n+1}$ *as the Mayer problem with the cost functional* $J(u(\cdot))$

$$J(u(\cdot)) = h_0(x(T)) + x_{n+1}(T) \tag{22.56}$$

where the function $h_0(x)$ *does not depend on the last coordinate* $x_{n+1}(t)$*, that is,*

$$\frac{\partial}{\partial x_{n+1}} h_0(x) = 0 \tag{22.57}$$

Summary 22.2. *From the relations above it follows that the Mayer problem with the cost function (22.56) is equivalent to the initial optimization control problem (22.50) in the Bolza form.*

There exist two principal approaches to solving *optimal control problems*:

- the first one is the *maximum principle* (MP) of L. Pontryagin (Boltyanski *et al.* 1956; Pontryagin *et al.* 1969 (translated from Russian));

- and the second one is the *dynamic programing method* (DPM) of R. Bellman (1957).

We will touch on both of them below.

22.6 Maximum principle

The maximum principle is a basic instrument to derive a set of *necessary conditions* which should satisfy any optimal control. As an optimal control problem may be regarded as an optimization problem in the corresponding infinite dimensional (Hilbert or, in general, Banach) space, the necessary conditions (resembling the *Kuhn–Tucker conditions* in the finite-dimensional optimization) take place. They are known as the *maximum principle* which is really a milestone in modern optimal control theory. It states that any dynamic system closed by an optimal control strategy or, simply, by an optimal control is a Hamiltonian (with the doubled dimension) system given by a system of the forward–backward ordinary differential equations and, in addition, an optimal control maximizes the function called Hamiltonian. Its mathematical importance consists of the following fact: the maximization of the Hamiltonian with respect to a control variable given in a finite-dimensional space looks and really is much easier than the original optimization problem formulated in an infinite-dimensional space. The key idea of the original version of the maximum principle comes from classical variations calculus. To derive the main MP formulation, first one needs to perturb slightly an optimal control using the so-called needle-shape (spike) variations and, second, to consider the first-order term in a Taylor expansion with respect to this perturbation. Tending perturbations to zero, some variation inequalities may be obtained. Then the final result follows directly from duality.

22.6.1 Needle-shape variations

Let $(x^*(\cdot), u^*(\cdot))$ be the given optimal pair for the problem (22.52) and $M_\varepsilon \subseteq [0, T]$ be a measurable set of the time interval with Lebesgue measure $|M_\varepsilon| = \varepsilon > 0$. Let now $u(\cdot) \in \mathcal{U}_{admis}[0, T]$ be any given admissible control.

Definition 22.7. *Define the following control*

$$u^\varepsilon(t) := \begin{cases} u^*(t) & \text{if} \quad t \in [0, T] \setminus M_\varepsilon \\ u(t) \in \mathcal{U}_{admis}[0, T] & \text{if} \quad t \in M_\varepsilon \end{cases} \tag{22.58}$$

Evidently $u^\varepsilon(\cdot) \in \mathcal{U}_{admis}[0, T]$. *Below* $u^\varepsilon(\cdot)$ *is referred to as a* ***needle-shape*** *or* ***spike variation*** *of the optimal control* $u^*(t)$.

The next lemma plays a key role in proving the basic MP theorem.

Lemma 22.8. (The variational equation) *Let* $x^\varepsilon(\cdot) := x(\cdot, u^\varepsilon(\cdot))$ *be the solution of (22.52) for the plant model given by (22.40) under the control* $u^\varepsilon(\cdot)$ *and* $\Delta^\varepsilon(\cdot)$ *be the*

solution to the following differential equation

$$\begin{aligned}\dot{\Delta}^{\varepsilon}(t) &= \frac{\partial}{\partial x} f(x^*(t), u^*(t), t)\,\Delta^{\varepsilon}(t) \\ &\quad + [f(x^*(t), u^{\varepsilon}(t), t) - f(x^*(t), u^*(t), t)]\,\chi_{M_{\varepsilon}}(t) \\ \Delta^{\varepsilon}(0) &= 0\end{aligned} \tag{22.59}$$

where $\chi_{M_{\varepsilon}}(t)$ *is the characteristic function of the set* M_{ε}*, that is*

$$\chi_{M_{\varepsilon}}(t) := \begin{cases} 1 & if \quad t \in M_{\varepsilon} \\ 0 & if \quad t \notin M_{\varepsilon} \end{cases} \tag{22.60}$$

Then

$$\left.\begin{aligned} &\max_{t\in[0,T]} \|x^{\varepsilon}(t) - x^*(t)\| = O(\varepsilon) \\ &\max_{t\in[0,T]} \|\Delta^{\varepsilon}(t)\| = O(\varepsilon) \\ &\max_{t\in[0,T]} \|x^{\varepsilon}(t) - x^*(t) - \Delta^{\varepsilon}(t)\| = o(\varepsilon) \end{aligned}\right\} \tag{22.61}$$

and the following variational equations hold
(a) for the cost function given in Bolza form (22.41)

$$\begin{aligned} J(u^{\varepsilon}(\cdot)) - J(u^*(\cdot)) &= \left(\frac{\partial}{\partial x} h_0(x^*(T)), \Delta^{\varepsilon}(T)\right) \\ &+ \int_{t=0}^{T} \left\{ \left(\frac{\partial}{\partial x} h(x^*(t), u^*(t), t), \Delta^{\varepsilon}(t)\right) \right. \\ &\left. + [h(x^*(t), u^{\varepsilon}(t), t) - h(x^*(t), u^*(t), t)]\,\chi_{M_{\varepsilon}}(t) \right\} dt + o(\varepsilon) \end{aligned} \tag{22.62}$$

(b) for the cost function given in Mayer form (22.44)

$$J(u^{\varepsilon}(\cdot)) - J(u^*(\cdot)) = \left(\frac{\partial}{\partial x} h_0(x^*(T)), \Delta^{\varepsilon}(T)\right) + o(\varepsilon) \tag{22.63}$$

Proof. Define $\delta^{\varepsilon}(t) := x^{\varepsilon}(t) - x^*(t)$. Then assumption (A2) (22.46) for any $t \in [0, T]$ implies

$$\|\delta^{\varepsilon}(t)\| \le \int_{s=0}^{t} L\,\|\delta^{\varepsilon}(s)\|\,ds + K\varepsilon \tag{22.64}$$

which, by the Gronwall lemma 19.4, leads to the first relation in (22.61). Define

$$\eta^{\varepsilon}(t) := x^{\varepsilon}(t) - x^*(t) - \Delta^{\varepsilon}(t) = \delta^{\varepsilon}(t) - \Delta^{\varepsilon}(t) \tag{22.65}$$

Then

$$\begin{aligned}\dot{\eta}^{\varepsilon}(t) &= [f(x^{\varepsilon}(t), u^{\varepsilon}(t), t) - f(x^{*}(t), u^{*}(t), t)] \\ &\quad - \frac{\partial}{\partial x} f(x^{*}(t), u^{*}(t), t)\, \Delta^{\varepsilon}(t) \\ &\quad - [f(x^{*}(t), u^{\varepsilon}(t), t) - f(x^{*}(t), u^{*}(t), t)]\, \chi_{M_{\varepsilon}}(t) \\ &= \int_{\theta=0}^{1} \left[\frac{\partial}{\partial x} f(x^{*}(t) + \theta\delta^{\varepsilon}(t), u^{\varepsilon}(t), t) - \frac{\partial}{\partial x} f(x^{*}(t), u^{*}(t), t)\right] d\theta\, \delta^{\varepsilon}(t) \\ &\quad - [f(x^{*}(t), u^{\varepsilon}(t), t) - f(x^{*}(t), u^{*}(t), t)]\, \chi_{M_{\varepsilon}}(t) \\ &\quad + \frac{\partial}{\partial x} f(x^{*}(t), u^{*}(t), t)\, \eta^{\varepsilon}(t)\end{aligned} \tag{22.66}$$

Integrating the last identity (22.66) and in view of (A2) (22.46) and (A3) (22.47), we obtain

$$\begin{aligned}\|\eta^{\varepsilon}(t)\| &\le \int_{s=0}^{t} \int_{\theta=0}^{1} [\bar{\omega}(\theta\, \|\delta^{\varepsilon}(s)\| + d(u^{\varepsilon}(s), u^{*}(s)))]\, \|\delta^{\varepsilon}(s)\|\, d\theta ds \\ &\quad + \int_{s=0}^{t} \bar{\omega}(d(u^{\varepsilon}(s), u^{*}(s)))\, \chi_{M_{\varepsilon}}(s)\, ds + \int_{s=0}^{t} \frac{\partial}{\partial x} f(x^{*}(s), u^{*}(s), s)\, \eta^{\varepsilon}(s)\, ds \\ &\le \text{Const} \cdot \varepsilon \cdot o(1) + \text{Const} \cdot \int_{s=0}^{t} \|\eta^{\varepsilon}(s)\|\, ds\end{aligned} \tag{22.67}$$

The last inequality in (22.67) by the Gronwall lemma directly implies the third relation in (22.61). The second relation follows from the first and third ones. The same manipulations lead to (22.62) and (22.63). □

22.6.2 Adjoint variables and MP formulation

The classical format of MP formulation gives a set of *first-order necessary conditions* for optimal pairs.

Theorem 22.11. (MP for Mayer form with a fixed horizon) *If under assumptions (A1)–(A3) a pair* $(x^{*}(\cdot), u^{*}(\cdot))$ *is optimal then there exist the vector functions* $\psi(t)$, *satisfying the system of the* ***adjoint equations***

$$\boxed{\dot{\psi}(t) = -\frac{\partial}{\partial x} f(x^{*}(t), u^{*}(t), t)^{\mathsf{T}}\, \psi(t) \quad a.e. \quad t \in [0, T]} \tag{22.68}$$

and nonnegative constants $\mu \geq 0$ and $\nu_l \geq 0$ $(l = 1, \ldots, L)$ such that the following four conditions hold:

(a) ***(the maximality condition)****: for almost all $t \in [0, T]$*

$$\boxed{H(\psi(t), x^*(t), u^*(t), t) = \max_{u \in U} H(\psi(t), x^*(t), u, t)} \tag{22.69}$$

where the Hamiltonian is defined by

$$\begin{aligned} &H(\psi, x, u, t) := \psi^{\mathsf{T}} f(x, u, t) \\ &t, x, u, \psi \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^n \end{aligned} \tag{22.70}$$

(b) ***(transversality condition)****: the equality*

$$\boxed{\psi(T) + \mu \frac{\partial}{\partial x} h_0(x^*(T)) + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)) = 0} \tag{22.71}$$

holds;

(c) ***(complementary slackness conditions)****: either the equality $g_l(x^*(T)) = 0$ holds, or $\nu_l = 0$, that is, for any $(l = 1, \ldots, L)$*

$$\boxed{\nu_l g_l(x^*(T)) = 0} \tag{22.72}$$

(d) ***(nontriviality condition)****: at least one of the numbers $|\psi(T)|$ and ν_l is distinct from zero, that is,*

$$\boxed{\|\psi(T)\| + \mu + \sum_{l=1}^{L} \nu_l > 0} \tag{22.73}$$

Proof. Let $\psi(t)$ be the solution of (22.68) corresponding to the terminal condition $\psi(T) = b$ and $\bar{t} \in [0, T]$. Define $M_\varepsilon := [\bar{t}, \bar{t} + \varepsilon] \subseteq [0, T]$. If $u^*(t)$ is an optimal control, then according to the Lagrange principle (see Theorem 21.12), formulated for a Banach space, there exist constants $\mu \geq 0$ and $\nu_l \geq 0$ $(l = 1, \ldots, L)$ such that for any $\varepsilon \geq 0$

$$\mathcal{L}(u^\varepsilon(\cdot), \mu, \nu) - \mathcal{L}(u^*(\cdot), \mu, \nu) \geq 0 \tag{22.74}$$

where

$$\mathcal{L}(u(\cdot), \mu, \nu) := \mu J(u(\cdot)) + \sum_{l=1}^{L} \nu_l g_l(x(T)) \tag{22.75}$$

Taking into account that $\psi(T) = b$ and $\Delta^\varepsilon(0) = 0$, by the differential chain rule, applied to the term $\psi(t)^{\mathsf{T}} \Delta^\varepsilon(t)$, and, in view of (22.59) and (22.68), we obtain

$$\begin{aligned} b^{\mathsf{T}} \Delta^\varepsilon(T) &= \psi(T)^{\mathsf{T}} \Delta^\varepsilon(T) - \psi(0)^{\mathsf{T}} \Delta^\varepsilon(0) \\ &= \int_{t=0}^{T} d(\psi(t)^{\mathsf{T}} \Delta^\varepsilon(t)) = \int_{t=0}^{T} \left(\dot{\psi}(t)^{\mathsf{T}} \Delta^\varepsilon(t) + \psi(t)^{\mathsf{T}} \dot{\Delta}^\varepsilon(t)\right) dt \end{aligned}$$

$$
\begin{aligned}
&= \int_{t=0}^{T} \Big[-\Delta^{\varepsilon}(t)^{\mathsf{T}} \frac{\partial}{\partial x} f(x^*(t), u^*(t), t)^{\mathsf{T}} \psi(t) \\
&\quad + \psi(t)^{\mathsf{T}} \frac{\partial}{\partial x} f(x^*(t), u^*(t), t)\, \Delta^{\varepsilon}(t) \\
&\quad + \psi(t)^{\mathsf{T}} [f(x^*(t), u^{\varepsilon}(t), t) - f(x^*(t), u^*(t), t)]\, \chi_{M_\varepsilon}(t) \Big] dt \\
&= \int_{t=0}^{T} \psi(t)^{\mathsf{T}} [f(x^*(t), u^{\varepsilon}(t), t) - f(x^*(t), u^*(t), t)]\, \chi_{M_\varepsilon}(t)\, dt
\end{aligned}
\tag{22.76}
$$

The variational equality (22.62) together with (22.74) and (22.76) implies

$$
\begin{aligned}
0 &\le \mathcal{L}(u^{\varepsilon}(\cdot), \mu, \nu) - \mathcal{L}(u^*(\cdot), \mu, \nu) \\
&= \mu \left(\frac{\partial}{\partial x} h_0(x^*(T)), \Delta^{\varepsilon}(T) \right) + b^{\mathsf{T}} \Delta^{\varepsilon}(T) \\
&\quad - \int_{t=0}^{T} \psi(t)^{\mathsf{T}} (f(x^*(t), u^{\varepsilon}(t), t) - f(x^*(t), u^*(t), t))\, \chi_{M_\varepsilon}(t)\, dt \\
&\quad + \sum_{l=1}^{L} \nu_l [g_l(x(T)) - g_l(x^*(T))] + o(\varepsilon) \\
&= \left(\mu \frac{\partial}{\partial x} h_0(x^*(T)) + b + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)), \Delta^{\varepsilon}(T) \right) \\
&\quad - \int_{t=\bar{t}}^{\bar{t}+\varepsilon} [\psi(t)^{\mathsf{T}} (f(x^*(t), u^{\varepsilon}(t), t) - f(x^*(t), u^*(t), t))]\, dt + o(\varepsilon) \\
&= \left(\mu \frac{\partial}{\partial x} h_0(x^*(T)) + b + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)), \Delta^{\varepsilon}(T) \right) \\
&\quad - \int_{t=\bar{t}}^{\bar{t}+\varepsilon} [H(\psi(t), x^*(t), u^{\varepsilon}(t), t) - H(\psi(t), x^*(t), u^*(t), t)]\, dt
\end{aligned}
\tag{22.77}
$$

1. Tending ε to zero from (22.77) it follows that

$$
0 \le \left(\mu \frac{\partial}{\partial x} h_0(x^*(T)) + b + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)), \Delta^{\varepsilon}(T) \right) |_{\varepsilon=0}
$$

which should be valid for any $\Delta^{\varepsilon}(T)\,|_{\varepsilon=0}$. This is possible only if (this can be proved by contradiction)

$$\mu \frac{\partial}{\partial x} h_0(x^*(T)) + b + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)) = 0 \tag{22.78}$$

which is equivalent to (22.71). So, the transversality condition is proven.

2. In view of (22.78), the inequality (22.77) may be simplified to the following one

$$0 \le -\int_{t=\bar{t}}^{\bar{t}+\varepsilon} [H(\psi(t), x^*(t), u^{\varepsilon}(t), t) - H(\psi(t), x^*(t), u^*(t), t)]\, dt \tag{22.79}$$

This inequality together with separability of metric space U directly leads to the maximality condition (22.69).

3. Suppose that (22.72) does not hold, that is, there exist an index l_0 and a multiplier $\tilde{\nu}_{l_0}$ such that $\nu_l g_l(x^*(T)) < 0$. This gives

$$\begin{aligned}\mathcal{L}(u^*(\cdot), \mu, \tilde{\nu}) &:= \mu J(u^*(\cdot)) + \sum_{l=1}^{L} \tilde{\nu}_l g_l(x^*(T)) \\ &= \mu J(u^*(\cdot)) + \tilde{\nu}_{l_0} g_{l_0}(x^*(T)) < \mu J(u^*(\cdot)) = \mathcal{L}(u^*(\cdot), \mu, \nu)\end{aligned}$$

It means that $u^*(\cdot)$ is not optimal control. We obtain the contradiction. So, the complementary slackness condition is proven too.

4. Suppose that (22.73) is not valid, i.e., $|\psi(T)| + \mu + \sum_{l=1}^{L} \nu_l = 0$. This implies $\psi(T) = 0, \mu = \nu_l = 0\,(l = 1, \ldots, L)$, and, hence, in view of (22.68) and by the Gronwall lemma 19.4, it follows that $\psi(t) = 0$ for all $t \in [0, T]$. So, $H(\psi(t), x(t), u(t), t) = 0$ for any $u(t)$ (not only for $u^*(t)$). This means that the application of any admissible control keeps the cost function unchangeable which corresponds to a trivial situation. So, the nontriviality condition is proven too. □

22.6.3 The regular case

In the, so-called, *regular case*, when $\mu > 0$ (this means that the nontriviality condition holds automatically), the variable $\psi(t)$ and constants ν_l may be normalized and change to $\tilde{\psi}(t) := \psi(t)/\mu$ and $\tilde{\nu}_l := \nu_l/\mu$. In this new variable the MP formulation looks as follows.

Theorem 22.12. (MP in the regular case) *If under assumptions (A1)–(A3) a pair* $(x^*(\cdot), u^*(\cdot))$ *is optimal, then there exists a vector function* $\tilde{\psi}(t)$ *satisfying the system of the* ***adjoint equations***

$$\frac{d}{dt}\tilde{\psi}(t) = -\frac{\partial}{\partial x} f(x^*(t), u^*(t), t)^{\mathsf{T}}\, \tilde{\psi}(t) \quad a.e.\ \ t \in [0, T]$$

and $\nu_l \geq 0$ $(l = 1, \ldots, L)$ *such that the following three conditions hold:*

(a) (***the maximality condition***): *for almost all* $t \in [0, T]$

$$H(\tilde{\psi}(t), x^*(t), u^*(t), t) = \max_{u \in U} H(\tilde{\psi}(t), x^*(t), u, t) \tag{22.80}$$

where the Hamiltonian is defined by

$$H(\psi, x, u, t) := \tilde{\psi}^{\mathsf{T}} f(x, u, t)$$
$$t, x, u, \psi \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^n$$

(b) (***transversality condition***): *for every* $\alpha \in \mathcal{A}$, *the equalities*

$$\tilde{\psi}(T) + \frac{\partial}{\partial x} h_0(x^*(T)) + \sum_{l=1}^{L} \tilde{\nu}_l \frac{\partial}{\partial x} g_l(x^*(T)) = 0$$

hold;

(c) (***complementary slackness conditions***): *either the equality* $g_l(x^*(T)) = 0$ *holds, or* $\nu_l = 0$, *that is, for any* $(l = 1, \ldots, L)$ $\nu_l g_l(x^*(T)) = 0$.

Remark 22.2. *This means that without loss of generality we may put* $\mu = 1$. *It may be shown that the regularity property takes place if the vectors* $\frac{\partial}{\partial x} g_l(x^*(T))$ *are linearly independent. The verification of this property is usually not so simple a task.*

22.6.4 Hamiltonian form and constancy property

Corollary 22.3. (Hamiltonian for the Bolza problem) *Hamiltonian for the Bolza problem has the form*

$$H(\psi, x, u, t) := \psi^{\mathsf{T}} f(x, u, t) - \mu h(x(t), u(t), t)$$
$$t, x, u, \psi \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^n \tag{22.81}$$

Proof. It follows from (22.53)–(22.57). Indeed, since the Mayer's form representation $(\dot{x}_{n+1}(t) = h(x(t), u(t), t))$ implies $\dot{\psi}_{n+1}(t) = 0$, then $\psi_{n+1}(T) = -\mu$. □

Corollary 22.4. (Hamiltonian form) *Equations (22.40) and (22.68) may be represented in the, so-called, Hamiltonian form (the forward–backward ODE form):*

$$\left.\begin{aligned} \dot{x}^*(t) &= \frac{\partial}{\partial \psi} H(\psi(t), x^*(t), u^*(t), t), \quad x^*(0) = x_0 \\ \dot{\psi}(t) &= -\frac{\partial}{\partial x} H(\psi(t), x^*(t), u^*(t), t) \\ \psi(T) &= -\mu \frac{\partial}{\partial x} h_0(x^*(T)) - \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)) \end{aligned}\right\} \tag{22.82}$$

Proof. It directly follows from comparison of the right-hand side of (22.70) with (22.40) and (22.68). □

Corollary 22.5. (Constancy property) *For stationary systems when in (22.40), (22.41)*

$$f = f(x(t), u(t)), \quad h = h(x(t), u(t)) \tag{22.83}$$

it follows that for all $t \in [t_0, T]$

$$H(\psi(t), x^*(t), u^*(\psi(t), x^*(t))) = \text{const} \tag{22.84}$$

Proof. One can see that in this case the Hamiltonian $H = H(\psi(t), x(t), u(t))$ does not depend on t directly, that is, $\frac{\partial}{\partial t} H = 0$. Hence, $u^*(t)$ is a function of $\psi(t)$ and $x^*(t)$ only, i.e., $u^*(t) = u^*(\psi(t), x^*(t))$. Denote

$$H(\psi(t), x^*(t), u^*(\psi(t), x^*(t))) := \tilde{H}(\psi(t), x^*(t))$$

Then (22.82) becomes

$$\dot{x}(t) = \frac{\partial}{\partial \psi} \tilde{H}(\psi(t), x^*(t)), \quad \dot{\psi}(t) = -\frac{\partial}{\partial x} \tilde{H}(\psi(t), x^*(t))$$

which implies

$$\frac{d}{dt} \tilde{H}(\psi(t), x^*(t)) = \frac{\partial}{\partial \psi} \tilde{H}(\psi(t), x^*(t))^{\mathsf{T}} \dot{\psi}(t) + \frac{\partial}{\partial x} \tilde{H}(\psi(t), x^*(t))^{\mathsf{T}} \dot{x}(t) = 0$$

and hence $\tilde{H}(\psi(t), x^*(t)) = \text{const}$ for any $t \in [t_0, T]$. □

22.6.5 Nonfixed horizon optimal control problem and zero property

Consider the following generalization of the optimal control problem (22.40), (22.44), (22.50) permitting terminal time to be free. In view of this, the optimization problem may be formulated in the following manner: *minimize*

$$J(u(\cdot)) = h_0(x(T), T) \tag{22.85}$$

over $u(\cdot) \in \mathcal{U}_{\text{admis}}[0, T]$ and $T \geq 0$ *with the terminal set* $\mathcal{M}(T)$ *given by*

$$\mathcal{M}(T) = \{x(T) \in \mathbb{R}^n : g_l(x(T), T) \leq 0 \quad (l = 1, \ldots, L)\} \tag{22.86}$$

Theorem 22.13. (MP for non fixed horizon case) *If under assumptions (A1)–(A3) the pair* $(T^*, u^*(\cdot))$ *is a solution of the problem (22.85), (22.86) and* $x^*(t)$ *is the corresponding optimal trajectory, then there exist the vector functions* $\psi(t)$*, satisfying the system of the adjoint equations (22.68), and nonnegative constants* $\mu \geq 0$ *and* $\nu_l \geq 0$

($l = 1, \ldots, L$) such that all four conditions of the previous Theorem 22.11 are fulfilled and, in addition, the following condition to the terminal time holds:

$$\boxed{\begin{aligned} &H(\psi(T), x(T), u(T), T) := \psi^{\intercal}(t)\, f(x(T), u(T-0), T) \\ &= \mu \frac{\partial}{\partial T} h_0(x^*(T), T) + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial T} g_l(x^*(T), T) \end{aligned}} \tag{22.87}$$

Proof. Since $(T^*, u^*(\cdot))$ is a solution of the problem then evidently $u^*(\cdot)$ is a solution of the problem (22.40), (22.44), (22.50) with the fixed horizon $T = T^*$ and, hence, all four properties of the Theorem 22.11 with $T = T^*$ should be fulfilled. Let us find the additional condition to the terminal time T^* which should be satisfied too.

(a) Consider again the needle-shape variation defined as

$$u^{\varepsilon}(t) := \begin{cases} u^*(t) & \text{if} \quad t \in [0, T^*] \setminus (M_\varepsilon \wedge (T^* - \varepsilon, T^*]) \\ u(t) \in \mathcal{U}_{\text{admis}}[0, T^*] & \text{if} \quad t \in M_\varepsilon \subseteq [0, T^* - \varepsilon) \\ u(t) \in \mathcal{U}_{\text{admis}}[0, T^*] & \text{if} \quad t \in [T^* - \varepsilon, T^*] \end{cases} \tag{22.88}$$

Then, for $\mathcal{L}(u(\cdot), \mu, \nu, T)$,

$$\mathcal{L}(u(\cdot), \mu, \nu, T) := \mu J(u(\cdot), T) + \sum_{l=1}^{L} \nu_l g_l(x(T), T) \tag{22.89}$$

it follows that

$$\begin{aligned} 0 &\le \mathcal{L}(u^{\varepsilon}(\cdot), \mu, \nu, T^* - \varepsilon) - \mathcal{L}(u^*(\cdot), \mu, \nu, T^*) \\ &= \mu h_0(x(T^* - \varepsilon), T^* - \varepsilon) + \sum_{l=1}^{L} \nu_l g_l(x(T^* - \varepsilon), T^* - \varepsilon) \\ &\quad - \mu h_0(x^*(T^*), T^*) - \sum_{l=1}^{L} \nu_l g_l(x^*(T^*), T^*) \end{aligned}$$

Hence, applying the transversality condition (22.71) we obtain:

$$\begin{aligned} 0 \le & -\varepsilon \left(\mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \right) + o(\varepsilon) \\ & -\varepsilon \left(\mu \frac{\partial}{\partial x} h_0(x(T^*), T^*) \right. \\ & \left. + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x(T^*), T^*),\ f(x(T^*), u^*(T^* - 0), T^*) \right) \\ = & -\varepsilon \left(\mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \right) \\ & + \varepsilon\, \psi^{\intercal}(T^*) f(x(T^*), u^*(T^* - 0), T^*) + o(\varepsilon) \\ = & -\varepsilon \left(\mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \right) \\ & + \varepsilon\, H(\psi(T^*), x^*(T^*), u^*(T - 0), T^*) + o(\varepsilon) \end{aligned}$$

which, by dividing by ε and tending ε to zero, implies

$$H(\psi(T^*), x^*(T^*), u^*(T-0), T^*) \geq \mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \tag{22.90}$$

(b) Analogously, for the needle–shape variation

$$u^{\varepsilon}(t) := \begin{cases} u^*(t) & \text{if} \quad t \in [0, T^*) \setminus M_{\varepsilon} \\ u(t) \in \mathcal{U}_{admis}[0, T^*] & \text{if} \quad t \in M_{\varepsilon} \\ u^*(T^*-0) & \text{if} \quad t \in [T^*, T^*+\varepsilon] \end{cases} \tag{22.91}$$

it follows that

$$\begin{aligned} 0 &\leq \mathcal{L}(u^{\varepsilon}(\cdot), \mu, \nu, T^*+\varepsilon) - \mathcal{L}(u^*(\cdot), \mu, \nu, T^*) \\ &= \varepsilon \left(\mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \right) \\ &\quad - \varepsilon H(\psi(T^*), x^*(T^*), u^*(T-0), T^*) + o(\varepsilon) \end{aligned}$$

and

$$H(\psi(T^*), x^*(T^*), u^*(T-0), T^*) \leq \mu \frac{\partial}{\partial T} h_0(x(T^*), T^*) + \nu_l \frac{\partial}{\partial T} g_l(x(T^*), T^*) \tag{22.92}$$

Combining (22.88) and (22.91), we obtain (22.87). Theorem is proven. □

Corollary 22.6. (Zero property) *If under the conditions of Theorem 22.13 the functions* $h_0(x, T)$, $g_l(x, T)$ $(l = 1, \ldots, L)$ *do not depend on* T *directly, that is,*

$$\frac{\partial}{\partial T} h_0(x, T) = \frac{\partial}{\partial T} g_l(x, T) = 0 \quad (l = 1, \ldots, L)$$

then

$$H(\psi(T^*), x^*(T^*), u^*(T-0), T^*) = 0 \tag{22.93}$$

If, in addition, the stationary case is considered (see (22.83)), then (22.93) holds for all $t \in [0, T^*]$, *that is,*

$$H(\psi(t), x^*(t), u^*(\psi(t), x^*(t))) = 0 \tag{22.94}$$

Proof. The result directly follows from (22.84) and (22.93). □

22.6.6 *Joint optimal control and parametric optimization problem*

Consider the nonlinear plant given by

$$\left.\begin{aligned} \dot{x}_a(t) &= f(x_a(t), u(t), t; a), \quad \text{a.e.} \quad t \in [0, T] \\ x_a(0) &= x_0 \end{aligned}\right\} \tag{22.95}$$

at the fixed horizon T, where $a \in \mathbb{R}^p$ is a vector of parameters which also can be selected to optimize the functional (22.44) which in this case is

$$J(u(\cdot), a) = h_0(x_a(T)) \tag{22.96}$$

(A4) It will be supposed that the right-hand side of (22.95) is differentiable on a at all $a \in \mathbb{R}^p$.

In view of this OCP is formulated as follows:

$$\begin{aligned} &\text{Minimize} \quad J(u(\cdot), a) \quad (22.96) \\ &\text{over} \quad u(\cdot) \in \mathcal{U}_{admis}[0, T] \quad \text{and} \quad a \in \mathbb{R}^p \end{aligned} \tag{22.97}$$

Theorem 22.14. (Joint OC and parametric optimization) *If under assumptions (A1)–(A3) and (A4) the pair $(u^*(\cdot), a^*)$ is a solution of the problem (22.85), (22.86) and $x^*(t)$ is the corresponding optimal trajectory, then there exist the vector functions $\psi(t)$ satisfying the system of the adjoint equations (22.68) with $x^*(t)$, $u^*(t)$, a^* and nonnegative constants $\mu \geq 0$ and $\nu_l \geq 0$ $(l = 1, \ldots, L)$ such that all four conditions of Theorem 22.11 are fulfilled and, in addition, the following condition to the optimal parameter holds:*

$$\int_{t=0}^{T} \frac{\partial}{\partial a} H(\psi(t), x^*(t), u^*(t), t; a^*) dt = 0 \tag{22.98}$$

Proof. For this problem $\mathcal{L}(u(\cdot), \mu, \nu, a)$ is defined by

$$\mathcal{L}(u(\cdot), \mu, \nu, a) := \mu h_0(x(T)) + \sum_{l=1}^{L} \nu_l g_l(x(T)) \tag{22.99}$$

Introduce the matrix $\Delta^a(t) = \dfrac{\partial}{\partial a} x^*(t) \in \mathbb{R}^{n \times p}$, called the *matrix of sensitivity (with respect to parameter variations)*, which satisfies the following differential equation:

$$\begin{aligned} \dot{\Delta}^a(t) &= \frac{d}{dt}\frac{\partial}{\partial a} x^*(t) = \frac{\partial}{\partial a}\dot{x}^*(t) = \frac{\partial}{\partial a} f(x^*(t), u^*(t), t; a^*) \\ &= \frac{\partial}{\partial a} f(x^*(t), u^*(t), t; a^*) \\ &\quad + \frac{\partial}{\partial x} f(x^*(t), u^*(t), t; a^*)\, \Delta^a(t), \quad \Delta^a(0) = 0 \end{aligned} \tag{22.100}$$

In view of this and using (22.68), it follows that

$$
\begin{aligned}
0 &\le \mathcal{L}(u^*(\cdot), \mu, \nu, a) - \mathcal{L}(u^*(\cdot), \mu, \nu, a^*) \\
&= (a - a^*)^\intercal \Delta^a(T)^\intercal \left(\mu \frac{\partial}{\partial x} h_0(x^*(T)) + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l(x^*(T)) \right) \\
&\quad + o(\|a - a^*\|) = (a - a^*)^\intercal \Delta^a(T)^\intercal \psi(T) + o(\|a - a^*\|) \\
&= (a - a^*)^\intercal [\Delta^a(T)^\intercal \psi(T) - \Delta^a(0)^\intercal \psi(0)] + o(\|a - a^*\|) \\
&= (a - a^*)^\intercal \int_{t=0}^{T} d[\Delta^a(t)^\intercal \psi(t)] + o(\|a - a^*\|) \\
&= (a - a^*)^\intercal \int_{t=0}^{T} \Big[-\Delta^a(t)^\intercal \frac{\partial}{\partial x} f(x^*(t), u^*(t), t; a^*)^\intercal \psi(t) \\
&\quad + \Delta^a(t)^\intercal \frac{\partial}{\partial x} f(x^*(t), u^*(t), t; a^*)^\intercal \psi(t) \\
&\quad + \frac{\partial}{\partial a} f(x^*(t), u^*(t), t; a^*)^\intercal \psi(t) \Big] dt + o(\|a - a^*\|) \\
&= (a - a^*)^\intercal \int_{t=0}^{T} \frac{\partial}{\partial a} f(x^*(t), u^*(t), t; a^*)^\intercal \psi(t) dt + o(\|a - a^*\|)
\end{aligned}
$$

But this inequality is possible for any $a \in \mathbb{R}^p$ in a small neighborhood of a^* if and only if the relation (22.98) holds (which may be proved by contradiction). Theorem is proven. □

22.6.7 *Sufficient conditions of optimality*

The necessary and sufficient conditions of the constrained concave optimization problem on $x \in X \subseteq \mathbb{R}^n$ is (see (21.80))

$$(\partial f(x^*), x - x^*) \le 0$$

which should be valid for all $x \in X$ (X is supposed to be a convex set and $f(x)$ is *concave* on X).

Here we will also need an additional assumption concerning the control region.

(A4) The control domain U is supposed to be a *convex body* (i.e., it is convex and has a nonempty interior).

Lemma 22.9. (on a mixed subgradient) *Let φ be a convex (or concave) function on $\mathbb{R}^n \times U$ where U is a convex body. Assuming that $\varphi(x, u)$ is differentiable in x and*

is continuous in (x, u)*, the following inclusion turns out to be valid for any* $(x^*, u^*) \in \mathbb{R}^n \times U$*:*

$$\{(\varphi_x(x^*, u^*), r) \mid r \in \partial_u \varphi(x^*, u^*)\} \subseteq \partial_{x,u} \varphi(x^*, u^*) \tag{22.101}$$

Proof. For any $y \in \mathbb{R}^n$, in view of the convexity of φ and its differentiability on x, it follows that

$$\varphi(x^* + y, u^*) - \varphi(x^*, u^*) \geq (\varphi_x(x^*, u^*), y) \tag{22.102}$$

Similarly, in view of the convexity of φ in u, there exists a vector $r \in \mathbb{R}^r$ such that for any $x^*, y \in \mathbb{R}^n$ and any $\bar{u} \in U$

$$\varphi(x^* + y, u^* + \bar{u}) - \varphi(x^* + y, u^*) \geq (r, \bar{u}) \tag{22.103}$$

So, taking into account the previous inequalities (22.102)–(22.103), we derive

$$\begin{aligned} &\varphi(x^* + y, u^* + \bar{u}) - \varphi(x^*, u^*) \\ &\quad = [\varphi(x^* + y, u^* + \bar{u}) - \varphi(x^* + y, u^*)] \\ &\qquad + [\varphi(x^* + y, u^*) - \varphi(x^*, u^*)] \geq (r, \bar{u}) + (\varphi_x(x^*, u^*), y) \end{aligned} \tag{22.104}$$

Then, by the definition of subgradient (21.69), it means that

$$(\varphi_x(x^*, u^*); r) \subseteq \partial d_{x,u} \varphi(x^*, u^*)$$

The concavity case is very similar if we note that $(-\varphi)$ is convex. □

Now we are ready to formulate the central result of this subsection.

Theorem 22.15. (Sufficient condition of optimality) *Let, under assumptions (A1)–(A4), the pair* $(x^*(\cdot), u^*(\cdot))$ *be an admissible pair and* $\psi(t)$ *be the corresponding adjoint variable satisfying (22.68). Assume that*

1. $h_0(x)$ *and* $g_l(x)$ $(l = 1, \ldots, L)$ *are* ***convex****;*
2. $H(\psi(t), x, u, t)$ *is* ***concave*** *in* (x, u) *for any fixed* $t \in [0, T]$ *and any* $\psi(t) \in \mathbb{R}^n$.

Then this pair $(x^*(\cdot), u^*(\cdot))$ *is optimal in the sense of the cost functional* $J(u(\cdot)) = h_0(x(T))$ *(22.44) if*

$$H(\psi(t), x^*(t), u^*(t), t) = \max_{u \in U} H(\psi(t), x^*(t), u, t) \tag{22.105}$$

at almost all $t \in [0, T]$.

Proof. By (22.105) and in view of the criterion of optimality (21.80), it follows for any $u \in U$ that

$$(\partial_u H(\psi(t), x^*(t), u^*(t), t), u - u^*(t)) \leq 0 \tag{22.106}$$

Then, by concavity of $H(\psi(t), x, u, t)$ in (x, u), for any admissible pair (x, u) and applying the integration operation, in view of (22.106) we get

$$\begin{aligned}
&\int_{t=0}^{T} H(\psi(t), x(t), u, t)dt - \int_{t=0}^{T} H(\psi(t), x^*(t), u^*(t), t)dt \\
&\quad \le \int_{t=0}^{T} \left[\left(\frac{\partial}{\partial x} H(\psi(t), x^*(t), u^*(t), t), x(t) - x^*(t) \right) \right. \\
&\quad \left. + \left(\partial_u H(\psi(t), x^*(t), u^*(t), t), u - u^*(t)\right) \right] dt \\
&\quad \le \int_{t=0}^{T} \left(\frac{\partial}{\partial x} H(\psi(t), x^*(t), u^*(t), t), x(t) - x^*(t) \right) dt
\end{aligned} \tag{22.107}$$

Let us introduce the "sensitivity" process $\delta(t) := x(t) - x^*(t)$ which evidently satisfies

$$\begin{aligned}
&\dot{\delta}(t) = \eta(t) \text{ a.e. } t \in [0, T], \delta(0) = 0 \\
&\eta(t) := f(x(t), u(t), t) - f(x^*(t), u^*(t), t)
\end{aligned} \tag{22.108}$$

Then, in view of (22.68) and (22.107), it follows that

$$\begin{aligned}
&\frac{\partial}{\partial x} h_0(x^*(T))^{\intercal} \delta(T) = -\left[\psi(T)^{\intercal} \delta(T) - \psi(0)^{\intercal} \delta(0)\right] \\
&\quad = -\int_{t=0}^{T} d\left[\psi(t)^{\intercal} \delta(t)\right] = \int_{t=0}^{T} \frac{\partial}{\partial x} H(\psi(t), x^*(t), u^*(t), t)^{\intercal} \delta(t)\, dt \\
&\quad - \int_{t=0}^{T} \psi(t)^{\intercal} \left(f(x(t), u(t), t) - f(x^*(t), u^*(t), t)\right) dt \\
&\quad \ge \int_{t=0}^{T} \left[H(\psi(t), x(t), u, t) - H(\psi(t), x^*(t), u^*, t)\right] dt \\
&\quad - \int_{t=0}^{T} \psi(t)^{\intercal} \left(f(x(t), u(t), t) - f(x^*(t), u^*(t), t)\right) dt = 0
\end{aligned} \tag{22.109}$$

The convexity of $h_0(x)$ and $g_l(x)$ $(l = 1, \ldots, L)$ and the complementary slackness condition yield $\left(\frac{\partial}{\partial x} g_l(x^*(T)), \delta(T) \right) \ge 0$ and, hence,

$$\begin{aligned}
&\frac{\partial}{\partial x} h_0 (x^*(T))\, \delta (T) \\
&\quad \le \left[\frac{\partial}{\partial x} h_0 (x^*(T)) + \sum_{l=1}^{L} \nu_l \frac{\partial}{\partial x} g_l (x^*(T)) \right]^{\mathsf{T}} \delta (T) \\
&\quad \le h_0 (x(T)) - h_0 (x^*(T)) + \sum_{l=1}^{L} \nu_l g_l (x^*(T)) \\
&\quad = h_0 (x(T)) - h_0 (x^*(T))
\end{aligned} \tag{22.110}$$

Combining (22.109) with (22.110), we derive

$$\begin{aligned}
&J (u(\cdot)) - J (u^*(\cdot)) \\
&\quad = h_0 (x(T)) - h_0 (x^*(T)) \ge \frac{\partial}{\partial x} h_0 (x^*(T))\, \delta (T) \ge 0
\end{aligned}$$

and, since $u(\cdot)$ is arbitrarily admissible, the desired result follows. □

Remark 22.3. *Notice that to check the concavity property of $H(\psi(t), x, u, t)$ (22.70) in (x, u) for any fixed $t \in [0, T]$ and any $\psi(t) \in \mathbb{R}^n$ is not a simple task since it depends on the sign of the $\psi_i(t)$ components. So, the theorem given above may be applied directly practically only for a very narrow class of particular problems where the concavity property may be analytically checked.*

Example 22.5. *Consider the following variation calculus problem:*

$$\begin{aligned}
&\int_{t=0}^{T} |\dot{x}(t)|\, dt \to \inf_{x \in C^1[0,T]} \\
&\dot{x}(t) \ge a > 0, \quad x(0) = 0, \quad x(T) = \xi
\end{aligned} \tag{22.111}$$

It can be represented as an optimal control problem. Indeed, denoting $\dot{x} := u$, the initial problem (22.111) can be represented as

$$\begin{aligned}
&\int_{t=0}^{T} |u(t)|\, dt \to \inf_{u \in C[0,T]} \\
&\dot{x}(t) = u(t), \quad u(t) \ge a > 0, \quad x(0) = 0, \quad x(T) = \xi
\end{aligned} \tag{22.112}$$

According to (22.81), the corresponding Hamiltonian function is

$$H(\psi, x, u, t) := \psi u - \mu |u(t)|$$

where $\psi = \psi(t)$ satisfies the following adjoint ODE (22.68) $\dot{\psi}(t) = 0$ with the transversality condition $\psi(T) = 0$, which gives $\psi(t) = 0$ for all $t \in [0, T]$. The nontriviality condition (22.73) implies $\mu > 0$. So, by the maximality condition (22.80), it follows that

$$\arg\max_{u \ge a} H(\psi, x, u, t) = \arg\min_{u \ge a} \mu |u(t)| = a$$

This leads to the following form for the optimal control: $u(t) = a$. The corresponding optimal curve is $x(t) = at + c$. The boundary conditions imply $c = 0$, $aT = \xi$. So, finally, we may conclude that the initial problem (22.111) has the unique solution $x(t) = at$ if and only if the terminal value $x(T) = \xi > 0$ is equal to $\xi = aT$. In any other cases the solution does not exist.

Example 22.6. *Consider the following variation calculus problem*

$$\begin{aligned} &T \to \inf_{x \in C^2[0,T]} \\ &|\ddot{x}(t)| \le 2 \qquad (22.113) \\ &x(0) = 1, \quad x(T) = -1, \quad \dot{x}(0) = \dot{x}(T) = 0 \end{aligned}$$

Let us introduce $u(t) := \ddot{x}(t) \in C[-1, T]$, $x_1(t) := x(t)$ and $x_2(t) := \dot{x}(t)$. Then the initial problem (22.113) can be represented as the following optimal control problem:

$$\begin{aligned} &T \to \inf_{u \in C[0,T]} \\ &\dot{x}_1 = x_2, \quad \dot{x}_2 = u, \quad |u(t)| \le 2 \qquad (22.114) \\ &x_1(0) = 1, \quad x_1(T) = -1, \quad x_2(0) = x_2(T) = 0 \end{aligned}$$

By (22.81), the corresponding Hamiltonian function is

$$H(\psi, x, u, t) := \psi_1 x_2 + \psi_2 u - \mu, \quad \mu \ge 0$$

since $T = \int_{t=0}^{T} h\,dt$ with $h = 1$. Here $\psi = \psi(t)$ satisfies the following system of the adjoint ODE (22.68)

$$\dot{\psi}_1(t) = 0, \quad \dot{\psi}_2(t) = -\psi_1$$

which solution is a ramp function

$$\begin{aligned} &\psi_1(t) = c_1 \\ &\psi_2(t) = -c_1 t + c_2,\ c_1, c_2 = \text{const} \end{aligned}$$

So, by the maximality condition (22.80), it follows that

$$\begin{aligned} u^*(t) &= \arg\max_{|u| \le 2} H(\psi, x, u, t) \\ &= \arg\max_{|u| \le 2} H(\psi_2 u) = 2 \operatorname{sign} \psi_2 = 2 \operatorname{sign}(c_2 - c_1 t) \end{aligned}$$

The corresponding optimal curve is

$$\begin{aligned} x_1(t) &= 2 \int_{s=0}^{t} \int_{\tau=0}^{s} \operatorname{sign}(c_2 - c_1 \tau)\, d\tau\, ds + c_3 t + c_4 \\ x_2(t) &= \int_{\tau=0}^{t} \operatorname{sign}(c_2 - c_1 \tau)\, d\tau + c_3 \end{aligned}$$

By the boundary conditions it follows that

$$c_4 = 1, \quad c_3 = 0$$

$$-1 = 2 \int_{s=0}^{T} \left[\int_{\tau=0}^{s} \operatorname{sign}(c_2 - c_1\tau)\, d\tau \right] ds + 1$$

$$0 = \int_{\tau=0}^{T} \operatorname{sign}(c_2 - c_1\tau)\, d\tau$$

Changing the time scale as $\tau' := \tau/T$ *leads to*

$$0 = \int_{\tau'=0}^{1} \operatorname{sign}\left(c_2 - c_1'\tau'\right) d\tau', \quad c_1' = c_1 T$$

which implies $2c_2 = c_1'$. *Since* $\operatorname{sign}(ab) = \operatorname{sign}(a)\operatorname{sign}(b)$

$$-1 = \int_{s=0}^{T} \left[\int_{\tau=0}^{s} \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau \right] ds$$

$$= \int_{s=0}^{T} \left[\int_{\tau=0}^{T} \chi_{s \geq \tau} \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau \right] ds$$

$$= \int_{\tau=0}^{T} \left(\int_{s=0}^{T} \chi_{s \geq \tau} ds \right) \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau = \int_{\tau=0}^{T} (T - \tau) \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau$$

$$= \int_{\tau=0}^{T/2} (T - \tau) \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau + \int_{\tau=T/2}^{T} [T - \tau] \operatorname{sign}\left[c_1 \left(\frac{T}{2} - \tau \right) \right] d\tau$$

$$\operatorname{sign} c_1 \left(\int_{\tau=0}^{T/2} (T - \tau)\, d\tau - \int_{\tau=T/2}^{T} (T - \tau)\, d\tau \right) = \frac{T^2}{4} \operatorname{sign} c_1$$

This leads to the following conclusion

$$c_1 < 0, \quad T = 2$$

Many other interesting examples can be found in Alexeev *et al.* (1984).

22.7 Dynamic programing

The dynamic programing method is another powerful approach to solving optimal control problems. It provides *sufficient conditions* for testing if some control is optimal or not. The basic idea of this approach consists of considering a family of optimal

control problems with different initial conditions (times and states) and of obtaining some relationships among them via the, so-called, *Hamilton–Jacoby–Bellman equation* (HJB) which is a nonlinear first-order partial differential equation. The optimal control can be designed by maximization (or minimization) of the generalized Hamiltonian involved in this equation. If this HJB equation is solvable (analytically or even numerically) then the corresponding optimal controllers turn out to be given by a nonlinear feedback depending on the optimized plant nonlinearity as well as the solution of the corresponding HJB equation. Such approach actually provides the solutions to the whole family of optimization problems, and, in particular, to the original problem. Such a technique is called "invariant embedding". The major drawback of the classical HJB method is that it requires that this partial differential equation admits a smooth enough solution. Unfortunately this is not the case even for some very simple situations. To overcome this problem the so-called *viscosity solutions* have been introduced (Crandall & Lions 1983). These solutions are some sort of nonsmooth solutions with a key function to replace the conventional derivatives by set-valued super/sub-differentials maintaining the uniqueness of solutions under very mild conditions. These approaches not only save the DPM as a mathematical method, but make it a powerful tool in tackling optimal control. In this section we do not touch on this approach. But we will discuss the gap between necessary (MP) and sufficient (DPM) conditions.

22.7.1 Bellman's principle of optimality

Claim 22.1. (Bellman's principle (BP) of optimality) "Any tail of an optimal trajectory is optimal too."[2]

In other words, if some trajectory in the phase space connects the initial $x(0)$ and terminal $x(T)$ points and is optimal in the sense of some cost functional, then the sub-trajectory, connecting any intermediate point $x(t')$ of the same trajectory with the same terminal point $x(T)$, should also be optimal (see Fig. 22.2).

22.7.2 Sufficient conditions for BP fulfilling

Theorem 22.16. (Sufficient condition for BP fulfilling) *Let*

1. *the performance index (a cost functional) $J(u(\cdot))$ with $u(\cdot) \in \mathcal{U}_{admis}[0, T]$ be* ***separable*** *for any time $t' \in (0, T)$ such that*

$$\boxed{J(u(\cdot)) = J_1(u_1(\cdot), \quad J_2(u_2(\cdot)))} \tag{22.115}$$

where $u_1(\cdot)$ is the control within the time interval $[0, t')$ called the ***initial control strategy*** *and $u_2(\cdot)$ is the control within the time interval $[t', T]$ called the* ***terminal control strategy****;*

[2] Bellman's principle of optimality, formulated in Bellman (1960), is as follows: "An optimal policy has the property that whatever the initial state and the initial decisions it must constitute an optimal policy with regards to the state resulting from the first decision."

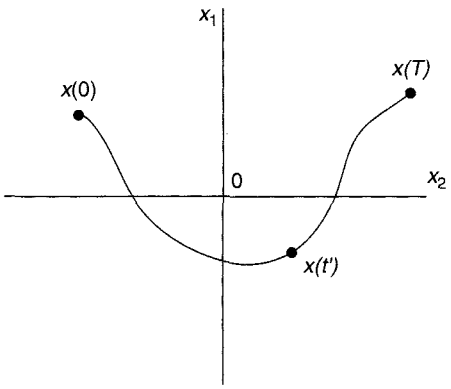

Fig. 22.2. Illustration of Bellman's principle of optimality.

2. *the functional* $J_1(u_1(\cdot), J_2(u_2(\cdot)))$ *is* ***monotonically nondecreasing with respect to its second argument*** $J_2(u_2(\cdot))$, *that is,*

$$\boxed{\begin{array}{l} J_1(u_1(\cdot), J_2(u_2(\cdot))) \geq J_1\left(u_1(\cdot), J_2\left(u_2'(\cdot)\right)\right) \\ \text{if} \quad J_2(u_2(\cdot)) \geq J_2\left(u_2'(\cdot)\right) \end{array}} \tag{22.116}$$

Then Bellman's principle of optimality takes place for this functional.

Proof. For any admissible control strategies $u_1(\cdot)$, $u_2(\cdot)$ the following inequality holds

$$\begin{aligned} J^* &:= \inf_{u \in \mathcal{U}_{\text{admis}}[0,T]} J(u(\cdot)) \\ &= \inf_{u_1 \in \mathcal{U}_{\text{admis}}[0,t'),\ u_2 \in \mathcal{U}_{\text{admis}}[t',T]} J_1(u_1(\cdot), J_2(u_2(\cdot))) \\ &\leq J_1(u_1(\cdot), J_2(u_2(\cdot))) \end{aligned} \tag{22.117}$$

Select

$$u_2(\cdot) = \arg \inf_{u_2 \in \mathcal{U}_{\text{admis}}[t',T]} J_2(u_2(\cdot)) \tag{22.118}$$

Then (22.117) and (22.118) imply

$$J^* \leq J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{\text{admis}}[t',T]} J_2(u_2(\cdot))\right) \tag{22.119}$$

So,

$$u_1(\cdot) = \arg \inf_{u_1 \in \mathcal{U}_{\text{admis}}[t',T]} J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{\text{admis}}[t',T]} J_2(u_2(\cdot))\right) \tag{22.120}$$

leads to

$$J^* \leq \inf_{u_1 \in \mathcal{U}_{\text{admis}}[t',T]} J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{\text{admis}}[t',T]} J_2(u_2(\cdot))\right) \tag{22.121}$$

Since $J_1(u_1(\cdot), J_2(u_2(\cdot)))$ is monotonically nondecreasing with respect to the second argument, from (22.121) we obtain

$$\begin{aligned}
&\inf_{u_1 \in \mathcal{U}_{admis}[t',T]} J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{admis}[t',T]} J_2(u_2(\cdot))\right) \\
&\quad \leq \inf_{u_1 \in \mathcal{U}_{admis}[t',T]} \inf_{u_2 \in \mathcal{U}_{admis}[t',T]} J_1(u_1(\cdot), J_2(u_2(\cdot))) \\
&\quad = \inf_{u \in \mathcal{U}_{admis}[0,T]} J(u(\cdot)) = J^*
\end{aligned} \tag{22.122}$$

Combining (22.121) and (22.122), we finally derive that

$$J^* = \inf_{u_1 \in \mathcal{U}_{admis}[t',T]} J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{admis}[t',T]} J_2(u_2(\cdot))\right) \tag{22.123}$$

This proves the desired result. □

Summary 22.3. *In strict mathematical form this fact may be expressed as follows: under the assumptions of the theorem above for any time* $t' \in (0, T)$

$$\begin{aligned}
&\inf_{u \in \mathcal{U}_{admis}[0,T]} J(u(\cdot)) \\
&\quad = \inf_{u_1 \in \mathcal{U}_{admis}[t',T]} J_1\left(u_1(\cdot), \inf_{u_2 \in \mathcal{U}_{admis}[t',T]} J_2(u_2(\cdot))\right)
\end{aligned} \tag{22.124}$$

Corollary 22.7. *For the cost functional*

$$J(u(\cdot)) := h_0(x(T)) + \int_{t=0}^{T} h(x(t), u(t), t)\, dt$$

given in the Bolza form (22.41) Bellman's principle holds.

Proof. For any $t' \in (0, T)$ from (22.41) obviously it follows that

$$J(u(\cdot)) = J_1(u_1(\cdot)) + J_2(u_2(\cdot)) \tag{22.125}$$

where

$$\begin{aligned}
J_1(u_1(\cdot)) &:= \int_{t=0}^{t'} h(x(t), u_1(t), t)\, dt \\
J_2(u_2(\cdot)) &:= h_0(x(T)) + \int_{t=t'}^{T} h(x(t), u_2(t), t)\, dt
\end{aligned} \tag{22.126}$$

The representation (22.125) evidently yields the validity (22.115) and (22.116) for this functional. □

22.7.3 Invariant embedding

22.7.3.1 System description and basic assumptions

Let $(s, y) \in [0, T) \times \mathbb{R}^n$ be "an initial time and state pair" to the following controlled system over $[s, T]$:

$$\left.\begin{aligned} &\dot{x}(t) = f(x(t), u(t), t), \quad \text{a.e.} \quad t \in [s, T] \\ &x(s) = y \end{aligned}\right\} \tag{22.127}$$

where $x \in \mathbb{R}^n$ is its state vector, and $u \in \mathbb{R}^r$ is the control that may run over a given control region $U \subset \mathbb{R}^r$ with the cost functional in the Bolza form

$$J(s, y; u(\cdot)) := h_0(x(T)) + \int_{t=s}^{T} h(x(t), u(t), t)\, dt \tag{22.128}$$

containing the integral term as well as the terminal one and with *the terminal set* $\mathcal{M} \subseteq \mathbb{R}^n$ given by the inequalities (22.42). Here, as before, $u(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$. For $s = 0$ and $y = x_0$ this plant coincides with the original one given by (22.40).

Suppose also that assumption (A1) is accepted and, instead of (A2), its small modification holds:

(A2′) The maps

$$\left.\begin{aligned} &f : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R}^n \\ &h : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R} \\ &h_0 : \mathbb{R}^n \times U \times [0, T] \to \mathbb{R} \\ &g_l : \mathbb{R}^n \to \mathbb{R}^n \quad (l = 1, \ldots, L) \end{aligned}\right\} \tag{22.129}$$

are **uniformly continuous** in (x, u, t) including t (before in (A2) they were assumed to be only measurable) and there exists a constant L such that for $\varphi = f(x, u, t)$, $h(x, u, t)$, $h_0(x, u, t)$, $g_l(x)$ $(l = 1, \ldots, L)$ the following inequalities hold:

$$\left.\begin{aligned} &\left\|\varphi(x, u, t) - \varphi(\hat{x}, \hat{u}, t)\right\| \le L \left\|x - \hat{x}\right\| \\ &\forall t \in [0, T], \ x, \hat{x} \in \mathbb{R}^n, \ u \in U \\ &\|\varphi(0, u, t)\| \le L \quad \forall u, t \in U \times [0, T] \end{aligned}\right\} \tag{22.130}$$

It is evident that under assumptions (A1)–(A2′) for any $(s, y) \in [0, T) \times \mathbb{R}^n$ and any $u(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$ the optimization problem

$$J(s, y; u(\cdot)) \to \min_{u(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]} \tag{22.131}$$

formulated for the plant (22.127) and for the cost functional $J(s, y; u(\cdot))$ (22.128), admits a unique solution $x(\cdot) := x(\cdot, s, y, u(\cdot))$ and the functional (22.128) is well defined.

Definition 22.8. (The value function) *The function* $V(s, y)$ *defined for any* $(s, y) \in [0, T) \times \mathbb{R}^n$ *as*

$$\left.\begin{aligned} V(s, y) &:= \inf_{u(\cdot)\in\mathcal{U}_{\text{admis}}[s,T]} J(s, y; u(\cdot)) \\ V(T, y) &= h_0(y) \end{aligned}\right\} \tag{22.132}$$

is called the ***value function*** *of the optimization problem (22.131).*

22.7.3.2 Dynamic programing equation in the integral form

Theorem 22.17. *Under assumptions (A1)–(A2′) for any* $(s, y) \in [0, T) \times \mathbb{R}^n$ *the following relation holds*

$$\begin{aligned} V(s, y) = \inf_{u(\cdot)\in\mathcal{U}_{\text{admis}}[s,T]} \Bigg\{ &\int_{t=s}^{\hat{s}} h(x(t, s, y, u(\cdot)), u(t), t)\, dt \\ &+ V\left(\hat{s}, x\left(\hat{s}, s, y, u(\cdot)\right)\right) \Bigg\} \qquad \forall \hat{s} \in [s, T] \end{aligned} \tag{22.133}$$

Proof. The result follows directly from BP of optimality (22.124), but, in view of the great importance of this result, we present the proof again, using the concrete form of the Bolza cost functional (22.128). Denoting the right-hand side of (22.133) by $\bar{V}(s, y)$ and taking into account the definition (22.132), for any $u(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$ we have

$$\begin{aligned} V(s, y) &\le J(s, y; u(\cdot)) \\ &= \int_{t=s}^{\hat{s}} h(x(t, s, y, u(\cdot)), u(t), t)\, dt + J\left(\hat{s}, x(\hat{s}); u(\cdot)\right) \end{aligned}$$

and, taking infimum over $u(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$, it follows that

$$V(s, y) \le \bar{V}(s, y) \tag{22.134}$$

Hence, for any $\varepsilon > 0$ there exists a control $u_\varepsilon(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$ such that for $x_\varepsilon(\cdot) := x(\cdot, s, y, u_\varepsilon(\cdot))$

$$\begin{aligned} V(s, y) + \varepsilon &\ge J(s, y; u_\varepsilon(\cdot)) \\ &\ge \int_{t=s}^{\hat{s}} h(x(t, s, y, u_\varepsilon(\cdot)), u_\varepsilon(t), t)\, dt + V\left(\hat{s}, x_\varepsilon\left(\hat{s}\right)\right) \ge \bar{V}(s, y) \end{aligned} \tag{22.135}$$

Tending $\varepsilon \to 0$ the inequalities (22.134), (22.135) imply the result (22.133) of this theorem. □

Finding a solution $V(s, y)$ to equation (22.133), we would be able to solve the origin optimal control problem putting $s = 0$ and $y = x_0$. Unfortunately, this equation is very difficult to handle because of overcomplicated operations involved on its right-hand side. That's why in the next subsection we will explore this equation further, trying to get another equation for the function $V(s, y)$ with a simpler and more practically used form.

22.7.4 Hamilton–Jacoby–Bellman equation

To simplify the sequent calculations and following Young & Zhou (1999) we will consider the original optimization problem without any terminal set, that is, $\mathcal{M} = \mathbb{R}^n$. This may be expressed with the constraint function equal to

$$g(x) := 0 \cdot \|x\|^2 - \varepsilon \leq 0 \quad (\varepsilon > 0) \tag{22.136}$$

which is true for any $x \in \mathbb{R}^n$. Slater's condition (21.88) is evidently valid (also for any $x \in \mathbb{R}^n$). So, we deal here with the regular case. Denote by $C^1\left([0, T) \times \mathbb{R}^n\right)$ the set of all continuously differentiable functions $v : [0, T) \times \mathbb{R}^n \to \mathbb{R}$.

Theorem 22.18. (The HJB equation) *Suppose that under assumptions (A1)–(A2′) the value function $V(s, y)$ (22.132) is continuously differentiable, that is, $V \in C^1\left([0, T) \times \mathbb{R}^n\right)$. Then $V(s, y)$ is a solution to the following terminal value problem of a first-order partial differential equation, named below the* ***Hamilton–Jacoby–Bellman (HJB) equation*** *associated with the original optimization problem (22.131) without terminal set ($\mathcal{M} = \mathbb{R}^n$):*

$$\left.\begin{aligned} &-\frac{\partial}{\partial t} V(t, x) + \sup_{u \in U} H\left(-\frac{\partial}{\partial x} V(t, x), x(t), u(t), t\right) = 0 \\ &(t, x) \in [0, T) \times \mathbb{R}^n, \quad V(T, x) = h_0(x), \quad x \in \mathbb{R}^n \end{aligned}\right\} \tag{22.137}$$

where

$$\begin{aligned} &H(\psi, x, u, t) := \psi^{\mathsf{T}} f(x, u, t) - h(x(t), u(t), t) \\ &t, x, u, \psi \in [0, T] \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^n \end{aligned} \tag{22.138}$$

is the same as in (22.81) with $\mu = 1$ corresponding to the regular optimization problem.

Proof. Fixing $u(t) \equiv u \in U$, by (22.133) with $\hat{s} \downarrow s$ we obtain

$$\frac{V(s, y) - V\left(\hat{s}, x\left(\hat{s}, s, y, u(\cdot)\right)\right)}{\hat{s} - s} - \frac{1}{\hat{s} - s} \int_{t=s}^{\hat{s}} h(x(t, s, y, u(\cdot)), u(t), t)\, dt \leq 0$$

which implies

$$-\frac{\partial}{\partial t} V(s, y) - \frac{\partial}{\partial x} V(s, y)^{\mathsf{T}} f(s, y, u) - h(s, u, t) \leq 0$$

resulting in the following inequality

$$0 \geq -\frac{\partial}{\partial t} V(s, y) + \sup_{u \in U} H\left(-\frac{\partial}{\partial x} V(t, x), x(t), u(t), t\right) \tag{22.139}$$

On the other hand, for any $\varepsilon > 0$ and s, closed to $\hat{s}$, there exists a control $u(\cdot) := u_{\varepsilon,\hat{s}}(\cdot) \in \mathcal{U}_{\text{admis}}[s, T]$ for which

$$V(s, y) + \varepsilon(\hat{s} - s) \geq \int_{t=s}^{\hat{s}} h(x(t, s, y, u(\cdot)), u(t), t)\, dt + V(\hat{s}, x(\hat{s})) \tag{22.140}$$

Since $V \in C^1([0, T) \times \mathbb{R}^n)$, the last inequality leads to the following

$$\begin{aligned}
-\varepsilon &\leq -\frac{V(\hat{s}, x(\hat{s})) - V(s, y)}{\hat{s} - s} - \frac{1}{\hat{s} - s} \int_{t=s}^{\hat{s}} h(x(t, s, y, u(\cdot)), u(t), t)\, dt \\
&= \frac{1}{\hat{s} - s} \int_{t=s}^{\hat{s}} \Big[-\frac{\partial}{\partial t} V(t, x(t, s, y, u(\cdot))) \\
&\quad - \frac{\partial}{\partial x} V(t, x(t, s, y, u(\cdot)))^{\mathsf{T}} f(t, x(t, s, y, u(\cdot)), u) \\
&\quad - h(x(t, s, y, u(\cdot)), u(t), t) \Big] dt \\
&= \frac{1}{\hat{s} - s} \int_{t=s}^{\hat{s}} \Big[-\frac{\partial}{\partial t} V(t, x(t, s, y, u(\cdot))) \\
&\quad + H(-\frac{\partial}{\partial x} V(t, x(t, s, y, u(\cdot))), x(t, s, y, u(\cdot)), u(t), t) \Big] dt \\
&\leq \frac{1}{\hat{s} - s} \int_{t=s}^{\hat{s}} \Big[-\frac{\partial}{\partial t} V(t, x(t, s, y, u(\cdot))) \\
&\quad + \sup_{u \in U} H(-\frac{\partial}{\partial x} V(t, x(t, s, y, u(\cdot))), x(t, s, y, u(\cdot)), u(t), t) \Big] dt
\end{aligned} \tag{22.141}$$

which for $\hat{s} \downarrow s$ gives

$$-\varepsilon \leq -\frac{\partial}{\partial t} V(s, y) + \sup_{u \in U} H(-\frac{\partial}{\partial x} V(s, y), y, u, s) \tag{22.142}$$

Here the uniform continuity property of the functions f and h has been used, namely,

$$\lim_{t \downarrow s} \sup_{y \in \mathbb{R}^n, u \in U} \|\varphi(t, y, u) - \varphi(s, y, u)\| = 0, \quad \varphi = f, h \tag{22.143}$$

Combining (22.139) and (22.142) when $\varepsilon \to 0$ we obtain (22.137). □

The theorem below, representing the *sufficient conditions* of *optimality*, is known as the *verification rule*.

Theorem 22.19. (The verification rule) *Accept the following assumptions:*

1. *Let* $u^*(\cdot) := u^*\left(t, x, \dfrac{\partial}{\partial x} V(t,x)\right)$ *be a solution to the following optimization problem*

$$H(-\frac{\partial}{\partial x} V(t,x), x, u, t) \to \sup_{u \in U} \tag{22.144}$$

with fixed values x, t *and* $\dfrac{\partial}{\partial x} V(t,x)$;

2. *Suppose that we can obtain the solution* $V(t,x)$ *to the HJB equation*

$$\left.\begin{aligned} &-\frac{\partial}{\partial t} V(t,x) + H(-\frac{\partial}{\partial x} V(t,x), x, u^*(\cdot), t) = 0 \\ &V(T,x) = h_0(x), \quad (t,x) \in [0,T) \times \mathbb{R}^n \end{aligned}\right\} \tag{22.145}$$

which for any $(t,x) \in [0,T) \times \mathbb{R}^n$ *is unique and smooth, that is,* $V \in C^1\left([0,T) \times \mathbb{R}^n\right)$;

3. *Suppose that for any* $(s,x) \in [0,T) \times \mathbb{R}^n$ *there exists (a.e.* $t \in [s,T]$*) a solution* $x^*(s,x)$ *to the following ODE (ordinary differential equation)*

$$\begin{aligned} \dot{x}^*(t) &= f\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right), t\right) \\ x^*(s) &= x \end{aligned} \tag{22.146}$$

Then with $(s,x) = (0, x_0)$ *the pair*

$$\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right)\right) \tag{22.147}$$

is optimal, that is, $u^*\left(t, x^*(t), \dfrac{\partial}{\partial x} V(t, x^*(t))\right)$ *is an* ***optimal control.***

Proof. The relations (22.138) and (22.145) imply

$$\begin{aligned} \frac{d}{dt} V(t, x^*(t)) &= -\frac{\partial}{\partial t} V(t, x^*(t)) \\ &\quad + \frac{\partial}{\partial x} V(t, x^*(t))^{\intercal} f\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right), t\right) \\ &= -h\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right), t\right) \end{aligned} \tag{22.148}$$

Integrating this equality by t within $[s,T]$ leads to the following relation

$$\begin{aligned} &V(T, x^*(T)) - V(s, x^*(s)) \\ &= -\int_{t=s}^{T} h\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right), t\right) dt \end{aligned}$$

which, in view of the identity $V(T, x^*(T)) = h_0(x^*(T))$, is equal to the following one

$$V(s, x^*(s)) = h_0(x^*(T)) + \int_{t=s}^{T} h\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right), t\right) dt \tag{22.149}$$

By (22.133), this last equation means exactly that

$$\left(x^*(t), u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right)\right)$$

is an optimal pair and $u^*\left(t, x^*(t), \frac{\partial}{\partial x} V(t, x^*(t))\right)$ is an optimal control. □

22.8 Linear quadratic optimal control

22.8.1 Nonstationary linear systems and quadratic criterion

Consider in this section the dynamic plants (22.40) in their partial representation when at each time $t \in [0, T]$ the right-hand side of the mathematical model is a linear function with respect to the state vector $x(t)$ and the control action $u(t)$ as well, namely, for almost all $t \in [0, T]$

$$\left.\begin{array}{l} \dot{x}(t) = A(t)\, x(t) + B(t)\, u(t) + d(t) \\ x(0) = x_0 \end{array}\right\} \tag{22.150}$$

Here the functional matrices $A(t) \in \mathbb{R}^{n\times n}$ and $B(t) \in \mathbb{R}^{n\times r}$ are supposed to be bounded almost everywhere and the *shifting vector function* $d(t) \in \mathbb{R}^n$ is quadratically integrable, that is,

$$A(\cdot) \in \mathcal{L}^{\infty}\left(0, T; \mathbb{R}^{n\times n}\right), \quad B(\cdot) \in \mathcal{L}^{\infty}\left(0, T; \mathbb{R}^{n\times r}\right), \quad d(\cdot) \in \mathcal{L}^{2}(0, T; \mathbb{R}^{n}) \tag{22.151}$$

The admissible control is assumed to be quadratically integrable on $[0, T]$ and the terminal set $\mathcal{M}$ coincides with all space $\mathbb{R}^n$ (no terminal constraints), i.e.,

$$\mathcal{U}_{\text{admis}}[0, T] := \left\{u(\cdot) : u(\cdot) \in \mathcal{L}^{2}(0, T; \mathbb{R}^{r}),\ \mathcal{M} = \mathbb{R}^{n}\right\} \tag{22.152}$$

The cost functional is considered in the form (22.41) with quadratic functions inside, that is,

$$J(u(\cdot)) = \frac{1}{2} x^{\mathsf{T}}(T)\, G x(T) + \frac{1}{2} \int_{t=0}^{T} \left[x^{\mathsf{T}}(t)\, Q(t)\, x(t) + 2u^{\mathsf{T}}(t)\, S(t)\, x(t) + u^{\mathsf{T}}(t)\, R(t)\, u(t)\right] dt \tag{22.153}$$

where

$$G \in \mathbb{R}^{n\times n}, \quad Q(\cdot) \in \mathcal{L}^{\infty}\left(0, T; \mathbb{R}^{n\times n}\right)$$
$$S(\cdot) \in \mathcal{L}^{\infty}\left(0, T; \mathbb{R}^{n\times r}\right), \quad R(\cdot) \in \mathcal{L}^{\infty}\left(0, T; \mathbb{R}^{r\times r}\right) \tag{22.154}$$

such that for almost all $t \in [0, T]$

$$G \geq 0, \quad Q(t) \geq 0, \quad R(t) \geq \delta I, \quad \delta > 0 \tag{22.155}$$

Note that all coefficients (except G) in (22.150) and (22.153) are dependent on time t.

22.8.2 Linear quadratic problem

Problem 22.3. (Linear quadratic (LQ) problem) *For the dynamic model (22.150) find an admissible control $u^*(\cdot) \in \mathcal{U}_{admis}[0, T]$ such that*

$$J(u^*(\cdot)) = \inf_{u(\cdot)\in\mathcal{U}_{admis}[0,T]} J(u(\cdot)) \tag{22.156}$$

where the cost function $J(u(\cdot))$ is given by (22.153).

We will refer to this problem as *the linear quadratic optimal control problem* (LQ).

22.8.3 Maximum principle for LQ problem

22.8.3.1 MP formulation

Theorem 22.20. (Maximum principle for LQ problem) *If a pair $(x^*(t), u^*(\cdot))$ is optimal, then*

1. *there exists a solution $\psi(t)$ to the following ODE on $[0, T]$*

$$\left.\begin{aligned} \dot{\psi}(t) &= -A^{\mathsf{T}}(t)\psi(t) + Q(t)x^*(t) + S^{\mathsf{T}}(t)u^*(t) \\ \psi(T) &= -Gx^*(T) \end{aligned}\right\} \tag{22.157}$$

2. *the optimal control $u^*(\cdot) \in \mathcal{U}_{admis}[0, T]$ is as follows*

$$u^*(t) = R^{-1}(t)\left[B^{\mathsf{T}}(t)\psi(t) - S(t)x^*(t)\right] \tag{22.158}$$

Proof. Since in this problem we have no terminal conditions, we deal with the regular case and may take $\mu = 1$. Then by (22.81) and (22.82) it follows that

$$H(\psi, x, u, t) := \psi^{\mathsf{T}}\left[A(t)x + B(t)u + d(t)\right] - \frac{1}{2}x^{\mathsf{T}}(t)Q(t)x(t) - u^{\mathsf{T}}(t)S(t)x(t) - \frac{1}{2}u^{\mathsf{T}}(t)R(t)u(t) \tag{22.159}$$

So,

$$\begin{aligned}\dot{\psi}(t) &= -\frac{\partial}{\partial x} H(\psi(t), x^*(t), u^*(t), t) \\ &= -A^\intercal(t)\,\psi(t) + Q(t)\,x^*(t) + S^\intercal(t)\,u^*(t) \\ \psi(T) &= -\frac{\partial}{\partial x} h_0(x^*(T)) = -Gx^*(T)\end{aligned}$$

which proves claim (1) (22.157) of this theorem. Besides, by MP implementation, we have

$$u^*(t) \in \operatorname{Arg}\min_{u\in\mathbb{R}^r} H(\psi, x^*, u, t)$$

or, equivalently,

$$\frac{\partial}{\partial u} H(\psi, x^*, u^*, t) = B^\intercal(t)\,\psi(t) - R(t)\,u^*(t) - S(t)\,x^*(t) = 0 \tag{22.160}$$

which leads to claim (2) (22.158). □

22.8.4 Sufficiency condition

Theorem 22.21. (on the sufficiency of MP) *If the control $u^*(t)$ is as in (22.158) and*

$$Q(t) - S(t)\,R^{-1}(t)\,S^\intercal(t) \geq 0 \tag{22.161}$$

then it is a unique optimal one.

Proof. It follows directly from Theorem 22.15 on the sufficient conditions of optimality. The uniqueness is the result of equation (22.160) which has a unique solution if $R(t) \geq \delta I$ a.e. $t \in [0, T]$, $\delta > 0$ (22.155). Besides, the Hessian of the function $H(\psi, x, u, t)$ (22.159) is as follows

$$\left\| \begin{matrix} \frac{\partial^2}{\partial x^2} H(\psi, x, u, t) & \frac{\partial^2}{\partial x \partial u} H(\psi, x, u, t) \\ \frac{\partial^2}{\partial u \partial x} H(\psi, x, u, t) & \frac{\partial^2}{\partial u^2} H(\psi, x, u, t) \end{matrix} \right\| = -\left\| \begin{matrix} Q(t) & S(t) \\ S^\intercal(t) & R(t) \end{matrix} \right\|$$

Let us show that $\left\| \begin{matrix} Q(t) & S(t) \\ S^\intercal(t) & R(t) \end{matrix} \right\| \geq 0$. A symmetric block-matrix $\begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\intercal & M_{22} \end{bmatrix}$ with $M_{22} > 0$ is nonnegative definite, that is,

$$\begin{bmatrix} M_{11} & M_{12} \\ M_{12}^\intercal & M_{22} \end{bmatrix} \geq 0$$

if and only if

$$M_{11} \geq 0, \quad M_{11} - M_{12} M_{22}^{-1} M_{12}^\intercal \geq 0$$

So, by the assumption (22.161) of the theorem we have that

$$\left\| \begin{array}{cc} \dfrac{\partial^2}{\partial x^2} H(\psi, x, u, t) & \dfrac{\partial^2}{\partial x \partial u} H(\psi, x, u, t) \\ \dfrac{\partial^2}{\partial u \partial x} H(\psi, x, u, t) & \dfrac{\partial^2}{\partial u^2} H(\psi, x, u, t) \end{array} \right\| \leq 0$$

This means that the function $H(\psi, x, u, t)$ is concave (not obligatory strictly) on (x, u) for any fixed $\psi(t)$ and any $t \in [0, T]$. □

Corollary 22.8. *If $S(t) \equiv 0$, then the control $u^*(t)$ (22.158) is always uniquely optimal.*

Proof. Under this assumption the inequality (22.161) always holds. □

22.8.5 Riccati differential equation and feedback optimal control

22.8.5.1 Riccati differential equation

Let us introduce the symmetric matrix function $P(t) = P^{\mathsf{T}}(t) \in C^1\left(0, T; \mathbb{R}^{n \times n}\right)$ and the vector function $p(t) \in C^1(0, T; \mathbb{R}^n)$ which satisfy (a.e. $t \in [0, T]$) the following ODE:

$$\left.\begin{aligned} -\dot{P}(t) &= P(t) A(t) + A^{\mathsf{T}}(t) P(t) + Q(t) \\ &\quad - \left[B^{\mathsf{T}}(t) P(t) + S(t)\right]^{\mathsf{T}} R^{-1}(t) \left[B^{\mathsf{T}}(t) P(t) + S(t)\right] \\ &= P(t) \tilde{A}(t) + \tilde{A}(t)^{\mathsf{T}} P(t) - P(t) \left[B(t) R^{-1}(t) B^{\mathsf{T}}(t)\right] P(t) + \tilde{Q}(t) \\ P(T) &= G \end{aligned}\right\} \tag{22.162}$$

with

$$\left.\begin{aligned} \tilde{A}(t) &= A(t) - B(t) R^{-1}(t) S(t) \\ \tilde{Q}(t) &= Q(t) - S^{\mathsf{T}}(t) R^{-1}(t) S(t) \end{aligned}\right\} \tag{22.163}$$

and

$$\left.\begin{aligned} -\dot{p}(t) &= \left[\left(A(t) - B(t) R^{-1}(t) S(t)\right)^{\mathsf{T}}\right. \\ &\quad \left. - P(t) B(t) R^{-1}(t) B^{\mathsf{T}}(t)\right] p(t) + P(t) d(t) \\ p(T) &= 0 \end{aligned}\right\} \tag{22.164}$$

Definition 22.9. *We refer to ODE (22.162) as the **Riccati differential equation** and $p(t)$ is referred to as the **shifting vector** associated with the problem (22.156).*

22.8.6 Linear feedback control

Theorem 22.22. (on a linear feedback control) *Assume that*

$$P(t) = P^{\mathsf{T}}(t) \in C^1\left(0, T; \mathbb{R}^{n \times n}\right)$$

is a solution of (22.162) and

$$p(t) \in C^1(0, T; \mathbb{R}^n)$$

verifies (22.164). Then the optimal control $u^*(\cdot) \in \mathcal{U}_{\text{admis}}[0, T]$ *for the problem (22.156) has the linear feedback form*

$$u^*(t) = -R^{-1}(t)\left[\left(B^\intercal(t) P(t) + S(t)\right) x^*(t) + B^\intercal(t) p(t)\right] \tag{22.165}$$

and the optimal cost function $J(u^*(\cdot))$ *is as follows*

$$\begin{aligned} J(u^*(\cdot)) = &\frac{1}{2} x_0^\intercal P(0) x_0 + p^\intercal(0) x_0 \\ &+ \frac{1}{2}\int_{t=0}^{T} \left[2p^\intercal(t) d(t) - \left\|R^{-1/2}(t) B^\intercal(t) p(t)\right\|^2\right] dt \end{aligned} \tag{22.166}$$

Proof.

1. Let us try to find the solution of ODE (22.157) in the form

$$\psi(t) = -P(t) x^*(t) - p(t) \tag{22.167}$$

The direct substitution of (22.167) into (22.157) leads to the following identity (t will be suppressed for simplicity):

$$\begin{aligned} &\left(Q - S^\intercal R^{-1} S\right) x^* - \left(A^\intercal - S^\intercal R^{-1} B^\intercal\right)\left[-P(t) x^* - p\right] = \dot{\psi} \\ &= -\dot{P} x^* - P\left[Ax^* + B(u^*) + d\right] - \dot{p} = -\dot{P} x^* \\ &-P\left[Ax^* + BR^{-1}\left[B^\intercal(-Px^* - p) - Sx^*\right] + d\right] - \dot{p} \\ &= -\dot{P} x^* - P\left(A - BR^{-1}\left[B^\intercal P + S\right]\right) x^* + PBR^{-1} p - Pd - \dot{p} \end{aligned}$$

This yields

$$\begin{aligned} 0 = &\left(\dot{P}(t) + P(t) A(t) + A^\intercal(t) P(t) + Q(t)\right. \\ &\left. - \left[B^\intercal(t) P(t) + S(t)\right]^\intercal R^{-1}(t)\left[B^\intercal(t) P(t) + S(t)\right]\right) x^* \\ \dot{p}(t) + &\left[\left(A(t) - B(t) R^{-1}(t) S(t)\right)^\intercal\right. \\ &\left. - P(t) B(t) R^{-1}(t) B^\intercal(t)\right] p(t) + P(t) d(t) \end{aligned} \tag{22.168}$$

But, in view of (22.162) and (22.164), the right-hand side of (22.168) is identically zero. The transversality condition $\psi(T) = -Gx^*(T)$ in (22.162) implies

$$\psi(T) = -P(T) x^*(T) - p(T) = -Gx^*(T)$$

which holds for any $x^*(T)$ if $P(T) = G$ and $p(T) = 0$.

2. To prove (22.166) let us apply the chain integration rule for $x^{\mathsf{T}}(t)\,P(t)\,x(t)$ and for $p^{\mathsf{T}}(t)\,x(t)$, respectively. In view of (22.150) and (22.162) we obtain

$$\begin{aligned}
&x^{\mathsf{T}}(T)\,P(T)\,x(T)-x^{\mathsf{T}}(s)\,P(s)\,x(s)\\
&\quad=x^{*\mathsf{T}}(T)\,Gx^{*}(T)-x^{*\mathsf{T}}(s)\,P(s)\,x^{*}(s)\\
&\quad=\int_{t=s}^{T}\frac{d}{dt}\left[x^{\mathsf{T}}(t)\,P(t)\,x(t)\right]dt=\int_{t=s}^{T}\left[2x^{\mathsf{T}}(t)\,P(t)\,\dot{x}(t)+x^{\mathsf{T}}(t)\,\dot{P}(t)\,x(t)\right]dt\\
&\quad=\int_{t=s}^{T}\{x^{\mathsf{T}}(t)\left(\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]R^{-1}(t)\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]^{\mathsf{T}}\right.\\
&\qquad\left.-Q(t)\right)x(t)+2u^{*\mathsf{T}}(t)\,B^{\mathsf{T}}(t)\,P(t)\,x(t)+2d^{\mathsf{T}}(t)\,P(t)\,x(t)\}\,dt
\end{aligned}\tag{22.169}$$

and, applying (22.164),

$$\begin{aligned}
&p^{\mathsf{T}}(T)\,x(T)-p^{\mathsf{T}}(s)\,x(s)=-p^{\mathsf{T}}(s)\,x(s)\\
&\quad=\int_{t=s}^{T}\frac{d}{dt}\left[p^{\mathsf{T}}(t)\,x(t)\right]dt=\int_{t=s}^{T}\left[\dot{p}^{\mathsf{T}}(t)\,x(t)+p^{\mathsf{T}}(t)\,\dot{x}(t)\right]dt\\
&\quad=\int_{t=s}^{T}\{x^{\mathsf{T}}(t)\left(\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]R^{-1}(t)\,B(t)\,p(t)-P(t)\,d(t)\right)\\
&\qquad+p^{\mathsf{T}}(t)\left[B(t)\,u^{*}(t)+d(t)\right]\}\,dt
\end{aligned}\tag{22.170}$$

Summing (22.169) and (22.170) and denoting

$$\begin{aligned}
&J^{*}(s,x(s)):=\frac{1}{2}x^{\mathsf{T}}(s)\,P(s)\,x(s)\\
&\quad+\frac{1}{2}\int_{t=s}^{T}\left[x^{\mathsf{T}}(t)\,Q(t)\,x(t)+u^{*\mathsf{T}}(t)\,R(t)\,u^{*}(t)+2u^{*\mathsf{T}}(t)\,S(t)\,x(t)\right]
\end{aligned}$$

we get

$$\begin{aligned}
&J^{*}(s,x(s))-\frac{1}{2}x^{\mathsf{T}}(s)\,P(s)\,x(s)-p^{\mathsf{T}}(s)\,x(s)\\
&\quad=\frac{1}{2}\int_{t=s}^{T}\{u^{*\mathsf{T}}(t)\,R(t)\,u^{*}(t)\\
&\qquad+x^{\mathsf{T}}(t)\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]R^{-1}(t)\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]^{\mathsf{T}}x(t)\\
&\qquad+2x^{\mathsf{T}}(t)\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]^{\mathsf{T}}u^{*}(t)\\
&\qquad+2x^{\mathsf{T}}(t)\left[P(t)\,B(t)+S^{\mathsf{T}}(t)\right]^{\mathsf{T}}R^{-1}(t)\,B^{\mathsf{T}}(t)\,p(t)
\end{aligned}$$

$$+ 2u^{*\mathsf{T}}(t) B^{\mathsf{T}}(t) p(t) + 2p^{\mathsf{T}}(t) d(t)\Big\} dt$$

$$= \frac{1}{2}\int_{t=s}^{T} \Big\{\big\|R^{-1/2}(t)\left[R(t) u^*(t) + [P(t) B(t) + S^{\mathsf{T}}(t)]^{\mathsf{T}} x^{\mathsf{T}}(t)\right]$$

$$+ B^{\mathsf{T}}(t) p(t)\|^2 - \big\|R^{-1/2}(t) B^{\mathsf{T}}(t) p(t)\big\|^2 + 2p^{\mathsf{T}}(t) d(t)\Big\} dt \tag{22.171}$$

which, taking $s = 0$, $x(s) = x_0$, and in view of

$$R(t) u^*(t) + [P(t) B(t) + S^{\mathsf{T}}(t)]^{\mathsf{T}} x^{\mathsf{T}}(t) = 0$$

yields (22.166). Theorem is proven. □

Theorem 22.23. (on the uniqueness of the optimal control) *The optimal control* $u^*(\cdot) \in \mathcal{U}_{\text{admis}}[0, T]$ *is* ***unique*** *if and only if the corresponding Riccati differential equation (22.162) has a unique solution* $P(t) \geq 0$ *on* $[0, T]$.

Proof.

1. *Necessity*. Assume that $u^*(\cdot) \in \mathcal{U}_{\text{admis}}[0, T]$ is unique and is given by (22.165). But this is possible only if $P(t)$ is uniquely defined ($p(t)$ will be uniquely defined automatically). So, the corresponding Riccati differential equation (22.162) should have a unique solution $P(t) \geq 0$ on $[0, T]$.
2. *Sufficiency*. If the corresponding Riccati differential equation (22.162) has a unique solution $P(t) \geq 0$ on $[0, T]$, then, by the previous theorem, $u^*(\cdot)$ is uniquely defined by (22.165) and the dynamics $x^*(t)$ is given by

$$\begin{aligned}\dot{x}^*(t) = &\left[A(t) - B(t) R^{-1}(t) (B^{\mathsf{T}}(t) P(t) + S(t))\right] x^*(t)\\ &- B(t) R^{-1}(t) B^{\mathsf{T}}(t) p(t) + d(t)\end{aligned} \tag{22.172}$$

So, the uniqueness of (22.165) follows from the uniqueness of the solution of ODE (22.172). □

22.8.7 Stationary systems on the infinite horizon

22.8.7.1 Stationary systems and the infinite horizon cost function

Let us consider a stationary linear plant given by the following ODE

$$\left.\begin{aligned}&\dot{x}(t) = Ax(t) + Bu(t), \quad t \in [0, \infty]\\ &x(0) = x_0, \ A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times r}\end{aligned}\right\} \tag{22.173}$$

supplied by the quadratic cost function in the Lagrange form, namely,

$$J(u(\cdot)) = \int_{t=0}^{\infty} [x^{\mathsf{T}}(t) Qx(t) + u^{\mathsf{T}}(t) Ru(t)] dt \tag{22.174}$$

where $0 \leq Q = Q^{\mathsf{T}} \in \mathbb{R}^{n\times n}$, and $0 < R = R^{\mathsf{T}} \in \mathbb{R}^{r\times r}$ are the weighting matrices.

The problem is as before: *find a control* $u^*(\cdot)$ *minimizing* $J(u(\cdot))$ *over all controls within the class of admissible control strategies consisting of all* $u(\cdot)$ *such that*

- *the solution of (22.173) exists;*
- $u(\cdot) \in L^2(0, \infty; \mathbb{R}^r)$ *(otherwise, the criterion (22.174) does not exist).*

We will try to solve this problem by two methods: the, so-called, *direct method* and DPM.

22.8.7.2 Direct method

Let us introduce the function $V : \mathbb{R}^n \longmapsto \mathbb{R}$ as follows

$$V(x) := x^{\mathsf{T}} P x \tag{22.175}$$

where the matrix P is a symmetric matrix $P = P^{\mathsf{T}} \in \mathbb{R}^{n\times n}$. Then, in view of (22.173), we obtain

$$\dot{V}(x(t)) = 2x^{\mathsf{T}}(t) P\dot{x}(t) = 2x^{\mathsf{T}}(t) P[Ax(t) + Bu(t)]$$

The integration of this equation leads to the following:

$$V(x(T)) - V(x(0)) = x^{\mathsf{T}}(T) Px(T) - x_0^{\mathsf{T}} P x_0$$
$$= \int_{t=0}^{T} 2x^{\mathsf{T}}(t) P[Ax(t) + Bu(t)]\, dt$$

Adding and subtracting the terms $x^{\mathsf{T}}(t) Qx(t)$ and $u^{\mathsf{T}}(t) Ru(t)$, the last identity may be rewritten in the following form

$$x^{\mathsf{T}}(T) Px(T) - x_0^{\mathsf{T}} P x_0$$
$$= \int_{t=0}^{T} \left(2x^{\mathsf{T}}(t) P[Ax(t) + Bu(t)] + x^{\mathsf{T}}(t) Qx(t) + u^{\mathsf{T}}(t) Ru(t)\right) dt$$
$$- \int_{t=0}^{T} [x^{\mathsf{T}}(t) Qx(t) + u^{\mathsf{T}}(t) Ru(t)]\, dt = \int_{t=0}^{T} \left(x^{\mathsf{T}}(t)[PA + A^{\mathsf{T}}P + Q]x(t)\right.$$
$$\left. + 2\left(R^{-1/2}B^{\mathsf{T}}Px(t)\right)^{\mathsf{T}} R^{1/2}u(t) + \left\|R^{1/2}u(t)\right\|^2\right) dt$$
$$- \int_{t=0}^{T} [x^{\mathsf{T}}(t) Qx(t) + u^{\mathsf{T}}(t) Ru(t)]\, dt$$
$$= \int_{t=0}^{T} \left(x^{\mathsf{T}}(t)\left[PA + AP + Q - PBR^{-1}B^{\mathsf{T}}P\right]x(t)\right.$$
$$\left. + \left\|R^{-1/2}B^{\mathsf{T}}Px(t) + R^{1/2}u(t)\right\|^2\right) dt - \int_{t=0}^{T} [x^{\mathsf{T}}(t) Qx(t) + u^{\mathsf{T}}(t) Ru(t)]\, dt$$

which implies

$$\int_{t=0}^{T} \left[x^{\mathsf{T}}(t)\, Qx(t) + u^{\mathsf{T}}(t)\, Ru(t)\right] dt = x_0^{\mathsf{T}} P x_0 - x^{\mathsf{T}}(T)\, Px(T)$$
$$+ \int_{t=0}^{T} \left\| R^{-1/2} B^{\mathsf{T}} Px(t) + R^{1/2} u(t) \right\|^2 dt$$
$$+ \int_{t=0}^{T} x^{\mathsf{T}}(t) \left[PA + A^{\mathsf{T}} P + Q - PBR^{-1} B^{\mathsf{T}} P\right] x(t)\, dt \tag{22.176}$$

Selecting (if it is possible) the matrix P as a solution to the following *matrix Riccati equation*

$$PA + A^{\mathsf{T}} P + Q - PBR^{-1} B^{\mathsf{T}} P = 0 \tag{22.177}$$

from (22.176) we get

$$\int_{t=0}^{T} \left[x^{\mathsf{T}}(t)\, Qx(t) + u^{\mathsf{T}}(t)\, Ru(t)\right] dt = x_0^{\mathsf{T}} P x_0 - x^{\mathsf{T}}(T)\, Px(T)$$
$$+ \int_{t=0}^{T} \left\| R^{-1/2} B^{\mathsf{T}} Px(t) + R^{1/2} u(t) \right\|^2 dt$$
$$\geq x_0^{\mathsf{T}} P x_0 - x^{\mathsf{T}}(T)\, Px(T) \tag{22.178}$$

Theorem 22.24. *If for the system (22.173) the pair (A, B) is* ***stabilizable*** *and the pair $\left(Q^{1/2}, A\right)$ is* ***observable*** *then the optimal control $u^*(t)$ minimizing (22.174) is given by*

$$\boxed{u^*(t) = -R^{-1} B^{\mathsf{T}} Px(t)} \tag{22.179}$$

where P is the unique positive definite solution of (22.177) making the closed-loop system asymptotically stable. Moreover, the minimal value of the cost functional is

$$\boxed{J(u(\cdot)) = \int_{t=0}^{\infty} \left[x^{\mathsf{T}}(t)\, Qx(t) + u^{\mathsf{T}}(t)\, Ru(t)\right] dt = x_0^{\mathsf{T}} P x_0} \tag{22.180}$$

Proof. First, notice that the dynamic system (22.173) closed by an optimal control should be stable. Indeed, suppose that there exists at least one unstable mode of the controlled system. Then, by the observability of the pair (C, A), it follows that the vector $y(t) = Cx(t)$

($Q := C^{\top}C$) should tend to infinity and, hence, $\int_{t=0}^{T} x^{\top}(t)\,Qx(t)\,dt$ tends to infinity with $T \to \infty$ that cannot correspond to an optimal control. So, the term $x^{\top}(T)\,Px(T)$ has to tend to zero and can be disregarded. By Theorem 10.8 there exists the unique positive definite solution P of (22.177) which makes the closed-loop system stable since

$$\dot{x}(t) = Ax(t) + Bu^*(t) = \left(A - BR^{-1}B^{\top}P\right)x(t) = A_{\text{closed}}x(t)$$

where A_{closed} is defined in (10.26). Taking this P in (22.179) and in view of (22.178) under $u(\cdot) = u^*(\cdot)$ the equality (22.180) holds. Theorem is proven. □

Does the optimal control stabilize nonobservable systems?

Proposition 22.1. *The optimal control does not necessarily stabilize a nonobservable linear stationary system.*

The next simple example illustrates the statement given above.

Example 22.7. *Consider the following linear second-order controllable time invariant system*

$$\dot{x}_1(t) = x_2(t), \quad \dot{x}_2(t) = x_1(t) + u, \quad x_1(0) = x_{10}, \quad x_2(0) = x_{20}$$

Let the cost functional be $J(u(\cdot)) = \int_{t=0}^{\infty}\left[x^{\top}(t)\,Qx(t) + u^2(t)\right]dt$ *with* $x(t) := \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix}$ *and* $Q := \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \geq 0$. *In our case this system can be represented in the form (22.173) with* $A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ *and* $B = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. *Notice that this system is unobservable since the column rank of the observability matrix* $\mathcal{O} = \begin{bmatrix} C \\ CA \end{bmatrix} = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ *is incomplete and equal to* 1. *The statement of the proposition becomes evident if we define* $y(t) := x_1(t) - x_2(t)$ *and represent this system as*

$$\begin{aligned} &\dot{y}(t) = -y(t) - u \\ &J(u(\cdot)) = \int_{t=0}^{\infty}\left[y^2(t) + u^2(t)\right]dt \end{aligned} \tag{22.181}$$

with

$$\dot{x}_2(t) = y(t) + x_2(t) + u$$

According to Theorem 22.24 the optimal control in (22.181) is $u^*(t) = py(t)$ *where* $p = \sqrt{2} - 1$. *As it is expected, in the optimal system* $y(t) \to 0$ *as* $t \to \infty$. *Evidently, the second component* $x_2(t) \to \infty$, *and, hence, the optimal system is unstable which proves the above proposition.* ***This effect appears due to instability of the unobservable state component*** $x_2(t)$*!*

The revealed fact permits to indicate the class of stable unobservable optimal systems. If the pair (C, A) is unobservable then there exists a nonsingular linear transformation $Tx = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$ such that the system (22.173) can be represented in the, so-called, *canonical observability form*

$$\begin{aligned} &\dot{x}_1 = A_{11} + B_1 u, \ \dot{x}_2 = A_{21} x_1 + A_{22} x_2 + B_2 u \\ &y = C_1 x_1 \end{aligned} \tag{22.182}$$

with x_1 and x_2 being the observable and unobservable state components. Let us show how this transformation can be found. Select T in the form $T = \begin{pmatrix} v \\ w \end{pmatrix}$ where the matrix $v \in \mathbb{R}^{k \times n}$ consists of k basis row vectors of the observability matrix (9.63)

$$\mathcal{O} := \begin{bmatrix} C \\ CA \\ CA^2 \\ \vdots \\ CA^{n-1} \end{bmatrix}$$

Since the system is unobservable then $k < n$. The matrix $w \in \mathbb{R}^{(n-k) \times n}$ is an arbitrary one such that $\det T \neq 0$. Denote

$$T^{-1} := (N_1 \, N_2), \quad N_1 \in \mathbb{R}^{n \times k}$$

The identity $TT^{-1} = I$ implies

$$I = \begin{pmatrix} v \\ w \end{pmatrix} (N_1 \, N_2) = \begin{pmatrix} vN_1 & vN_2 \\ wN_1 & wN_2 \end{pmatrix}$$

and

$$vN_2 = 0 \tag{22.183}$$

Since vA and vC are in the same basis there exist matrices L_A and L_C such that

$$vA = L_A v, \quad C = L_C v$$

Then

$$\begin{aligned} &\begin{pmatrix} \dot{x}_1 \\ \dot{x}_2 \end{pmatrix} = TAT^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + TBu \\ &y = CT^{-1} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \end{aligned}$$

In view of (22.183) and the relations

$$TAT^{-1} = \begin{pmatrix} v \\ w \end{pmatrix} A \begin{pmatrix} N_1 & N_2 \end{pmatrix}$$

$$= \begin{pmatrix} L_A v \\ wA \end{pmatrix} \begin{pmatrix} N_1 & N_2 \end{pmatrix} = \begin{pmatrix} L_A v N_1 & 0 \\ wAN_1 & wAN_2 \end{pmatrix}$$

$$CT^{-1} = \begin{pmatrix} L_C v N_1 & L_C v N_2 \end{pmatrix} = \begin{pmatrix} CN_1 & 0 \end{pmatrix}$$

we finally get (22.182) and the cost functional (22.174) becomes

$$J(u(\cdot)) = \int_{t=0}^{\infty} [x_1^{\mathsf{T}}(t) Q_1 x_1(t) + u^{\mathsf{T}}(t) Ru(t)] dt$$

where $Q_1 = C_1^{\mathsf{T}} C_1$ and the pair (C_1, A_{11}) is observable. Then, by Theorem 22.24, the optimal control is

$$u^*(t) = -R^{-1} B_1^{\mathsf{T}} P_1 x_1(t)$$

with P_1 being the positive-definite solution to the reduced order Riccati equation

$$P_1 A_{11} + A_{11}^{\mathsf{T}} P_1 + Q - P_1 B_1 R^{-1} B_1^{\mathsf{T}} P_1 = 0$$

which makes the first subsystem (with respect to x_1) stable. It is evident that the optimal system (22.173) is stable if the matrix A_{22} is Hurwitz (stable). According to the PBH test 9.1, such systems are called *detectable*. Finally we may formulate the following claim.

Claim 22.2. *The linear time invariant system (22.173) optimal in the sense of the cost functional (22.174) is stable if and only if this system is* ***stabilizable*** *and* ***detectable****.*

22.8.7.3 DPM approach

Consider the following HBJ equation (22.137):

$$\left.\begin{aligned} &-\bar{h} + \sup_{u \in U} H\left(-\frac{\partial}{\partial x} V(x), x, u\right) = 0 \\ &x \in \mathbb{R}^n, \ \bar{h} = \text{const}, \ V(0) = 0 \end{aligned}\right\} \tag{22.184a}$$

with

$$\begin{aligned} &H(\psi, x, u) := \psi^{\mathsf{T}}(Ax + Bu) - x^{\mathsf{T}} Qx - u^{\mathsf{T}} Ru \\ &x, u, \psi \in [0, \infty] \times \mathbb{R}^n \times \mathbb{R}^r \times \mathbb{R}^n \end{aligned} \tag{22.185}$$

Theorem 22.25. (Verification rule for LQ-problem) *If the control u^* is a maximizing vector for (22.185) with some $\bar{h} = \text{const}$, that is,*

$$u^* = -\frac{1}{2} R^{-1} B^{\mathsf{T}} \frac{\partial}{\partial x} V(x)$$

where $V(x)$ is a solution to the following HJ equation:

$$-\frac{\partial}{\partial x}V^{\top}(x)\,Ax - x^{\top}Qx + \frac{1}{4}\frac{\partial}{\partial x}V^{\top}(x)\,BR^{-1}B^{\top}\frac{\partial}{\partial x}V(x) = \bar{h}$$

and the closed-loop system is stable, then such u^* is an optimal control.

Proof. It is evident that only admissible control may be stabilizing (if not, the cost function does not exist). By (22.185) for any stabilizing $u(\cdot)$ it follows that

$$H(-\frac{\partial}{\partial x}V(x^*), x^*, u^*) = \bar{h},\ H(-\frac{\partial}{\partial x}V(x), x, u) \le \bar{h}$$

and, hence,

$$H(-\frac{\partial}{\partial x}V(x), x, u) \le H(-\frac{\partial}{\partial x}V(x^*), x^*, u^*)$$

which, after integration, leads to the following inequality

$$\int_{t=0}^{\infty}\left[-\frac{\partial}{\partial x}V^{\top}(x)(Ax+Bu) - x^{\top}Qx - u^{\top}Ru\right]dt$$
$$\le \int_{t=0}^{\infty}\left[-\frac{\partial}{\partial x}V^{\top}(x^*)(Ax^*+Bu^*) - x^{*\top}Qx^* - u^{*\top}Ru^*\right]dt$$

or, equivalently,

$$\int_{t=0}^{T}[x^{*\top}Qx^* + u^{*\top}Ru^*]\,dt \le \int_{t=0}^{T}[x^{\top}Qx + u^{\top}Ru]\,dt + \int_{t=0}^{\infty}d(V(x) - V(x^*))$$
$$\overset{V(x(0))=V^{\top}(x^*(0))}{=} \int_{t=0}^{T}[x^{\top}Qx + u^{\top}Ru]\,dt + V(x(T)) - V^{\top}(x^*(T))$$

Within the class of stabilizing strategies we have

$$V(x(T)) - V^{\top}(x^*(T)) \underset{T\to\infty}{\to} 0$$

which, in view of the last inequality, shows that $u^*(\cdot)$ is an optimal control. □

Try to find the solution to (22.185) as $V(x) = x^{\top}Px$ with $P = P^{\top} \ge 0$. This implies $\frac{\partial}{\partial x}V(x) = 2Px$, and, hence,

$$-2x^{\top}P\left(Ax - BR^{-1}B^{\top}Px\right) - x^{\top}Qx - x^{\top}PBR^{-1}B^{\top}Px$$
$$= x^{\top}\left(-PA - A^{\top}P - Q + PBR^{-1}B^{\top}P\right)x = 0$$

The last equation is identically fulfilled for any $x \in \mathbb{R}^n$ if P is the solution to the same Riccati matrix equation as in (22.177) for a stabilizable and observable system. So, finally, the optimal control is $u^*(t) = -R^{-1}B^{\top}Px(t)$ which naturally coincides with (22.179).

22.9 Linear-time optimization

22.9.1 General result

For this problem the *cost functional* is

$$\boxed{J(u(\cdot)) = T} \tag{22.186}$$

It can be obtained from the Bolza form functional

$$J(u(\cdot)) := h_0(x(T)) + \int_{t=0}^{T} h(x(t), u(t), t)\, dt$$

if we put $h_0(x) \equiv 0$, $h(x, u, t) \equiv 1$. Then for a linear plant, given by (22.150), the Hamiltonian (22.81) is

$$H(\psi, x, u, t) := \psi^{\top}[A(t)x + B(t)u + d(t)] - \mu \tag{22.187}$$

and, hence, the maximality condition (22.69) becomes as follows:

$$\begin{aligned} u^*(t) &\in \operatorname*{Argmax}_{u \in U} \psi^{\top}(t)[A(t)x(t) + B(t)u(t) + d(t)] \\ &= \operatorname*{Argmax}_{u \in U} \psi^{\top}(t) B(t) u(t) = \operatorname*{Argmax}_{u \in U} [B^{\top}(t)\psi(t)]^{\top} u(t) \\ &= \operatorname*{Argmax}_{u \in U} \sum_{k=1}^{r} [B^{\top}(t)\psi(t)]_k u_k(t) \end{aligned} \tag{22.188}$$

Theorem 22.26. (on linear time-optimal control) *If the set U of the admissible control values is a polytope defined by*

$$U := \{u \in \mathbb{R}^r : u_k^- \le u_k(t) \le u_k^+,\ k = 1, \ldots, r\} \tag{22.189}$$

then the optimal control (22.188) is as follows

$$u_k^*(t) = \begin{cases} u_k^+ & \text{if} \quad [B^{\top}(t)\psi(t)]_k > 0 \\ u_k^- & \text{if} \quad [B^{\top}(t)\psi(t)]_k < 0 \\ \text{any}\ \bar{u} \in U & \text{if} \quad [B^{\top}(t)\psi(t)]_k = 0 \end{cases} \tag{22.190}$$

and it is unique.

Proof. Formula (22.190) follows directly from (22.188), (22.189) and the uniqueness is the consequence of the theorem on the sufficient condition of the optimality which demands the concavity (and not obligatory strict) of the Hamiltonian with respect to (x, u) for any fixed ψ, which is evidently fulfilled for the Hamiltonian function (22.187) which is linear on x and u. □

22.9.2 Theorem on n-intervals for stationary linear systems

Consider in detail the partial case of linear systems (22.150) when the matrices of the system are constant, that is, $A(t) = A$, $B(t) = B$. For this case the result below has been obtained in Feldbaum (1953) and is known as the *theorem on n-intervals*. But first, let us prove an axillary lemma.

Lemma 22.10. *If $\lambda_1, \lambda_2, \ldots, \lambda_m$ are real numbers and $f_1(t), \ldots, f_m(t)$ are the polynomials with real coefficients and having the orders $k_1, \ldots, k_m$, correspondingly. Then the function*

$$\varphi(t) = \sum_{i=1}^{m} f_i(t)\, e^{\lambda_i t} \tag{22.191}$$

has a number of real roots which does not exceed

$$n_0 := k_1 + \ldots + k_m + m - 1 \tag{22.192}$$

Proof. To prove this result let us use the induction method.

1. For $m = 1$ the lemma is true. Indeed, in this case the function $\varphi(t) = f_1(t)\, e^{\lambda_1 t}$ has the number of roots coinciding with k_1 since $e^{\lambda_1 t} > 0$ for any t.
2. Suppose that this lemma is valid for $m - 1 > 0$. Then let us prove that it holds for m. Multiplying (22.191) by $e^{-\lambda_m t}$ we obtain

$$\varphi(t)\, e^{-\lambda_m t} = \sum_{i=1}^{m-1} f_i(t)\, e^{(\lambda_i - \lambda_m)t} + f_m(t) \tag{22.193}$$

Differentiation by t the relation (22.193) $(k_m + 1)$-times implies

$$\frac{d^{(k_m+1)}}{dt^{(k_m+1)}}\left(\varphi(t)\, e^{-\lambda_m t}\right) = \sum_{i=1}^{m-1} \tilde{f}_i(t)\, e^{(\lambda_i - \lambda_m)t} := \varphi_{k_m+1}(t)$$

where $\tilde{f}_i(t)$ are the polynomials of the same order as $f_i(t)$. By the supposition before, the function $\varphi_1(t)$ has a number of roots which do not exceed

$$n_{k_m+1} := k_1 + \ldots + k_{m-1} + m - 2$$

Since between two roots of continuously differentiable function, there is at least one root of its derivative, then the function $\varphi_{k_m}(t) := \dfrac{d^{k_m}}{dt^{k_m}}\left(\varphi(t)\, e^{-\lambda_m t}\right)$ will have $n_{k_m} = n_{k_m+1} + 1$. Continuing this process, finally we get that the function $\varphi_0(t) := \varphi(t)\, e^{-\lambda_m t}$ will have

$$\begin{aligned} n_0 &= n_1 + 1 = n_2 + 2 = \ldots = n_{k_m+1} + (k_m + 1) \\ &= (k_1 + \ldots + k_{m-1} + m - 2) + (k_m + 1) = k_1 + \ldots + k_m + m - 1 \end{aligned}$$

And, since $e^{-\lambda_m t} > 0$ always, we may conclude that $\varphi(t)$ has the same number of roots as $\varphi_0(t)$. Lemma is proven. □

Now we are ready to prove the main result of this section.

Theorem 22.27. (Feldbaum 1953) *If the matrix $A \in \mathbb{R}^{n\times n}$ has only real eigenvalues, then the number of switches of any component of the optimal control (22.190) does not exceed $(n-1)$, that is, a number of the intervals, where each component of the optimal program (22.190) is constant, does not exceed n.*

Proof. Let $\lambda_1, \lambda_2, \ldots, \lambda_m$ be the different eigenvalues of the matrix A and $r_1, r_2, \ldots, r_m$ are their multiplicity numbers, correspondingly. Then a general solution of the adjoint system of equations $\dot{\psi}(t) = -A^{\mathrm{T}}\psi(t)$ may be represented as

$$\psi_i(t) = \sum_{j=1}^{m} p_{ij}(t)\, e^{-\lambda_j t},\ i = 1, \ldots, n \tag{22.194}$$

where $p_{ij}(t)$ are polynomials on t whose order does not exceed $(r_j - 1)$. Substituting (22.194) into (22.188) implies

$$\begin{aligned} u_k^*(t) &= \frac{u_k^+}{2}\left[1 + \operatorname{sign}\left(\sum_{i=1}^{n} b_{ik}\psi_i(t)\right)\right] + \frac{u_k^-}{2}\left[1 - \operatorname{sign}\left(\sum_{i=1}^{n} b_{ik}\psi_i(t)\right)\right] \\ &= \frac{u_k^+}{2}\left[1 + \operatorname{sign}\left(\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{m} p_{ij}(t)\, e^{-\lambda_j t}\right)\right] \\ &\quad + \frac{u_k^-}{2}\left[1 - \operatorname{sign}\left(\sum_{i=1}^{n} b_{ik}\sum_{j=1}^{m} p_{ij}(t)\, e^{-\lambda_j t}\right)\right] \\ &= \frac{u_k^+}{2}\left[1 + \operatorname{sign}\left(\sum_{j=1}^{m} \tilde{p}_{kj}(t)\, e^{-\lambda_j t}\right)\right] + \frac{u_k^-}{2}\left[1 - \operatorname{sign}\left(\sum_{j=1}^{m} \tilde{p}_{kj}(t)\, e^{-\lambda_j t}\right)\right] \end{aligned}$$

where $\tilde{p}_{kj}(t)$ are the polynomials on t, whose order does not exceed $(r_j - 1)$, equal to

$$\tilde{p}_{kj}(t) := \sum_{i=1}^{n} b_{ik} p_{ij}(t) \tag{22.195}$$

Now the number of switches is defined by the number of the roots of the polynomials (22.195). Applying directly the lemma above, we obtain that the polynomial function $\sum_{j=1}^{m} \tilde{p}_{kj}(t)\, e^{-\lambda_j t}$ has a number of real roots which do not exceed

$$(r_1 - 1) + (r_2 - 1) + \ldots + (r_k - 1) + (k - 1) = r_1 + \ldots + r_k - 1 = n - 1$$

Theorem is proven. □

23 $\mathbb{H}_2$ and $\mathbb{H}_\infty$ Optimization

Contents

In this chapter we will present the material following Francis (1987), Zhou *et al.* (1996), Curtain & Zwart (1995) and Poznyak (1991).

23.1 $\mathbb{H}_2$-optimization

23.1.1 Kalman canonical decompositions

The class of *finite dimensional linear time invariant dynamic systems* consists of systems described by the following ODE with constant coefficients:

$$\boxed{\begin{array}{c}\dot{x}(t) = Ax(t) + Bu(t), \quad x(t_0) = x_0 \\ \\ y(t) = Cx(t) + Du(t)\end{array}} \tag{23.1}$$

where $x(t) \in \mathbb{R}^n$ is associated with the *system state*, $u(t) \in \mathbb{R}^r$ is the *input*, and $y(t) \in \mathbb{R}^p$ is the *system output*. A, B, C and D are appropriately dimensioned real constant matrices. If $r = p = 1$, then a dynamic system (23.1) is called SISO (single input–single output), otherwise it is called MIMO (multiple input–multiple output). In compact form (23.1) can be rewritten as

$$\boxed{\begin{pmatrix}\dot{x} \\ y\end{pmatrix} = \begin{bmatrix}A & B \\ C & D\end{bmatrix}\begin{pmatrix}x \\ u\end{pmatrix}} \tag{23.2}$$

where $\begin{bmatrix}A & B \\ C & D\end{bmatrix}$ will be referred to as *a state space realization*. The corresponding transfer matrix $G(s)$ from u to y which connected their Laplace transformations $U(s)$ and $Y(s)$ (with zero-initial conditions) is defined by

$$\boxed{Y(s) = G(s)\,U(s)} \tag{23.3}$$

and is equal to

$$\boxed{G(s) = C(sI - A)^{-1} B + D} \tag{23.4}$$

Proposition 23.1. *The transfer matrix $G(s)$ is not changed under any nonsingular coordinate transformation $\tilde{x} = Tx$ $(\det T \neq 0)$ which converts (23.1) into*

$$\begin{aligned} \frac{d}{dt}\tilde{x}(t) &= TAT^{-1}\tilde{x}(t) + TBu(t) \\ y(t) &= CT^{-1}\tilde{x}(t) + Du(t) \end{aligned} \tag{23.5}$$

Proof. Evidently,

$$G(s) = C(sI - A)^{-1} B + D = CT^{-1}\left(sI - TAT^{-1}\right)^{-1} TB + D \qquad \square$$

Proposition 23.2. *The corresponding controllability $\tilde{\mathcal{C}}$ and observability $\tilde{\mathcal{O}}$ matrices are related to the original ones $\mathcal{C}$ (9.55) and $\mathcal{O}$ (9.63) by*

$$\boxed{\tilde{\mathcal{C}} = T\mathcal{C}, \quad \tilde{\mathcal{O}} = \mathcal{O}T} \tag{23.6}$$

This implies that the controllability and observability properties are invariant under the similarity (nonsingular) coordinate transformations.

Proof. It follows directly from the definitions (9.55) and (9.63). $\square$

The next theorems, known as the *Kalman decompositions*, show (see the details in Zhou *et al.* (1996)) that any linear system (23.1) can be transformed by a similarity transformation into a system having two groups of the coordinates such that one of them is obligatory controllable, or observable, or both properties hold simultaneously.

Theorem 23.1. (on the controllable canonical form) *If the controllability matrix has rank $k_c < n$, then there exists a similarity transformation*

$$\tilde{x} = \begin{pmatrix} \tilde{x}_c \\ \tilde{x}_{unc} \end{pmatrix} = Tx$$

such that

$$\frac{d}{dt}\begin{pmatrix} \tilde{x}_c \\ \tilde{x}_{unc} \end{pmatrix} = \begin{bmatrix} \tilde{A}_c & \tilde{A}_{12} \\ 0 & \tilde{A}_{unc} \end{bmatrix}\begin{pmatrix} \tilde{x}_c \\ \tilde{x}_{unc} \end{pmatrix} + \begin{bmatrix} \tilde{B}_c \\ 0 \end{bmatrix} u$$

$$y = \begin{bmatrix} \tilde{C}_c & \tilde{C}_{unc} \end{bmatrix}\begin{pmatrix} \tilde{x}_c \\ \tilde{x}_{unc} \end{pmatrix} + Du$$

where $\tilde{A}_c \in \mathbb{R}^{k_c \times k_c}$ and the pair $\left(\tilde{A}_c, \tilde{B}_c\right)$ is controllable (see Criteria 9.8). Moreover,

$$G(s) = C(sI - A)^{-1} B + D = \tilde{C}_c\left(sI - \tilde{A}_c\right)^{-1} \tilde{B}_c + D$$

Proof. Since the pair (A, B) is uncontrollable and the rank of the controllability matrix $\mathcal{C}$ is equal to $k_c < n$, there exist k_c linearly independent columns, say, $(v_1, \ldots, v_{k_c})$ of $\mathcal{C}$ such that $v_i := BA^{j_i}$ $(j_i \in [0, n-1])$. Adding any linearly independent (among themselves and with $(v_1, \ldots, v_{k_c})$) vectors $(v_{k_c+1}, \ldots, v_n)$ one can form the matrix

$$Q := \begin{bmatrix} v_1 \cdots v_{k_c} & v_{k_c+1} \cdots v_n \end{bmatrix} \tag{23.7}$$

which is nonsingular by the construction. Then the matrix

$$T = Q^{-1} \tag{23.8}$$

will give the desired decomposition. Indeed, since by the Cayley–Hamilton theorem 3.1 any vector v_i can be represented as a linear combination of the columns of $\mathcal{C}$, which implies

$$\begin{aligned} AT^{-1} &= \begin{bmatrix} Av_1 \cdots Av_{k_c} & Av_{k_c+1} \cdots Av_n \end{bmatrix} \\ &= \begin{bmatrix} v_1 \cdots v_{k_c} & v_{k_c+1} \cdots v_n \end{bmatrix} \begin{bmatrix} \tilde{A}_c & \tilde{A}_{12} \\ 0 & \tilde{A}_{unc} \end{bmatrix} = Q \begin{bmatrix} \tilde{A}_c & \tilde{A}_{12} \\ 0 & \tilde{A}_{unc} \end{bmatrix} \end{aligned}$$

By the same way, each column of the matrix B is a linear combination of vectors $(v_1, \ldots, v_{k_c})$, which also leads to the following relation

$$B = Q \begin{bmatrix} \tilde{B}_c \\ 0 \end{bmatrix} = T^{-1} \begin{bmatrix} \tilde{B}_c \\ 0 \end{bmatrix}$$

Notice also that $\mathcal{C}$ can be represented as

$$\mathcal{C} = T^{-1} \begin{bmatrix} \tilde{B}_c & \tilde{A}_c\tilde{B}_c & \cdots & \left(\tilde{A}_c\right)^{k_c-1}\tilde{B}_c & \cdots & \left(\tilde{A}_c\right)^{n-1}\tilde{B}_c \\ 0 & 0 & \cdots & 0 & \cdots & 0 \end{bmatrix}$$

Again, by the Cayley–Hamilton theorem 3.1 any matrix $\left(\tilde{A}_c\right)^i$ with $i > k_c$ can be represented as a linear combination of the matrices $\left(\tilde{A}_c\right)^j$ $(j = 1, \ldots, k_c)$, which is why

$$\operatorname{rank} \begin{bmatrix} \tilde{B}_c & \tilde{A}_c\tilde{B}_c & \cdots & \left(\tilde{A}_c\right)^{k_c-1}\tilde{B}_c \end{bmatrix} = k_c$$

So, the pair $\left(\tilde{A}_c, \tilde{B}_c\right)$ is controllable. Theorem is proven. □

Corollary 23.1. *According to Theorem 23.1 the state space* $\{\tilde{x}\}$ *may be partitioned in two orthogonal subspaces* $\left\{\begin{pmatrix} \tilde{x}_c \\ 0 \end{pmatrix}\right\}$ *and* $\begin{pmatrix} 0 \\ \tilde{x}_{unc} \end{pmatrix}$ *where the first subspace is controllable*

from the input and the second one is completely uncontrollable from the input. Moreover, since

$$x = T^{-1}\tilde{x} = \begin{bmatrix} v_1 & \cdots & v_{k_c} & v_{k_c+1} & \cdots & v_n \end{bmatrix} \begin{pmatrix} \tilde{x}_c \\ \tilde{x}_{unc} \end{pmatrix}$$

it follows that the controllable subspace is the span of the vectors v_i $(i = 1, \ldots, k_c)$, *or equivalently* $\operatorname{Im} \mathcal{C}$.

Theorem 23.2. (on the observable canonical form) *If the observability matrix has rank* $k_o < n$, *then there exists a similarity transformation*

$$\tilde{x} = \begin{pmatrix} \tilde{x}_o \\ \tilde{x}_{uno} \end{pmatrix} = Tx$$

such that

$$\frac{d}{dt}\begin{pmatrix} \tilde{x}_0 \\ \tilde{x}_{uno} \end{pmatrix} = \begin{bmatrix} \tilde{A}_o & 0 \\ \tilde{A}_{21} & \tilde{A}_{uno} \end{bmatrix} \begin{pmatrix} \tilde{x}_o \\ \tilde{x}_{uno} \end{pmatrix} + \begin{bmatrix} \tilde{B}_o \\ \tilde{B}_{uno} \end{bmatrix} u$$

$$y = \begin{bmatrix} \tilde{C}_o & 0 \end{bmatrix} \begin{pmatrix} \tilde{x}_o \\ \tilde{x}_{uno} \end{pmatrix} + Du$$

where $\tilde{A}_o \in \mathbb{R}^{k_o \times k_o}$ *and the pair* $\left(\tilde{C}_o, \tilde{A}_o\right)$ *is observable (see Criteria 9.10). Moreover,*

$$G(s) = C(sI - A)^{-1}B + D = \tilde{C}_o\left(sI - \tilde{A}_o\right)^{-1}\tilde{B}_o + D$$

Proof. By duality of the controllability and observability properties (see Criterion 6 in Theorem 9.10) the proof of this theorem can be converted to the proof of the previous one. □

Combining the two above theorems one can get the following joint result.

Theorem 23.3. (The Kalman canonical decomposition) *The state vector* x *of any finite dimensional linear time invariant dynamic system, given by (23.1), may be transformed by a nonsingular transformation* T $(\det T \neq 0)$ *into the new states*

$$\boxed{\tilde{x} = \begin{pmatrix} \tilde{x}_{c,o} \\ \tilde{x}_{c,uno} \\ \tilde{x}_{unc,o} \\ \tilde{x}_{unc,uno} \end{pmatrix} = Tx}$$

such that

$$\frac{d}{dt}\begin{pmatrix}\tilde{x}_{c,o}\\ \tilde{x}_{c,uno}\\ \tilde{x}_{unc,o}\\ \tilde{x}_{unc,uno}\end{pmatrix} = \begin{bmatrix}\tilde{A}_{c,o} & 0 & \tilde{A}_{13} & 0\\ \tilde{A}_{21} & \tilde{A}_{c,uno} & \tilde{A}_{23} & \tilde{A}_{24}\\ 0 & 0 & \tilde{A}_{unc,o} & 0\\ 0 & 0 & \tilde{A}_{43} & \tilde{A}_{unc,uno}\end{bmatrix}\begin{pmatrix}\tilde{x}_{c,o}\\ \tilde{x}_{c,uno}\\ \tilde{x}_{unc,o}\\ \tilde{x}_{unc,uno}\end{pmatrix} + \begin{bmatrix}\tilde{B}_{c,o}\\ \tilde{B}_{c,uno}\\ 0\\ 0\end{bmatrix} u$$

$$y = \begin{bmatrix}\tilde{C}_o & 0\end{bmatrix}\begin{bmatrix}\tilde{C}_{c,o} & 0 & \tilde{C}_{unc,o} & 0\end{bmatrix}\begin{pmatrix}\tilde{x}_{c,o}\\ \tilde{x}_{c,uno}\\ \tilde{x}_{unc,o}\\ \tilde{x}_{unc,uno}\end{pmatrix} + Du$$

where the vector $\tilde{x}_{c,o}$ *is controllable and observable,* $\tilde{x}_{c,uno}$ *is controllable but unobservable,* $\tilde{x}_{unc,o}$ *is uncontrollable but observable, and, finally,* $\tilde{x}_{unc,uno}$ *is both uncontrollable and unobservable. Moreover,*

$$G(s) = C(sI - A)^{-1}B + D = \tilde{C}_{c,o}\left(sI - \tilde{A}_{co}\right)^{-1}\tilde{B}_{co} + D$$

23.1.2 Minimal and balanced realizations

Criteria for the minimality of transfer matrix realizations

Definition 23.1. *A state space realization* $\begin{bmatrix}A & B\\ C & D\end{bmatrix}$ *of the transfer matrix function* $G(s)$ *is said to be a* ***minimal realization*** *of* $G(s)$ *if the matrix* A *has the smallest possible dimension. Sometimes, this minimal dimension of* A *is called the* ***McMillan degree*** *of* $G(s)$.

Lemma 23.1. (The criterion of minimality of a realization) *A state space realization* $\begin{bmatrix}A & B\\ C & D\end{bmatrix}$ *of the transfer matrix function* $G(s)$ *is* ***minimal*** *if and only if the pair* (A, B) *is controllable and the pair* (C, A) *is observable.*

Proof.

1. *Necessity.* First, show that if $\begin{bmatrix}A & B\\ C & D\end{bmatrix}$ is minimal then the pair (A, B) is controllable and the pair (C, A) is observable. On the contrary, supposing that (A, B) is uncontrollable and/or (C, A) is unobservable, by Theorem 23.3 there exists another realization with a smaller McMillan degree that contradicts the minimality of the considered realization. This fact proves necessity.

2. *Sufficiency.* Let now the pair (A, B) be controllable and the pair (C, A) be observable. Suppose that the given realization is not minimal and there exists another realization $\begin{bmatrix} \tilde{A} & \tilde{B} \\ \tilde{C} & D \end{bmatrix}$ which is minimal with order $n_{\min} < n$. Since by Theorem 23.3

$$G(s) = C(sI - A)^{-1} B + D = \tilde{C}\left(sI - \tilde{A}\right)^{-1} \tilde{B} + D$$

for any $i = 0, 1, \ldots$ one has $CA^i B = \tilde{C}\tilde{A}^i \tilde{B}$ which implies

$$\mathcal{OC} = \tilde{\mathcal{O}}\tilde{\mathcal{C}} \tag{23.9}$$

By the controllability and observability assumptions

$$\operatorname{rank}(\mathcal{O}) = \operatorname{rank}(\mathcal{O}) = n$$

and, hence, by the Sylvester inequality (2.24) we also have that $\operatorname{rank}(\mathcal{OC}) = n$. By the same reasons,

$$\operatorname{rank}\left(\tilde{\mathcal{O}}\right) = \operatorname{rank}\left(\tilde{\mathcal{C}}\right) = k = \operatorname{rank}\left(\tilde{\mathcal{O}}\tilde{\mathcal{C}}\right)$$

which contradicts the identity $\operatorname{rank}(\mathcal{OC}) = \operatorname{rank}\left(\tilde{\mathcal{O}}\tilde{\mathcal{C}}\right)$ resulting from (23.9). Sufficiency is proven. □

Corollary 23.2. *If* $\begin{bmatrix} A_i & B_i \\ C_i & D_i \end{bmatrix}$ $(i = 1, 2)$ *are two minimal realizations with the controllability* $\mathcal{C}_i$ *and observability* $\mathcal{O}_i$ *matrices respectively, then there exists the unique nonsingular coordinate transformation*

$$\boxed{\begin{array}{c} x^{(2)} = Tx^{(1)} \\ T = \left(\mathcal{O}_2^{\mathsf{T}}\mathcal{O}_2\right)^{-1}\mathcal{O}_2^{\mathsf{T}}\mathcal{O}_1 \quad or \quad T^{-1} = \mathcal{C}_1\mathcal{C}_2^{\mathsf{T}}\left(\mathcal{C}_2\mathcal{C}_2^{\mathsf{T}}\right)^{-1} \end{array}} \tag{23.10}$$

such that in the compact forms presentation (23.2) the corresponding matrices are related as

$$\boxed{A_2 = TA_1T^{-1}, \quad B_2 = TB_1, \quad C_2 = C_1T^{-1}} \tag{23.11}$$

Proof. It directly follows from (23.9) and (23.5). □

Balanced realization for a transfer matrix

In spite of the fact that there are infinitely many different state space realizations for a given transfer matrix, some particular realizations turn out to be very useful for control engineering practice. First, let us prove the following lemma on the relation of the structure of a state space realization with the solutions of the corresponding matrix Riccati equations.

Lemma 23.2. *Let* $\left[\begin{matrix} A & B \\ C & D \end{matrix}\right]$ *be a state space realization of a (not necessarily stable) transfer matrix* $G(s)$. *Suppose that there exists symmetric matrices*

$$\boxed{P = \begin{bmatrix} P_1 & 0 \\ 0 & 0 \end{bmatrix} \quad \text{and} \quad Q = \begin{bmatrix} Q_1 & 0 \\ 0 & 0 \end{bmatrix}} \tag{23.12}$$

with P_1, Q_1 *nonsingular, that is,* $P_1 > 0$ *and* $Q_1 > 0$, *such that*

$$\boxed{\begin{aligned} AP + PA^\mathsf{T} + BB^\mathsf{T} &= 0 \\ AQ + QA^\mathsf{T} + C^\mathsf{T}C &= 0 \end{aligned}} \tag{23.13}$$

(in fact, P *and* Q *are the controllability (9.54) and observability (9.62) grammians, respectively).*

1. *If the partition of the state space realization, compatible with* P, *is* $\begin{bmatrix} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ C_1 & C_2 & D \end{bmatrix}$, *then* $\begin{bmatrix} A_{11} & B_1 \\ C_1 & D \end{bmatrix}$ *is also the realization of* $G(s)$, *and, moreover, the pair* (A_{11}, B_1) *is controllable,* A_{11} *is stable and* $P_1 > 0$ *satisfies the following matrix Lyapunov equation*

$$\boxed{A_{11}P_1 + P_1A_{11}^\mathsf{T} + B_1B_1^\mathsf{T} = 0} \tag{23.14}$$

2. *If the partition of the state space realization, compatible with* Q, *is* $\begin{bmatrix} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ C_1 & C_2 & D \end{bmatrix}$, *then* $\begin{bmatrix} A_{11} & B_1 \\ C_1 & D \end{bmatrix}$ *is also the realization of* $G(s)$, *and, moreover, the pair* $(C_1, A_{11},)$ *is observable,* A_{11} *is stable and* $Q_1 > 0$ *satisfies the following matrix Lyapunov equation*

$$\boxed{A_{11}^\mathsf{T}Q_1 + Q_1A_{11} + C_1^\mathsf{T}C_1 = 0} \tag{23.15}$$

Proof.

1. Substituting (23.12) into (23.13) implies

$$0 = AP + PA^\mathsf{T} + BB^\mathsf{T} = \begin{bmatrix} A_{11}P_1 + P_1A_{11}^\mathsf{T} + B_1B_1^\mathsf{T} & P_1A_{21}^\mathsf{T} + B_1B_2^\mathsf{T} \\ A_{21}P_1 + B_2B_1^\mathsf{T} & B_2B_2^\mathsf{T} \end{bmatrix}$$

which, since P_1 is nonsingular, gives $B_2 = 0$ and $A_{21} = 0$. Hence,

$$\begin{bmatrix} A_{11} & A_{12} & B_1 \\ A_{21} & A_{22} & B_2 \\ C_1 & C_2 & D \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & B_1 \\ 0 & A_{22} & 0 \\ C_1 & C_2 & D \end{bmatrix}$$

and, by Lemma 2.2, one has

$$G(s) = C(sI - A)^{-1}B + D$$

$$= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} (sI - A_{11})^{-1} & -(sI - A)^{-1}A_{12}(sI - A_{22})^{-1} \\ 0 & (sI - A_{22})^{-1} \end{bmatrix} \begin{bmatrix} B_1 \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} C_1 & C_2 \end{bmatrix} \begin{bmatrix} (sI - A_{11})^{-1}B_1 \\ 0 \end{bmatrix} = C_1(sI - A_{11})^{-1}B_1$$

and, hence, $\begin{bmatrix} A_{11} & B_1 \\ C_1 & D \end{bmatrix}$ is also a realization. From Lemma 9.1, it follows that the pair (A_{11}, B_1) is controllable and A_{11} is stable if and only if $P_1 > 0$.

2. The second part of the theorem results from duality and can be proven following the analogous procedure. □

Definition 23.2. *A minimal* $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *state space realization of a transfer matrix* $G(s)$ *is said to be* ***balanced****, if two grammians* P *and* Q *are equal, that is,*

$$\boxed{P = Q} \tag{23.16}$$

Proposition 23.3. (The construction of a balanced realization) *Let* $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *be a minimal realization of* $G(s)$*. Then the following procedure leads to a balanced realization:*

1. *Using (23.13), compute the controllability* $P > 0$ *and the observability grammians* $Q > 0$.
2. *Using the Cholesky factorization (4.31), find matrix* R *such that*

$$P = R^{\top}R$$

3. *Diagonalize* $RQR^{\top}$ *getting*

$$RQR^{\top} = U\Sigma^2U^{\top}$$

4. *Let* $T = R^{\top}U\Sigma^{-1/2}$ *and obtain new* P_{bal} *and* Q_{bal} *as*

$$P_{bal} := TPT^{\top} = (T^{\top})^{-1}QT^{-1} := Q_{bal} = \Sigma \tag{23.17}$$

Proof. The validity of this construction follows from Theorem (7.4) if $A = P$ and $B = Q$. Taking into account that for minimal realization $A > 0$ and $B > 0$, we get (23.17). □

Corollary 23.3.

$$\boxed{P_{bal}Q_{bal} = \Sigma^2 = \operatorname{diag}(\sigma_1^2, \ldots, \sigma_n^2)} \tag{23.18}$$

where, with the decreasing order number, $\sigma_1 \geq \ldots \geq \sigma_n$ *are called the* ***Hankel singular values*** *of a time invariant linear system with transfer matrix* $G(s)$.

23.1.3 $\mathbb{H}_2$ *norm and its computing*

It was mentioned in sections 18.1.8 and 18.1.9 that the **Lebesgue space** $\mathbb{L}_2^{m\times k}$ (or simply $\mathbb{L}_2$) consists of all quadratically integrable complex $(m \times k)$ matrices, i.e.,

$$\boxed{\begin{array}{c} \mathbb{L}_2^{m\times k} := \left\{ F : \mathbb{C} \to \mathbb{C}^{m\times k} \mid \right. \\ \left. \|F\|^2_{\mathbb{L}_2^{m\times k}} := \dfrac{1}{2\pi} \displaystyle\int_{\omega=-\infty}^{\infty} \mathrm{tr}\{F(j\omega)\, F^{\sim}(j\omega)\}\, d\omega < \infty \right\} \\ (\text{with} \quad F^{\sim}(j\omega) := F^{\mathsf{T}}(-j\omega)) \end{array}} \tag{23.19}$$

$\mathbb{L}_2$ space is (see (18.18)) a *Hilbert space* with the scalar (inner) product defined by

$$\boxed{\langle X, Y\rangle_{\mathbb{L}_2} := \frac{1}{2\pi} \int_{\omega=-\infty}^{\infty} \mathrm{tr}\{X(j\omega)\, Y^{\sim}(j\omega)\}\, d\omega} \tag{23.20}$$

The **Hardy space** $\mathbb{H}_2^{m\times k}$ (or, simply, $\mathbb{H}_2$) is the subspace of $\mathbb{L}_p^{m\times k}$ consisting of all quadratically integrable complex $(m \times k)$ matrices with only regular (holomorphic) (see Definition 17.2) elements on the open half-plane $\mathrm{Re}\, s > 0$. Evidently, $\mathbb{H}_2$ is also a Hilbert space with the same scalar product (23.20).

Lemma 23.3.

$$\boxed{\mathbb{L}_2 = \mathbb{H}_2 \oplus \mathbb{H}_2^{\perp}} \tag{23.21}$$

such that if $X \in \mathbb{H}_2$ *and* $Y \in \mathbb{H}_2^{\perp}$ *then*

$$\boxed{\langle X, Y\rangle_{\mathbb{L}_2} = 0} \tag{23.22}$$

Proof. It is a direct consequence from Lemma 18.1 on the orthogonal complement of a subset of a Hilbert space. □

The next theorems state the relation between $L_2^{m\times k}[0, \infty)$ and $\mathbb{H}_2^{m\times k}$.

Theorem 23.4. *If* $f(t), g(t) \in L_2^{m\times k}[0,\infty)$ *and their* ***Laplace transformation*** *(17.73) are* $F(p), G(p) \in \mathbb{H}_2^{m\times k}$, *then the following identities hold:*

1.

$$\boxed{\begin{aligned}\langle f, g\rangle_{L_2} &:= \int_{t=0}^{\infty} \operatorname{tr}\{f(t)\, g^{\mathsf{T}}(t)\}\, dt \\ &= \langle F, G\rangle_{\mathbb{H}_2} := \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{F(j\omega)\, G^{\sim}(j\omega)\right\} d\omega\end{aligned}} \tag{23.23}$$

where

$$G^{\sim}(j\omega) := G^{\mathsf{T}}(-j\omega) \tag{23.24}$$

2.

$$\boxed{\begin{aligned}\|f\|_{L_2} &:= \left(\int_{t=0}^{\infty} \operatorname{tr}\{f(t)\, f^{\mathsf{T}}(t)\}\, dt\right)^{1/2} \\ &= \|F\|_{\mathbb{H}_2} := \left(\frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{F(j\omega)\, G^{\sim}(j\omega)\right\} d\omega\right)^{1/2}\end{aligned}} \tag{23.25}$$

The identities (23.23) and (23.25) will be referred to as the ***generalized Parseval's identities.***

Proof. It is a direct consequence of the Plancherel theorem 17.18 and its Corollary 17.14. □

Remark 23.1. *It is obvious from the manipulations above that* L_2*-norm* $\|f\|_{L_2}$ *is finite if and only if the corresponding transfer matrix* $F(p)$ *is strictly proper, i.e.,* $F(\infty) = 0$.

Sure, $\|F\|_{\mathbb{H}_2}$ can be computed, in principle, directly from its definition (23.25). But there exist two other possibilities to realize this computation.

1. **The first computational method**
   By the residue Theorem 17.5, $\|F\|_{\mathbb{H}_2}^2$ is equal to the sum of the residues of $\operatorname{tr}\left\{F(j\omega)\, G^{\sim}(j\omega)\right\}$ at its poles a_k $(k = 1, \ldots, n)$ in the left half-plane of the complex plane C, i.e.,

$$\boxed{\begin{aligned}\|F\|_{\mathbb{H}_2}^2 &= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{F(j\omega)\,\tilde{G}(j\omega)\right\} d\omega \\ &= i\sum_{k=1}^{n} \operatorname{res}\left(\operatorname{tr}\left\{F\tilde{G}\right\}\right)(a_k)\end{aligned}} \tag{23.26}$$

2. **The second computational method**
   It turns out to be useful in many applications to have an alternative characterization of $\|F\|_{\mathbb{H}_2}^2$ using the advantages of the state space representation.

Theorem 23.5. *Let a transfer matrix $G(s)$ have a state space realization* $\left[\begin{smallmatrix} A & B \\ C & 0 \end{smallmatrix}\right]$ *with the matrix A stable (Hurwitz). Then $\|G\|_{\mathbb{H}_2}^2$ can be computed as follows:*

$$\boxed{\|G\|_{\mathbb{H}_2}^2 = \operatorname{tr}\{B^\top Q B\} = \operatorname{tr}\{CPC^\top\}} \tag{23.27}$$

where P is the controllability (9.54) and Q is the observability (9.62) grammians, respectively, which can be obtained from the following matrix Lyapunov equations

$$\boxed{\begin{aligned} AP + PA^\top + BB^\top &= 0 \\ A^\top Q + QA + C^\top C &= 0\end{aligned}} \tag{23.28}$$

Proof. Since A is stable it follows that

$$g(t) = \mathcal{L}^{-1}(G) = \begin{cases} Ce^{At}B & \text{if} \quad t \geq 0 \\ 0 & \text{if} \quad t < 0 \end{cases}$$

and, by Parseval's identity (23.25) and the Lyapunov Lemma 9.1, we have

$$\begin{aligned}\|G\|_{\mathbb{H}_2}^2 = \|g\|_{L_2}^2 &= \int_{t=0}^{\infty} \operatorname{tr}\{g(t)\,g^\top(t)\}\,dt \\ &= \int_{t=0}^{\infty} \operatorname{tr}\left\{B^\top e^{A^\top t} C^\top C e^{At} B\right\} dt \\ &= \operatorname{tr}\left\{B^\top \left(\int_{t=0}^{\infty} e^{A^\top t} C^\top C e^{At} dt\right) B\right\} = \operatorname{tr}\{B^\top P B\} \\ &= \int_{t=0}^{\infty} \operatorname{tr}\left\{Ce^{At} BB^\top e^{A^\top t} C^\top\right\} dt \\ &= \operatorname{tr}\left\{C\left(\int_{t=0}^{\infty} e^{At} BB^\top e^{A^\top t} dt\right) C^\top\right\} = \operatorname{tr}\{CQC^\top\}\end{aligned}$$

which proves the theorem. □

Example 23.1. *If the state space realization of* $G(p)$ *is* $\left[\begin{array}{cc} -1 & 1 \\ 2 & 0 \end{array}\right]$ *then by (23.28)* $P = 1/2$ *and* $Q = 2$. *So, by (23.27),* $\|G\|^2_{\mathbb{H}_2} = 2$.

Remark 23.2. *To compute the norm* $\|G\|^2_{\mathbb{RL}_2}$ *it is possible to use the following procedure:*

1. ***Separation**: represent* $G(p)$ *as*

$$\boxed{G(p) = G_+(p) + G_-(p)} \tag{23.29}$$

where $G_+(p) \in \mathbb{RH}_2$, *i.e., it contains only stable elements, and* $G_-(p) \in \mathbb{RH}_2^\perp$.

2. ***Representation**:*

$$\boxed{\|G\|^2_{\mathbb{RL}_2} = \|G_+\|^2_{\mathbb{H}_2} + \|G_-\|^2_{\mathbb{H}_2}} \tag{23.30}$$

3. ***Calculation**: using state space representations of* $G_+(s)$ *and* $G_-(-s)$ *(which correspond to a stable system) calculate* $\|G_+\|^2_{\mathbb{H}_2}$, $\|G_-(s)\|^2_{\mathbb{H}_2} = \|G_-(-s)\|^2_{\mathbb{H}_2}$ *and, finally,* $\|G\|^2_{\mathbb{RL}_2}$ *applying (23.30).*

23.1.4 $\mathbb{H}_2$ optimal control problem and its solution

Consider a linear dynamic system given by

$$\boxed{\begin{array}{c} \dot{x}(t) = Ax(t) + B\tilde{u}(t), \quad x(0) = x_0 \\ A \in \mathbb{R}^{n\times n}, \quad B \in \mathbb{R}^{n\times r} \end{array}} \tag{23.31}$$

Problem 23.1. *The problem, called* ***LQR*** *(linear quadratic regulation), consists of finding a feedback control* $\tilde{u}(t) = Kx(t) \in L_2^r[0,\infty)$ *which*

1. *provides the property* $x(t) \in L_2^n[0,\infty)$;
2. *minimizes the quadratic performance index*

$$\boxed{J(u(\cdot)) := \int_{t=0}^{\infty} \begin{pmatrix} x(t) \\ \tilde{u}(t) \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} Q & \tilde{S} \\ \tilde{S}^{\mathsf{T}} & R \end{pmatrix} \begin{pmatrix} x(t) \\ \tilde{u}(t) \end{pmatrix} dt} \tag{23.32}$$

where it is supposed that

$$\boxed{\begin{pmatrix} Q & \tilde{S} \\ \tilde{S}^{\mathsf{T}} & R \end{pmatrix} \geq 0, \quad Q = Q^{\mathsf{T}} \geq 0, \quad R = R^{\mathsf{T}} > 0} \tag{23.33}$$

Denote

$$\boxed{u(t) := R^{1/2}\tilde{u}(t)} \tag{23.34}$$

permits to represent (23.32) as follows

$$\boxed{J(u(\cdot)) = \int_{t=0}^{\infty} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} Q & S \\ S^{\mathsf{T}} & I \end{pmatrix} \begin{pmatrix} x(t) \\ u(t) \end{pmatrix} dt, \quad S = \tilde{S}R^{-1/2}}$$

and the factorization

$$\begin{pmatrix} Q & S \\ S^{\mathsf{T}} & I \end{pmatrix} = \begin{pmatrix} C \\ D \end{pmatrix} (C\ D)$$

$$Q = C^{\mathsf{T}}C, \quad I = D^{\mathsf{T}}D, \quad S = C^{\mathsf{T}}D$$

leads to

$$\begin{pmatrix} x \\ u \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} Q & S \\ S^{\mathsf{T}} & I \end{pmatrix} \begin{pmatrix} x \\ u \end{pmatrix} = \|Cx + Du\|^2$$

Defining

$$\boxed{z(t) := Cx(t) + Du(t)} \tag{23.35}$$

we can reformulate Problem 23.1 as the following L_2 problem:

Problem 23.2. (LQR-L_2 optimization)

$$\boxed{\begin{array}{c} \|z\|_{L_2}^2 \to \min\limits_{u \in L_2^n[0,\infty)} \\ \\ \dot{x} = Ax + Bu \\ z = Cx + Du \\ u = Kx \end{array}} \tag{23.36}$$

Under a fixed feedback $u = Kx$ the given linear controlled system can be represented as an uncontrolled system with a singular input:

$$\boxed{\begin{array}{c} \dot{x} = A_K x + x_0 \delta(t), \quad x(0) = 0 \\ z = C_K x \\ A_K := A + BK, \quad C_K := C + DK \end{array}} \tag{23.37}$$

The associated transfer matrix $G_K(s)$ from the singular input $x_0\delta(t)$ to the "output" z is

$$\boxed{G_K(s) = C_K(Is - A_K)^{-1}} \tag{23.38}$$

with the state space realization $\left[\begin{array}{cc} A_K & I \\ C_K & 0 \end{array}\right]$.

Theorem 23.6. (Zhou *et al.* 1996) *If the pair* (A, B) *is controllable and* (C, A) *is detectable, then the solution of the LQR problem is given by*

$$\boxed{K = K^* := -(B^\intercal X + D^\intercal C)} \tag{23.39}$$

and the corresponding optimal performance index $J(u(\cdot))$ *(23.32) is*

$$\boxed{J(u(\cdot)) = \|z\|_{L_2}^2 = \|G_{K^*}(s)\, x_0\|_{\mathbb{H}_2}^2 = x_0^\intercal X x_0} \tag{23.40}$$

where X *is the stabilizing solution of the following matrix Riccati equation*

$$\boxed{\begin{array}{l}(A - BD^\intercal C)\, X + X (A - BD^\intercal C)^\intercal \\ \quad -XBB^\intercal X + (D^\intercal C)^\intercal (D^\intercal C) = 0\end{array}} \tag{23.41}$$

(In fact, X *is the observability grammian of* (C_{K^*}, A_{K^*}) *satisfying the matrix Lyapunov equation*

$$A_{K^*}^\intercal X + X A_{K^*} - C_{K^*}^\intercal C_{K^*} = 0 \tag{23.42}$$

coinciding with (23.41)).

Proof. First, notice that under the conditions of this theorem and by Theorem 10.7 the Riccati equation (23.41) has the unique stabilizing nonnegative definite solution X. If $K = K^*$ is fixed, then the relation (23.40) results from the Plancherel theorem 17.18 and the formula (23.27) if $B = I$. To prove the inequality $\|G_K(s)\, x_0\|_{\mathbb{L}_2}^2 \geq \|G_{K^*}(s)\, x_0\|_{\mathbb{L}_2}^2$ for any stabilizing feedback $u = Kx$, let us consider in (23.36)

$$u(t) = K^* x(t) + v(t)$$

which gives

$$\begin{aligned}\dot{x} &= A_{K^*} x + Bv, \quad x(0) = x_0 \\ z &= C_{K^*} x + Dv\end{aligned}$$

or, equivalently,

$$\begin{aligned}\dot{x} &= A_K x + x_0 \delta(t) + Bv, \quad x(0) = 0 \\ z &= C_{K^*} x + Dv\end{aligned}$$

Applying the Laplace transformation to this relation, in the frequency domain we have

$$\begin{aligned}Z(s) &= C_{K^*} (Is - A_{K^*})^{-1} [x_0 + BV(s)] + DV(s) \\ &= G_{K^*}(s)\, x_0 + U(s)\, V(s)\end{aligned} \tag{23.43}$$

where

$$V(s) := \mathcal{L}\{v\}, \quad U(s) := [G_{K^*}(s)\, B + D] \in \mathbb{RH}_\infty$$

(since A_{K^*} is a stable matrix). Using (23.42) it is not difficult to check that

$$U^{\sim}(s)\,U(s) = I, \quad U^{\sim}(s)\,G_{K^{\sim}}(s) \in \mathbb{RH}_\infty^{\perp}$$

Indeed, the space realizations of $U(s)$ and $U^{\sim}(s)$ are

$$U(s) \Longrightarrow \begin{bmatrix} A_{K^{\sim}} & B \\ C_{K^*} & D \end{bmatrix}, \quad U^{\sim}(s) \Longrightarrow \begin{bmatrix} -A_{K^*}^{\mathsf{T}} & -C_{K^*}^{\mathsf{T}} \\ B^{\mathsf{T}} & D^{\mathsf{T}} \end{bmatrix}$$

So,

$$U^{\sim}(s)\,U(s) \Longrightarrow \begin{bmatrix} -A_{K^*}^{\mathsf{T}} & -C_{K^*}^{\mathsf{T}}C_{K^*} & -C_{K^*}^{\mathsf{T}}D \\ 0 & A_{K^*} & B \\ B^{\mathsf{T}} & D^{\mathsf{T}}C_{K^*} & I \end{bmatrix}$$

Define the matrix

$$T := \begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}$$

which is nonsingular for any $X \geq 0$. Then, the application of this similarity state transformation (which does not change $U(s)$) to the state vector leads to the following state space realization:

$$\begin{bmatrix} -A_{K^*}^{\mathsf{T}} & -C_{K^*}^{\mathsf{T}}C_{K^*} & -C_{K^*}^{\mathsf{T}}D \\ 0 & A_{K^*} & B \\ B^{\mathsf{T}} & D^{\mathsf{T}}C_{K^*} & I \end{bmatrix} \overset{T}{\Longrightarrow} \begin{bmatrix} -A_{K^*}^{\mathsf{T}} & 0 & 0 \\ 0 & A_{K^*} & B \\ B^{\mathsf{T}} & 0 & I \end{bmatrix}$$

which gives $U^{\sim}(s)\,U(s) = I$. Also

$$U^{\sim}(s)\,G_{K^*}(s) \Longrightarrow \begin{bmatrix} -A_{K^*}^{\mathsf{T}} & 0 & -X \\ 0 & A_{K^*} & I \\ B^{\mathsf{T}} & 0 & 0 \end{bmatrix} \overset{T}{\Longrightarrow} \begin{bmatrix} -A_{K^*}^{\mathsf{T}} & -X \\ B^{\mathsf{T}} & 0 \end{bmatrix}$$

which is equivalent to $U^{\sim}(s)\,G_{K^*}(s) \in \mathbb{RH}_\infty^{\perp}$. Taking these properties into account and in view of (23.43) we obtain

$$Z(s) = G_{K^*}(s)\,x_0 + U(s)\,V(s)$$

and

$$\begin{aligned} \|Z(s)\|_{\mathbb{L}_2}^2 &= \|G_{K^*}(s)\,x_0 + U(s)\,V(s)\|_{\mathbb{L}_2}^2 \\ &= \|G_{K^*}(s)\,x_0\|_{\mathbb{L}_2}^2 + \|U(s)\,V(s)\|_{\mathbb{L}_2}^2 + 2\,\langle G_{K^*}(s)\,x_0,\, U(s)\,V(s)\rangle_{\mathbb{L}_2} \\ &= \|G_{K^*}(s)\,x_0\|_{\mathbb{L}_2}^2 + \|U(s)\,V(s)\|_{\mathbb{L}_2}^2 + 2\,\langle U^*(s)\,G_{K^*}(s)\,x_0,\, V(s)\rangle_{\mathbb{L}_2} \\ &= \|G_{K^*}(s)\,x_0\|_{\mathbb{L}_2}^2 + \|U(s)\,V(s)\|_{\mathbb{L}_2}^2 \geq \|G_{K^*}(s)\,x_0\|_{\mathbb{L}_2}^2 \end{aligned}$$

where the equality is attained when $U(s)\,V(s) = 0$ for any $V(s)$ which is possible if and only if $U(s) \equiv 0$. By the stability of $G_{K^*}(s)$ we get that $\|G_{K^*}(s)\,x_0\|^2_{\mathbb{L}_2} = \|G_{K^*}(s)\,x_0\|^2_{\mathbb{H}_2}$. Theorem is proven. □

23.2 $\mathbb{H}_\infty$-optimization

23.2.1 $\mathbb{L}_\infty$, $\mathbb{H}_\infty$ *norms*

As it has already been mentioned in section 18.1,

1. the *Lebesgue space* $\mathbb{L}_\infty^{m\times k}$ is the space of all complex matrices bounded (almost everywhere) on the imaginary axis elements, i.e.,

$$\boxed{\begin{array}{l}\mathbb{L}_\infty^{m\times k} := \left\{F : \mathbb{C} \to \mathbb{C}^{m\times k} \mid \right. \\ \left. \|F\|_{\mathbb{L}_\infty^{m\times k}} := \underset{\omega\in(-\infty,\infty)}{ess\,\sup}\ \bar{\sigma}\,(F\,(i\omega)) < \infty\right\}\end{array}} \tag{23.44}$$

   where

$$\boxed{\begin{array}{l}\bar{\sigma}\,(F\,(i\omega)) := \lambda_{\max}^{1/2}\,\{F\,(i\omega)\,F^*\,(i\omega)\} \\ \qquad\qquad = \lambda_{\max}^{1/2}\,\{F^*\,(i\omega)\,F\,(i\omega)\}\end{array}} \tag{23.45}$$

   is the *largest singular value* of the matrix $F(i\omega)$. The space $\mathbb{L}_\infty^{m\times k}$ with the norm $\|F\|_{\mathbb{L}_\infty^{m\times k}}$ (23.44) is a *Banach* space;
2. the *rational subspace* of $\mathbb{L}_\infty^{m\times k}$, denoted by $\mathbb{RL}_\infty^{m\times k}$, consists of all proper and (with real coefficients) **rational transfer matrices**, defined on $\mathbb{C}$, with no poles on the imaginary axis;
3. the *Hardy spaces* $\mathbb{H}_\infty^{m\times k}$ and $\mathbb{RH}_\infty^{m\times k}$ are closed subspaces of the corresponding Lebesgue spaces $\mathbb{L}_\infty^{m\times k}$ and $\mathbb{RL}_\infty^{m\times k}$ containing complex matrices with only regular (holomorphic) (see Definition 17.2) elements on the open half-plane $\mathrm{Re}\,s > 0$. The $\mathbb{H}_\infty^{m\times k}$ norm is defined as

$$\boxed{\begin{array}{l}\mathbb{H}_\infty^{m\times k} := \left\{F : \mathbb{C} \to \mathbb{C}^{m\times k} \mid \right. \\ \left. \|F\|_{\mathbb{H}_\infty^{m\times k}} := \underset{\omega\in(-\infty,\infty)}{ess\,\sup}\ \bar{\sigma}\,(F\,(i\omega)) < \infty\right\}\end{array}} \tag{23.46}$$

4. the *rational subspace* of $\mathbb{H}_\infty^{m\times k}$, denoted by $\mathbb{RH}_\infty^{m\times k}$, consists of all proper and *rational* ***stable*** *transfer matrices with real coefficients.*

An engineering interpretation of the H_∞ norm $\|F\|_{\mathbb{H}_\infty^{m\times k}}$ (23.46) of a scalar transfer function is the distance in the complex plane $\mathbb{C}$ from the origin to the farthest point on the Nyquist plot ($x := \mathrm{Re}F\,(i\omega)$, $y := \mathrm{Im}F\,(i\omega)$) of F, and it also appears as the peak value in the Bode magnitude plot of $|F(i\omega|$.

Example 23.2. *For*

$$F\,(s) = \frac{1-s}{(1+s)\,(1+2s)}$$

(in fact, $F(s) \in \mathbb{RH}_2 \subset \mathbb{RH}_\infty$*) we have*

$$F(i\omega) = \mathrm{Re}F(i\omega) + i\mathrm{Im}F(i\omega)$$

$$\mathrm{Re}F(i\omega) = \frac{1+3\omega^2}{(1+\omega^2)(1+4\omega^2)}$$

$$\mathrm{Im}F(i\omega) = \frac{-2\omega(2-\omega^2)}{(1+\omega^2)(1+4\omega^2)}$$

$$|F(i\omega)| = \frac{\sqrt{(1+3\omega^2)^2 + 4\omega^2(2-\omega^2)^2}}{(1+\omega^2)(1+4\omega^2)}$$

The Nyquist plot is given in Fig. 23.1, and the Bode magnitude plot is depicted in Fig. 23.2.

Example 23.3. *For*

$$F(s) = \frac{s-1}{s+1} \in \mathbb{RH}_\infty$$

it follows that

$$F(i\omega)\,\tilde{F}(i\omega) = \frac{i\omega - 1}{i\omega + 1}\frac{-i\omega - 1}{-i\omega + 1} = 1 = \bar{\sigma}^2(F(i\omega))$$

$$\|F\|_{\mathbb{H}_\infty} = \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \bar{\sigma}(F(i\omega)) = 1$$

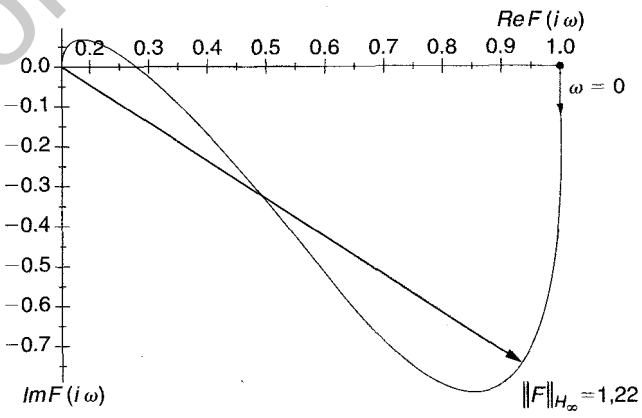

Fig. 23.1. The Nyquist plot of $F(i\omega)$.

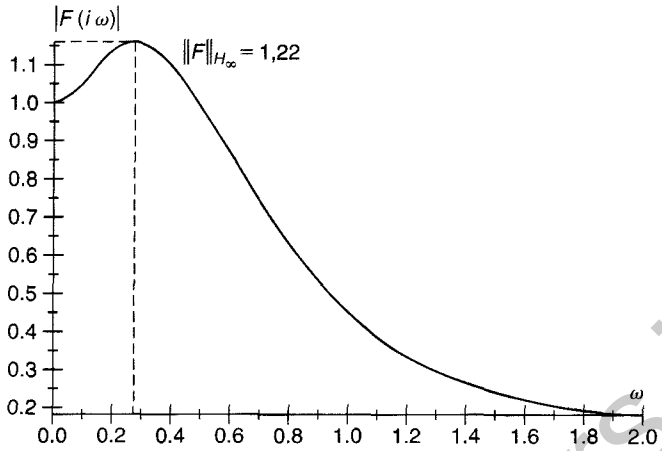

Fig. 23.2. The Bode magnitude plot $|F(i\omega)|$.

Example 23.4. *For*

$$F(s) = \frac{1}{s+1}\begin{pmatrix} s & -s \\ 0 & 1 \end{pmatrix} \in \mathbb{RH}_\infty$$

we have

$$F(i\omega)\,F^{\sim}(i\omega) = \frac{1}{1+\omega^2}\begin{pmatrix} i\omega & -i\omega \\ 0 & 1 \end{pmatrix}\begin{pmatrix} -i\omega & 0 \\ i\omega & 1 \end{pmatrix}$$

$$= \frac{1}{1+\omega^2}\begin{pmatrix} 2\omega^2 & -i\omega \\ i\omega & 1 \end{pmatrix}$$

and

$$\bar{\sigma}^2(F(i\omega)) = \frac{1+2\omega^2+\sqrt{1+4\omega^4}}{2[1+\omega^2]} = \frac{1/2+\omega^2+\sqrt{1/4+\omega^4}}{1+\omega^2}$$

So,

$$\|F\|_{\mathbb{H}_\infty}^2 = \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \bar{\sigma}^2(F(i\omega))$$

$$= 1 + \max_{\omega\in(-\infty,\infty)} \frac{\sqrt{1/4+\omega^4} - 1/2}{1+\omega^2} = 2$$

The following inequality turns out to be important in the considerations below.

Lemma 23.4. *[on a relation between* $\mathbb{L}_2$ *and* $\mathbb{L}_\infty$ *norms] For any* $g(s) \in \mathbb{L}_2^k$ *and any* $G(s) \in \mathbb{L}_\infty^{m\times k}$

$$\boxed{\|G(s)\,g(s)\|_{\mathbb{L}_2^m} \le \|G(s)\|_{\mathbb{L}_\infty^{m\times k}}\,\|g(s)\|_{\mathbb{L}_2^k}} \tag{23.47}$$

Proof. By the definition (23.19) it follows that

$$\begin{aligned}
\|G(s)\,g(s)\|^2_{\mathbb{L}_2^m} &:= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\left[g^{\sim}(i\omega)G^{\sim}(i\omega)\,G(i\omega)\,g(i\omega)\right]d\omega \\
&\le \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\left[g^{\sim}(i\omega)\lambda\left(\max G^{\sim}(i\omega)\,G(i\omega)\right)g(i\omega)\right]d\omega \\
&\le \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\operatorname{tr}\left\{G^{\sim}(i\omega)\,G(i\omega)\right\}g^{\sim}(i\omega)\,g(i\omega)d\omega \\
&\le \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\left[\operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)}\operatorname{tr}\left\{G^{\sim}(i\omega)\,G(i\omega)\right\}\right]g^{\sim}(i\omega)\,g(i\omega)d\omega \\
&\le \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\|G(s)\|_{\mathbb{L}_\infty^{m\times k}}\,g^{\sim}(i\omega)g(i\omega)d\omega
\end{aligned}$$

which proves the lemma. □

23.2.2 *Laurent, Toeplitz and Hankel operators*

Main definitions

Definition 23.3. *For* $G \in \mathbb{L}_\infty^{m\times k}$ *we may define the* ***Laurent*** *(or,* ***multiplication) operator***

$$\boxed{\Lambda_G : \mathbb{L}_2^k \to \mathbb{L}_2^m} \tag{23.48}$$

acting as

$$\Lambda_G F := GF \in \mathbb{L}_2^m \text{ if } F \in \mathbb{L}_2^k$$

Lemma 23.5. Λ_G *is a linear bounded operator, that is,*

$$\boxed{\|\Lambda_G F\|_{\mathbb{L}_2^m} \le \|G\|_{\mathbb{L}_\infty^{m\times k}}\,\|F\|_{\mathbb{L}_2^k}} \tag{23.49}$$

Proof. Directly, by the definition (23.19) and in view of (18.32), characterizing the operator norm, it follows that

$$\begin{aligned}
\|\Lambda_G F\|_{\mathbb{L}_2^m} &:= \int_{\omega=-\infty}^{\infty} \|\Lambda_G F(j\omega)\|_2 \, d\omega \\
&= \int_{\omega=-\infty}^{\infty} \|F(j\omega)\|_2 \frac{\|\Lambda_G F(j\omega)\|_2}{\|F(j\omega)\|_2} d\omega \\
&\le \int_{\omega=-\infty}^{\infty} \|F(j\omega)\|_2 \left(ess\sup_{\omega\in(-\infty,\infty)} \frac{\|\Lambda_G F(j\omega)\|_2}{\|F(j\omega)\|_2} \right) d\omega \\
&= \left(ess\sup_{\omega\in(-\infty,\infty)} \frac{\|\Lambda_G F(j\omega)\|_2}{\|F(j\omega)\|_2} \right) \int_{\omega=-\infty}^{\infty} \|F(j\omega)\|_2 \, d\omega = \|G\|_{\mathbb{L}_\infty^{m\times k}} \|F\|_{\mathbb{L}_2^m}
\end{aligned}$$

which completes the proof. □

By the orthogonal decomposition of the Hilbert space $\mathbb{L}_2^k$ (see (18.23)) it follows that

$$\boxed{\mathbb{L}_2^k = \left(\mathbb{H}_2^k\right)^\perp \oplus \mathbb{H}_2^k} \tag{23.50}$$

where $\left(\mathbb{H}_2^k\right)^\perp$ is the orthogonal completion in $\mathbb{L}_2^k$ of $\mathbb{H}_2^k$, that is, $\left(\mathbb{H}_2^k\right)^\perp$ is given by

$$\begin{aligned}
\left(\mathbb{H}_2^k\right)^\perp := \{ & F : \mathbb{C}^- \to \mathbb{C}^k, \ F \text{ is holomorphic (see Definition 17.2)} \\
\text{and} \quad & \|F\|^2_{\mathbb{L}_p^{m\times k}} := \sup_{\zeta<0} \int_{\omega=-\infty}^{\infty} \|F(\zeta + j\omega)\|^2 \, d\omega \}
\end{aligned} \tag{23.51}$$

In view of the decomposition (23.50) the Laurent operator Λ_G (23.48) can be represented in the "block form"

$$\boxed{\Lambda_G = \begin{bmatrix} \Lambda_G^{11} & \Lambda_G^{12} \\ \Lambda_G^{21} & \Lambda_G^{22} \end{bmatrix}} \tag{23.52}$$

where its "projections" Λ_G^{ij} $(i, j = 1, 2)$ act as

$$\boxed{\begin{array}{c} \Lambda_G^{11} : \left(\mathbb{H}_2^k\right)^\perp \to \left(\mathbb{H}_2^m\right)^\perp \\ \Lambda_G^{12} : \mathbb{H}_2^k \to \left(\mathbb{H}_2^m\right)^\perp \\ \Lambda_G^{21} : \left(\mathbb{H}_2^k\right)^\perp \to \mathbb{H}_2^m \\ \Lambda_G^{22} : \mathbb{H}_2^k \to \mathbb{H}_2^m \end{array}} \tag{23.53}$$

so that

$$\begin{pmatrix} \left(\mathbb{H}_2^m\right)^\perp \\ \mathbb{H}_2^m \end{pmatrix} \Longleftarrow \begin{bmatrix} \Lambda_G^{11} & \Lambda_G^{12} \\ \Lambda_G^{21} & \Lambda_G^{22} \end{bmatrix} \begin{pmatrix} \left(\mathbb{H}_2^k\right)^\perp \\ \mathbb{H}_2^k \end{pmatrix}$$

Definition 23.4.
(a) The ***Hankel operator*** Γ_G *associated with* $G \in \mathbb{L}_\infty^{m\times k}$ *is defined by*

$$\boxed{\Gamma_G := \Lambda_G^{21} : \left(\mathbb{H}_2^k\right)^\perp \to \mathbb{H}_2^m} \tag{23.54}$$

that is,

$$\boxed{\Gamma_G F = \Pi\left(\Lambda_G F(-s)\right) \in \mathbb{H}_2^m \text{ for } F(s) \in \mathbb{H}_2^k} \tag{23.55}$$

where Π *is the orthogonal projection operator from* $\mathbb{L}_2^m$ *onto* $\mathbb{H}_2^m$.
(b) The ***Toeplitz operator*** T_G *associated with* $G \in \mathbb{L}_\infty^{m\times k}$ *is defined by*

$$\boxed{\Theta_G := \Lambda_G^{22} : \mathbb{H}_2^k \to \mathbb{H}_2^m} \tag{23.56}$$

that is,

$$\boxed{\Theta_G F = \Pi\left(\Lambda_G F(s)\right) \in \mathbb{H}_2^m \text{ for } F(s) \in \mathbb{H}_2^k} \tag{23.57}$$

Example 23.5. (See Curtain & Zwart (1995)) *Consider* $G(s) = \dfrac{1}{s+a}$ *with* $\operatorname{Re} a > 0$. *Any* $F \in \mathbb{H}_2$ *can be represented as*

$$F(s) = F(a) + (s-a)X(s) \quad \text{for any} \quad s \in C^+ := \{s \in \mathbb{C} \mid \operatorname{Re} s > 0\}$$

where $X(s) \in \mathbb{H}_2$. *So, we have*

$$\Gamma_G F = \Pi\left(\Lambda_G F(-s)\right) = \Pi\left(\frac{1}{s+a} F(-s)\right)$$

$$= \Pi\left(\frac{1}{s+a}\left[F(a) + (-s-a)X(-s)\right]\right) = \frac{F(a)}{s+a}$$

Properties of Hankel operator Γ_G
Proposition 23.4. *The Hankel operator (23.54) associated with* $G \in \mathbb{L}_\infty^{m\times k}$ *has the following properties:*

1.

$$\boxed{\|\Gamma_G\| \le \|G\|_{\mathbb{L}_\infty^{m\times k}}} \tag{23.58}$$

2. if $G_1, G_2 \in \mathbb{L}_\infty^{m\times k}$ *then*

$$\boxed{\Gamma_{G_1+G_2} = \Gamma_{G_1} + \Gamma_{G2}} \tag{23.59}$$

Proof.

1. By (23.55) and (23.49) it follows that

$$\|\Gamma_G F\|_{\mathbb{H}_2^m} = \|\Pi\left(\Lambda_G F(-s)\right)\|_{\mathbb{H}_2^m}$$

$$\leq \|\Lambda_G F(-s)\|_{\mathbb{H}_2^m} \leq \|G\|_{\mathbb{L}_\infty^{m\times k}} \|F\|_{\mathbb{L}_2^k}$$

which implies

$$\|\Gamma_G\| = \sup_{F\in_{\mathbb{L}_2^k}} \frac{\|\Gamma_G F\|_{\mathbb{H}_2^m}}{\|F\|_{\mathbb{L}_2^k}} \leq \|G\|_{\mathbb{L}_\infty^{m\times k}}$$

2. The property (23.59) easily follows from the fact that the Laurent (multiplication) operator Λ_G (23.48) satisfies a similar relation, namely,

$$\Lambda_{G_1+G_2} = \Lambda_{G_1} + \Lambda_{G2}$$

Proposition is proven. □

Remark 23.3. *If $G(s) \in \mathbb{RL}_\infty^{m\times k}$ then $G(s)$ can be decomposed into a "strictly causal part" $G_c(s) \in \mathbb{RH}_2^{m\times k}$ and an "anticausal part" $G_{untc}(s)$, where $G_{untc}(s) \in \left(\mathbb{RH}_2^{m\times k}\right)^\perp$, such that for all $s \in \mathbb{C}$*

$$\boxed{G(s) = G_c(s) + G_{untc}(s) + G(\infty)} \tag{23.60}$$

Hence, one can check that if $F \in \left(\mathbb{H}_2^k\right)^\perp$ then

$$\Gamma_G F = \Pi\left(\Lambda_G F(-s)\right) = \Pi\left(\left[G_c(s) + G_{untc}(s) + G(\infty)\right] F(-s)\right)$$

$$= \Pi\left(G_c(s) F(-s)\right) + \Pi\left(\left[G_{untc}(s) + G(\infty)\right] F(-s)\right) = \Pi\left(G_c(s) F(-s)\right)$$

or, shortly,

$$\boxed{\Gamma_G F = \Pi\left(G_c(s) F(-s)\right) = \Gamma_{G_c} F} \tag{23.61}$$

*that is, **the Hankel operator** Γ_G associated with $G(s) \in \mathbb{RL}_\infty^{m\times k}$ **depends only on the strictly causal part** $G_c(s)$ of $G(s)$.*

Remark 23.4. *Particularly, if $G(s)$ is antistable, i.e., $G^\sim(s) \in \mathbb{RH}_\infty^{m\times k}$, then $G_c(s) = 0$ and*

$$\boxed{\Gamma_G = 0} \tag{23.62}$$

Hankel operator in the time domain

Here we will introduce the time domain Hankel operator which has a natural relation with the corresponding Hankel operator Γ_G (23.54) defined in the frequency domain.

Definition 23.5. *Let* $g(t) \in L_1^{m\times k}[0,\infty)$. *Then the* ***time domain Hankel operator*** Γ_g *is defined by*

$$\boxed{\begin{aligned}&\Gamma_g : L_2^k[0,\infty) \to L_2^m[0,\infty)\\ &\left(\Gamma_g u\right)(t) := \int_{\tau=0}^{\infty} g(t+\tau)\,u(\tau)\,d\tau, \quad t \geq 0\end{aligned}} \tag{23.63}$$

Proposition 23.5. *1. For* $t \geq 0$

$$\boxed{\begin{aligned}\left(\Gamma_g u\right)(t) &= \int_{\tau=-\infty}^{\infty} \tilde{g}(\tau)\,\tilde{u}(t-\tau)\,d\tau\\ &= \int_{\tau=-\infty}^{0} \tilde{g}(t-\tau)\,\tilde{u}(\tau)\,d\tau\end{aligned}} \tag{23.64}$$

where

$$\tilde{g}(\tau) := \begin{cases} g(\tau) & \text{if} \quad \tau \geq 0\\ 0 & \text{if} \quad \tau < 0\end{cases}$$

$$\tilde{u}(\tau) = \begin{cases} 0 & \text{if} \quad \tau \geq 0\\ u(-\tau) & \text{if} \quad \tau < 0\end{cases}$$

2.

$$\boxed{\left\|\Gamma_g\right\| \leq \int_{\tau=0}^{\infty} \|g(\tau)\|\,d\tau} \tag{23.65}$$

Proof.

1. Indeed,

$$\begin{aligned}\left(\Gamma_g u\right)(t) &= \int_{\tau=0}^{\infty} g(t+\tau)\, u(\tau)\, d\tau = \int_{s=t}^{\infty} g(s)\, u(s-t)\, ds \\ &= \int_{s=t}^{\infty} g(s)\, \tilde{u}(t-s)\, ds = \int_{s=0}^{\infty} g(s)\, \tilde{u}(t-s)\, ds \\ &= \int_{s=0}^{\infty} \tilde{g}(s)\, \tilde{u}(t-s)\, ds = \int_{s=-\infty}^{\infty} \tilde{g}(s)\, \tilde{u}(t-s)\, ds \\ &\overset{t-s=\tau}{=} \int_{s=-\infty}^{\infty} \tilde{g}(t-\tau)\, \tilde{u}(\tau)\, d\tau = \int_{s=-\infty}^{0} \tilde{g}(t-\tau)\, \tilde{u}(\tau)\, d\tau\end{aligned}$$

2. Define the operator $\tilde{\Gamma}_g : L_2^k[0,\infty) \to L_2^m(-\infty,\infty)$ by the relation

$$\left(\tilde{\Gamma}_g u\right)(t) := \int_{s=-\infty}^{\infty} \tilde{g}(s)\, \tilde{u}(t-s)\, ds$$

Then

$$\begin{aligned}\left\|\tilde{\Gamma}_g u\right\|^2_{L_2^m(-\infty,\infty)} &= \int_{t=-\infty}^{\infty} \left\|\left(\tilde{\Gamma}_g u\right)(t)\right\|^2 dt \\ &= \int_{t=-\infty}^{\infty} \left\| \int_{s=0}^{\infty} g(s)\, \tilde{u}(t-s)\, ds \right\|^2 dt \le \int_{t=-\infty}^{\infty} \left(\int_{s=0}^{\infty} \|g(s)\|\, \|\tilde{u}(s-t)\|\, ds \right)^2 dt \\ &= \int_{t=-\infty}^{\infty} \int_{s=0}^{\infty} \int_{s'=0}^{\infty} \|g(s)\|\, \left\|g\left(s'\right)\right\|\, \|\tilde{u}(s-t)\|\, \left\|\tilde{u}\left(s'-t\right)\right\| ds\, ds'\, dt \\ &= \int_{s=0}^{\infty} \int_{s'=0}^{\infty} \|g(s)\|\, \left\|g\left(s'\right)\right\| \left[\int_{t=-\infty}^{\infty} \|\tilde{u}(s-t)\|\, \left\|\tilde{u}\left(s'-t\right)\right\| dt \right] ds\, ds' \\ &= \int_{s=0}^{\infty} \int_{s'=0}^{\infty} \|g(s)\|\, \left\|g\left(s'\right)\right\|\, \|u\|^2_{L_2^k[0,\infty)}\, ds\, ds' = \|g\|^2_{L_1^{m\times k}[0,\infty)}\, \|u\|^2_{L_2^k[0,\infty)}\end{aligned}$$

The result (23.65) follows then from the inequalities

$$\int_{t=0}^{\infty} \left\|\left(\Gamma_g u\right)(t)\right\|^2 dt \le \int_{t=-\infty}^{\infty} \left\|\left(\tilde{\Gamma}_g u\right)(t)\right\|^2 dt \le \|g\|^2_{L_1^{m\times k}[0,\infty)}\, \|u\|^2_{L_2^k[0,\infty)}$$

The proposition is proven. □

The relation (23.64) permits to interpret the Hankel operator Γ_g, associated with the transition function g, as the map from the past inputs $u(t)\,|_{t<0}$ to the further output $y(t)\,|_{y\ge 0}$, that is, $\left(y(t)\,|_{y\ge 0}\right) = \Gamma_g\left(u(t)\,|_{t<0}\right)$ (see Fig. 23.3).

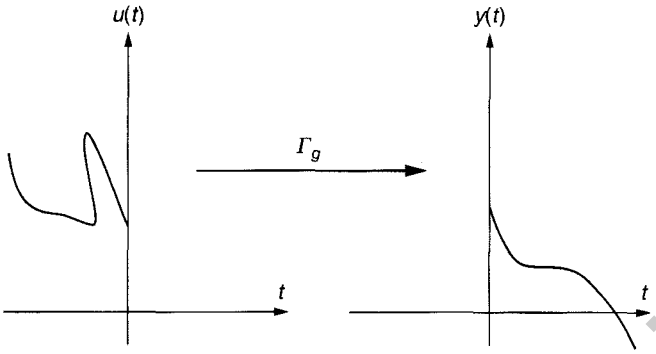

Fig. 23.3. Time domain interpreation of the Hankel operator.

Lemma 23.6. *The frequency domain Hankel operator* Γ_G *(23.54) is the Laplace-transformed version of the time domain operator* Γ_g *(23.63) where* $g(t)$ *is the inverse (bilateral) Laplace transformation of* $G(s)$*, that is, if* $U(s) = \mathcal{L}\{u\}(s)$ *then*

$$\boxed{\mathcal{L}\{\Gamma_g u\} = \Gamma_G U} \tag{23.66}$$

and

$$\boxed{\|\Gamma_g\| = \|\Gamma_G\|} \tag{23.67}$$

Proof. From (23.64) by the property (17.95) we have

$$(\Gamma_g u)(t) = \int_{\tau=-\infty}^{\infty} \tilde{g}(\tau)\,\tilde{u}(t-\tau)\,d\tau = \int_{s=-\infty}^{\infty} \tilde{g}(t-s)\,\tilde{u}(s)\,ds$$

and, hence,

$$\mathcal{L}\{\Gamma_g u\} = \mathcal{L}\left\{\int_{\tau=-\infty}^{\infty} \tilde{g}(\tau)\,\tilde{u}(t-\tau)\,d\tau\right\}$$

$$= \mathcal{L}\{\tilde{g}\}\,\mathcal{L}\{\tilde{u}\} = \Pi(\Lambda_G U(-s)) = \Gamma_G U$$

The equality (23.67) is true because of the isomorphism property between L_2 and $\mathbb{L}_2$ spaces (see the Plancherel theorem 17.18). □

The norm of Hankel operator Γ_G acting from $\left(\mathbb{RH}_2^k\right)^\perp$ onto $\mathbb{RH}_2^m$

Suppose that $G(s) \in \mathbb{RH}_\infty^{m\times k}$ has a minimal state space realization $\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]$ with A stable. Then, according to (9.54), (9.62) and Lemma 9.1, the controllability G_c and observability G_o grammians are given by

$$\boxed{G_c := \int_{\tau=0}^{\infty} e^{A\tau} B B^\top e^{A^\top \tau} d\tau, \quad G_o := \int_{\tau=0}^{\infty} e^{A^\top \tau} C^\top C e^{A\tau} d\tau} \tag{23.68}$$

and satisfy the matrix Lyapunov equations

$$\boxed{AG_c + G_c A^\top = -BB^\top, \quad A^\top G_o + G_o A = -C^\top C} \tag{23.69}$$

Let us define also the *controllability* Ψ_c and the *observability* Ψ_o operators which are defined, respectively, as

$$\boxed{\begin{array}{c} \Psi_c : L_2^k(-\infty, 0] \to \mathbb{C}^n \\ \\ \Psi_c u(t) := \displaystyle\int_{\tau=-\infty}^{0} e^{-A\tau} B u(\tau)\, d\tau \end{array}} \tag{23.70}$$

and

$$\boxed{\begin{array}{c} \Psi_0 : \mathbb{C}^n \to L_2^m[0, \infty) \\ \\ \Psi_o x_0 := C e^{At} x_0, \quad t \geq 0 \end{array}} \tag{23.71}$$

Lemma 23.7. *For any $G(s) \in \mathbb{RH}_\infty^{m\times k}$ with a minimal state space realization $\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]$ we have*

$$\boxed{\Gamma_g = \Psi_o \Psi_c} \tag{23.72}$$

and for any $z \in \mathbb{C}^m$

$$\boxed{\begin{array}{l} \Psi_c \Psi_c^* z = G_c z \\ \\ \Psi_o^* \Psi_o z = G_o z \end{array}} \tag{23.73}$$

Proof. This fact can be easily checked directly using the representation of Γ_g in the form (23.64). Indeed, assuming that $x(-\infty) = 0$ and in view of (23.64), we get

$$(\Psi_o \Psi_c u)(t) = Ce^{At} \int_{\tau=-\infty}^{0} e^{-A\tau} Bu(\tau)\, d\tau$$
$$= \int_{\tau=-\infty}^{0} Ce^{A(t-\tau)} Bu(\tau)\, d\tau = \int_{\tau=-\infty}^{0} \tilde{g}(t-\tau)\, \tilde{u}(\tau)\, d\tau = (\Gamma_g u)(t)$$

The relation (23.73) results from the definitions (23.68), (23.70) and (23.71) which completes the proof. □

Theorem 23.7. (on the norm of the Hankel operator) *For any* $G(s) \in \mathbb{RH}_\infty^{m\times k}$ *with a minimal state space realization* $\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]$

1. the operators $\Gamma_g^*\Gamma_g$, $\Gamma_G^*\Gamma_G$ *and the matrix* $G_c G_o$ *have the same positive eigenvalues;*
2.

$$\boxed{\|\Gamma_g\| = \|\Gamma_G\| = \sqrt{\lambda_{\max}(G_c G_o)}} \tag{23.74}$$

Proof.

1*(a)* If σ^2 is an eigenvalue of $\Gamma_g^*\Gamma_g$ corresponding to an eigenvector (function) $0 \neq u \in L_2^k(-\infty, 0]$, then by definition

$$\Gamma_g^*\Gamma_g u = \Psi_c^*\Psi_o^*\Psi_o\Psi_c u = \sigma^2 u$$

and in view of (23.73), after the pre-multiplication of the last identity by Ψ_c and defining $z := \Psi_c u$, it follows that

$$\Psi_c\Gamma_g^*\Gamma_g u = \Psi_c\Psi_c^*\Psi_o^*\Psi_o z = G_c G_o z = \sigma^2 \Psi_c u = \sigma^2 z$$

So, σ^2 is an eigenvalue of $G_c G_o$. Since both matrices G_c and G_o are strictly positive it follows that $\sigma^2 > 0$.

1*(b)* To show that the operator $\Gamma_G^*\Gamma_G$ has the same eigenvalues let us rewrite equations (23.69) in the following form

$$\begin{aligned} -(sI - A)\, G_c + G_c\,(sI + A^\top) &= -BB^\top \\ (sI + A^\top)\, G_o - G_o\,(sI - A) &= -C^\top C \end{aligned} \tag{23.75}$$

Pre- and post-multiplying the first equation in (23.75) by $C(sI - A)^{-1}$ and $(sI + A^\top)^{-1} v$, respectively, where

$$v := \frac{1}{\sqrt{\lambda_{\max}(G_c G_o)}} G_o w \tag{23.76}$$

and w is an eigenvector corresponding to $\lambda_{\max}(G_c G_o)$, namely,

$$G_c G_o w = \lambda_{\max}(G_c G_o)\, w$$

we get

$$\begin{aligned} &-CG_c(sI + A^\intercal)^{-1} v + C(sI - A)^{-1} G_c v \\ &\quad = -C(sI - A)^{-1} BB^\intercal (sI + A^\intercal)^{-1} v \end{aligned} \tag{23.77}$$

Notice that

$$-CG_c(sI + A^\intercal)^{-1} v = -CG_c(sI - (-A)^\intercal)\, v \in \left(\mathbb{RH}_2^m\right)^\perp$$

$$C(sI - A)^{-1} G_c v \in \mathbb{RH}_2^m$$

Define two vector functions

$$\boxed{\begin{aligned} &\mathsf{f}(s) := C(sI - A)^{-1} w \in \mathbb{RH}_2^m \\ &\mathsf{g}(s) := -B^\intercal (sI + A^\intercal)^{-1} v \in \left(\mathbb{RH}_2^k\right)^\perp \end{aligned}} \tag{23.78}$$

which will be referred to as the *Schmidt pair* of vectors corresponding to the transfer matrix $G(s) \in \mathbb{RH}_\infty^{m\times k}$ with a minimal state space realization $\left[\begin{array}{cc} A & B \\ C & 0 \end{array}\right]$. Then

$$G_c v = \sqrt{\lambda_{\max}(G_c G_o)}\, w$$

and

$$C(sI - A)^{-1} G_c v = \sqrt{\lambda_{\max}(G_c G_o)}\, C(sI - A)^{-1} w = \sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{f}(s)$$

Notice that the right-hand side in (23.77) is

$$-C(sI - A)^{-1} BB^\intercal (sI + A^\intercal)^{-1} v = C(sI - A)^{-1} B\mathsf{g}(s) = G(s)\, \mathsf{g}(s)$$

which implies

$$-CG_c(sI + A^\intercal)^{-1} v + \sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{f}(s) = G(s)\, \mathsf{g}(s)$$

Projecting this equality to $\mathbb{RH}_2^m$ and in view of (23.54) we get

$$\boxed{\sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{f}(s) = \Pi G(s)\, \mathsf{g}(s) := \Gamma_G \mathsf{g}(s)} \tag{23.79}$$

Analogously, pre- and post-multiplying the second equation in (23.75) by $w^\intercal (sI + A^\intercal)^{-1}$ and $(sI - A)^{-1} B$, respectively, we get

$$w^\intercal G_o (sI - A)^{-1} B - w^\intercal (sI + A^\intercal)^{-1} G_o B$$

$$= -w^\intercal (sI + A^\intercal)^{-1} C^\intercal C (sI - A)^{-1} B = \mathsf{f}^\intercal(-s) G(s) = \tilde{\mathsf{f}}(s) G(s)$$

Since by definition $G_o w = \sqrt{\lambda_{\max}(G_c G_o)} v$ the last equality becomes as follows

$$\sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{g}^\intercal(-s) - w^\intercal (sI + A^\intercal)^{-1} G_o B = \tilde{\mathsf{f}}(s) G(s)$$

Projecting this identity to $\mathbb{RH}_2^m$ we obtain

$$\sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{g}^\intercal(-s) = \sqrt{\lambda_{\max}(G_c G_o)}\, \tilde{\mathsf{g}}(s)$$

$$= \Pi\left(\tilde{\mathsf{f}}(s) G(s)\right) := \tilde{\mathsf{f}}(s) \Gamma_G$$

Taking the conjugation operation from both sides of this equality we have

$$\boxed{\Gamma_G^* \mathsf{f}(s) = \sqrt{\lambda_{\max}(G_c G_o)}\, \mathsf{g}(s)} \tag{23.80}$$

The relations (23.79) and (23.80) lead to the following identity

$$\Gamma_G^* \Gamma_G \mathsf{g}(s) = \sqrt{\lambda_{\max}(G_c G_o)}\, \Gamma_G^* \mathsf{f}(s) = \lambda_{\max}(G_c G_o)\, \mathsf{g}(s) \tag{23.81}$$

This means that $\mathsf{g}(s)$ is the eigenvector of $\Gamma_G^* \Gamma_G$ corresponding to the eigenvalue $\lambda_{\max}(G_c G_o)$. Evidently, the other eigenvectors are

$$\mathsf{g}_i(s) := -B^\intercal (sI + A^\intercal)^{-1} v_i, \quad v_i = \frac{1}{\sqrt{\lambda_{\max}(G_c G_o)}} G_o w_i$$

$$G_c G_o w_i = \lambda_i w_i \quad (i = 1, \ldots, n)$$

such that $\lambda_{i_0} = \lambda_{\max}(G_c G_o)$ for an index i_0. This shows that

$$\lambda_{\max}\left(\Gamma_G^* \Gamma_G\right) = \lambda_{\max}(G_c G_o)$$

and, hence,

$$\left\|\Gamma_G^* \Gamma_G\right\| := \sup_{U \in \left(\mathbb{RH}_2^k\right)^\perp} \frac{U^* \Gamma_G^* \Gamma_G U}{\tilde{U} U} = \lambda_{\max}\left(\Gamma_G^* \Gamma_G\right) = \lambda_{\max}(G_c G_o) \tag{23.82}$$

2. (23.74) follows from (23.82) and the relation (23.67). Theorem is proven. □

Corollary 23.4. *For any* $s \in \mathbb{C}$

$$\boxed{\tilde{\mathsf{f}}(s) \mathsf{f}(s) = \tilde{\mathsf{g}}(s) \mathsf{g}(s)} \tag{23.83}$$

Proof. It results directly from (23.79) and (23.80) since

$$\mathsf{f}^{\sim}(s)\mathsf{f}(s) = \frac{1}{\sqrt{\lambda_{\max}(G_c G_o)}}\mathsf{f}^{\sim}(s)\left(\Gamma_G \mathsf{g}(s)\right)$$

$$= \frac{1}{\sqrt{\lambda_{\max}(G_c G_o)}}\left(\mathsf{f}^{\sim}(s)\Gamma_G\right)\mathsf{g}(s) = \mathsf{g}^{\sim}(s)\mathsf{g}(s)$$

Corollary is proven. □

Example 23.6. *For*

$$G(s) = \frac{1-s}{1+s} = G_c(s) + G(\infty)$$

$$G_c(s) = \frac{2}{1+s}, \quad G(\infty) = -1$$

in view of the relation (23.61) $\Gamma_G = \Gamma_{G_c}$ *it follows that the minimal state space realization of* $G_c(s)$ *is* $\left[\begin{matrix} -1 & 1 \\ 2 & 0 \end{matrix}\right]$. *So*

$$A = -1, \quad B = 1 \quad \text{and} \quad C = 2$$

and, hence, by (23.69)

$$G_c = \frac{B^2}{2|A|} = \frac{1}{2}, \quad G_o = \frac{C^2}{2|A|} = 2$$

Using (23.74) we get

$$\|\Gamma_G\| = \left\|\Gamma_{G_c}\right\| = \sqrt{G_c G_o} = 1$$

23.2.3 Nehari problem in $\mathbb{RL}_\infty^{m\times k}$

Nehari problem formulation

The *Nehari problem* deals with the approximation of a transfer matrix $G(s) \in \mathbb{RL}_\infty^{m\times k}$ by an anticausal transfer matrix $X \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ where the approximation is done with respect to $\mathbb{L}_\infty$ norm. It is naturally formulated in frequency domain terms in the following way:

Given a matrix-valued function $G(s) \in \mathbb{RL}_\infty^{m\times k}$, find the $\mathbb{L}_\infty$ distance of G from the set of unstable matrix valued functions $X(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ and define one of $X_0(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ where this minimal distance is achieved, that is, for the given $G(s) \in \mathbb{RL}_\infty^{m\times k}$ calculate

$$\boxed{\gamma_{opt} = \operatorname{dist}\left(G, \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp\right) := \inf_{X(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \|G(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}}} \tag{23.84}$$

and find any $X_0(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$, referred to as an **optimal element**, such that

$$\boxed{\|G(s) - X_0(s)\|_{\mathbb{L}_\infty^{m\times k}} = \gamma_{opt}} \tag{23.85}$$

Even this problem is formulated in Banach space $\mathbb{RL}_\infty^{m\times k}$, it admits a precise and elegant solution in terms of the Hankel operator Γ_G of the matrix-valued function $G(s)$.

Theorem 23.8. (Nehari 1957) *Let* $G(s) \in \mathbb{RL}_\infty^{m\times k}$. *Then*

$$\boxed{\begin{aligned} \gamma_{opt} &= \inf_{X(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \|G(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}} = \\ &\inf_{X(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \|G_c(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}} = \|\Gamma_G\| = \left\|\Gamma_{G_c}\right\| \end{aligned}} \tag{23.86}$$

where $G_c \in \mathbb{RH}_2^{m\times k}$ *is the causal part* $G(s)$. *If there exists an optimal element* $X_0(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ *such that (23.85) holds, then it should satisfy the identity*

$$\boxed{[G(s) - X_0(s)]\,\mathsf{g}(s) = \Gamma_{G_c}\mathsf{g}(s)} \tag{23.87}$$

where $\mathsf{g}(s)$ *is defined by (23.78) for a minimal state-space realization of* $G_c(s)$.

Proof. By (23.58), (23.59), (23.61) and (23.62) for any $X(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ it follows that

$$\|G(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}} \geq \|\Gamma_{G-X}\| = \|\Gamma_G - \Gamma_X\| = \|\Gamma_G\| = \left\|\Gamma_{G_c}\right\|$$

which states the inequality

$$\gamma_{opt} = \inf_{X(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \|G(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}} \geq \|\Gamma_G\| \tag{23.88}$$

Suppose that there exists $X_0(s) \in \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp$ such that

$$\begin{aligned} \|G(s) - X_0(s)\|_{\mathbb{L}_\infty^{m\times k}}^2 = \|\Gamma_G\|^2 = \left\|\Gamma_{G_c}\right\|^2 &= \sup_{U\in\left(\mathbb{RH}_2^k\right)^\perp} \frac{U^{\sim}\Gamma_{G_c}^*\Gamma_{G_c}U}{U^{\sim}U} \\ &= \lambda_{\max}\left(\Gamma_{G_c}^*\Gamma_{G_c}\right) \overset{(23.81)}{=} \frac{\mathsf{g}^{\sim}(s)\Gamma_{G_c}^*\Gamma_{G_c}\mathsf{g}(s)}{\mathsf{g}^{\sim}(s)\mathsf{g}(s)} \end{aligned} \tag{23.89}$$

Denote

$$h(s) := [G(s) - X_0(s)]\,\mathsf{g}(s)$$

and consider

$$\left\|h(s)-\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2=\|h(s)\|_{\mathbb{L}_2^m}^2+\left\|\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2 -2\left\langle h(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m\mathbb{L}_2^m}^2 \tag{23.90}$$

Using the presentation

$$h(s)=\Pi h(s)+\Pi^{\perp}h(s)$$
$$\Pi h(s)\in\mathbb{RH}_2^m,\quad \Pi^{\perp}h(s)\in\left(\mathbb{RH}_2^m\right)^{\perp}$$

and remembering that

$$\mathsf{g}(s)\in\left(\mathbb{RH}_2^k\right)^{\perp},\quad X_0(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^{\perp},\quad X_0(s)\,\mathsf{g}(s)\in\left(\mathbb{RH}_\infty^m\right)^{\perp}$$

we derive

$$\begin{aligned}\left\langle h(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}&=\left\langle \Pi h(s)+\Pi^{\perp}h(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}\\&=\left\langle \Pi h(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}=\left\langle \Pi\left[G(s)-X_0(s)\right]\mathsf{g}(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}\\&=\left\langle \Pi\left[G(s)-X_0(s)\right]\mathsf{g}(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}=\left\langle \Pi G(s)\mathsf{g}(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}\\&=\left\langle \Gamma_G\mathsf{g}(s),\Gamma_{G_c}\mathsf{g}(s)\right\rangle_{\mathbb{L}_2^m}=\left\|\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2\end{aligned}$$

So, (23.90) becomes

$$\left\|h(s)-\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2=\|h(s)\|_{\mathbb{L}_2^m}^2-\left\|\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2$$

which, by (23.47) and (23.89), implies

$$\begin{aligned}\left\|h(s)-\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2&=\|h(s)\|_{\mathbb{L}_2^m}^2-\left\|\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2\le\|G(s)-X_0(s)\|_{\mathbb{L}_\infty^{m\times k}}^2\\&\quad-\|\Gamma_G\mathsf{g}(s)\|_{\mathbb{L}_2^n}^2\le\|\Gamma_G\|^2-\left\|\Gamma_{G_c}\mathsf{g}(s)\right\|_{\mathbb{L}_2^m}^2\\&=\lambda_{\max}(G_cG_o)-\lambda_{\max}(G_cG_o)=0\end{aligned}$$

This means that if (23.89) holds then obligatory (23.87) holds too. Theorem is proven. □

Corollary 23.5. *In a multidimensional case (when $k+m>2$) the Nehari problem (23.84) has infinitely many solutions, and at least one of them, referred to as* ***the central optimal element****, is given by*

$$\boxed{X_0(s)=G(s)-\sqrt{\lambda_{\max}(G_cG_o)}\frac{\mathsf{f}(s)\tilde{\mathsf{g}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}} \tag{23.91}$$

Proof. The direct substitution of (23.91) into (23.87) and (23.86) shows that $X_0(s)$ given by (23.91) is a solution of the Nehari problem. Indeed,

$$[G(s) - X_0(s)]\,\mathsf{g}(s) = \sqrt{\lambda_{\max}(G_c G_o)}\,\frac{\mathsf{f}(s)\tilde{\mathsf{g}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\,\mathsf{g}(s)$$

$$= \sqrt{\lambda_{\max}(G_c G_o)}\,\mathsf{f}(s) = \Gamma_{G_c}\mathsf{g}(s)$$

and, by (23.79) and (23.80),

$$\|G(s) - X_0(s)\|^2_{\mathbb{L}_\infty^{m\times k}} = \lambda_{\max}(G_c G_o)\left\|\frac{\mathsf{f}(s)\tilde{\mathsf{g}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\right\|^2_{\mathbb{L}_\infty^{m\times k}}$$

$$= \lambda_{\max}(G_c G_o)\ \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}\left(\frac{\mathsf{f}(s)\tilde{\mathsf{g}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\,\frac{\mathsf{g}(s)\tilde{\mathsf{f}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\right)$$

$$= \lambda_{\max}(G_c G_o)\ \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}\left(\frac{\mathsf{f}(s)\tilde{\mathsf{f}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\right)$$

$$\lambda_{\max}(G_c G_o)\ \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)}\left(\frac{\tilde{\mathsf{f}}(s)\mathsf{f}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\right) = \lambda_{\max}(G_c G_o) = \|\Gamma_{G_c}\|^2$$

which proves the desired result. □

Corollary 23.6. *The transfer matrix* $[G(s) - X_0(s)]$ *with* $X_0(s)$ *given by (23.91) is an* ***all-pass transfer matrix****, that is,*

$$\boxed{\operatorname{tr}\left([G(s) - X_0(s)]^{\sim}\,[G(s) - X_0(s)]\right) = \operatorname*{const}_s} \tag{23.92}$$

Proof. Indeed, by (23.91), we have

$$\operatorname{tr}\left([G(s) - X_0(s)]^{\sim}\,[G(s) - X_0(s)]\right)$$

$$= \lambda_{\max}(G_c G_o)\operatorname{tr}\left(\frac{\mathsf{g}(s)\tilde{\mathsf{f}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\,\frac{\mathsf{f}(s)\tilde{\mathsf{g}}(s)}{\tilde{\mathsf{g}}(s)\mathsf{g}(s)}\right) = \lambda_{\max}(G_c G_o)$$

□

Corollary 23.7. (Adamjan *et al.* 1971) *In the scalar case* $(m = k = 1)$ *the Nehari problem (23.84) has the* ***unique solution*** *given by*

$$\boxed{X_0(s) = G(s) - \sqrt{\lambda_{\max}(G_c G_o)}\,\frac{\mathsf{f}(s)}{\mathsf{g}(s)}} \tag{23.93}$$

Proof. This fact follows directly from (23.87) since it is the unique central element. Indeed,

$$[G(s) - X_0(s)]\,\mathsf{g}(s) = \Gamma_{G_c}\mathsf{g}(s) = \sqrt{\lambda_{\max}(G_c G_o)}\,\mathsf{f}(s)$$

and in the scalar case

$$G(s) - X_0(s)\,\mathsf{g}(s) = \frac{\Gamma_{G_c}\mathsf{g}(s)}{\mathsf{g}(s)} = \sqrt{\lambda_{\max}(G_c G_o)}\,\frac{\mathsf{f}(s)}{\mathsf{g}(s)}$$

□

Example 23.7. *Let*

$$G(s) = \frac{(1+s)^2(5+s)}{(1+10s)(s-1)(s-5)}$$

It can be represented as

$$G(s) = G_c(s) + G_{untc}(s) + G(\infty)$$

$$G_c(s) = \frac{a}{1+10s}, \quad a = 0.70749$$

$$G_{untc}(s) = \frac{b+cs}{(s-1)(s-5)}, \quad b = 0.96257, \quad c = 1.2193$$

$$G(\infty) = 0.1$$

The minimal state-space realization of $G_c(s)$ *is*

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix} = \begin{bmatrix} -0.1 & 0.1 \\ a & 0 \end{bmatrix}$$

which gives

$$G_c = 0.05, \quad G_0 = 5a^2$$

$$\left\|\Gamma_{G_c(s)}\right\| = \sqrt{G_c G_0} = 0.5a = 0.35375$$

$$w = 1, \quad v = 10a, \quad \mathsf{f}(s) = \frac{a}{s+0.1}, \quad \mathsf{g}(s) := -\frac{a}{s-0.1}$$

and, finally,

$$X_0(s) = G(s) - \sqrt{\lambda_{\max}(G_c G_o)}\frac{\mathsf{f}(s)}{\mathsf{g}(s)}$$

$$= 0.45374 + \frac{b + cs}{(s-1)(s-5)} = \frac{0.45374s^2 - 1.5031s + 3.2313}{(s-1)(s-5)}.$$

Remark 23.5. *The problem dealing with the approximation of a transfer matrix* $G(s) \in \mathbb{RL}_\infty^{m\times k}$ *by a causal transfer matrix* $X(s) \in \mathbb{RH}_\infty^{m\times k}$*, where the approximation is done with respect to* $\mathbb{L}_\infty$ *norm, is* ***equivalent*** *to the Nehari problem (23.84) where the given transfer matrix* $G(s) \in \mathbb{RL}_\infty^{m\times k}$ *is changed to*

$$\boxed{\bar{G}(s) := G(-s) \in \mathbb{RL}_\infty^{m\times k}} \tag{23.94}$$

Indeed, changing variable s *to* $(-s)$ *it follows that*

$$\begin{aligned}
\operatorname{dist}\left(G, \mathbb{RH}_\infty^{m\times k}\right) &:= \inf_{X(s)\in\mathbb{RH}_\infty^{m\times k}} \|G(s) - X(s)\|_{\mathbb{L}_\infty^{m\times k}} \\
&= \inf_{X(-s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \|G(-s) - X(-s)\|_{\mathbb{L}_\infty^{m\times k}} \\
&= \inf_{X(-s):=\bar{X}(s)\in\left(\mathbb{RH}_\infty^{m\times k}\right)^\perp} \left\|\bar{G}(s) - \bar{X}(s)\right\|_{\mathbb{L}_\infty^{m\times k}} \\
&= \operatorname{dist}\left(\bar{G}, \left(\mathbb{RH}_\infty^{m\times k}\right)^\perp\right)
\end{aligned} \tag{23.95}$$

23.2.4 Model-matching (MMP) problem

A controlled system is said to be *robust* if it possesses a guaranteeing working quality in spite of the presence of some uncertain factors (usually related to environment perturbations) which may affect it during a normal regime. Formally, the problem of synthesis of robust controlled systems belongs to the class of the, so-called, min–max optimization problems where *max* is taken over the set of uncertainties or disturbances and *min* is taken over the set of admissible controllers. In this subsection we will consider one of the most important min–max control problems and show its close relationship to the Nehari problem discussed above.

MMP problem formulation

Consider a multi-connected linear system which block scheme is presented in Fig. 23.4. Suppose that all blocks in Fig. 23.4 are stable, that is,

$$T_1 \in \mathbb{RH}_\infty^{m_1 \times k_1}, \quad T_2 \in \mathbb{RH}_\infty^{m_2 \times k_2}, \quad T_3 \in \mathbb{RH}_\infty^{m_3 \times k_3}, \quad Q \in \mathbb{RH}_\infty^{m_q \times k_q}$$

$$k_1 = k_2, \quad m_2 = k_q, \quad m_q = k_3, \quad m_3 = m_1$$

Problem 23.3. (MMP) *The model-matching problem (MMP) consists of finding a stable block with the transfer matrix $Q(s)$ which for the worst external perturbation ξ of a bounded energy, i.e.,*

$$\boxed{\xi \in L_2^{k_1} : \int_{t=0}^{\infty} \|\xi(t)\|^2 dt \le C < \infty} \tag{23.96}$$

would provide the "best approximation" of the given plant with the transfer matrix $T_1(s)$ by the model with the transfer function $T_2(s)\, Q(s)\, T_3(s)$, namely, the MMP problem is

$$\boxed{\begin{aligned} J(Q) := \sup_{\xi \in L_2^{k_1} : \int_{t=0}^{\infty} \|\xi(t)\|^2 dt \le C} \int_{t=0}^{\infty} \|\varepsilon(t)\|^2 dt \to \inf_{Q \in \mathbb{RH}_\infty^{m_q \times k_q}} \\ \varepsilon(t) := y(t) - \hat{y}(t) \end{aligned}} \tag{23.97}$$

where

$$y(t) := \mathcal{L}^{-1}\{T_1(s)\, \Xi(s)\}$$

$$\hat{y}(t) := \mathcal{L}^{-1}\{[T_2(s)\, Q(s)\, T_3(s)]\, \Xi(s)\}$$

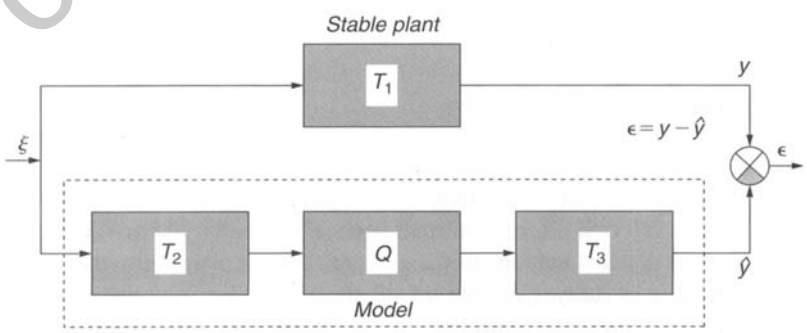

Fig. 23.4. The block scheme for the MMP problem.

are the inverse Laplace transformations (17.79) of the corresponding vector function defined on C and

$$\Xi(s) := \mathcal{L}\{\xi(t)\}$$

is the Laplace transformation (17.74) of $\xi(t)$.

One can see that it is a min–max optimization problem.

The equivalent MMP problem formulation in the frequency domain

Lemma 23.8. (on the equivalency) *The MMP problem (23.97) in the time domain is equivalent to the following MMP problem in the frequency domain:*

$$\boxed{\begin{aligned} J(Q) &= C\,\|T_1(s) - T_2(s)\,Q(s)\,T_3(s)\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}} \\ &\to \inf_{Q\in\mathbb{RH}_\infty^{m_q\times k_q}} \end{aligned}} \tag{23.98}$$

Proof. By Parseval's identity (17.107) it follows that

$$\begin{aligned} J(Q) &= \sup_{\Xi:\frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\|\Xi(i\omega)\|^2 d\omega\le C} \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty}\|E(i\omega)\|^2 d\omega \\ &= \sup_{\Xi:\|\Xi\|^2_{\mathbb{L}_2^{k_1}}\le C}\|E\|^2_{\mathbb{L}_2^{k_1}} = \sup_{\Xi:\|\Xi\|^2_{\mathbb{L}_2^{k_1}}\le C}\left\|[T_1(s) - T_2(s)\,Q(s)\,T_3(s)]\,\Xi(s)\right\|^2_{\mathbb{L}_2^{k_1}} \\ &= C\,\|T_1(s) - T_2(s)\,Q(s)\,T_3(s)\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}} \end{aligned}$$

where $E(i\omega) := \mathcal{L}\{\varepsilon(t)\}$ is the Laplace transformation (17.74) of $\varepsilon(t)$. But by (23.47)

$$\begin{aligned} &\left\|[T_1(s) - T_2(s)\,Q(s)\,T_3(s)]\,\Xi(s)\right\|^2_{\mathbb{L}_2^{k_1}} \\ &\quad\le \|T_1(s) - T_2(s)\,Q(s)\,T_3(s)\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}}\,\|\Xi(s)\|^2_{\mathbb{L}_2^{k_1}} \end{aligned} \tag{23.99}$$

and the equality is attained. Indeed, if

$$\omega_0 := \arg\max \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \bar{\sigma}(T_1(i\omega) - T_2(i\omega)\,Q(i\omega)\,T_3(i\omega))$$

is finite, then the equality in (23.99) is attained when

$$\Xi(i\omega)\,\tilde{\Xi}(i\omega) = 2\pi\delta(\omega - \omega_0)\,I_{k_1\times k_1}$$

If $|\omega_0| = \infty$, then $\underset{\omega\in(-\infty,\infty)}{ess\sup}\ \bar{\sigma}\left(T_1(i\omega) - T_2(i\omega)\,Q(i\omega)\,T_3(i\omega)\right)$ can be approximated with any desired accuracy since the functional

$$[T_1(s) - T_2(s)\,Q(s)\,T_3(s)]$$

is bounded everywhere in the right semi-plane of $\mathbb{C}$. In view of this,

$$\begin{aligned}
&\sup_{\Xi:\|\Xi\|^2_{\mathbb{L}_2^{k_1}}\le C} \left\|[T_1(s) - T_2(s)\,Q(s)\,T_3(s)]\,\Xi(s)\right\|^2_{\mathbb{L}_2^{k_1}} \\
&= \underset{\omega\in(-\infty,\infty)}{ess\sup}\ \bar{\sigma}\left(T_1(i\omega) - T_2(i\omega)\,Q(i\omega)\,T_3(i\omega)\right) \sup_{\Xi:\|\Xi\|^2_{\mathbb{L}_2^{k_1}}\le C} \|\Xi(s)\|^2_{\mathbb{L}_2^{k_1}} \\
&= C\,\|T_1(s) - T_2(s)\,Q(s)\,T_3(s)\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}}
\end{aligned} \tag{23.100}$$

which completes the proof. □

Inner–outer factorization

Definition 23.6. *A transfer $(m \times k)$ matrix $N(s)$ is called*

1. ***inner*** *if for all $s \in \mathbb{C}$*

$$\boxed{N(s) \in \mathbb{RH}_\infty^{m\times k}} \tag{23.101}$$

i.e., $N(s)$ is ***stable****, and*

$$\boxed{N^{\sim}(s)\,N(s) = N^{\mathsf{T}}(-s)\,N(s) = I_{k\times k}} \tag{23.102}$$

2. ***co-inner*** *if for all $s \in \mathbb{C}$*

$$\boxed{N(s) \in \mathbb{RH}_\infty^{m\times k}} \tag{23.103}$$

i.e., $N(s)$ is ***stable****, and*

$$\boxed{N(s)\,N^{\sim}(s) = N(s)\,N^{\mathsf{T}}(-s) = I_{m\times m}} \tag{23.104}$$

Remark 23.6. *Observe that for $N(s)$ to be* ***inner*** *it must be "tall", i.e., the number of rows should be more or equal to the number of columns. Sure that by duality, for $N(s)$ to be* ***co-inner*** *(or, equivalently, for $N^{\mathsf{T}}(-s)$ to be inner) the number of columns should be more or equal to the number of rows.*

Example 23.8. *In the scalar case $(m = k = 1)$ the following functions are inners:*

1, $\dfrac{a-s}{a+s}$, $\dfrac{1-as+bs^2}{1+as+bs^2}$ *(a is any real number)*

Proposition 23.6. *If* $N(s)$ *is inner and* $F(s) \in \mathbb{L}_2^k$ *then*

$$\boxed{\|NF\|_{\mathbb{L}_2^k} = \|F\|_{\mathbb{L}_2^k}} \tag{23.105}$$

that is, the multiplication of any vector $F(s) \in \mathbb{L}_2^k$ *by an inner* N *does not change the* $\mathbb{L}_2^k$ *norm of* F. *Analogously, if* $\tilde{N}(s)$ *is co-inner and* $F(s) \in \mathbb{L}_2^k$ *then*

$$\boxed{\left\|F\tilde{N}\right\|_{\mathbb{L}_2^k} = \|F\|_{\mathbb{L}_2^k}} \tag{23.106}$$

Proof. Indeed,

$$\begin{aligned}\|NF\|_{\mathbb{L}_2^k}^2 &:= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{N(j\omega)F(j\omega)F^{\sim}(j\omega)N^{\sim}(j\omega)\right\}d\omega \\ &= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{N^{\sim}(j\omega)N(j\omega)F(j\omega)F^{\sim}(j\omega)\right\}d\omega \\ &= \frac{1}{2\pi}\int_{\omega=-\infty}^{\infty} \operatorname{tr}\left\{F(j\omega)F^{\sim}(j\omega)\right\}d\omega = \|F\|_{\mathbb{L}_2^k}^2\end{aligned}$$

For the co-inner case the proof is similar. □

The next lemma presents the state-space characterization of inner transfer functions.

Lemma 23.9. (Zhou *et al.* 1996) *Suppose that a transfer matrix* $N(s) \in \mathbb{RH}_\infty^{m\times k}$ *has a state-space realization* $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *and* $X = X^\top \geq 0$ *satisfies the matrix Lyapunov equation*

$$\boxed{A^\top X + X^\top A + C^\top C = 0} \tag{23.107}$$

Then the following properties hold:
(a) the relation

$$\boxed{D^\top C + B^\top X = 0} \tag{23.108}$$

implies

$$\boxed{N^{\sim}(s)N(s) = D^\top D} \tag{23.109}$$

(b) if the pair (A, B) *is controllable and (23.109) holds then (23.108) holds too.*

Proof. Notice that the state-space realization of $N^{\sim}(s)\,N(s)$ is

$$\begin{bmatrix} A & 0 & B \\ -C^{\intercal}C & -A^{\intercal} & -C^{\intercal}D \\ D^{\intercal}C & B^{\intercal} & D^{\intercal}D \end{bmatrix} \tag{23.110}$$

Define the matrix

$$T := \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix}$$

which is nonsingular for any $X \geq 0$. Notice that

$$T^{-1} = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}$$

Then, the application of this similarity state transformation (which does not change the corresponding transfer matrix) to the state vector leads to the following state-space realization and corresponds to the multiplication of (23.110) by T and post-multiplication by T^{-1} yields that $N^{\sim}(s)\,N(s)$ also has the state-space realization

$$\begin{bmatrix} A & 0 & B \\ -(A^{\intercal}X + X^{\intercal}A + C^{\intercal}C) & -A^{\intercal} & -(XB^{\intercal} + C^{\intercal}D) \\ B^{\intercal}X + D^{\intercal}C & B^{\intercal} & D^{\intercal}D \end{bmatrix}$$

$$= \begin{bmatrix} A & 0 & B \\ 0 & -A^{\intercal} & -(XB^{\intercal} + C^{\intercal}D) \\ B^{\intercal}X + D^{\intercal}C & B^{\intercal} & D^{\intercal}D \end{bmatrix}$$

$$= \begin{bmatrix} B^{\intercal}X + D^{\intercal}C & B^{\intercal} \end{bmatrix} \left[\left(sI - \begin{matrix} A & 0 \\ 0 & -A^{\intercal} \end{matrix} \right) \right]^{-1} \begin{pmatrix} B \\ -(XB^{\intercal} + C^{\intercal}D) \end{pmatrix}$$

$$\begin{bmatrix} B^{\intercal}X + D^{\intercal}C & B^{\intercal} \end{bmatrix} \begin{pmatrix} (sI - A)^{-1} & 0 \\ 0' & (sI + A^{\intercal})^{-1} \end{pmatrix}$$

$$\times \begin{pmatrix} B \\ -(XB^{\intercal} + C^{\intercal}D) \end{pmatrix} + D^{\intercal}D$$

Then (a) and (b) follow easily. □

One can see that adding the simple condition $D^{\intercal}D = I$ provides that $N^{\sim}(s)\,N(s) = I$. The next corollary of the theorem above evidently states the more exact result.

Corollary 23.8. *Suppose $N(s)$ has a state-space realization $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ which is minimal and stable (A is Hurwitz). Let G_o be the corresponding observability grammian. Then $N(s)$ is an inner if and only if*

1.

$$B^\intercal G_o + D^\intercal C = 0$$

2.

$$D^\intercal D = I \tag{23.111}$$

Definition 23.7. *We say that for* $G(s) \in \mathbb{LH}_\infty^{m\times k}$ *there exist*
(a) a ***right co-prime factorization*** *(RCF) if*

$$\boxed{G(s) = N(s)\,M^{-1}(s)} \tag{23.112}$$

where $N(s) \in \mathbb{RH}_\infty^{m\times p}$ *is an inner and* $M(s) \in \mathbb{RH}_\infty^{k\times p}$ *(i.e.,* $M(s)$ *is stable);*
(b) a ***left co-prime factorization*** *(LCF) if*

$$\boxed{G(s) = \tilde{M}^{-1}(s)\,\tilde{N}(s)} \tag{23.113}$$

where $\tilde{N}(s) \in \mathbb{RH}_\infty^{p\times k}$ *is a co-inner and* $\tilde{M}(s) \in \mathbb{RH}_\infty^{p\times m}$ *(i.e.,* $\tilde{M}(s)$ *is stable)*

Remark 23.7. *It is not difficult to show (see Zhou et al. (1996)) that*
(a) two matrices $N(s)$ *and* $M(s)$ *in (23.112) are right-co-prime over* $\mathbb{RH}_\infty$ *if they have the same number of columns and there exist matrices* X *and* Y *(of the corresponding size) in* $\mathbb{RH}_\infty$ *such that*

$$\boxed{\begin{bmatrix} X & Y \end{bmatrix}\begin{bmatrix} M \\ N \end{bmatrix} = XM + YN = I} \tag{23.114}$$

(b) two matrices $\tilde{M}(s)$ *and* $\tilde{N}(s)$ *in (23.113) are left-co-prime over* $\mathbb{RH}_\infty$ *if they have the same number of rows and there exist matrices* $\tilde{X}$ *and* $\tilde{Y}$ *(of the corresponding size) in* $\mathbb{RH}_\infty$ *such that*

$$\boxed{\begin{bmatrix} \tilde{M} & \tilde{N} \end{bmatrix}\begin{bmatrix} \tilde{X} \\ \tilde{Y} \end{bmatrix} = \tilde{M}\tilde{X} + \tilde{N}\tilde{Y} = I} \tag{23.115}$$

The relations (23.114) and (23.115) are often called the ***Bezout identities****.*

The next theorems show when such factorizations exist.

Theorem 23.9. (Zhou *et al.* 1996) *Suppose* $G(s) \in \mathbb{LH}_\infty^{m\times k}$ *and* $m \geq k$*. Then there exists an RCF* $G(s) = N(s)\,M^{-1}(s)$ *such that* $N(s)$ *is an inner if and only if*

$$\boxed{\tilde{G}(i\omega)\,G(i\omega) = G^\intercal(-i\omega)\,G(i\omega) > 0} \tag{23.116}$$

for all $\omega \in [-\infty, \infty]$ *(including at* $\pm\infty$*). Furthermore, if* $G(s) \in \mathbb{RH}_\infty^{m\times k}$ *and its state-space realization* $\begin{bmatrix} A & B \\ C & D \end{bmatrix}$ *has the* ***stabilizable pair*** (A, B) *such that the transfer matrix* $G(i\omega) = C(i\omega - A)^{-1}B + D$ *has the full column rank for all* $\omega \in [-\infty, \infty]$*, then a particular state-space realization of the desired RCF* $\begin{pmatrix} M(s) \\ N(s) \end{pmatrix} \in \mathbb{RH}_\infty^{k\times p+m\times p}$ *is*

$$\boxed{\begin{bmatrix} A+BF & BR^{-1/2} \\ F & R^{-1/2} \\ C+DF & DR^{-1/2} \end{bmatrix}} \tag{23.117}$$

where

$$\boxed{\begin{aligned} &R = D^\top D > 0 \\ &F = -R^{-1}(B^\top X + D^\top C) \end{aligned}} \tag{23.118}$$

and $X \geq 0$ *is a solution of the following Riccati equation*

$$\boxed{\begin{aligned} &(A - BR^{-1}D^\top C)X + X(A - BR^{-1}D^\top C)^\top \\ &\quad - XBR^{-1}B^\top X - C^\top(I - DR^{-1}D^\top)C = 0 \end{aligned}} \tag{23.119}$$

Proof. (a) *Necessity.* Suppose $G = NM^{-1}$ with N an inner and $M(s) \in \mathbb{RH}_\infty^{k\times p}$. Then

$$\begin{aligned} \tilde{G}(i\omega)G(i\omega) &= \left(\tilde{M^{-1}}(i\omega)\right)\tilde{N}(i\omega)N(i\omega)M^{-1}(i\omega) \\ &= \left(\tilde{M^{-1}}(i\omega)\right)M^{-1}(i\omega) > 0 \end{aligned}$$

for any $\omega \in [-\infty, \infty]$ since $M(s) \in \mathbb{RH}_\infty^{k\times p}$, that is, $M(s)$ is "stable".

(b) *Sufficiency.* First notice that if $G = NM^{-1}$ is an RCF, then $G = (NZ)(MZ)^{-1}$ is also RCF for any nonsingular matrix $Z \in \mathbb{R}^{p\times p}$. Suppose now that N has its state-space realization as

$$\begin{bmatrix} A+BF & BZ \\ C+DF & DZ \end{bmatrix}$$

For N to be an inner, as it follows from Corollary 23.8, we should have

$$(DZ)^\top DZ = I$$

Select F in such a manner that the following matrix identities have been fulfilled:

$$\begin{aligned} &(BZ)^\top X + (DZ)^\top(C + DF) = 0 \\ &(A+BF)^\top X + X(A+BF) + (C+DF)^\top(C+DF) = 0 \end{aligned} \tag{23.120}$$

Take $Z = R^{-1/2}$ where $R := D^\top D > 0$ by the assumption of the theorem. Then the first equation in (23.120) becomes

$$\left(BR^{-1/2}\right)^\top X + \left(DR^{-1/2}\right)^\top (C + DF)$$

$$= R^{-1/2}B^\top X + R^{-1/2}D^\top (C + DF) = 0$$

which implies

$$R^{-1/2}D^\top DF = -R^{-1/2}(B^\top X + D^\top C)$$

Then, pre-multiplication of the last equation by $R^{-1/2}$ gives

$$F = -R^{-1}(B^\top X + D^\top C)$$

Substituting this F into the second equation in (23.120) gives exactly (23.119). The existence of a nonnegative solution $X \geq 0$ for (23.119) guarantees the fulfilling of (23.116).[1] □

Remark 23.8. *Notice that if* $G(s) = N(s)M^{-1}(s) \in \mathbb{RH}_\infty^{m\times k}$, *then* $M^{-1}(s) \in \mathbb{RH}_\infty^{p\times k}$ *and* $M^{-1}(s)$ *is called an* ***outer****, so*

$$\boxed{G(s) = G_{inn}(s)\,G_{out}(s)} \tag{23.121}$$

where

$$\boxed{G_{inn}(s) := N(s), \quad G_{out}(s) = M^{-1}(s)}$$

The factorization (23.121) is referred to as ***an inner–outer factorization****. In an analogous way the* ***co-outer*** *can be introduced such that the following* ***co-inner–co-outer factorization*** *takes place:*

$$\boxed{\begin{array}{l} G(s) = \tilde{M}^{-1}(s)\,\tilde{N}(s) = \tilde{G}_{out}(s)\,\tilde{G}_{inn}(s) \\ \tilde{G}_{inn}(s) := \tilde{N}(s), \quad \tilde{G}_{out}(s) = \tilde{M}^{-1}(s) \end{array}} \tag{23.122}$$

Remark 23.9. *The analogue result is valid for LCF (23.113), namely, supposing* $G(s) \in \mathbb{LH}_\infty^{m\times k}$ *and* $m \leq k$, *one concludes that there exists an RCF* $G(s) = \tilde{M}^{-1}(s)\,\tilde{N}(s)$ *where* $\tilde{N}(s) \in \mathbb{RH}_\infty^{p\times k}$ *is a co-inner and* $\tilde{M}(s) \in \mathbb{RH}_\infty^{p\times m}$ *(i.e.,* $\tilde{M}(s)$ *is stable) if and only if*

$$\boxed{G(i\omega)\,\tilde{G}(i\omega) = G(i\omega)\,G^\top(-i\omega) > 0} \tag{23.123}$$

for all $\omega \in [-\infty, \infty]$ *(including at* $\pm\infty$*).*

[1] For details see Theorem 13.19 in Zhou *et al.* (1996).

Lemma 23.10. *For any* $F(s) \in \mathbb{LH}_\infty^{m\times k}$ *and any inner* $N(s) \in \mathbb{RH}_\infty^{p\times m}$ *(or any co-inner* $\tilde{N}(s) \in \mathbb{RH}_\infty^{k\times p}$*)*

$$\boxed{\|F\|_{\mathbb{L}_\infty^{m\times k}} = \|N(s)F\|_{\mathbb{L}_\infty^{m\times k}} = \left\|F\tilde{N}(s)\right\|_{\mathbb{L}_\infty^{m\times k}}} \tag{23.124}$$

that is, the pre-multiplication by an inner or the post-multiplication by a co-inner preserves the $\mathbb{L}_\infty$ *norm of any transfer matrix.*

Proof. It follows directly from the definition (23.44) since

$$\|NF\|_{\mathbb{L}_\infty^{m\times k}} := \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \bar{\sigma}\left(N(i\omega)F(i\omega)\right)$$

$$= \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}^{1/2}\left\{F^{\sim}(i\omega)N^{\sim}(i\omega)N(i\omega)F(i\omega)\right\}$$

$$= \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}^{1/2}\left\{F^{\sim}(i\omega)F(i\omega)\right\} = \|F\|_{\mathbb{L}_\infty^{m\times k}}$$

and

$$\left\|F\tilde{N}\right\|_{\mathbb{L}_\infty^{m\times k}} := \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \bar{\sigma}\left(F(i\omega)\tilde{N}(i\omega)\right)$$

$$= \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}^{1/2}\left\{F(i\omega)\tilde{N}(i\omega)\tilde{N}^{\sim}(i\omega)F^{\sim}(i\omega)\right\}$$

$$= \operatorname*{ess\,sup}_{\omega\in(-\infty,\infty)} \lambda_{\max}^{1/2}\left\{F(i\omega)F^{\sim}(i\omega)\right\} = \|F\|_{\mathbb{L}_\infty^{m\times k}}$$

□

Now we are ready to give the solution to the MMP problem (23.98) showing its equivalence to the Nehari problem (23.95).

Theorem 23.10. *The MMP problem (23.98) is equivalent to the Nehari problem (23.95) with*

$$\boxed{\begin{array}{c} X(s) = M_2^{-1}(s)\,Q(s)\,\tilde{M}_3^{-1}(s) \in \mathbb{RH}_\infty^{p_2\times p_1} \\ \\ G(s) = N_2^{\sim}(s)\,T_1(s)\,\tilde{N}_3^{\sim}(s) \end{array}} \tag{23.125}$$

and one of its solutions is

$$X(s) = M_2(s)\left[\tilde{N_2}(-s)\,T_1(-s)\,\tilde{N_3}(-s) - \sqrt{\lambda_{\max}(G_c G_o)}\,\frac{\mathsf{f}(-s)\tilde{\mathsf{g}}(-s)}{\tilde{\mathsf{g}}(-s)\mathsf{g}(-s)}\right]\tilde{M}_3(s) \tag{23.126}$$

where $\mathsf{f}(s)$ *and* $\tilde{\mathsf{g}}(s)$ *is the Schmidt pair (23.78) for* $\tilde{N_2}(s)\,T_1(s)\,\tilde{N_3}(s)$ *and* G_c, G_o *its grammians (23.68).*

Proof. Using RCF and LCF for $T_2(s) = N_2(s)\,M_2^{-1}(s)$ and for $T_3(s) = \tilde{M}_3^{-1}(s)\,\tilde{N}_3(s)$, respectively, and applying the property (23.124), we get

$$\begin{aligned} J(Q) &= C\,\|T_1(s) - T_2(s)\,Q(s)\,T_3(s)\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}} \\ &= C\left\|T_1(s) - N_2(s)\,M_2^{-1}(s)\,Q(s)\,\tilde{M}_3^{-1}(s)\,\tilde{N}_3(s)\right\|^2_{\mathbb{RH}_\infty^{m_1\times k_1}} \\ &= C\left\|\tilde{N_2}(s)\,T_1(s)\,\tilde{N_3}(s) - M_2^{-1}(s)\,Q(s)\,\tilde{M}_3^{-1}(s)\right\|^2_{\mathbb{RH}_\infty^{p_2\times p_1}} \end{aligned}$$

which is exactly (23.95) with (23.125). The solution (23.126) results from (23.126). □

23.2.5 *Some control problems converted to MMP*

$\mathbb{H}_\infty$–control robust with respect to external perturbations–Problem formulation

Problem 23.4. *For the linear system given in the frequency domain by Fig. 23.5*

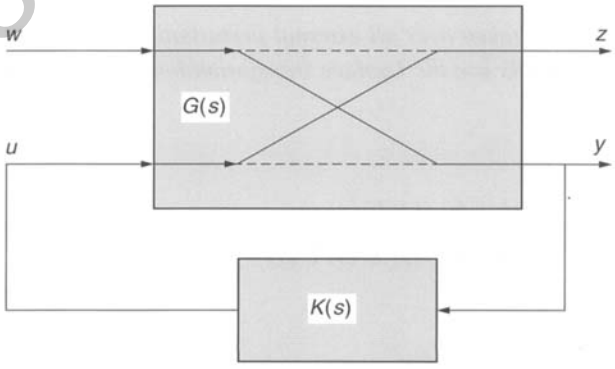

Fig. 23.5. The block scheme of a linear system with an external input perturbation and an internal feedback.

$$\boxed{\begin{array}{l} G(s) := \begin{bmatrix} G_{wz}(s) & G_{uz}(s) \\ G_{wy}(s) & G_{uy}(s) \end{bmatrix} \\ \\ \begin{pmatrix} Z(s) \\ Y(s) \end{pmatrix} = G(s) \begin{pmatrix} W(s) \\ U(s) \end{pmatrix} \\ \\ U(s) := \mathcal{L}\{u\}, \quad Y(s) := \mathcal{L}\{y\} \\ Z(s) := \mathcal{L}\{z\}, \quad W(s) := \mathcal{L}\{w\} \end{array}} \tag{23.127}$$

($z(t) \in \mathbb{R}^n$ is associated with the part of coordinates forming a cost function, $u(t) \in \mathbb{R}^r$ is the controlled input, $y(t) \in \mathbb{R}^p$ is the system output which may be used for feedback designing and $w(t) \in \mathbb{R}^p$ is an external perturbation) ***design a feedback (dynamic) controller*** *with a transfer matrix $K(s)$ given by*

$$\boxed{U(s) = K(s)[Y(s) + W(s)]} \tag{23.128}$$

such that

1. *$K(s)$ is proper and stable, i.e.,*

$$\boxed{K(s) \in \mathbb{RH}_\infty^{r \times p}} \tag{23.129}$$

2. *it minimizes the following min–max criterion*

$$\boxed{J(K) \to \min_{K(s) \in \mathbb{RH}_\infty^{r \times p}}} \tag{23.130}$$

$$J(K) := \sup_{W \in \mathbb{L}_2^n : \|W\|_{\mathbb{L}_2^n}^2 \le c_w} \|Z\|_{\mathbb{H}_2^{n+r}}^2$$

where supremum is taken over all external perturbations of a bounded energy. Here the large symbols are the Laplace transformation of the small ones.

Notice that

$$\begin{aligned} Z(s) &= G_{wz}(s) W(s) + G_{uz}(s) U(s) \\ &= G_{wz}(s) W(s) + G_{uz}(s) K(s) Y(s) \end{aligned}$$

and

$$\begin{aligned} Y(s) &= G_{wy}(s) W(s) + G_{uy}(s) U(s) \\ &= G_{wy}(s) W(s) + G_{uy}(s) K(s) Y(s) \end{aligned}$$

which leads to the following relations

$$Y(s) = [I - G_{uy}(s) K(s)]^{-1} G_{wy}(s) W(s)$$

$$Z(s) = \left(G_{wy}(s) + G_{uy}(s) K(s) [I - G_{uy}(s) K(s)]^{-1} G_{wy}(s)\right) W(s)$$

$$U(s) = K(s) [I - G_{uy}(s) K(s)]^{-1} G_{wy}(s) W(s)$$

Remark 23.10. *1. The cost functional* $J(K)$ *(23.130) exists (finite) if and only if the closed-loop system is stable, that is, when* $Z(s) \in \mathbb{L}_2^{n+r}$ *and* $Y(s) \in \mathbb{L}_2^p$ *if* $W(s) \in \mathbb{L}_2^n$.

2. By Lemma 23.4 it is possible if and only if the transfer functions from $W(s)$ *to* $Z(s)$ *and* $Y(s)$ *are stable, namely when*

$$G_{wy}(s) + G_{uy}(s) K(s) [I - G_{uy}(s) K(s)]^{-1} G_{wy}(s) \in \mathbb{RH}_\infty^{(n+r)\times n}$$

$$[I - G_{uy}(s) K(s)]^{-1} G_{wy}(s) \in \mathbb{RH}_\infty^{p\times n}$$

Parametrization of all stabilizing stationary feedbacks

Problem 23.5. *Characterize the set of* ***all stabilizing feedbacks*** $K(s)$ *which provides the stability for* $K(s)$ *itself as well as for* $[I - G_{uy}(s) K(s)]^{-1}$ *supposing that the transfer matrices*

$$G_{wy}(s) \text{ and } G_{uy}(s)$$

are ***stable****, that is, we will try to find*

$$\boxed{K(s) \in \mathbb{RH}_\infty^{r\times p}}$$

such that

$$\boxed{[I - G_{uy}(s) K(s)]^{-1} \in \mathbb{RH}_\infty^{p\times p}} \tag{23.131}$$

providing

$$G_{wy}(s) \in \mathbb{RH}_\infty^{p\times p}, \quad G_{uy}(s) \in \mathbb{RH}_\infty^{p\times r}$$

Proposition 23.7. $K(s)$ *stabilizes* $G(s)$ *if and only if it stabilizes* $G_{uy}(s)$.

Proof. It follows directly from (23.131). □

Theorem 23.11. (Youla parametrization, 1961) *The set of all (proper real rational) transfer matrices $K(s)$ stabilizing $G_{uy}(s)$ is parametrized by the formula*

$$\boxed{\begin{aligned} K(s) &= [Y(s) - M(s)Q(s)][X(s) - N(s)Q(s)]^{-1} \\ &= \left[\tilde{X}(s) - Q(s)\tilde{N}(s)\right]^{-1}\left[\tilde{Y}(s) - Q(s)\tilde{M}(s)\right] \end{aligned}} \tag{23.132}$$

where the matrix $Q(s) \in \mathbb{RH}_\infty$ (provided that two inverted matrices exist) and all other matrices are also from $\mathbb{RH}_\infty$ and define the, so-called, ***double-co-prime factorization*** *of $G_{uy}(s)$, namely,*

$$\boxed{\begin{aligned} &G_{uy}(s) = N(s)M^{-1}(s) = \tilde{M}^{-1}(s)\tilde{N}(s) \\ &\begin{bmatrix} \tilde{X}(s) & -\tilde{Y}(s) \\ -\tilde{N}(s) & \tilde{M}(s) \end{bmatrix}\begin{bmatrix} M(s) & Y(s) \\ N(s) & X(s) \end{bmatrix} = I \end{aligned}} \tag{23.133}$$

Proof. It can be found in Francis (1987) as well as in Zhou *et al.* (1996). □

Corollary 23.9. *If $K(s)$ is as in (23.132), then*

$$\boxed{\begin{aligned} [I - G_{uy}(s)K(s)]^{-1} &= M(s)\left(\tilde{X}(s) - Q(s)\tilde{N}(s)\right) \\ K(s)[I - G_{uy}(s)K(s)]^{-1} &= N(s)\left(\tilde{X}(s) - Q(s)\tilde{N}(s)\right) \end{aligned}} \tag{23.134}$$

Proof. Substituting $G_{uy}(s) = N(s)M^{-1}(s) = \tilde{M}^{-1}(s)\tilde{N}(s)$ and using the second identity in (23.133) imply the desired relations (23.134). □

The solution of $\mathbb{H}_\infty$ – linear robust control problem

Theorem 23.12. *The problem 23.4 is equivalent to the model matching problem (23.98) with*

$$\boxed{\begin{aligned} T_1(s) &= G_{wy}(s) + G_{uy}(s)N(s)\tilde{X}(s)G_{wy}(s) \\ T_2(s) &= G_{uy}(s)N(s) \\ T_3(s) &= \tilde{N}(s)G_{wy}(s) \end{aligned}} \tag{23.135}$$

Proof. The cost functional (23.130) can be represented as

$$J(K) := \sup_{W \in \mathbb{L}_2^n : \|W\|_{\mathbb{L}_2^n}^2 \le c_w} \|Z\|_{\mathbb{H}_2^{n+r}}^2 = \sup_{W \in \mathbb{L}_2^n : \|W\|_{\mathbb{L}_2^n}^2 \le c_w} \|TW\|_{\mathbb{H}_2^{n+r}}^2$$

where (omitting the argument)

$$T = G_{wy} + G_{uy} K \left[I - G_{uy} K \right]^{-1} G_{wy}$$

which in view of (23.134) becomes

$$T = G_{wy} + G_{uy} N \left(\tilde{X} - Q \tilde{N} \right) G_{wy}$$

$$= \left[G_{wy} + G_{uy} N \tilde{X} G_{wy} \right] - \left[G_{uy} N \right] Q \left[\tilde{N} G_{wy} \right]$$

which completes the proof. □

Robust filtering problem: consider a linear system which block scheme is given at Fig. 23.6.

Problem 23.6. *Find a filter* $F(s) \in \mathbb{RH}_\infty^{m \times n}$ *if it exists which minimizes*

$$\boxed{J(F) := \sup_{w : \|w\|^2_{L_2^k[0,\infty)} \le c} \frac{\left\| x - \hat{x} \right\|^2_{L_2^n[0,\infty)}}{\|w\|^2_{L_2^k[0,\infty)}}} \tag{23.136}$$

Here $w(t) \in L_2^k[0,\infty)$ *is an external noise,* $x(t) \in L_2^n[0,\infty)$ *is a state vector,* $y(t) \in L_2^m[0,\infty)$ *is an output of the system and* $\hat{x}(t) \in L_2^n[0,\infty)$ *is a state estimate. For physical reasons it is supposed that* ***all transfer matrices are stable****, that is,*

$$\boxed{G(s) \in \mathbb{RH}_\infty^{k \times n}, \quad F(s) \in \mathbb{RH}_\infty^{n \times m}, \quad H(s) \in \mathbb{RH}_\infty^{n \times m}}$$

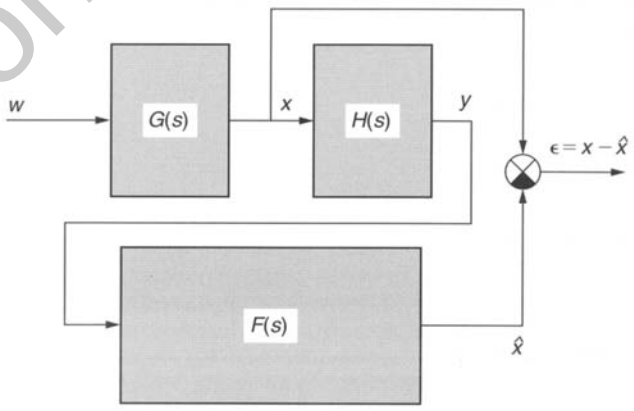

Fig. 23.6. Filtering problem illustration.

As before, in view of (17.107) and (23.100), we have

$$J(F) = \sup_{W:\|W\|^2_{\mathbb{L}_2^k}\le c} \frac{\left\|X(s) - \hat{X}(s)\right\|^2_{\mathbb{L}_2^n}}{\|W\|^2_{\mathbb{L}_2^k}}$$

$$= \sup_{W:\|W\|^2_{\mathbb{L}_2^k}\le c} \|X(s) - F(s)\,Y(s)\|^2_{\mathbb{L}_2^n}$$

$$= \sup_{W:\|W\|^2_{\mathbb{L}_2^k}\le c} \left\|[G(s) - F(s)\,H(s)]\,W(s)\right\|^2_{\mathbb{L}_2^n}$$

$$= c\,\|G(s) - F(s)\,H(s)\|^2_{\mathbb{RH}_\infty^{k\times n}}$$

Using the left-co-prime factorization (23.113)

$$\boxed{H(s) = \tilde{M}^{-1}(s)\tilde{N}(s)}$$

and applying the property (23.106), we derive

$$J(F) = c\left\|G(s) - F(s)\,\tilde{M}^{-1}(s)\tilde{N}(s)\right\|^2_{\mathbb{RH}_\infty^{k\times n}}$$

$$= c\left\|\left[G(s) - F(s)\,\tilde{M}^{-1}(s)\tilde{N}(s)\right]\tilde{N}^{\sim}(s)\right\|^2_{\mathbb{RH}_\infty^{k\times n}} \tag{23.137}$$

$$= c\left\|G(s)\,\tilde{N}^{\sim}(s) - F(s)\,\tilde{M}^{-1}(s)\right\|^2_{\mathbb{RH}_\infty^{k\times n}}$$

which leads to the following result.

Theorem 23.13. *The* ***robust filtering problem*** *(23.6) is equivalent to the model matching problem (23.98) with*

$$\boxed{\begin{array}{c} T_1(s) = G(s)\,\tilde{N}^{\sim}(s) \\ T_2(s) = I,\ Q(s) = F(s) \\ T_3(s) = \tilde{M}^{-1}(s) \end{array}} \tag{23.138}$$

Proof. It results from (23.137). □

BIBLIOGRAPHY

Adamjan, V., Arov, D. & Krein, M. (1971), "Analytic properties of Schmidt pairs for a Hankel operator and the generalized Schur-takagi problem", *Math. USSR Sbornik* **15**, 31–73 (in Russian).

Albert, A. (1972), *Regression and Moore-Penrose Pseudoinverse*, Vol. 94 of *Mathematics in Science and Engineering*, Academic Press, New York, London.

Alexeev, V., Galeev, E. & Tihomirov, V. (1984), *Collection of Problems on Optimization*, Nauka, Moscow (in Russian).

Alexeev, V., Tikhomirov, V. & Fomin, S. (1979), *Optimal Control*, Nauka, Moscow (in Russian).

Antosiewicz, H. (1958), "A survey of Lyapunov's second method", *Control of Nonlinear Oscillations* **4**, 141–166.

Apostol, T. (1974), *Mathematical Analysis*, Addison-Wesley Series in Mathematics, 2nd edn, Addison-Wesley, Massachusetts.

Arrow, K. J., Hurwics, L. & Uzawa, H. (1958), *Studies in Linear and Non-linear Programming*, Stanford University Press.

Aubin, J. (1979), *Mathematical Methods of Game and Economic Theory*, North-Holland Publishing Company.

Barbashin, E. & Krasovskii, N. (1952), "On asymptotic global stability of motions", *Doklady Academy of Sciences, USSR* **86**(3), 453–458 (in Russian).

Bellman, R. (1957), *Dynamic Programming*, Princeton University Press, Princeton, New Jersey.

Bellman, R. (1960), *Introduction to Matrix Analysis*, McGraw-Hill Book Company, Inc., New York, USA.

Bihari, I. (1956), "A generalization of a lemma of Bellman and its applications to uniqueness problems of differential equations", *Acta Math. Sci. Hungar.* **4**(7), 71–94.

Boltyanski, V., Gamkrelidze, R. & Pontryagin, L. (1956), "On the theory of optimal processes", *Doklady AN USSR* **110**(1), 7–10 (in Russian).

Boyd, S., El Chaoui, L., Feron, E. & Balakrishnan, V. (1994), *Linear Matrix Inequalities in System and Control Theory*, Vol. 15 of *SIAM Studies in Applied Mathematics*, SIAM, Philadelphia, Pennsylvania, USA.

Chetaev, N. (1965), *Stability of Motions*, Nauka, Moscow (in Russian).

Chung, K. (1954), "On stochastic approximation method", *Annals of Mathematical Statistics* **25**(3), 463–483.

Cline, R. (1964), "Representations for the generalized inverse of a partitioned matrix", *SIAM J. Applied Mathematics* **12**, 588–600.

Cline, R. (1965), "Representations for the generalized inverse of sums of matrices", *SIAM J. Applied Mathematics* **2**, 99–114.

Crandall, M. & Lions, P. (1983), "Viscosity solution of Hamilton–Jacobi equations", *Trans. Amer. Math. Soc.* **277**, 1–42.

Curtain, R. & Zwart, H. (1995), *An Introduction to Infinite-Dimensional Linear System Theory*, Vol. 21 of *Texts in Applied Mathematics*, Springer-Verlag, New York.

Daneš, J. (1972), "A geometric theorem useful in nonlinear functional analysis", *Bulletino U.M.I.* **6**(4), MR 47:5678, 369–375.

Datta, B. N. (2004), *Numerical Methods for Linear Control Systems (Design and Analysis)*, Elsevier Inc. (Academic Press), San Diego, California, USA.

Dax, A. (1997), "An elementary proof of Farkaš lemma", *SIAM Review* **39**(3), 503–507.

El'sgol'ts, L. (1961), *Differential Equations*, International Monographs on Advanced Mathematics and Physics (translated from Russian edition, 1957), Hindustan Publishing Corporation (India), Delhi.

Farkaš, J. (1902), "Uber die Theorie der einfachen Ungleichungen", *J. Reine Angew. Math.* **124**, 1–24.

Feldbaum, A. (1953), "Optimal processes in systems of automatic control", *Avtomatika i Telemehanika* **14**(6), 712–728 (in Russian).

Filippov, A. (1988), *Differential Equations with Discontinuous Right-hand Sides*, Kluwer Academic Publishers, Dordrecht.

Finsler, P. (1937), "Uber das vorkommen definiter und semi-definiter formen in scharen quadratischer formen", *Comentarii Mathematici Helvetici* **9**, 192–199.

Fuchs, B. & Shabat, B. (1964), *Functions of a Complex Analysis*, Pergamon Press (translated from Russian, Fizmatgiz, Moscow, 1959), Oxford.

Gantmacher, F. (1990), *The Theory of Matrices*, 4th edn, Chelsea Publishing Company, New York.

Gel'fand, I. & Fomin, S. (1961), *Variation Calculus*, Fizmatgiz, Moscow, Russia (in Russian).

Gelig, A., Leonov, G. & Yakubovich, V. (1978), *Stability of Nonlinear Systems with Nonunique Steady State*, Nauka, Moscow, Russia (in Russian).

Greville, T. (1960), "Some applications of the pseudoinverse of a matrix", *SIAM Review* **2**, 15–22.

Gronwall, T. (1919), "Note on the derivatives with respect to a parameter of the solutions of a system of differential equations", *Ann. Math.* **2**(20), 292–296.

Guillemin, E. (1949), *The Mathematics of Circuit Analysis: Extensions to the Mathematical Training of Electrical Engineers*, John Wiley and Sons, New York.

Halanay, A. (1966), *Differential Equations: Stability. Oscillations. Time Lag*, Academic Press, New York.

Hartman, P. (2002), *Ordinary Differential Equations*, Vol. 38 of *Classics in Applied Mathematics*, 2nd edn, SIAM, Philadelphia.

Hautus, M. & Silverman, L. (1983), "System structure and singular control", *System Structure and Singular Control* **50**, 369–402.

Highan, N. (1996), *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia.

Hinrichsen, D. & Pritchard, A. (2005), *Mathematical Systems Theory I: Modelling, State Space Analysis, Stability and Robustness*, Vol. 48 of *Texts in Applied Mathematic*, Springer, Berlin.

Ivanov, V. & Faldin, N. (1981), *Theory of Optimal Automatic Control Systems*, Nauka, Moscow, Russia (in Russian).

Kharitonov, V. (1978), "Asymptotic stability of a family of linear differential equations", *Differential Equations* **14**(11), 2086–2088 (in Russian).

Krasovskii, N. (1952), "On stability of dynamic motions", *Applied Mathematics and Mechanics* **16**(5), 3–24 (in Russian).

Kuhn, H. & Tucker, A. (1951), "Nonlinear programming", *in* University of California Press, ed., *Proceedings of the Second Berkeley Symposium on Math. Statistics and Probability*, Berkeley, CA, pp. 481–492.

Lancaster, P. (1969), *Theory of Matrices*, Academic Press, New York.

Lankaster, P. & Tismenetsky, M. (1985), *The Theory of Matrices*, Computer Science and Applied Mathematics, 2nd edn, Academic Press, Inc., Orlando, Florida, 32887.

Lavrentiev, M. & Shabat, B. (1987), *Methods of Theory of Functions of Complex Variables*, Nauka, Moscow, Russia (in Russian).

Lefschetz, S. (1965), *Stability of Nonlinear Control Systems*, Academic Press, New York.

Lurie, A. & Postnikov, V. (1944), "To the theory of control systems", *Problems of Mathematics and Mechanics (PMM)* **8**(3).

Lyapunov, A. (1892), *General Problem of a Movement Stability,* Doctor thesis, Leningrad-Moscow (in Russian) (the original 1897).

Lyapunov, A. (1907), "Problèm Le général de la stabilité de mouvement" (translation of the paper published in *Comm. Soc. Math. Krakow* 1893, reprinted in *Ann. Math. Studies*, 17, Princeton, 1949), *Ann. Fac. Sci., Toulose* **2**, 203–474.

Marcus, M. & Minc, H. (1992), *A Survey of Matrix Theory and Matrix Inequality*, 2nd edn, Dover Publications, Inc., New York.

Nazin, A. & Poznyak, A. (1986), *Adaptive Choice of Variants, Theoretical Backgrounds of Technical Cybernetics*, Nauka, Moscow, Russia (in Russian).

Nehari, Z. (1957), "On bounded bilinear forms", *Annals of Mathematics* **15**(1), 153–162.

Nesterov, Y. & Nemirovsky, A. (1994), *Interior-point Polynomial Methods in Convex Programming*, Vol. 13 of *Studies in Applied Mathematics*, SIAM, Philadelphia, Pennsylvania, USA.

Pliss, V. (1964), *Nonlocal Problems of Oscillation Theory*, Nauka, Moscow-Leningrad (in Russian).

Polyak, B. (1987), *Introduction to Optimization*, Translations series in Mathematics and Engineering, Optimization Software, Inc., Publications Division, New York.

Polyak, B. & Sherbakov, P. (2002), *Robust Stability and Control*, Nauka, Moscow, Russia.

Polyak, B. & Tsypkin, Y. Z. (1990), "Frequency-domain criteria for robust stability and aperiodicity of linear systems", *Automation and Remote Control* **51**(9), 1192–1200. Translated from "Automaticay Telemechanica" (in Russian), 1990, v.9, pp. 45–54.

Pontryagin, L. S., Boltyansky, V. G., Gamkrelidze, R. V. & Mishenko, E. F. (1969 (translated from Russian)), *Mathematical Theory of Optimal Processes*, Interscience, New York.

Popov, V. (1961), "Global asymptotic stability of nonlinear systems of automatic control", *Automation and Remote Control* (Translation from "Automatica i Telemechanica") **22**(8), 961–979.

Poznyak, A. (1991), *Lectures on Basics of Robust Control (H-inf Theory)*, MPhTI (Moscow Physical Technical Institute), Moscow (in Russian).

Rasvan, V. (1975), *Absolute Stability of Time Lag Control Systems*, Ed. Academiei, Bucharest (in Romanian, Russian revised edition by Nauka, Moscow, 1983).

Riesz, F. & Nagy, B. (1978 (original in French, 1955)), *Functional Analysis*, Frederick Ungar, New York.

Rockafellar, R. (1970), *Convex Analysis*, Princeton University Press.

Royden, H. (1968), *Real Analysis*, 2nd edn, MacMillan Publishing Co., Inc., New York.

Rudin, W. (1973), *Real and Complex Analysis*, McGraw-Hill, New York.

Rudin, W. (1976), *Principle of Mathematical Analysis*, International Series in Pure and Applied Mathematics, 3rd edn, McGraw-Hill, New York.

Schlaepfer, F. & Schweppe, F. (1972), "Continuous-time state estimation under disturbances bounded by convex sets", *IEEE Transactions on Automatic Control* **AC-17**(2), 197–205.

Trenogin, V. (1980), *Functional Analysis*, Nauka, Moscow, Russia (in Russian).

Troutman, J. (1996), *Variation Calculus and Optimal Control (Optimization with Elementary Covexity)*, 2nd edn, Springer-Verlag, New York.

Utkin, V. (1992), *Sliding Modes in Control and Optimization*, Springer-Verlag.

Utkin, V., Guldner, J. & Shi, J. (1999), *Sliding Modes in Electromechanical Systems*, Sliding Modes in Electromechanical Systems, London.

Vidyasagar, M. (1993), *Nonlinear Systems Analysis*, Prentice Hall, Englewood Cliffs, New Jersey 07632.

Wintner, A. (1945), "The nonlocal existence problem of ordinary differential equations", *American Journal of Mathematics* **67**.

Yakubovich, V. (1973), "Auto-oscillation frequency conditions in nonlinear systems with a single stationary nonlinearity", *Siberian Mathematical Journal* **14**(5), 1100–1129 (in Russian).

Yoshida, K. (1979), *Functional Analysis*, Narosa Publishing House, New Delhi.

Young, J. & Zhou, X. Y. (1999), *Stochastic Controls: Hamiltonian Systems and HJB Equations*, Springer-Verlag, New York.

Zeidler, E. (1986), *Nonlinear Functional Analysis: Fixed-Point Theorems*, Vol. 1, Springer-Verlag, New York.

Zeidler, E. (1995), *Applied Functional Analysis*, Vol. 108 of *Applied Mathematical Sciences*, Springer-Verlag, New York.

Zhou, et al., K., Doyle, J. & Glover, K. (1996), *Robust and Optimal Control*, Prentice Hall, Upper Saddle River, New Jersey.

Zubov, V. (1962), *Mathematical Methods for the Study of Automatic Control Systems*, Pergamon Press, New York.

Zubov, V. (1964), *Methods of A.M. Lyapunov and Their Applications*, Noordhoff, Groningen.

INDEX

Index Terms	Links

D

E

F

G

H

I

N

P

S

Index Terms	Links
Series	
alternating	371
Fourier	429
geometric	372
Laurent	425
nth partial sum	368
sum	368
Taylor	420
telescopic	369
Series convergence	
Abel's test	375
Dirichlet's test	375
integral test	373
Set	
Borel	303
bounded	257
bounded above	231
closed	257
closure	258
compact	260
complement	257
connected	257
dense	257
diameter	263
elementary	297
finitely μ-measurable	300
low bound	231
μ-measurable	300
open	257
relative compact	469
separable	625
strictly separable	625

T

W