Texts in Applied Mathematics 68

Armin Iske

Approximation Theory and Algorithms for Data Analysis

Springer

Texts in Applied Mathematics

Volume 68

Editors-in-chief

S. S. Antman, University of Maryland, College Park, USA
A. Bloch, University of Michigan, Public University, City of Michigan, USA
A. Goriely, Universiyty of Oxford, Oxford, UK
L. Greengard, New York University, New York, USA
P. J. Holmes, Princeton University, Princeton, USA

Series editors

J. Bell, Lawrence Berkeley National Lab, Berkeley, USA
R. Kohn, New York University, New York, USA
P. Newton, University of Southern California, Los Angeles, USA
C. Peskin, New York University, New York, USA
R. Pego, Carnegie Mellon University, Pittsburgh, USA
L. Ryzhik, Stanford University, Stanford, USA
A. Singer, Princeton University, Princeton, USA
A. Stevens, Max-Planck-Institute for Mathematics, Leipzig, Germany
A. Stuart, University of Warwick, Coventry, UK
T. Witelski, Duke University, Durham, USA
S. Wright, University of Wisconsin, Madison, USA

The mathematization of all sciences, the fading of traditional scientific boundaries, the impact of computer technology, the growing importance of computer modeling and the necessity of scientific planning all create the need both in education and research for books that are introductory to and abreast of these developments. The aim of this series is to provide such textbooks in applied mathematics for the student scientist. Books should be well illustrated and have clear exposition and sound pedagogy. Large number of examples and exercises at varying levels are recommended. TAM publishes textbooks suitable for advanced undergraduate and beginning graduate courses, and complements the Applied Mathematical Sciences (AMS) series, which focuses on advanced textbooks and research-level monographs.

More information about this series at http://www.springer.com/series/1214

Armin Iske

Approximation Theory and Algorithms for Data Analysis

Springer

Armin Iske
Department of Mathematics
University of Hamburg
Hamburg, Germany

ISSN 0939-2475 ISSN 2196-9949 (electronic)
Texts in Applied Mathematics
ISBN 978-3-030-05227-0 ISBN 978-3-030-05228-7 (eBook)
https://doi.org/10.1007/978-3-030-05228-7

Library of Congress Control Number: 2018963282

Mathematics Subject Classification (2010): 41-XX, 42-XX, 65-XX, 94A12

Original German edition published by Springer-Verlag GmbH, Heidelberg, 2017. Title of German edition: Approximation.
© Springer Nature Switzerland AG 2018
This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.
The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This textbook offers an elementary introduction to the theory and numerics of approximation methods, combining classical topics of approximation with selected recent advances in mathematical signal processing, and adopting a constructive approach, in which the development of numerical algorithms for data analysis plays an important role.

Although the title may suggest otherwise, this textbook is not a result of the current hype on *big data science*. Nevertheless, both classical and contemporary topics of approximation include the analysis and representation of functions (e.g. signals), where suitable mathematical tools (e.g. Fourier transforms) are essential for the analysis and synthesis of the data. As such, the subject of *data analysis* is a central topic within approximation, as we will discuss in further detail.

Prerequisites. This textbook is suitable for undergraduate students who have a sound background in linear algebra and analysis. Further relevant basics on numerical methods are provided in Chapter 2, so that this textbook can be used by students attending a course on numerical mathematics. For others, the material in Chapter 2 offers a welcome review of basic numerical methods. The text of this work is suitable for courses, seminars, and distance learning programs on approximation.

Contents and Standard Topics. The central theme of approximation is the characterization and construction of best approximations in normed linear spaces. Readers are introduced to this standard topic (in Chapter 3), before approximation in Euclidean spaces (in Chapter 4) and Chebyshev approximation (in Chapter 5) are addressed. These are followed by asymptotic results concerning the approximation of univariate continuous functions by algebraic and trigonometric polynomials (in Chapter 6), where the asymptotic behaviour of Fourier partial sums is of primary importance. The core topics of Chapters 3-6 should be an essential part of any introductory course on approximation theory.

More Advanced Topics. Chapters 7-9 discuss more advanced topics and address recent developments in modern approximation and its relevant applications. To this end, Chapter 7 explains the basic concepts of signal approximation using Fourier and wavelet methods. This is followed by a more comprehensive introduction to multivariate approximation by meshfree positive definite kernels in Chapter 8. The material in Sections 8.4-8.5 provides more recent results concerning relevant aspects of convergence, stability, and update strategies for kernel-based approximation. Moreover, Section 8.6 presents basic facts on kernel-based learning. Lastly, Chapter 9 focuses on mathematical methods of computerized tomography, exploring this important application field from the viewpoint of approximation. In particular, new convergence results concerning the approximation of bivariate functions from Radon data are proven in Section 9.4.

For those who have studied Chapters 3-6, any of the three more advanced topics in Chapters 7-9 could seamlessly be included in an introductory course on approximation. Nevertheless, it is strongly recommended that readers first study the Fourier basics presented in Sections 7.1-7.4, since much of the subsequent material in Chapters 8-9 relies on Fourier techniques.

Exercises and Problem Solving. Active participation in exercise classes is generally an essential requirement for the successful completion of a mathematics course, and a (decent) course on approximation is certainly no exception. As such, each of the Chapters 3-9 includes a separate section with exercises. To enhance learning, readers are strongly encouraged to work on these exercises, which have different levels of complexity and difficulty. Some of the exercise problems are suitable for group work in class, while others should be assigned for homework. Although a number of the exercise problems may appear difficult, they can be solved using the techniques explained in this book. Further hints and comments are available on the website

`www.math.uni-hamburg.de/home/iske/approx.en.html`.

Biographical Data. To allow readers to appreciate the historical context of the presented topics and their developments, we decided to provide footnotes, where we refer to those whose names are linked with the corresponding results, definitions, and terms. For a better overview, we have also added a name index. The listed biographical data mainly relies on the online archive *MacTutor History of Mathematics* [55] and on the free encyclopedia *Wikipedia* [73], where more detailed information can be found.

Acknowledgement. The material of this book has grown over many years out the courses on approximation and mathematical signal processing that I taught at the universities of Hamburg (Germany), Lund (Sweden), and Padua (Italy). I thank the participating students for their constructive feedback, which has added great didactical value to this textbook. Moreover, I would like to thank my (post)doctoral students Dr Adeleke Bankole, Dr Matthias Beckmann, Dr Benedikt Diederichs, and Niklas Wagner for their careful proofreading. Additional comments and suggestions from Dr Matthias Beckmann and Dr Benedikt Diederichs concerning conceptional and didactical aspects as well as the technical details of the presentation are gratefully appreciated. Last but not least, I would like to thank Dr Martin Peters (SpringerSpektrum, Heidelberg) for his support and encouragement, which led to the initiation of the book project.

Hamburg, October 2018

Armin Iske
iske@math.uni-hamburg.de

Table of Contents

1 Introduction

Contemporary applications in computational science and engineering, as well as in finance, require powerful mathematical methods to analyze big data sets. Due to the rapidly growing complexity of relevant application data at limited computational (hardware) capacities, efficient numerical algorithms are required for the simulation of complex systems with only a few parameters. Both the parameter identification and the data assimilation are based on high-performance computational methods to *approximately* represent mathematical functions.

This textbook gives an introduction to the theory and the numerics of approximation methods, where the approximation of *real-valued* functions

$$f : \Omega \longrightarrow \mathbb{R}$$

on compact parameter domains $\Omega \subset \mathbb{R}^d$, $d \geq 1$, plays a central role.

But we do not only restrict ourselves to the approximation of functions. We rather work with the more general assumption, where f lies in a linear space $\mathcal{F}$, i.e., $f \in \mathcal{F}$. For the construction of a concrete approximation method for elements $f \in \mathcal{F}$, we first fix a suitable subset $\mathcal{S} \subset \mathcal{F}$, from which we seek to select an *approximation* $s^* \in \mathcal{S}$ to f. But the selection of s^* requires particular care. For the "best possible" representation of f by $s^* \in \mathcal{S}$ we are interested in the selection of a *best approximation* $s^* \in \mathcal{S}$ to f,

$$s^* \approx f,$$

i.e., an element $s^* \in \mathcal{S}$ which is among all elements $s \in \mathcal{S}$ *closest* to f.

In this introduction, we first explain important concepts and notions of approximation, but only very briefly. Later in this chapter, we discuss concrete examples for relevant function spaces $\mathcal{F}$ and suitable subsets $\mathcal{S} \subset \mathcal{F}$. For further motivation and outlook, we finally sketch selected questions and results of approximation, which we will later address in more detail.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_1

1.1 Preliminaries, Definitions and Notations

For the construction of best approximations to $f \in \mathcal{F}$, we necessarily need to measure distances between f and its *approximations* $s \in \mathcal{S}$. To this end, we introduce a *norm* for $\mathcal{F}$, where throughout this text we assume that $\mathcal{F}$ is a linear space (i.e., vector space) over the real numbers $\mathbb{R}$ or over the complex numbers $\mathbb{C}$.

Definition 1.1. *For a linear space $\mathcal{F}$, a mapping $\|\cdot\| : \mathcal{F} \longrightarrow [0,\infty)$ is said to be a* `norm` *for $\mathcal{F}$, if the following properties are satisfied.*

(a) $\|u\| = 0$, if and only if $u = 0$ (definiteness)
(b) $\|\alpha u\| = |\alpha|\|u\|$ for all $u \in \mathcal{F}$ and all $\alpha \in \mathbb{R}$ (or, $\alpha \in \mathbb{C}$) (homogeneity)
(c) $\|u + v\| \le \|u\| + \|v\|$ for all $u, v \in \mathcal{F}$ (triangle inequality).

In this case, $\mathcal{F}$ with the norm $\|\cdot\|$, $(\mathcal{F}, \|\cdot\|)$, is called a `normed space`. ◯

For the approximation of functions, *infinite-dimensional* linear spaces $\mathcal{F}$ are of particular interest. Let us make one relevant example: For a compact domain $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$,

$$\mathscr{C}(\Omega) := \{u : \Omega \longrightarrow \mathbb{R} \,|\, u \text{ continuous on } \Omega\}$$

denotes the linear space of all continuous functions on Ω. Recall that $\mathscr{C}(\Omega)$ is a linear space of infinite dimension. When equipped with the *maximum norm* $\|\cdot\|_\infty$, defined as

$$\|u\|_\infty := \max_{x\in\Omega} |u(x)| \qquad \text{for } u \in \mathscr{C}(\Omega),$$

$\mathscr{C}(\Omega)$ is a *normed linear function space* (or, in short: a *normed space*). The normed space $\mathscr{C}(\Omega)$, equipped with the maximum norm $\|\cdot\|_\infty$, is *complete*, i.e., $(\mathscr{C}(\Omega), \|\cdot\|_\infty)$ is a *Banach space*. We note this important result as follows.

Theorem 1.2. *For compact $\Omega \subset \mathbb{R}^d$, $(\mathscr{C}(\Omega), \|\cdot\|_\infty)$ is a Banach space.* □

Further important examples for norms on $\mathscr{C}(\Omega)$ are the p-norms $\|\cdot\|_p$, $1 \le p < \infty$, defined as

$$\|u\|_p := \left(\int_\Omega |u(x)|^p \,\mathrm{d}x\right)^{1/p} \qquad \text{for } u \in \mathscr{C}(\Omega).$$

Example 1.3. For $1 \le p < \infty$, $(\mathscr{C}(\Omega), \|\cdot\|_p)$ is a normed space. ♢

We remark that the case $p = 2$ is of particular interest: In this case, the 2-norm $\|\cdot\|_2$ on $\mathscr{C}(\Omega)$ is generated by the *inner product* $(\cdot,\cdot)$,

$$(u, v) := \int_\Omega u(x)v(x) \,\mathrm{d}x \qquad \text{for } u, v \in \mathscr{C}(\Omega),$$

via $\|\cdot\|_2 = (\cdot,\cdot)^{1/2}$, so that

$$\|u\|_2 = \sqrt{\int_\Omega |u(x)|^2 \,\mathrm{d}x} \qquad \text{for } u \in \mathscr{C}(\Omega).$$

To be more general, a linear space $\mathcal{F}$, equipped with an inner product $(\cdot,\cdot) : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{R}$ is with $\|\cdot\| := (\cdot,\cdot)^{1/2}$ is a normed space, in which case we say that $\mathcal{F}$ is a *Euclidean space*.

Example 1.4. The normed space $(\mathscr{C}(\Omega), \|\cdot\|_2)$ is a Euclidean space. ♢

We analyze the approximation in Euclidean spaces in detail in Chapter 4.

As we will prove, the *smoothness* of a target function f to be approximated takes influence on the resulting *approximation quality*, where we quantify the smoothness of f by its *differentiation order* $k \in \mathbb{N}_0$. For this reason, the linear subspaces

$$\mathscr{C}^k(\Omega) = \{u : \Omega \longrightarrow \mathbb{R} \,|\, u \text{ has } k \text{ continuous derivatives on } \Omega\} \subset \mathscr{C}(\Omega)$$

are of particular interest. The function spaces $\mathscr{C}^k(\Omega)$ form a nested sequence

$$\mathscr{C}^\infty(\Omega) \subset \mathscr{C}^{k+1}(\Omega) \subset \mathscr{C}^k(\Omega) \subset \mathscr{C}^{k-1}(\Omega) \subset \cdots \subset \mathscr{C}^1(\Omega) \subset \mathscr{C}^0(\Omega) = \mathscr{C}(\Omega)$$

of infinite-dimensional linear subsets of $\mathscr{C}(\Omega)$, where

$$\mathscr{C}^\infty(\Omega) := \bigcap_{k\in\mathbb{N}_0} \mathscr{C}^k(\Omega)$$

is the linear space of functions with arbitrary differentiation order on Ω.

For the construction of approximation methods, *finite-dimensional* linear subspaces $\mathcal{S} \subset \mathcal{F}$ are useful. To this end, let $\{s_1, \ldots, s_n\} \subset \mathcal{S}$ be a fixed basis of $\mathcal{S}$, where $n = \dim(\mathcal{S}) \in \mathbb{N}$. In this case, any $s \in \mathcal{S}$ can uniquely be represented by a linear combination

$$s = \sum_{j=1}^{n} c_j s_j$$

by n parameters $c_1, \ldots, c_n \in \mathbb{R}$. As we will see later, the assumption of finite dimension for $\mathcal{S}$ will help simplify our computations, especially for the coding and the evaluation of best approximations $s^* \in \mathcal{S}$ to f. But the finite dimension of $\mathcal{S}$ will also be useful in theory, in particular when it comes to discussing the existence of best approximations.

For the special case of *univariate* functions, i.e., for $\Omega = [a, b] \subset \mathbb{R}$ compact, we consider the approximation of continuous functions $f \in \mathscr{C}[a, b]$ by using algebraic polynomials. In this case, we choose $\mathcal{S} = \mathcal{P}_n$, for a fixed degree $n \in \mathbb{N}_0$, so that $\mathcal{P}_n$ is the linear space of all univariate polynomials of degree at most n.

For the representation of algebraic polynomials from $\mathcal{P}_n$ the monomial basis $\{1, x, x^2, \ldots, x^n\}$ is particularly popular, where in this case any $p \in \mathcal{P}_n$ is represented by a unique linear combination of the monomial form

$$p(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \qquad \text{for } x \in \mathbb{R}$$

with real coefficients $a_0, \ldots, a_n$. Note that $\dim(\mathcal{P}_n) = n + 1$.

Further relevant examples for infinite-dimensional linear spaces of univariate functions $\mathcal{F}$ are the 2π-periodic continuous functions,

$$\mathscr{C}_{2\pi} := \{u \in \mathscr{C}(\mathbb{R}) \,|\, u(x) = u(x + 2\pi) \text{ for all } x \in \mathbb{R}\} \subset \mathscr{C}(\mathbb{R}),$$

and their linear subspaces

$$\mathscr{C}^k_{2\pi} := \mathscr{C}^k(\mathbb{R}) \cap \mathscr{C}_{2\pi} \qquad \text{for } k \in \mathbb{N}_0 \cup \{\infty\}.$$

Example 1.5. For $k \in \mathbb{N}_0 \cup \{\infty\}$ and $1 \leq p < \infty$, the function space $\mathscr{C}^k_{2\pi}$, equipped with the p-norm

$$\|u\|_p := \left(\int_0^{2\pi} |u(x)|^p \, \mathrm{d}x \right)^{1/p} \qquad \text{for } u \in \mathscr{C}^k_{2\pi},$$

is a normed linear space. For $p = 2$ the function space $\mathscr{C}^k_{2\pi}$ is Euclidean by the inner product

$$(u, v) := \int_0^{2\pi} u(x) v(x) \, \mathrm{d}x \qquad \text{for } u, v \in \mathscr{C}^k_{2\pi}.$$

Finally, the function space $\mathscr{C}^k_{2\pi}$, equipped with the maximum norm

$$\|u\|_\infty := \max_{x \in [0, 2\pi]} |u(x)| \qquad \text{for } u \in \mathscr{C}^k_{2\pi}$$

is a Banach space. ♢

The approximation of functions from $\mathscr{C}_{2\pi}$ plays an important role in mathematical signal processing, where trigonometric polynomials of the form

$$T(x) = \frac{a_0}{2} + \sum_{j=1}^{n} [a_j \cos(jx) + b_j \sin(jx)] \qquad \text{for } x \in \mathbb{R}$$

with *Fourier coefficients* $a_0, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$ are used. In this case, we choose $\mathcal{S} = \mathcal{T}_n$, for $n \in \mathbb{N}_0$, where

$$\mathcal{T}_n = \operatorname{span}\{1, \sin(x), \cos(x), \ldots, \sin(nx), \cos(nx)\} \subset \mathscr{C}_{2\pi}$$

is the linear space of all *real-valued* trigonometric polynomials of degree at most $n \in \mathbb{N}_0$. Note that $\dim(\mathcal{T}_n) = 2n + 1$.

We will discuss further relevant examples for normed spaces $(\mathcal{F}, \|\cdot\|)$ and approximation spaces $\mathcal{S} \subset \mathcal{F}$ later. In this short introduction, we will only touch a few more important aspects of approximation for an outlook.

1.2 Basic Problems and Outlook

For the analysis of best approximations, the following questions are relevant.

- On given $f \in \mathcal{F}$, does there exist a best approximation $s^* \in \mathcal{S}$ for f?
- Is a best approximation s^* for f unique?
- Are there necessary/sufficient conditions for a best approximation s^*?
- How can we compute a best approximation s^* analytically or numerically?

The answers to the above questions depend on the properties of the linear space $\mathcal{F}$, its norm $\|\cdot\|$, as well as on the chosen approximation space $\mathcal{S} \subset \mathcal{F}$. We will give satisfying answers to the above questions. In Chapter 3 we first provide general answers that do not depend on the particular choices of $\mathcal{F}$, $\|\cdot\|$, and $\mathcal{S}$, but rather on their structural properties. Then we analyze the special case of the Euclidean norm (in Chapter 4) and that of the maximum norm $\|\cdot\|_\infty$, also referred to as *Chebyshev norm* (in Chapter 5).

Later in Chapter 6, we study the *asymptotic behaviour* of approximation methods. In that discussion, we ask the question, how *well* we can approximate a target $f \in \mathcal{F}$ by certain sequences of approximations to f. To further explain on this, suppose $f \in \mathcal{F}$, the norm $\|\cdot\|$, and the approximation space $\mathcal{S}$ are fixed. Then we can quantify the *approximation quality* by the *minimal distance*

$$\eta \equiv \eta(f, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|s - f\| = \|s^* - f\|$$

between f and $\mathcal{S}$. In relevant application scenarios, we wish to approximate f *arbitrarily well*. For fixed $\mathcal{S}$, this will not work, however, since in that case the minimal distance $\eta(f, \mathcal{S})$ is already the best possible.

Therefore, we work with nested sequences of approximation spaces

$$\mathcal{S}_0 \subset \mathcal{S}_1 \subset \ldots \subset \mathcal{S}_n \subset \mathcal{F} \qquad \text{for } n \in \mathbb{N}_0$$

where we also regard the corresponding sequence of minimal distances

$$\eta(f, \mathcal{S}_0) \geq \eta(f, \mathcal{S}_1) \geq \ldots \geq \eta(f, \mathcal{S}_n) \geq 0,$$

whose *asymptotic* behaviour we will analyze. Now if we wish to approximate $f \in \mathcal{F}$ arbitrarily well, then the minimal distances must necessarily be a zero sequence, i.e.,

$$\eta(f, \mathcal{S}_n) \longrightarrow 0 \quad \text{for } n \to \infty.$$

This leads us to the following fundamental question of approximation.

Question: Is there for any $f \in \mathcal{F}$ and any $\varepsilon > 0$ one $n \in \mathbb{N}$ satisfying

$$\eta(f, \mathcal{S}_n) = \|s_n^* - f\| < \varepsilon,$$

where $s_n^* \in \mathcal{S}_n$ is a best approximation to f from $\mathcal{S}_n$? ♢

If the answer to the above question is positive, then the union

$$\mathcal{S} = \bigcup_{n \geq 0} \mathcal{S}_n \subset \mathcal{F}$$

is called *dense* in $\mathcal{F}$ with respect to the norm $\|\cdot\|$, or, *dense subset* of $\mathcal{F}$.

We are particularly interested in the approximation of continuous functions, where we first study the approximation of *univariate* continuous functions (in Chapters 2-6), before we later turn to *multivariate* approximation (in Chapters 8-9).

For an outlook, we quote two classical results from univariate approximation, which are studied in Chapter 6. The following result of Weierstrass (dating back to 1885) is often referred to as the *"birth of approximation"*.

Theorem 1.6. (Weierstrass, 1885). *For a compact interval $[a, b] \subset \mathbb{R}$, the set of algebraic polynomials $\mathcal{P}$ is dense in $\mathscr{C}[a, b]$ with respect to the maximum norm $\|\cdot\|_\infty$. In other words, for any $f \in \mathscr{C}[a, b]$ and $\varepsilon > 0$ there is an algebraic polynomial p satisfying*

$$\|p - f\|_{\infty,[a,b]} = \max_{x \in [a,b]} |p(x) - f(x)| < \varepsilon.$$

□

The above version of the Weierstrass theorem is an algebraic one. But there is also a *trigonometric version* of the Weierstrass theorem, according to which the set of trigonometric polynomials $\mathcal{T}$ is dense in $\mathscr{C}_{2\pi}$ with respect to the maximum norm $\|\cdot\|_\infty$. We will prove both versions of the Weierstrass theorem in Section 6.1. Moreover, in Sections 6.3 and 6.4 we will, for $f \in \mathscr{C}_{2\pi}$, analyze decay rates for the minimal distances

$$\eta(f, \mathcal{T}_n) := \inf_{T \in \mathcal{T}_n} \|T - f\| \quad \text{and} \quad \eta_\infty(f, \mathcal{T}_n) := \inf_{T \in \mathcal{T}_n} \|T - f\|_\infty$$

with respect to both the Euclidean norm $\|\cdot\|$ and the maximum norm $\|\cdot\|_\infty$. The latter will lead us to the Jackson theorems, one of which is as follows.

Theorem 1.7. (Jackson). *For $f \in \mathscr{C}_{2\pi}^k$ we have*

$$\eta_\infty(f, \mathcal{T}_n) \leq \left(\frac{\pi}{2(n+1)}\right)^k \cdot \|f^{(k)}\|_\infty = \mathcal{O}\left(n^{-k}\right) \quad \text{for } n \to \infty.$$

□

From this result, we see that the *power* of the approximation method does not only depend on the approximation spaces $\mathcal{T}_n$ but also and essentially on the smoothness of the target f. Indeed, the following principle holds:

The smoother the target function $f \in \mathscr{C}_{2\pi}$, the faster is the convergence of the minimal distances $\eta(f, \mathcal{T}_n)$, or, $\eta_\infty(f, \mathcal{T}_n)$ to zero.

We will prove this and other classical results concerning the asymptotic behaviour of minimal distances in Chapter 6.

1.3 Approximation Methods for Data Analysis

Having studied classical topics of approximation (in Chapters 3-6) we will address more recent developments and trends of approximation. To this end, we develop and analyze specific approximation methods for data analysis, where relevant applications in signal processing play an important role.

We first introduce (in Chapter 7) basic concepts of Fourier analysis. Further, in Section 7.3 we prove the *Shannon sampling theorem*, Theorem 7.34, which is a fundamental result of signal theory. According to the Shannon sampling theorem, any *band-limited* signal $f \in \mathrm{L}^2(\mathbb{R})$ (i.e., f has limited frequency density) can be recovered *exactly* from its values taken on an infinite discrete sampling mesh of constant mesh width. Our proof of the Shannon sampling theorem will demonstrate the relevance and the fundamental importance of the introduced Fourier methods.

Further advanced topics of approximation are comprising *wavelets* and *kernel-based* methods for multivariate approximation. In this introductory text, however, we can only cover a few selected theoretical and numerical aspects of these multifaceted topics. Therefore, as regards wavelet methods (in Section 7.5) we restrict ourselves to an introduction of the *Haar wavelet*. Moreover, the subsequent discussion on basic concepts of kernel-based approximation (in Chapter 8) is based on *positive definite* kernels. Among our addressed applications in multivariate data analysis are kernel-based methods in *machine learning* (in Section 8.6). For further details on this subject, we refer to our references in the following Section 1.4.

Another important application is the approximation of bivariate signals in *computer tomography*, as addressed in Chapter 9, where we analyze theoretical aspects of this inverse problem rigorously from the viewpoint of approximation. This finally leads us to novel error estimates and convergence rates, as developed in Section 9.4. The constructive account taken here provides a new concept of evaluation methods for low-pass filters. We finally discuss the implementation of the filtered back projection formula (in Section 9.5).

1.4 Hints on Classical and More Recent Literature

Approximation theory is a vivid research area of mathematics with a long history [55]. More recent developments have provided powerful numerical approximation methods aiming to address challenging application problems in the relevant areas of data and computer science, natural science and engineering. This has led to a large variety of diverse contributions to the approximation literature by research monographs and publications that can hardly be overviewed. In fact, it is obviously impossible to cover all relevant aspects of approximation in broad width and depth in one textbook. For this elementary introduction, we decided to first treat selected theoretical aspects of classical approximation, before we turn to more recent concepts of numerical approximation.

For further reading, we refer to a selection of classical and contemporary sources on approximation theory and numerical methods. Likewise, the list of references cannot be complete, and in fact we can only give a few hints, although the selection of more recent texts on approximation is rather limited. As regards classical texts on approximation (from the second half of the last century) we refer to [11, 12, 19, 50, 56, 70]. Further material on more advanced topics, including *nonlinear* approximation, can be found in [9, 21, 43].

For a more recent introduction to approximation theory with accentuated links to relevant applications in geomathematics we refer to [51]. A modern account to approximation with pronounced algorithmic and numerical elements is provided in the modern teaching concept of [68].

Literature sources to more specific topics of approximation are dealing with spline approximations [20, 36, 64, 65], wavelets [14, 18, 49] and radial basis functions [10, 24, 25, 27, 38, 72]. Since spline approximation is a well-established topic in standard courses on numerical mathematics [57], we decided to omit a discussion on splines in this work.

2 Basic Methods and Numerical Algorithms

In this chapter, we discuss basic mathematical methods and numerical algorithms for interpolation and approximation of functions in one variable. The concepts and principles which we address here should already be known from numerical mathematics. Nevertheless, the material of this chapter will be necessary for our subsequent discussion. Therefore, a repetition of selected elements from numerical mathematics should be most welcome.

For the sake of preparation, let us first fix a few notations. We denote by $f : [a, b] \longrightarrow \mathbb{R}$ a continuous function on a compact interval $[a, b] \subset \mathbb{R}$, $f \in \mathscr{C}[a, b]$. Moreover,

$$X = \{x_0, x_1, \ldots, x_n\} \subset [a, b] \qquad \text{for } n \in \mathbb{N}_0$$

is a set of $|X| = n + 1$ pairwise distinct *interpolation knots*. We collect the function values $f_j = f(x_j)$ of f on X in one *data vector*,

$$f_X = (f_0, f_1, \ldots, f_n)^T \in \mathbb{R}^{n+1}.$$

For the approximation of f, we will later specify suitable linear subspaces of continuous functions, $\mathcal{S} \subset \mathscr{C}[a, b]$, of finite dimension $\dim(\mathcal{S}) \leq n + 1$.

We first consider *linear least squares approximation*. In this problem, we seek an approximation $s^* \in \mathcal{S}$ to f which minimizes among all $s \in \mathcal{S}$ the sum of pointwise square errors on X, so that

$$\sum_{x \in X} |s^*(x) - f(x)|^2 \leq \sum_{x \in X} |s(x) - f(x)|^2 \qquad \text{for all } s \in \mathcal{S}. \tag{2.1}$$

Moreover, we discuss numerical algorithms for *interpolation*, which could be viewed as a special case of linear least squares approximation. To this end, we first consider using algebraic polynomials, where $\mathcal{S} = \mathcal{P}_n$. To compute a solution $s \in \mathcal{P}_n$ of the interpolation problem $s_X = f_X$, i.e.,

$$s(x_j) = f(x_j) \qquad \text{for all } 0 \leq j \leq n, \tag{2.2}$$

we develop efficient and numerically stable algorithms. Finally, we address interpolation to periodic functions by using trigonometric polynomials, where $\mathcal{S} = \mathcal{T}_n$. This leads us directly to the *discrete Fourier transform* (DFT), which will be of primary importance later in this book. We show how the DFT can be computed efficiently by using the *fast Fourier transform* (FFT).

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_2

2.1 Linear Least Squares Approximation

The following discussion on *linear least squares approximation* leads us to a first concrete example of an approximation problem. As a starting point for our investigations, we regard the minimization problem (2.1), whose solution we wish to construct. To this end, we first fix a set $\mathcal{B} = \{s_1, \ldots, s_m\} \subset \mathscr{C}[a,b]$ of $m \leq n+1$ linearly independent continuous functions. This immediately leads us to the linear approximation space

$$\mathcal{S} = \operatorname{span}\{s_1, \ldots, s_m\} := \left\{ \sum_{j=1}^{m} c_j s_j \;\middle|\; c_1, \ldots, c_m \in \mathbb{R} \right\} \subset \mathscr{C}[a,b]$$

with basis $\mathcal{B}$ and of dimension $\dim(\mathcal{S}) = m$. In typical applications of linear least squares approximation the number $n+1$ of given function values in f_X is assumed to be much larger than the dimension m of $\mathcal{S}$. Indeed, we prefer to work with a simple model for $\mathcal{S}$ which in particular is generated by only a few basis functions $\mathcal{B}$. We use the notation $m \ll n+1$ to indicate that m is assumed to be much smaller than $n+1$.

But the following method for solving the minimization problem (2.1) can be applied for all $m \leq n+1$. We formulate the linear least squares approximation problem (2.1) more precisely as follows.

Problem 2.1. Compute from a given set $X = \{x_0, \ldots, x_n\} \subset [a,b]$ of $n+1$ pairwise distinct points and a data vector $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$ a continuous function $s^* \in \mathcal{S} = \operatorname{span}\{s_1, \ldots, s_m\}$, for $m \leq n+1$, which minimizes among all $s \in \mathcal{S}$ the pointwise sum of square errors on X, so that

$$\|s_X^* - f_X\|_2^2 \leq \|s_X - f_X\|_2^2 \qquad \text{for all } s \in \mathcal{S}. \tag{2.3}$$

□

To solve the minimization of Problem 2.1 we represent $s^* \in \mathcal{S}$ as a unique linear combination

$$s^* = \sum_{j=1}^{m} c_j^* s_j \tag{2.4}$$

of the basis functions in $\mathcal{B}$. Thereby, the linear least squares approximation problem can be reformulated as an equivalent minimization problem of the form

$$\|Bc - f_X\|_2^2 \longrightarrow \min_{c \in \mathbb{R}^m} !\,, \tag{2.5}$$

where the *design matrix*

$$B \equiv B_{\mathcal{B},X} := \begin{bmatrix} s_1(x_0) & \cdots & s_m(x_0) \\ \vdots & & \vdots \\ s_1(x_n) & \cdots & s_m(x_n) \end{bmatrix} \in \mathbb{R}^{(n+1)\times m}$$

contains all evaluations of the basis functions from $\mathcal{B}$ at the points in X. To solve the minimization problem (2.5), we regard for the multivariate function $F : \mathbb{R}^m \longrightarrow [0, \infty)$, defined as

$$F(c) = \|Bc - f_X\|_2^2 = (Bc - f_X)^T (Bc - f_X) = c^T B^T Bc - 2c^T B^T f_X + f_X^T f_X,$$

its gradient

$$\nabla F(c) = 2B^T Bc - 2B^T f_X$$

and its (constant) Hessian[1] matrix

$$\nabla^2 F(c) = 2B^T B.$$

Recall that any local minimum of F can be characterized via the solution of the linear equation system

$$B^T Bc = B^T f_X, \tag{2.6}$$

referred to as `Gaussian`[2] `normal equation`. If B has full rank, i.e., $\operatorname{rank}(B) = m$, then the symmetric matrix $B^T B$ is positive definite. In this case, the Gaussian normal equation (2.6) has a unique solution $c^* = (c_1^*, \ldots, c_m^*)^T \in \mathbb{R}^m$ satisfying

$$F(c^*) < F(c) \qquad \text{for all } c \in \mathbb{R}^m \setminus \{c^*\}.$$

The solution $c^* \in \mathbb{R}^m$ yields the sought coefficient vector for s^* in (2.4).

Hence, our first approximation problem, Problem 2.1, is solved.

However, our suggested solution via the Gaussian normal equation (2.6) is problematic from a numerical viewpoint: If B has full rank, then the spectral condition numbers of the matrices $B^T B$ and B are related by (cf. [57, Section 3.1])

$$\kappa_2(B^T B) = (\kappa_2(B))^2 .$$

Therefore, the spectral condition number $\kappa_2(B^T B)$ of the matrix $B^T B$ grows quadratically proportional to the reciprocal of the smallest singular value of B. For matrices B arising in relevant applications of linear least squares approximation, however, its smallest singular value is typically very small, whereby the condition number $\kappa_2(B^T B)$ of $B^T B$ is even worse. In fact, the condition number of linear least squares approximation problems is, especially for very small residuals $\|Bc - f_X\|_2$, very critical, so that a solution via the Gaussian normal equation (2.6) should be avoided for the sake of numerical stability (see [28, Section 6.2]). A more comprehensive error analysis on linear least squares approximation can be found in the textbook [7].

Instead of this, a numerically stable solution for the linear least squares approximation problem works with a QR factorization

[1] Ludwig Otto Hesse (1811-1874), German mathematician
[2] Carl Friedrich Gauss (1777-1855), German mathematician and astronomer

$$B = QR \tag{2.7}$$

of the design matrix $B \in \mathbb{R}^{(n+1)\times m}$, where $Q \in \mathbb{R}^{(n+1)\times(n+1)}$ is an orthogonal matrix and R is an upper triangular matrix of the form

$$R = \begin{bmatrix} S \\ \hline 0 \end{bmatrix} = \left[\begin{array}{ccc} s_{11} & \cdots & s_{1m} \\ & \ddots & \vdots \\ & & s_{mm} \\ \hline & 0 & \end{array}\right] \in \mathbb{R}^{(n+1)\times m}. \tag{2.8}$$

Note that matrix B has full rank, $\operatorname{rank}(B) = m$, if and only if no diagonal entry s_{kk}, $1 \le k \le m$, in the upper triangular matrix $S \in \mathbb{R}^{m\times m}$ vanishes.

A numerically stable solution for the minimization problem (2.5) relies on the alternative representation

$$F(c) = \|Bc - f_X\|_2^2 = \|QRc - f_X\|_2^2 = \|Rc - Q^T f_X\|_2^2, \tag{2.9}$$

where we use the *isometry* of the inverse $Q^{-1} = Q^T$ with respect to the Euclidean norm $\|\cdot\|_2$, i.e.,

$$\|Q^T y\|_2 = \|y\|_2 \qquad \text{for all } y \in \mathbb{R}^{n+1}.$$

Now the vector $Q^T f_X \in \mathbb{R}^{n+1}$ can be partitioned into two blocks $g \in \mathbb{R}^m$ and $h \in \mathbb{R}^{n+1-m}$, so that

$$Q^T f_X = \begin{bmatrix} g \\ h \end{bmatrix} \in \mathbb{R}^{n+1}. \tag{2.10}$$

Therefore, the representation for $F(c)$ in (2.9) can be rewritten as a sum of the form

$$F(c) = \|Sc - g\|_2^2 + \|h\|_2^2, \tag{2.11}$$

where we use the partitioning (2.8) for R and that in (2.10) for $Q^T f_X$. The minimum of $F(c)$ in (2.11) can finally be computed via the solution of the triangular linear system

$$Sc = g$$

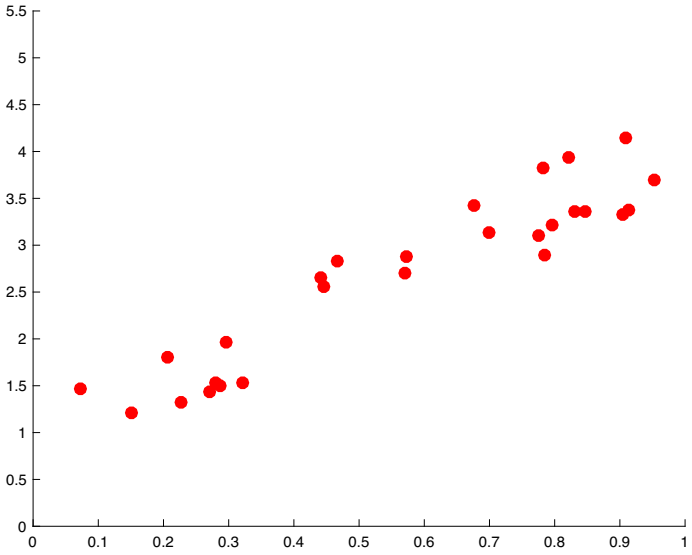
by a backward substitution. The solution c^* of this linear system is unique, if and only if B has full rank.

In conclusion, the described procedure provides a numerically stable algorithm to compute the solution c^* of the minimization problem (2.5), and this yields the coefficient vector c^* of s^* in (2.4). For the *approximation error*, we obtain

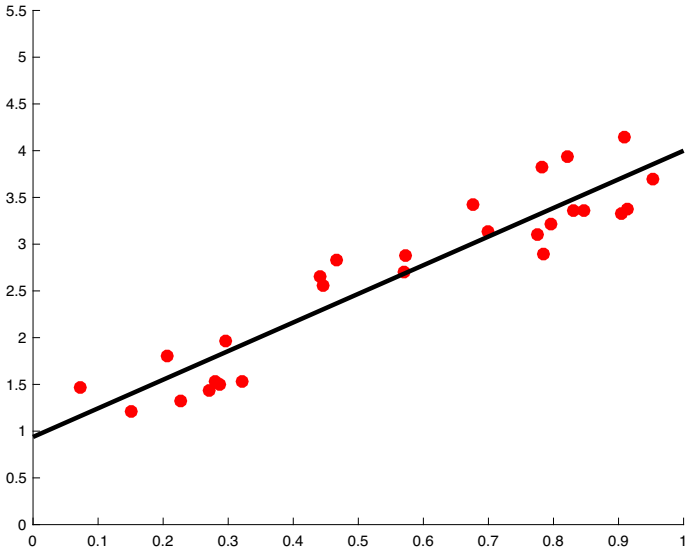
$$F(c^*) = \|Bc^* - f_X\|_2^2 = \|h\|_2^2.$$

This solves the linear least squares approximation problem, Problem 2.1.

For further illustration, we discuss one example of *linear regression*.



(a) noisy observations $(X, \tilde{f}_X)$



(b) regression line $s^* \in \mathcal{P}_1$

Fig. 2.1. (a) We take 26 noisy samples $\tilde{f}_X = f_X + \varepsilon_X$ from $f(x) = 1 + 3x$. (b) We compute the regression line $s^*(x) = c_0^* + c_1^* x$, with $c_0^* \approx 0.9379$ and $c_1^* \approx 3.0617$, by using linear least squares approximation (cf. Example 2.2).

Example 2.2. We assume a *linear model*, i.e., we approximate $f \in \mathscr{C}[a,b]$ by a linear function $s(x) = c_0 + c_1 x$, for $c_0, c_1 \in \mathbb{R}$. Moreover, we observe *noisy* measurements $\tilde{f}_X$, taken from f at $n+1$ sample points $X = \{x_0, x_1, \ldots, x_n\}$, so that

$$\tilde{f}(x_j) = f(x_j) + \varepsilon_j \qquad \text{for } 0 \le j \le n,$$

where ε_j is the error for the j-th sample. We collect the error terms in one vector $\varepsilon_X = (\varepsilon_0, \varepsilon_1, \ldots, \varepsilon_n)^T \in \mathbb{R}^{n+1}$. Figure 2.1 (a) shows an example for noisy observations $(X, \tilde{f}_X)$, where

$$\tilde{f}_X = f_X + \varepsilon_X.$$

We finally display the assumed linear relationship between the sample points X (of the input *data sites*) and the observations $\tilde{f}_X$ (of the *target object*). To this end, we fix the basis functions $s_1 \equiv 1$ and $s_2(x) = x$, so that $\operatorname{span}\{s_1, s_2\} = \mathcal{P}_1$. Now the solution of the resulting minimization problem in (2.5) has the form

$$\|Bc - \tilde{f}_X\|_2^2 \longrightarrow \min_{c \in \mathbb{R}^2} ! \tag{2.12}$$

with the design matrix $B \in \mathbb{R}^{(n+1)\times 2}$, where

$$B^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \end{bmatrix} \in \mathbb{R}^{2\times(n+1)}.$$

The minimization problem (2.12) can be solved via a QR decomposition of B in (2.7) to obtain a numerically stable solution $c^* = (c_0^*, c_1^*)^T \in \mathbb{R}^2$. The resulting regression line is given by $s^*(x) = c_0^* + c_1^* x$, see Figure 2.1 (b). $\diamondsuit$

2.2 Regularization Methods

Next, we develop a relevant extension of linear least squares approximation, Problem 2.1. To this end, we augment the *data error*

$$\eta_X(f, s) = \|s_X - f_X\|_2^2$$

of the cost function by a *regularization term*, given by a suitable *functional*

$$J : \mathcal{S} \longrightarrow \mathbb{R},$$

where $J(s)$ quantifies, for instance, the smoothness, the variation, the energy, or the oscillation of $s \in \mathcal{S}$. By the combination of the data error η_X and the regularization functional J, being balanced by a fixed *parameter* $\alpha > 0$, this leads us to an extension of linear least squares approximation, Problem 2.1, giving a *regularization method* that is described by the minimization problem

$$\|s_X - f_X\|_2^2 + \alpha J(s) \longrightarrow \min_{s \in \mathcal{S}} ! \tag{2.13}$$

By choice of the *regularization parameter* α, we aim to compromise between the approximation quality $\eta_X(f, s_\alpha^*)$ of a solution $s_\alpha^* \in \mathcal{S}$ for (2.13) and its *regularity*, being measured by $J(s_\alpha^*)$. We can further explain this as follows. On the one hand, for very small values α the error term $\eta_X(f, s)$ will be dominant over the regularization term $J(s)$, which improves the approximation quality of a solution s_α^* for (2.13). On the other hand, for very large values α the regularization term $J(s)$ will be dominant. In this case, however, we wish to avoid *overfitting* of a solution s_α^*. But the selection of $J : \mathcal{S} \longrightarrow \mathbb{R}$ requires particular care, where a problem-adapted choice of J relies on specific model assumptions from the application addressed. In applications of information technology, for instance, regularization methods are applied for smoothing, deblurring or denoising image and signal data (see [34]).

Now we discuss one relevant special case, where the functional $J : \mathcal{S} \longrightarrow \mathbb{R}$ is being defined by a symmetric positive definite matrix $A \in \mathbb{R}^{m \times m}$. To further explain the definition of J, for a fixed basis $\mathcal{B} = \{s_1, \ldots, s_m\}$ of $\mathcal{S}$, each element

$$s = \sum_{j=1}^{m} c_j s_j \in \mathcal{S}$$

is mapped onto the A-norm (i.e., the norm being induced by A)

$$\|c\|_A^2 := c^T A c \tag{2.14}$$

of its coefficients $c = (c_1, \ldots, c_m)^T \in \mathbb{R}^m$.

Starting from our above discussion on linear least squares approximation, this particular choice of J leads to a class of regularization methods termed `Tikhonov`[3] `regularization`. We formulate the problem of Tikhonov regularization as follows.

Problem 2.3. Under the assumption $\alpha > 0$ and $A \in \mathbb{R}^{m \times m}$ symmetric positive definite, compute on given function values $f_X \in \mathbb{R}^{n+1}$ and design matrix $B = (s_j(x_k))_{0 \le k \le n; 1 \le j \le m} \in \mathbb{R}^{(n+1) \times m}$, $m \le n+1$, a solution for the minimization problem

$$\|Bc - f_X\|_2^2 + \alpha \|c\|_A^2 \longrightarrow \min_{c \in \mathbb{R}^m} ! \tag{2.15}$$

□

Note that Problem 2.3 coincides for $\alpha = 0$ with the linear least squares approximation problem. As we show in the following, the minimization problem (2.15) of Tikhonov regularization has for any $\alpha > 0$ a unique solution, in particular for the case, where the design matrix B has no full rank. We further remark that the linear least squares approximation problem, Problem 2.1, has for $\operatorname{rank}(B) < m$ ambiguous solutions. However, as we will show,

[3] Andrey Nikolayevich Tikhonov (1906-1993), Russian mathematician

the solution $s_\alpha^* \in \mathcal{S}$ converges for $\alpha \searrow 0$ to a *norm minimal* solution s^* of linear least squares approximation.

Now we regard, for fixed $\alpha > 0$, the cost function $F_\alpha : \mathbb{R}^m \longrightarrow [0, \infty)$,

$$F_\alpha(c) = \|Bc - f_X\|_2^2 + \alpha \|c\|_A^2 = c^T \left(B^T B + \alpha A\right) c - 2c^T B^T f_X + f_X^T f_X,$$

its gradient

$$\nabla F_\alpha(c) = 2\left(B^T B + \alpha A\right) c - 2B^T f_X$$

and the (constant) positive definite Hessian matrix

$$\nabla^2 F_\alpha = 2\left(B^T B + \alpha A\right). \tag{2.16}$$

Note that the function F_α has one unique stationary point $c_\alpha^* \in \mathbb{R}^m$ satisfying the necessary condition $\nabla F_\alpha(c) = 0$. Therefore, c^* can be characterized as the unique solution of the minimization problem (2.15) via the unique solution of the linear system

$$\left(B^T B + \alpha A\right) c_\alpha^* = B^T f_X,$$

i.e., $c_\alpha^* = \left(B^T B + \alpha A\right)^{-1} B^T f_X$. Due to the positive definiteness of the Hessian $\nabla^2 F_\alpha$ in (2.16), c_α^* is a local minimum of F_α. Moreover, in this case F_α is convex, and so c_α^* is the unique minimum of F_α on $\mathbb{R}^m$.

Now we explain how c_α^* can be computed by a stable numerical algorithm. By the spectral theorem, there is an orthogonal matrix $U \in \mathbb{R}^{m \times m}$ satisfying

$$A = U \Lambda U^T,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^{m \times m}$ is a diagonal matrix containing the positive eigenvalues $\lambda_1, \ldots, \lambda_m > 0$ of A. This allows us to define the square root of A by letting

$$A^{1/2} := U \Lambda^{1/2} U^T,$$

where $\Lambda^{1/2} = \operatorname{diag}(\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_m}) \in \mathbb{R}^{m \times m}$. Note that the square root $A^{1/2}$ is, like A, also symmetric positive definite, and we have

$$\alpha \|c\|_A^2 = \alpha c^T A c = \left(\sqrt{\alpha} A^{1/2} c\right)^T \left(\sqrt{\alpha} A^{1/2} c\right) = \left\|\sqrt{\alpha} A^{1/2} c\right\|_2^2.$$

This implies

$$\|Bc - f_X\|_2^2 + \alpha \|c\|_A^2 = \|Bc - f_X\|_2^2 + \|\sqrt{\alpha} A^{1/2} c\|_2^2 = \left\| \begin{bmatrix} B \\ \sqrt{\alpha} A^{1/2} \end{bmatrix} c - \begin{bmatrix} f_X \\ 0 \end{bmatrix} \right\|_2^2.$$

Using the notation

$$B_\alpha = \begin{bmatrix} B \\ \sqrt{\alpha} A^{1/2} \end{bmatrix} \in \mathbb{R}^{((n+1)+m) \times m} \qquad \text{and} \qquad g_X = \begin{bmatrix} f_X \\ 0 \end{bmatrix} \in \mathbb{R}^{(n+1)+m},$$

we can reformulate (2.15) by the linear least squares approximation problem

$$\|B_\alpha c - g_X\|_2^2 \longrightarrow \min_{c\in\mathbb{R}^m}! \,, \tag{2.17}$$

whose solution c_α^* can be computed by a stable algorithm via a QR factorization of B_α, as already explained in the previous section for the numerical solution of the linear least squares approximation problem, Problem 2.1.

Next, we aim to characterize the asymptotic behaviour of s_α^* for $\alpha \searrow 0$ and for $\alpha \to \infty$. To this end, we first develop a suitable representation for the solution c_α^* to (2.15). Since A is symmetric positive definite, we have

$$\|Bc - f_X\|_2^2 + \alpha\|c\|_A^2 = \|BA^{-1/2}A^{1/2}c - f_X\|_2^2 + \alpha\|A^{1/2}c\|_2^2.$$

By using the notation

$$C = BA^{-1/2} \in \mathbb{R}^{(n+1)\times m} \qquad \text{and} \qquad b = A^{1/2}c \in \mathbb{R}^m$$

we can rewrite (2.15) as

$$\|Cb - f_X\|_2^2 + \alpha\|b\|_2^2 \longrightarrow \min_{b\in\mathbb{R}^m}! \tag{2.18}$$

For the solution of (2.18), we employ the singular value decomposition of C,

$$C = V\Sigma W^T,$$

where $V = (v_1, \ldots, v_{n+1}) \in \mathbb{R}^{(n+1)\times(n+1)}$ and $W = (w_1, \ldots, w_m) \in \mathbb{R}^{m\times m}$ are orthogonal, and where the matrix Σ has the form

$$\Sigma = \begin{bmatrix} \sigma_1 & & 0 \\ & \ddots & \\ & & \sigma_r \\ 0 & & 0 \end{bmatrix} \in \mathbb{R}^{(n+1)\times m}$$

with singular values $\sigma_1 \geq \ldots \geq \sigma_r > 0$, where $r = \operatorname{rank}(C) \leq m$.

But this implies

$$\begin{aligned} \|Cb - f_X\|_2^2 + \alpha\|b\|_2^2 &= \|V\Sigma W^T b - f_X\|_2^2 + \alpha\|b\|_2^2 \\ &= \|\Sigma W^T b - V^T f_X\|_2^2 + \alpha\|W^T b\|_2^2 \end{aligned}$$

and, moreover, we obtain for $a = W^T b \in \mathbb{R}^m$ the representation

$$\|Cb - f_X\|_2^2 + \alpha\|b\|_2^2 = \|\Sigma a - V^T f_X\|_2^2 + \alpha\|a\|_2^2$$

$$= \sum_{j=1}^{r}(\sigma_j a_j - v_j^T f_X)^2 + \sum_{j=r+1}^{n+1}(v_j^T f_X)^2 + \alpha\sum_{j=1}^{m} a_j^2.$$

For the minimization of this expression, we first let

$$a_j := 0 \qquad \text{for } r+1 \le j \le m,$$

so that it remains to solve the minimization problem

$$\sum_{j=1}^{r} \left[(\sigma_j a_j - v_j^T f_X)^2 + \alpha a_j^2 \right] \longrightarrow \min_{a_1,\ldots,a_r \in \mathbb{R}} ! \tag{2.19}$$

Since all terms of the cost function in (2.19) are non-negative, the minimization problem (2.19) can be split into the r independent subproblems

$$g_j(a_j) = (\sigma_j a_j - v_j^T f_X)^2 + \alpha a_j^2 \longrightarrow \min_{a_j \in \mathbb{R}} ! \qquad \text{for } 1 \le j \le r \tag{2.20}$$

with the scalar-valued cost functions $g_j : \mathbb{R} \longrightarrow \mathbb{R}$, for $1 \le j \le r$. Since

$$g_j'(a_j) = 2((\sigma_j^2 + \alpha)a_j - \sigma_j v_j^T f_X) \qquad \text{and} \qquad g_j''(a_j) = 2(\sigma_j^2 + \alpha) > 0$$

the function g_j is a convex parabola. The unique minimum a_j^* of g_j in (2.20) is given by

$$a_j^* = \frac{\sigma_j}{\sigma_j^2 + \alpha} v_j^T f_X \qquad \text{for all } 1 \le j \le r.$$

Therefore, for the unique solution b^* of (2.18) we have

$$b^* = W a^* = \sum_{j=1}^{r} \frac{\sigma_j}{\sigma_j^2 + \alpha} v_j^T f_X w_j.$$

From this we obtain

$$b^* \longrightarrow 0 \quad \text{for } \alpha \longrightarrow \infty$$

and

$$b^* \longrightarrow b_0^* := \sum_{j=1}^{r} \frac{1}{\sigma_j} v_j^T f_X w_j = C^+ f_X \quad \text{for } \alpha \searrow 0$$

for the asymptotic behaviour of b^*, where C^+ denotes the *pseudoinverse* of C. Therefore, b_0^* is the unique norm minimal solution of the linear least squares approximation problem

$$\|Cb - f_X\|_2^2 \longrightarrow \min_{b \in \mathbb{R}^m} !$$

Therefore, for the solution $c^* = A^{-1/2} b^*$ to (2.15) we get

$$c^* \longrightarrow 0 \quad \text{for } \alpha \longrightarrow \infty$$

and

$$c^* \longrightarrow c_0^* = A^{-1/2} b_0^* \quad \text{for } \alpha \searrow 0,$$

where $c_0^* \in \mathbb{R}^m$ denotes that solution of the linear least squares problem

$$\|Bc - f_X\|_2^2 \longrightarrow \min_{c \in \mathbb{R}^m}!$$

which minimizes the norm $\|\cdot\|_A$. For the solution $s_\alpha^* \in \mathcal{S}$ of (2.13), we obtain

$$s_\alpha^* \longrightarrow 0 \quad \text{for } \alpha \longrightarrow \infty$$

and

$$s_\alpha^* \longrightarrow s_0^* \quad \text{for } \alpha \searrow 0,$$

where $s_0^* \in \mathcal{S}$ is that solution for the linear least squares problem

$$\|s_X - f_X\|_2^2 \longrightarrow \min_{s \in \mathcal{S}}!$$

whose coefficients $c^* \in \mathbb{R}^m$ minimize the norm $\|\cdot\|_A$.

2.3 Interpolation by Algebraic Polynomials

In this section, we work with *algebraic polynomials* for the interpolation of a continuous function $f \in \mathscr{C}[a,b]$. Algebraic polynomials $p : \mathbb{R} \longrightarrow \mathbb{R}$ are often represented as linear combinations

$$p(x) = \sum_{k=0}^{n} a_k x^k = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{2.21}$$

of *monomials* $1, x, x^2, \ldots, x^n$ with coefficients $a_0, \ldots, a_n \in \mathbb{R}$. Recall that $n \in \mathbb{N}_0$ denotes the degree of p, provided that the leading coefficient $a_n \in \mathbb{R}$ does not vanish. We collect all algebraic polynomials of degree at most n in the linear space

$$\mathcal{P}_n := \operatorname{span}\{1, x, x^2, \ldots, x^n\} \qquad \text{for } n \in \mathbb{N}_0.$$

Now we consider the following interpolation problem.

Problem 2.4. Compute from a given set $X = \{x_0, \ldots, x_n\} \subset \mathbb{R}$ of $n+1$ pairwise distinct points and a data vector $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$ an algebraic polynomial $p \in \mathcal{P}_n$ satisfying $p_X = f_X$, i.e.,

$$p(x_j) = f_j \qquad \text{for all } 0 \le j \le n. \tag{2.22}$$

□

If we represent $p \in \mathcal{P}_n$ as a linear combination of monomials (2.21), then the interpolation conditions (2.22) lead to a linear system of the form

$$\begin{aligned} a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_n x_0^n &= f_0 \\ a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_n x_1^n &= f_1 \\ &\vdots \quad \vdots \\ a_0 + a_1 x_n + a_2 x_n^2 + \ldots + a_n x_n^n &= f_n, \end{aligned}$$

or, in the shorter matrix-vector notation,

$$V_X \cdot a = f_X \tag{2.23}$$

with coefficient vector $a = (a_0, \ldots, a_n)^T \in \mathbb{R}^{n+1}$ and the `Vandermonde`[4] `matrix`

$$V_X = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}. \tag{2.24}$$

Now the interpolation problem $p_X = f_X$ in (2.22) has a unique solution, if and only if the linear equation system (2.23) has a unique solution. Therefore, it remains to investigate the regularity of the Vandermonde matrix V_X, where we can rely on a classical result from linear algebra.

Theorem 2.5. *For the determinant of the Vandermonde matrix V_X, we have*

$$\det(V_X) = \prod_{0 \le j < k \le n} (x_k - x_j).$$

Proof. We prove the statement by induction on $n \in \mathbb{N}$, where we only use elementary properties of the determinant, in particular its linearity with respect to the matrix rows.

Initial step: For $n = 1$ we have $\det(V_X) = x_1 - x_0$ for $X = \{x_0, x_1\}$.

Induction hypothesis: Assume the statement is true for n points $\{x_1, \ldots, x_n\}$.

Induction step: For $X = \{x_0, x_1, \ldots, x_n\}$ we have

$$\det(V_X) = \det \begin{bmatrix} 1 & x_0 & \ldots & x_0^{n-1} & x_0^n \\ 1 & x_1 & \ldots & x_1^{n-1} & x_1^n \\ \vdots & \vdots & & \vdots & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} & x_n^n \end{bmatrix} = \det \begin{bmatrix} 1 & x_0 & \ldots & x_0^n \\ 0 & x_1 - x_0 & \ldots & x_1^n - x_0^n \\ \vdots & \vdots & & \vdots \\ 0 & x_n - x_0 & \ldots & x_n^n - x_0^n \end{bmatrix}$$

$$= \det \begin{bmatrix} x_1 - x_0 & \ldots & x_1^n - x_0^n \\ \vdots & & \vdots \\ x_n - x_0 & \ldots & x_n^n - x_0^n \end{bmatrix}$$

$$= \det \begin{bmatrix} x_1 - x_0 & x_1^2 - x_0 x_1 & \ldots & x_1^n - x_0 x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ x_n - x_0 & x_n^2 - x_0 x_n & \ldots & x_n^n - x_0 x_n^{n-1} \end{bmatrix}$$

$$= (x_1 - x_0) \cdot \ldots \cdot (x_n - x_0) \cdot \det \begin{bmatrix} 1 & x_1 & \ldots & x_1^{n-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \ldots & x_n^{n-1} \end{bmatrix}$$

[4] Alexandre-Théophile Vandermonde (1735-1796), French mathematician

$$= (x_1 - x_0) \cdot \ldots \cdot (x_n - x_0) \cdot \det(V_{X \setminus \{x_0\}})$$

$$= (x_1 - x_0) \cdot \ldots \cdot (x_n - x_0) \cdot \prod_{1 \le j < k \le n} (x_k - x_j)$$

$$= \prod_{0 \le j < k \le n} (x_k - x_j),$$

which already completes our proof. ■

We can conclude that for any set $X = \{x_0, x_1, \ldots, x_n\}$ of $n+1$ pairwise distinct interpolation points the Vandermonde matrix V_X is regular. This gives an answer to our initial question for the existence and uniqueness of a solution to Problem 2.4. We summarize our discussion on this as follows.

Corollary 2.6. *The interpolation problem (2.22), Problem 2.4, has a unique solution $p \equiv p_{f,X} \in \mathcal{P}_n$ whose coefficients with respect to its monomial representation in (2.21) are given by the solution of the linear system (2.23).* ■

We remark that the computation of the coefficient vector a in (2.23) is rather problematic. This is because heterogeneous distributions of interpolation points X typically yield ill-conditioned Vandermonde matrices V_X. Moreover, the computation of the coefficients in $a \in \mathbb{R}^{n+1}$ via (2.23) is too costly. Therefore, we prefer to avoid the linear system (2.23) for the solution of Problem 2.4, mainly for the sake of numerical stability and computational efficiency (see [28, Section 4.2]).

In fact, the above method for the solution to Problem 2.4 via (2.23) is rather naive. We remark that the solution to the interpolation problem (2.22) does not require a linear system at all. By choosing a suitable polynomial basis we can immediately give the solution to Problem 2.4.

To this end, we consider the `Lagrange`[5] `polynomials`

$$L_j(x) = \frac{(x - x_0) \cdot \ldots \cdot (x - x_{j-1}) \cdot (x - x_{j+1}) \cdot \ldots \cdot (x - x_n)}{(x_j - x_0) \cdot \ldots \cdot (x_j - x_{j-1}) \cdot (x_j - x_{j+1}) \cdot \ldots \cdot (x_j - x_n)}$$

$$= \prod_{\substack{k=0 \\ k \ne j}}^{n} \frac{x - x_k}{x_j - x_k} \qquad \text{for } 0 \le j \le n.$$

Any L_j is a polynomial of degree n, $L_j \in \mathcal{P}_n$, for $0 \le j \le n$, and we have

$$L_j(x_k) = \begin{cases} 1 & \text{for } k = j \\ 0 & \text{for } k \ne j \end{cases} \qquad \text{for all } 0 \le j, k \le n.$$

Therefore, the Lagrange polynomials $L_0, \ldots, L_n$ are a basis of the polynomial space $\mathcal{P}_n$. Moreover, the solution $p \in \mathcal{P}_n$ to the interpolation problem (2.22) is in its `Lagrange representation` given as

[5] Joseph-Louis Lagrange (1736-1813), mathematician and astronomer

$$p(x) = f_0 L_0(x) + \ldots + f_n L_n(x) = \sum_{j=0}^{n} f_j L_j(x). \tag{2.25}$$

Now let us make a first and very simple example.

Example 2.7. For two distinct interpolation points in $X = \{x_0, x_1\}$ the two corresponding Lagrange polynomials $L_0, L_1 \in \mathcal{P}_1$ are

$$L_0(x) = \frac{x_1 - x}{x_1 - x_0} \quad \text{and} \quad L_1(x) = \frac{x - x_0}{x_1 - x_0} \qquad \text{for } x \in \mathbb{R}.$$

Note that $L_0 + L_1 \equiv 1$. The Lagrange representation of the unique interpolation polynomial $p \in \mathcal{P}_1$ satisfying $p_X = f_X$, for given function values $f_X = (f_0, f_1)^T \in \mathbb{R}^2$, is given by the affine combination

$$p(x) = \frac{x_1 - x}{x_1 - x_0} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1 \qquad \text{for } x \in \mathbb{R}$$

of the function values f_0 and f_1. ♢

Now let us turn to a concrete interpolation problem.

Example 2.8. In this example, we consider interpolating the trigonometric function $f(x) = \cos(x)$ by a cubic polynomial for the set of interpolation points $X = \{0, \pi, 3\pi/2, 2\pi\}$. This yields the data vector $f_X = (1, -1, 0, 1)^T$, see Figure 2.2 (a). The cubic Lagrange polynomials for the points in X are

$$L_0(x) = \frac{(x-\pi)(x-3\pi/2)(x-2\pi)}{(0-\pi)(0-3\pi/2)(0-2\pi)} = -\frac{1}{3\pi^3} \cdot (x-\pi)\left(x - \frac{3}{2}\pi\right)(x-2\pi)$$

$$L_1(x) = \frac{(x-0)(x-3\pi/2)(x-2\pi)}{(\pi-0)(\pi-3\pi/2)(\pi-2\pi)} = \frac{2}{\pi^3} \cdot x\left(x - \frac{3}{2}\pi\right)(x-2\pi)$$

$$L_2(x) = \frac{(x-0)(x-\pi)(x-2\pi)}{(3\pi/2-0)(3\pi/2-\pi)(3\pi/2-2\pi)} = -\frac{8}{3\pi^3} \cdot x(x-\pi)(x-2\pi)$$

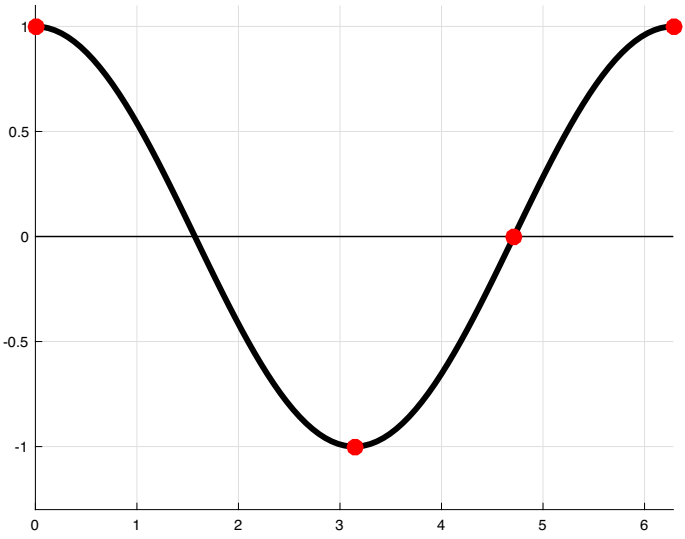
$$L_3(x) = \frac{(x-0)(x-\pi)(x-3\pi/2)}{(2\pi-0)(2\pi-\pi)(2\pi-3\pi/2)} = \frac{1}{\pi^3} \cdot x(x-\pi)\left(x - \frac{3}{2}\pi\right).$$

The function graphs of L_0, L_1, L_2, L_3 are shown in Figures 2.3 and 2.4.

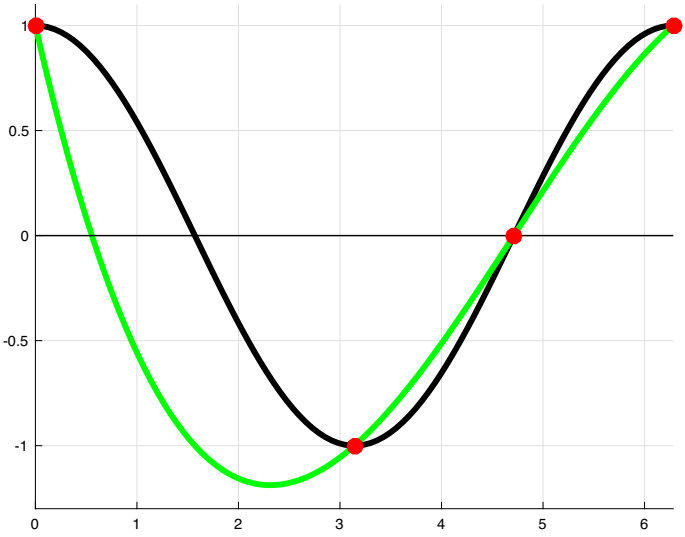
The unique solution to the interpolation problem $p_X = f_X$ is

$$\begin{aligned} p(x) &= L_0(x) - L_1(x) + L_3(x) \\ &= -\frac{1}{3\pi^3} \cdot (x-\pi)\left(x - \frac{3}{2}\pi\right)(x-2\pi) \\ &\quad - \frac{2}{\pi^3} \cdot x\left(x - \frac{3}{2}\pi\right)(x-2\pi) \\ &\quad + \frac{1}{\pi^3} \cdot x(x-\pi)\left(x - \frac{3}{2}\pi\right). \end{aligned}$$

The function graph of p is shown in Figure 2.2 (b). ♢

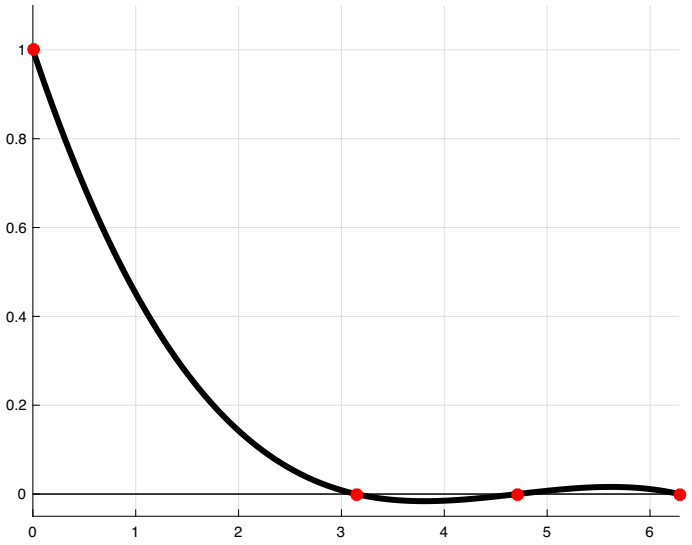


(a) $f(x) = \cos(x)$ and data X, f_X

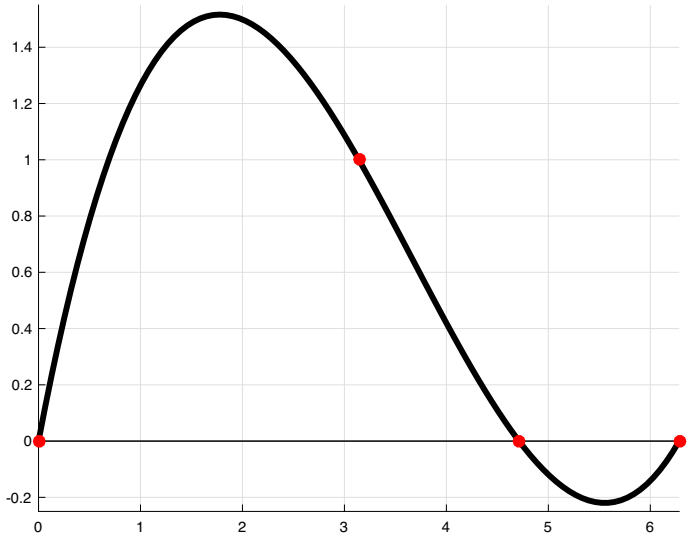


(b) Interpolant $p \in \mathcal{P}_3$ satisfying $p_X = f_X$

Fig. 2.2. For $X = \{0, \pi, 3\pi/2, 2\pi\}$ and $f_X = (1, -1, 0, 1)^T$ the cubic polynomial $p = L_0 - L_1 + L_3$ solves the interpolation problem $p_X = f_X$ from Example 2.8.

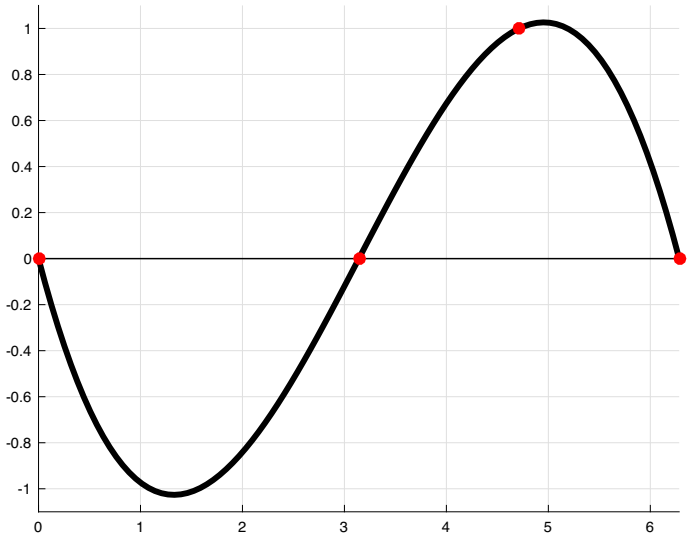


$$L_0(x) = -\tfrac{1}{3\pi^3} \cdot (x-\pi)(x-\tfrac{3}{2}\pi)(x-2\pi)$$

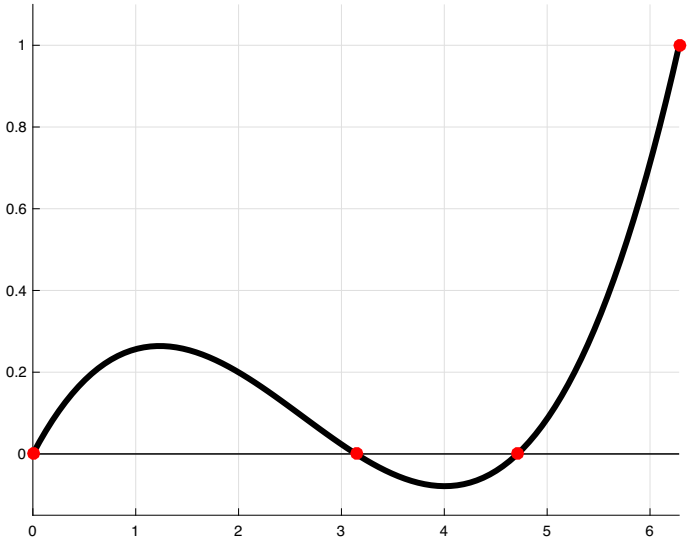


$$L_1(x) = \tfrac{2}{\pi^3} \cdot x(x-\tfrac{3}{2}\pi)(x-2\pi)$$

Fig. 2.3. The Lagrange polynomials $L_0, L_1 \in \mathcal{P}_3$ for $X = \{0, \pi, 3\pi/2, 2\pi\}$.



$$L_2(x) = -\tfrac{8}{3\pi^3} \cdot x(x-\pi)(x-2\pi)$$



$$L_3(x) = \tfrac{1}{\pi^3} \cdot x(x-\pi)(x-\tfrac{3}{2}\pi)$$

Fig. 2.4. The Lagrange polynomials $L_2, L_3 \in \mathcal{P}_3$ for $X = \{0, \pi, 3\pi/2, 2\pi\}$.

Although the Lagrange representation in (2.25) leads us directly to the solution of the interpolation problem (2.22), this *Lagrangian interpolation scheme* is not our preferred solution in practice. In fact, from a numerical viewpoint, the evaluation and updating of the interpolation polynomial in (2.25) is far too costly (see [28, Section 4.2] and [57, Section 8.2]).

An efficient and numerically stable method to evaluate interpolation polynomials is based on a recursive representation, which we explain in the following discussion. To this end, $p_{k,j} \in \mathcal{P}_j$ denotes, for $0 \le j \le k \le n$, the unique polynomial of degree at most j satisfying the interpolation conditions

$$p_{k,j}(x_\ell) = f_\ell \qquad \text{for all } k-j \le \ell \le k. \tag{2.26}$$

For fixed $x \in \mathbb{R}$ the values $p_{k,j}(x)$ can be computed recursively. This is done by using the *Aitken*[6] *lemma.*

Lemma 2.9. *For the interpolation polynomials $p_{k,j} \in \mathcal{P}_j$ satisfying (2.26) we have the recursion*

$$p_{k,0}(x) = f_k \quad \text{for } 0 \le k \le n$$

$$\begin{aligned} p_{k,j}(x) &= p_{k,j-1}(x) + \frac{x - x_k}{x_{k-j} - x_k}(p_{k-1,j-1}(x) - p_{k,j-1}(x)) \\ &= \frac{x_{k-j} - x}{x_{k-j} - x_k} p_{k,j-1}(x) + \frac{x - x_k}{x_{k-j} - x_k} p_{k-1,j-1}(x) \quad \text{for } k \ge j > 0. \end{aligned}$$

Proof. We prove the statement by induction on $j \ge 0$.

Initial step: For $j = 0$ we have $p_{k,0} \equiv f_k \in \mathcal{P}_0$, for $0 \le k \le n$.

Induction hypothesis: Suppose $p_{k,j-1} \in \mathcal{P}_{j-1}$ interpolates the data

$$(x_{k-j+1}, f_{k-j+1}), \ldots, (x_k, f_k)$$

and $p_{k-1,j-1} \in \mathcal{P}_{j-1}$ is the interpolation polynomial for the data

$$(x_{k-j}, f_{k-j}), \ldots, (x_{k-1}, f_{k-1}).$$

Induction step $(j-1 \longrightarrow j)$*:* Note that the right hand side of the recursion,

$$q(x) := p_{k,j-1}(x) + \frac{x - x_k}{x_{k-j} - x_k}(p_{k-1,j-1}(x) - p_{k,j-1}(x)),$$

is a polynomial of degree at most j, i.e., $q \in \mathcal{P}_j$. From the stated recursion and by using the induction hypothesis we can conclude that q, as well as $p_{k,j}$, interpolates the data

$$(x_{k-j}, f_{k-j}), \ldots, (x_k, f_k).$$

Therefore, we have $q \equiv p_{k,j}$ by uniqueness of the interpolant $p_{k,j}$. ∎

[6] Alexander Craig Aitken (1895-1967), New Zealand mathematician

By the recursion of the Aitken lemma, Lemma 2.9, we can, on given interpolation points X and function values f_X, recursively evaluate the unique interpolation polynomial $p \equiv p_{n,n} \in \mathcal{P}_n$ at any point $x \in \mathbb{R}$. To this end, we organize the values $p_{k,j} \equiv p_{k,j}(x)$, for $0 \le j \le k \le n$, in a triangular scheme as follows.

$$\begin{array}{lllll} f_0 = p_{0,0} & & & & \\ f_1 = p_{1,0} & p_{1,1} & & & \\ f_2 = p_{2,0} & p_{2,1} & p_{2,2} & & \\ \quad \vdots & \quad \vdots & \quad \vdots & \ddots & \\ f_n = p_{n,0} & p_{n,1} & p_{n,2} & \cdots & p_{n,n} \end{array}$$

The values in the first column of the triangular scheme are the given function values $p_{k,0} = f_k$, for $0 \le k \le n$. The values of the subsequent columns can be computed, according to the recursion in the Aitken lemma, from two values in the previous column. In this way, we can compute all entries of the triangular scheme, column-wise from left to right, and so we obtain the sought function value $p(x) = p_{n,n}$.

To compute the entry $p_{k,j}$ we merely need (besides the interpolation points x_{k-j} and x_k) the two entries $p_{k-1,j-1}$ and $p_{k,j-1}$ from the previous column. If we compute the entries in each column from the bottom to the top, then we can delete, in each step one entry, $p_{k,j-1}$, since $p_{k,j-1}$ is no longer needed in the subsequent computations.

This leads us to the *Neville*[7]*-Aitken algorithm*, Algorithm 1, giving a memory-efficient variant of the Aitken recursion in Lemma 2.9. The Neville-Aitken algorithm operates on the input data vector $f_X = (f_0, \ldots, f_n)^T$ recursively as shown in Algorithm 1.

Algorithm 1 Neville-Aitken algorithm

function NEVILLE-AITKEN(X,f_X, x)
  input: Interpolation points $X = \{x_0, \ldots, x_n\}$;
    Function values $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$;
    Evaluation point $x \in \mathbb{R}$ for $p \in \mathcal{P}_n$;
  for $j = 1, \ldots, n$ **do**
    for $k = n, \ldots, j$ **do**
      let

$$f_k := f_k + \frac{x - x_k}{x_{k-j} - x_k}(f_{k-1} - f_k);$$

    end for
  end for
  output: $p(x) = f_n$.
end function

[7] ERIC HAROLD NEVILLE (1889-1961), English mathematician

2.4 Divided Differences and the Newton Representation

Now we use the Aitken recursion in Lemma 2.9 to elaborate a suitable representation for interpolation polynomials. To this end, we consider for a fixed set $X = \{x_0, \ldots, x_n\}$ of interpolation points the *Newton*[8] *polynomials*

$$\omega_k(x) = \prod_{j=0}^{k-1}(x - x_j) \in \mathcal{P}_k \qquad \text{for } 0 \le k \le n. \tag{2.27}$$

For the Newton polynomials we have

$$\omega_k(x_\ell) = \begin{cases} 0 & \text{for } \ell < k, \\ \displaystyle\prod_{j=0}^{k-1}(x_\ell - x_j) \neq 0 & \text{for } \ell \ge k. \end{cases}$$

The Newton polynomials are obviously linearly independent, so that they are a basis for the polynomial space $\mathcal{P}_n$. Therefore, for any vector of function values $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$, the interpolation polynomial $p_n \in \mathcal{P}_n$ to f_X has *unique* `Newton coefficients` $b_0, b_1, \ldots, b_n \in \mathbb{R}$, so that

$$\begin{aligned} p_n(x) &= \sum_{k=0}^{n} b_k \omega_k(x) \\ &= b_0 + b_1(x - x_0) + \ldots + b_n(x - x_0) \cdot \ldots \cdot (x - x_{n-1}). \end{aligned} \tag{2.28}$$

The form of the polynomial p_n in (2.28) is called `Newton representation`.

Next, we turn to the computation of the Newton coefficients in (2.28). We start with the following scheme involving the function values of p_n on X.

$$\begin{aligned} f_0 &= p_n(x_0) = b_0 \\ f_1 &= p_n(x_1) = b_0 + b_1(x_1 - x_0) \\ f_2 &= p_n(x_2) = b_0 + b_1(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1) \\ &\vdots \quad \vdots \\ f_n &= p_n(x_n) = b_0 + \ldots + b_n(x_n - x_0) \cdot \ldots \cdot (x_n - x_{n-1}). \end{aligned}$$

Note that the Newton coefficients b_k of p_n can be determined recursively by

$$b_k = \frac{1}{\omega_k(x_k)} \left(f_k - \sum_{j=0}^{k-1} b_j \omega_j(x_k) \right) \qquad \text{for } k = 0, \ldots, n. \tag{2.29}$$

Further note that for the computation of b_k we only need the first $k+1$ data

[8] Sir Isaac Newton (1643-1727), English philosopher and scientist

$$(x_0, f_0), (x_1, f_1), \ldots, (x_k, f_k).$$

This gives the Newton representation an important advantage concerning updating: If we add one datum (x_{n+1}, f_{n+1}) to $X_n = \{x_0, \ldots, x_n\}$ and f_{X_n}, then it will be rather simple to update the interpolation polynomial p_n in (2.28). In fact, for the interpolation polynomial $p_{n+1} \in \mathcal{P}_{n+1}$ from data $X_{n+1} = \{x_0, \ldots, x_n, x_{n+1}\}$ and $f_{X_{n+1}}$ we have the representation

$$p_{n+1}(x) = p_n(x) + b_{n+1} \prod_{k=0}^{n} (x - x_k) = p_n(x) + b_{n+1}\omega_{n+1}(x),$$

where under the additional interpolation condition $p_{n+1}(x_{n+1}) = f_{n+1}$ we immediately get

$$b_{n+1} = \frac{f_{n+1} - p_n(x_{n+1})}{\omega_{n+1}(x_{n+1})}.$$

Now in order to develop a systematic scheme for computing the Newton coefficients of interpolation polynomials we introduce *divided differences*.

Definition 2.10. *On given data $X = \{x_0, \ldots, x_n\} \subset \mathbb{R}$ and $f_X \in \mathbb{R}^{n+1}$ let*

$$p(x) = \sum_{k=0}^{n} a_k x^k \in \mathcal{P}_n \tag{2.30}$$

be the unique interpolation polynomial satisfying $p_X = f_X$. Then, the leading coefficient $a_n \in \mathbb{R}$ of p in (2.30) is called the n-th `divided difference` *of f with respect to X, where we use the notation*

$$a_n = [x_0, \ldots, x_n](f). \tag{2.31}$$

The linear mapping $[x_0, \ldots, x_n] : \mathscr{C}(\mathbb{R}) \longrightarrow \mathbb{R}$ is referred to as the `divided difference operator`*, or the* `difference operator`*, with respect to X.* ◯

Before we discuss relevant properties of divided differences, we first make a remark for further clarification.

Remark 2.11. Note that the n-th divided difference $[x_0, \ldots, x_n](f)$ in Definition 2.10 is the leading coefficient of the interpolation polynomial p for f on X with respect to its *monomial representation* (2.30). We remark that the leading coefficient of p with respect to its *monomial representation* (2.30) coincides with the leading coefficient of p with respect to its *Newton representation* in (2.28) so that we have

$$p(x) = [x_0, \ldots, x_n](f)x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0 \tag{2.32}$$

with the coefficients $a_0, \ldots, a_{n-1}$ of the monomial representation of p and

$$p(x) = [x_0, \ldots, x_n](f)\omega_n(x) + b_{n-1}\omega_{n-1}(x) + \ldots + b_1\omega_1(x) + b_0 \tag{2.33}$$

with the coefficients $b_0, \ldots, b_{n-1}$ of the Newton representation. This property follows directly from the structure of the Newton polynomials $\omega_k \in \mathcal{P}_k$ in (2.27), with the leading Newton polynomial

$$\omega_n(x) = \prod_{j=0}^{n-1} (x - x_j) \in \mathcal{P}_n$$

as a product of the n linear factors $x - x_j$, for $0 \leq j \leq n-1$. Indeed, note that the leading coefficient of ω_n (with respect to the monomial basis) is one. Therefore, the leading coefficients of p, as in the Newton representation (2.33) and in the monomial representation (2.32), must be equal.

In hindsight, we could as well have introduced the n-th divided difference $[x_0, \ldots, x_n](f)$ in Definition 2.10 as the leading coefficient b_n of the interpolation polynomial p in its *Newton representation* in (2.28). Nevertheless, we have decided to follow the common standard from the literature. We finally remark that in the following recursive evaluation of $[x_0, \ldots, x_n](f)$ by *divided differences* (of a smaller order than n), the monomial representation (2.32) of the interpolation polynomial p will be quite useful. □

In our discussion so far, we have not made any assumptions on the ordering of the interpolation points in X. Since the interpolation polynomial p is, on given data f_X, always unique, we can conclude that the leading coefficient a_n of p in its monomial representation (2.30) is independent of the interpolation points' order in X. We formulate this observation as follows.

Proposition 2.12. *For $X = \{x_0, \ldots, x_n\}$ and $f_X \in \mathbb{R}^{n+1}$ the divided difference $[x_0, \ldots, x_n](f)$ is independent of the order of interpolation points $x_0, \ldots, x_n$ in X, i.e., for any permutation σ of the indices $\{0, \ldots, n\}$ we have*

$$[x_0, \ldots, x_n](f) = [x_{\sigma(0)}, \ldots, x_{\sigma(n)}](f).$$

■

As we show now, all coefficients in the Newton representation (2.28) of the interpolation polynomial p are divided differences.

Theorem 2.13. *For $X = \{x_0, \ldots, x_n\}$ and $f_X \in \mathbb{R}^{n+1}$,*

$$p(x) = \sum_{k=0}^{n} [x_0, \ldots, x_k](f) \cdot \omega_k(x) \in \mathcal{P}_n \tag{2.34}$$

is the unique interpolation polynomial satisfying $p_X = f_X$.

Proof. We prove the statement by induction on n.

Initial step: For $n = 0$ we have $p \equiv f_0 = [x_0](f)$ for $X = \{x_0\}$ and $f_0 \in \mathbb{R}$.

Induction hypothesis: Assume that, on given data $X = \{x_0, \ldots, x_{n-1}\}$ and $f_X \in \mathbb{R}^n$, $n \geq 1$,

$$p = \sum_{k=0}^{n-1} [x_0, \ldots, x_k](f) \cdot \omega_k \in \mathcal{P}_{n-1}$$

is the unique interpolation polynomial in $\mathcal{P}_{n-1}$ satisfying $p_X = f_X$.

Induction step $(n-1 \longrightarrow n)$*:* On given data $X = \{x_0, \ldots, x_n\}$ and $f_X \in \mathbb{R}^{n+1}$ the unique interpolation polynomial $p \in \mathcal{P}_n$ has, according to Remark 2.11, the representations (2.32) and

$$p(x) = [x_0, \ldots, x_n](f) \cdot \omega_n(x) + q_{n-1}(x) \tag{2.35}$$

with $q_{n-1} \in \mathcal{P}_{n-1}$, where the latter follows directly from (2.33). Since

$$q_{n-1}(x_k) = p(x_k) - [x_0, \ldots, x_n](f) \underbrace{\omega_n(x_k)}_{=0} = p(x_k),$$

for all $0 \leq k \leq n-1$, we see that the polynomial $q_{n-1} \in \mathcal{P}_{n-1}$ is the unique interpolation polynomial to f from $\mathcal{P}_{n-1}$ on the interpolation points $x_0, \ldots, x_{n-1}$. By the induction hypothesis, q_{n-1} has the representation

$$q_{n-1} = \sum_{k=0}^{n-1} [x_0, \ldots, x_k](f) \cdot \omega_k.$$

This in combination with (2.35) completes our proof already, since

$$p = [x_0, \ldots, x_n](f) \cdot \omega_n + \sum_{k=0}^{n-1} [x_0, \ldots, x_k](f) \cdot \omega_k = \sum_{k=0}^{n} [x_0, \ldots, x_k](f) \cdot \omega_k.$$

■

We finally turn to the computation of the divided differences. To this end, we rely on the Aitken recursion in Lemma 2.9.

Theorem 2.14. *For* $X = \{x_0, \ldots, x_n\}$ *and* $f_X \in \mathbb{R}^{n+1}$ *the recursion*

$$\begin{aligned} [x_j, \ldots, x_k](f) &= \frac{[x_{j+1}, \ldots, x_k](f) - [x_j, \ldots, x_{k-1}](f)}{x_k - x_j} && \text{for } 0 \leq j < k \leq n \\ [x_j](f) &= f(x_j) && \text{for } 0 \leq j \leq n \end{aligned}$$

holds.

Proof. For $n \geq k \geq j \geq 0$, let $p_{j,k} \in \mathcal{P}_{k-j}$ be the unique interpolation polynomial to f from $\mathcal{P}_{k-j}$ on the interpolation points $\{x_j, \ldots, x_k\}$. Moreover, for $k > j$, let $p_{j+1,k} \in \mathcal{P}_{k-j-1}$ be the unique interpolation polynomial to f on $\{x_{j+1}, \ldots, x_k\}$ and $p_{j,k-1} \in \mathcal{P}_{k-j-1}$ be the unique interpolation polynomial to f on $\{x_j, \ldots, x_{k-1}\}$.

Then, by the Aitken recursion in Lemma 2.9, we have the representation

$$p_{j,k}(x) = \frac{(x_j - x)p_{j+1,k}(x) - (x_k - x)p_{j,k-1}(x)}{x_j - x_k}. \tag{2.36}$$

If we compare the leading coefficients in (2.36), then we get

$$\begin{aligned}[x_j, \ldots, x_k](f) &= \frac{-[x_{j+1}, \ldots, x_k](f) + [x_j, \ldots, x_{k-1}](f)}{x_j - x_k} \\ &= \frac{[x_{j+1}, \ldots, x_k](f) - [x_j, \ldots, x_{k-1}](f)}{x_k - x_j}\end{aligned}$$

for $n \geq k > j \geq 0$. For $j = k$ we get $[x_j](f) = f(x_j)$, for $0 \leq j \leq n$. ■

Example 2.15. For $X = \{x_0, x_1\} \subset \mathbb{R}$ and $f_X = (f_0, f_1)^T \in \mathbb{R}^2$, the first order divided difference yields the difference quotient

$$[x_0, x_1](f) = \frac{[x_1](f) - [x_0](f)}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0}.$$

If f is differentiable at x_0, i.e., $f \in \mathscr{C}^1(x_0 - \varepsilon, x_0 + \varepsilon)$, for $\varepsilon > 0$, then we have

$$\lim_{x_1 \to x_0} [x_0, x_1](f) = f'(x_0).$$

Therefore, we allow coinciding interpolation points for $f \in \mathscr{C}^1$, where we let

$$[x, x](f) := f'(x).$$

♢

By the recursion in Theorem 2.14 we can view the n-th divided difference $[x_0, \ldots, x_n](f)$ as a discretization of the n-th derivative of $f \in \mathscr{C}^n$. We will be more precise on this observation later in this section.

Table 2.1. Organization of divided differences, on input data $X = \{x_0, \ldots, x_n\}$ and $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$, in a triangular scheme.

X	f_X				
x_0	f_0				
x_1	f_1	$[x_0, x_1](f)$			
x_2	f_2	$[x_1, x_2](f)$	$[x_0, x_1, x_2](f)$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	
x_n	f_n	$[x_{n-1}, x_n](f)$	$[x_{n-2}, x_{n-1}, x_n](f)$	$\cdots$	$[x_0, \ldots, x_n](f)$

On given points $X = \{x_0, \ldots, x_n\}$ and values $f_X = (f_0, \ldots, f_n)^T \in \mathbb{R}^{n+1}$ we can evaluate all divided differences $[x_j, \ldots, x_k](f)$, for $0 \leq j \leq k \leq n$,

by using the efficient and stable recursion of Theorem 2.14. To this end, we organize the divided differences in a triangular scheme, as shown in Table 2.1.

The organization of the data in Table 2.1 reminds us of the triangular scheme of the Neville-Aitken algorithm, Algorithm 1. In fact, to compute the Newton coefficients $[x_0, \ldots, x_k](f)$ in (2.34), we can (similarly as in Algorithm 1) process the data in Table 2.1 by a memory-efficient algorithm operating only on the data vector $f_X = (f_0, \ldots, f_n)^T$, see Algorithm 2.

Algorithm 2 Computation of Newton coefficients $[x_0, \ldots, x_k](f)$

function DIVIDED DIFFERENCES(X,f_X)
  input: interpolation points $X = \{x_0, x_1, \ldots, x_n\}$;
    function values $f_X = (f_0, f_1, \ldots, f_n)^T \in \mathbb{R}^{n+1}$;
  for $j = 1, \ldots, n$ **do**
    for $k = n, \ldots, j$ **do**
      let
$$f_k := \frac{f_k - f_{k-1}}{x_k - x_{k-j}};$$
    end for
  end for
  output: $(f_0, \ldots, f_n) = ([x_0](f), [x_0, x_1](f), \ldots, [x_0, \ldots, x_n](f))^T \in \mathbb{R}^{n+1}$.
end function

For further illustration, we make an example that is linked to Example 2.8.

Example 2.16. We consider interpolating the function $f(x) = \cos(x)$ on interpolation points $X_3 = \{0, \pi, 3\pi/2, 2\pi\}$. By $f_{X_3} = (1, -1, 0, 1)$ we get the following divided differences in the triangular scheme of Table 2.1, for $n = 3$.

X_3	f_{X_3}			
0	1			
π	-1	$-\frac{2}{\pi}$		
$\frac{3}{2}\pi$	0	$\frac{2}{\pi}$	$\frac{8}{3\pi^2}$	
2π	1	$\frac{2}{\pi}$	0	$-\frac{4}{3\pi^3}$

The Newton polynomials $\omega_0, \ldots, \omega_3$ for the point set X_3 are given by

$$\omega_0 \equiv 1, \quad \omega_1(x) = x, \quad \omega_2(x) = x(x-\pi), \quad \omega_3(x) = x(x-\pi)\left(x - \frac{3}{2}\pi\right).$$

Therefore, the cubic polynomial

$$p_3(x) = 1 - \frac{2}{\pi}x + \frac{8}{3\pi^2}x(x-\pi) - \frac{4}{3\pi^3}x(x-\pi)\left(x - \frac{3}{2}\pi\right) \tag{2.37}$$

is the unique interpolation polynomial in $\mathcal{P}_3$ satisfying $p_{X_3} = f_{X_3}$.

The leading coefficient of the interpolation polynomial p_3 in its Newton representation (2.37) coincides with that of its monomial representation (see Remark 2.11). The leading coefficient of p_3 with respect to its monomial representation can also be obtained by the sum of the coefficients of its Lagrange representation (see Example 2.8), i.e.,

$$-\frac{1}{3\pi^3} - \frac{2}{\pi^3} + \frac{1}{\pi^3} = -\frac{4}{3\pi^3}.$$

On the downside, the approximation quality of the cubic interpolation polynomial p_3 for f on X_3 is rather bad, see Figure 2.5 (a), where we find $\|p_3 - f\|_{\infty,[0,2\pi]} \approx 1.1104$ for the approximation error. To improve on the approximation quality we add one interpolation point $x_4 = \pi/4$ and so we obtain $X_4 = \{0, \pi, 3\pi/2, 2\pi, \pi/4\}$ for the updated set of interpolation points and $f_{X_4} = (1, -1, 0, 1, 1/\sqrt{2})$ for the updated data vector of function values. To compute the interpolation polynomial $p_4 \in \mathcal{P}_4$ we update the triangular scheme (see Table 2.1, for $n = 4$) as follows.

X_4	f_{X_4}				
0	1				
π	-1	$-\frac{2}{\pi}$			
$\frac{3}{2}\pi$	0	$\frac{2}{\pi}$	$\frac{8}{3\pi^2}$		
2π	1	$\frac{2}{\pi}$	0	$-\frac{4}{3\pi^3}$	
$\frac{\pi}{4}$	$\frac{1}{\sqrt{2}}$	$-\frac{4(1-\sqrt{2})}{7\sqrt{2}\pi}$	$\frac{8(2+5\sqrt{2})}{35\sqrt{2}\pi^2}$	$-\frac{32(2+5\sqrt{2})}{105\sqrt{2}\pi^3}$	$-\frac{16(16+5\sqrt{2})}{105\sqrt{2}\pi^4}$

Therefore, the *quartic* (i.e., degree four) polynomial

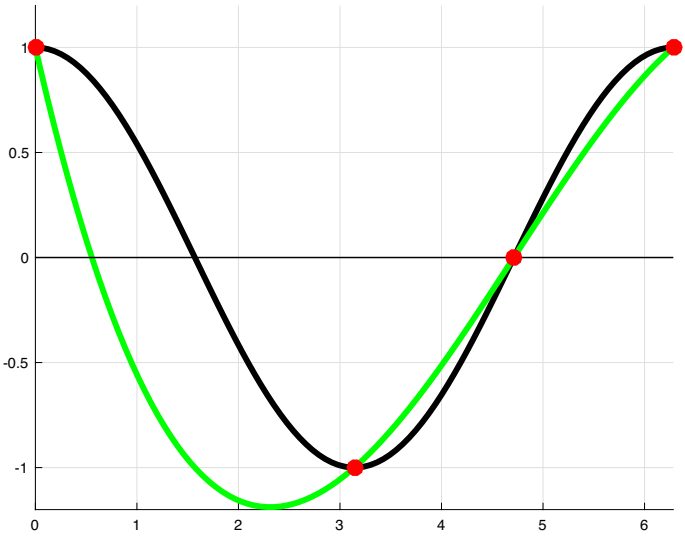
$$p_4(x) = p_3(x) - \frac{16(16+5\sqrt{2})}{105\sqrt{2}\pi^4} x(x-\pi)\left(x - \frac{3}{2}\pi\right)(x-2\pi)$$

is the unique interpolation polynomial in $\mathcal{P}_4$ satisfying $p_{X_4} = f_{X_4}$, where the approximation error $\|p_4 - f\|_{\infty,[0,2\pi]} \approx 0.0736$ of $p_4 \in \mathcal{P}_4$ is much smaller than that of $p_3 \in \mathcal{P}_3$, see Figure 2.5 (b). ♢

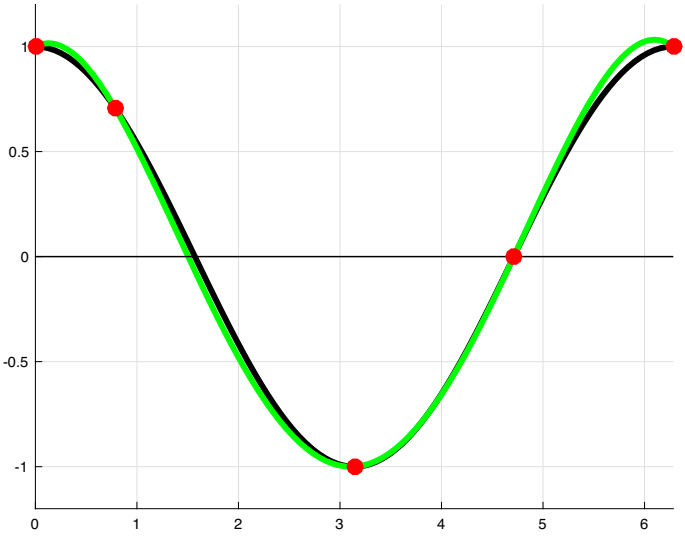
Next, we develop a very useful representation for divided differences, termed the *Hermite*[9]*-Genocchi*[10] *formula*, whereby divided differences can be viewed as mean values of derivatives of f over a simplex spanned by the interpolation points. In the following formulation for the Hermite-Genocchi formula, we regard the n-dimensional `standard simplex`

$$\Delta_n = \left\{ (\lambda_1, \ldots, \lambda_n)^T \in \mathbb{R}^n \,\middle|\, \lambda_k \geq 0 \text{ for } 1 \leq k \leq n \text{ and } \sum_{k=1}^{n} \lambda_k \leq 1 \right\}. \quad (2.38)$$

[9] Charles Hermite (1822-1901), French mathematician
[10] Angelo Genocchi (1817-1889), Italian mathematician



(a) p_3 with approximation error $\|p_3 - f\|_{\infty,[0,2\pi]} \approx 1.1104$



(b) p_4 with approximation error $\|p_4 - f\|_{\infty,[0,2\pi]} \approx 0.0736$

Fig. 2.5. (a) The cubic polynomial $p_3 \in \mathcal{P}_3$ interpolates the trigonometric function $f(x) = \cos(x)$ on $X_3 = \{0, \pi, 3\pi/2, 2\pi\}$. (b) The quartic polynomial $p_4 \in \mathcal{P}_4$ interpolates $f(x) = \cos(x)$ on $X_4 = \{0, \pi, 3\pi/2, 2\pi, \pi/4\}$ (see Example 2.16).

Theorem 2.17. *For $f \in \mathscr{C}^n$, $n \geq 1$, the* `Hermite-Genocchi formula`

$$[x_0, \ldots, x_n](f) = \int_{\Delta_n} f^{(n)} \left(x_0 + \sum_{k=1}^{n} \lambda_k (x_k - x_0) \right) \mathrm{d}\lambda,$$

holds, where Δ_n is the n-dimensional standard simplex (2.38) in $\mathbb{R}^n$.

Proof. We prove the Hermite-Genocchi formula by induction on n.
Initial step: For $n = 1$, we have $\Delta_1 = [0, 1]$ and so

$$\int_{\Delta_1} f' \left(x_0 + \lambda_1 (x_1 - x_0) \right) \mathrm{d}\lambda_1 = \frac{1}{x_1 - x_0} (f(x_1) - f(x_0)) = [x_0, x_1](f).$$

Induction hypothesis: Suppose the Hermite-Genocchi formula holds for $n \geq 1$.
Induction step $(n - 1 \longrightarrow n)$*:* For $\mathrm{d}\lambda = \mathrm{d}\lambda_1 \cdots \mathrm{d}\lambda_{n-1}$ we have

$$\begin{aligned}
&\int_{\Delta_n} f^{(n)} \left(x_0 + \sum_{k=1}^{n} \lambda_k (x_k - x_0) \right) \mathrm{d}\lambda \, \mathrm{d}\lambda_n \\
&= \int_{\Delta_{n-1}} \int_0^{1 - \sum_{k=1}^{n-1} \lambda_k} f^{(n)} \left(x_0 + \sum_{k=1}^{n-1} \lambda_k (x_k - x_0) + \lambda_n (x_n - x_0) \right) \mathrm{d}\lambda_n \, \mathrm{d}\lambda \\
&= \frac{1}{x_n - x_0} \int_{\Delta_{n-1}} \left[\int_{x_0 + \sum_{k=1}^{n-1} \lambda_k (x_k - x_0)}^{x_n + \sum_{k=1}^{n-1} \lambda_k (x_k - x_n)} f^{(n)}(z) \, \mathrm{d}z \right] \mathrm{d}\lambda \\
&= \frac{1}{x_n - x_0} \left[\int_{\Delta_{n-1}} f^{(n-1)} \left(x_n + \sum_{k=1}^{n-1} \lambda_k (x_k - x_n) \right) \mathrm{d}\lambda \right. \\
&\qquad \left. - \int_{\Delta_{n-1}} f^{(n-1)} \left(x_0 + \sum_{k=1}^{n-1} \lambda_k (x_k - x_0) \right) \mathrm{d}\lambda \right] \\
&= \frac{1}{x_n - x_0} \left([x_n, x_1, \ldots, x_{n-1}](f) - [x_0, \ldots, x_{n-1}](f) \right) \\
&= \frac{1}{x_n - x_0} \left([x_1, \ldots, x_n](f) - [x_0, \ldots, x_{n-1}](f) \right) \\
&= [x_0, \ldots, x_n](f).
\end{aligned}$$

■

Now we can state further properties of divided differences. The following results are direct consequences from the Hermite-Genocchi formula, Theorem 2.17, and the standard mean value theorem of integration.

Corollary 2.18. *The divided differences satisfy the following properties.*

(a) For $f \in \mathscr{C}^n$, $n \geq 0$, we have

$$[x_0, \ldots, x_n](f) = \frac{f^{(n)}(\tau)}{n!} \qquad \text{for some } \tau \in [x_{\min}, x_{\max}],$$

where $x_{\min} = \min_{0 \leq k \leq n} x_k$ and $x_{\max} = \max_{0 \leq k \leq n} x_k$.
For $x_0 = \ldots = x_n$, we have

$$[x_0, \ldots, x_n](f) = \frac{f^{(n)}(x_0)}{n!}.$$

(b) For $p \in \mathcal{P}_{n-1}$, we have $[x_0, \ldots, x_n](p) = 0$ for $n \geq 1$.

■

The discretization of higher order derivatives by divided differences is consistent with the standard product rule of differentiation. We show this by proving the *Leibniz*[11] *rule.*

Corollary 2.19. *For arbitrary points $x_0, \ldots, x_n$ and $f, g \in \mathscr{C}^n$, $n \in \mathbb{N}_0$, the* `Leibniz formula`

$$[x_0, \ldots, x_n](f \cdot g) = \sum_{j=0}^{n} [x_0, \ldots, x_j](f) \cdot [x_j, \ldots, x_n](g) \tag{2.39}$$

holds.

Proof. Suppose that $X = \{x_0, \ldots, x_n\}$ is a set of *pairwise distinct* points. Moreover, let $p_f \in \mathcal{P}_n$ be the unique interpolation polynomial for f on X and $p_g \in \mathcal{P}_n$ be the unique interpolation polynomial for g on X. Then, p_f and p_g have the representations

$$p_f = \sum_{k=0}^{n} [x_0, \ldots, x_k](f)\omega_k \qquad \text{and} \qquad p_g = \sum_{j=0}^{n} [x_j, \ldots, x_n](g)\widetilde{\omega}_j$$

with the Newton polynomials

$$\omega_k(x) = \prod_{\ell=0}^{k-1} (x - x_\ell) \in \mathcal{P}_k \quad \text{and} \quad \widetilde{\omega}_j(x) = \prod_{m=j+1}^{n} (x - x_m) \in \mathcal{P}_{n-j}$$

for $0 \leq k, j \leq n$, where we have used the independence of the divided differences on the order of the interpolation points in X (cf. Proposition 2.12).

Now the product

[11] Gottfried Wilhelm Leibniz (1646-1716), German philosopher and scientist

$$p := p_f \cdot p_g = \sum_{k,j=0}^{n} [x_0, \ldots, x_k](f)\, \omega_k \cdot [x_j, \ldots, x_n](g)\, \widetilde{\omega}_j \tag{2.40}$$

interpolates the function $f \cdot g$ on X. For the Newton polynomials ω_k and $\widetilde{\omega}_j$ we have

$$\omega_k(x_i) \cdot \widetilde{\omega}_j(x_i) = 0 \qquad \text{for all } 0 \le i \le n,$$

for $k > j$, so that the polynomial p in (2.40) has the representation

$$p = \sum_{\substack{k,j=0 \\ k \le j}}^{n} [x_0, \ldots, x_k](f) \cdot [x_j, \ldots, x_n](g)\, \omega_k \cdot \widetilde{\omega}_j\,.$$

Since $\omega_k \cdot \widetilde{\omega}_j \in \mathcal{P}_{n+k-j}$, for all $0 \le k, j \le n$, we have $p \in \mathcal{P}_n$. Therefore, p is the unique interpolation polynomial in $\mathcal{P}_n$ for $f \cdot g$ on X, and so we obtain the stated representation

$$[x_0, \ldots, x_n](f \cdot g) = \sum_{j=0}^{n} [x_0, \ldots, x_j](f) \cdot [x_j, \ldots, x_n](g) \tag{2.41}$$

for the case of *pairwise distinct* points $x_0, \ldots, x_n$.

By the Hermite-Genocchi formula, Theorem 2.17, the representation

$$[x_0, \ldots, x_m](h) = \int_{\Delta_m} h^{(m)} \left(x_0 + \sum_{k=1}^{m} \lambda_k (x_k - x_0) \right) \mathrm{d}\lambda \tag{2.42}$$

holds for $h \in \mathscr{C}^m$. Therefore, the divided differences $[x_0, \ldots, x_m](h)$ are for $h \in \mathscr{C}^m$ continuous in X, since the integrand $h^{(m)}$ in (2.42) is continuous in X. Since $f \cdot g \in \mathscr{C}^n$, we can conclude that the representation (2.41) holds for arbitrary point sets $X = \{x_0, \ldots, x_n\}$. ■

Remark 2.20. For coincident points $x_0 = \ldots = x_n$, the Leibniz formula (2.39), in combination with Corollary 2.18 (a), yields the identity

$$\frac{(f \cdot g)^{(n)}(x_0)}{n!} = \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!} \cdot \frac{g^{(n-j)}(x_0)}{(n-j)!}$$

and so

$$\begin{aligned} (f \cdot g)^{(n)}(x_0) &= \sum_{j=0}^{n} \frac{n!}{j!\,(n-j)!}\, f^{(j)}(x_0)\, g^{(n-j)}(x_0) \\ &= \sum_{j=0}^{n} \binom{n}{j} f^{(j)}(x_0)\, g^{(n-j)}(x_0), \end{aligned}$$

which is the standard product rule for higher derivatives. □

From Corollary 2.18 (a) we see that divided differences are also well-defined for *coincident* interpolation points, provided that f has sufficiently many derivatives. In particular, for the case of coincident interpolation points, all coefficients in the Newton representation (2.34) are in this case well-defined (cf. Example 2.15). Now we extend the problem of *Lagrange interpolation*, Problem 2.4, to the problem of *Hermite interpolation*. In the case of Hermite interpolation, the interpolation conditions contain not only point evaluations of f, but also derivative values of f. In this case, we require coincident interpolation points. To be more precise, we formulate the Hermite interpolation problem as follows.

Problem 2.21. Let $X = \{x_0, \ldots, x_n\}$ be a set of $n+1$ pairwise distinct interpolation points. Moreover, suppose we are given $N = \mu_0 + \mu_1 + \ldots + \mu_n$ *Hermite data*

$$f^{(\ell)}(x_k) \qquad \text{for } 0 \le \ell < \mu_k \text{ and } 0 \le k \le n \tag{2.43}$$

for $f \in \mathscr{C}^{m-1}$, where $m = \max_k \mu_k$ and $\mu_k \in \mathbb{N}$ for $0 \le k \le n$.

Then, the Hermite interpolation problem for (2.43) requires determining a polynomial $p \in \mathcal{P}_{N-1}$ satisfying the Hermite interpolation conditions

$$p^{(\ell)}(x_k) = f^{(\ell)}(x_k) \qquad \text{for } 0 \le \ell < \mu_k \text{ and } 0 \le k \le n. \tag{2.44}$$

□

Note that *Lagrange interpolation*, Problem 2.4, is by $\mu_k = 1$, $0 \le k \le n$, and $N = n+1$ a special case of Hermite interpolation. Further note that the Hermite data in (2.43) necessarily need to contain, for every interpolation point $x_k \in X$, *all* derivatives

$$f(x_k), f'(x_k), \ldots, f^{(\mu_k - 1)}(x_k) \qquad \text{for } k = 0, \ldots, n.$$

In the following solution to *Hermite interpolation*, Problem 2.21, we first add interpolation points to X, such that the resulting point set Y contains the interpolation points x_k multiple times, namely according to its multiplicity μ_k of the Hermite data in (2.44). This leads us to the *extended point set*

$$Y = \{\underbrace{x_0, \ldots, x_0}_{\mu_0\text{-fold}}, \underbrace{x_1, \ldots, x_1}_{\mu_1\text{-fold}}, \ldots, \underbrace{x_n, \ldots, x_n}_{\mu_n\text{-fold}}\} = \{y_0, \ldots, y_{N-1}\} \tag{2.45}$$

containing $N = \mu_0 + \mu_1 + \ldots + \mu_n$ interpolation points (including their multiplicities), where $x_0 = y_0 = \ldots = y_{\mu_0 - 1}$ and

$$x_k = y_{\mu_0 + \ldots + \mu_{k-1}} = \ldots = y_{\mu_0 + \ldots + \mu_k - 1} \qquad \text{for } 1 \le k \le n.$$

We can solve the Hermite interpolation problem, Problem 2.21, as follows.

Theorem 2.22. *The problem of Hermite interpolation, Problem 2.21, has a unique solution $p \in \mathcal{P}_{N-1}$. For the extended set of interpolation points $Y = \{y_0, \ldots, y_{N-1}\}$ in (2.45) and the divided differences*

$$[y_0, \ldots, y_k](f) \qquad \text{for } 0 \le k < N$$

the interpolation polynomial p has the Newton representation

$$p(x) = \sum_{k=0}^{N-1} [y_0, \ldots, y_k](f)\omega_k(x). \tag{2.46}$$

Proof. The linear mapping $L : \mathcal{P}_{N-1} \longrightarrow \mathbb{R}^N$, defined as

$$p \longmapsto L(p) = (p(x_0), \ldots, p^{(\mu_0-1)}(x_0), \ldots, p(x_n), \ldots, p^{(\mu_n-1)}(x_n))^T \in \mathbb{R}^N,$$

is injective, due to the fundamental theorem of algebra. Further, due to the dimension formula of linear algebra, L is also surjective, and so L is bijective.

The Newton representation (2.46) for p follows directly from our solution (2.34) to Lagrange interpolation, Problem 2.4, which holds in particular for the case of coincident interpolation points (with using our results on divided differences for coincident interpolation points). ■

Again, we can organize the divided differences of the Newton representation (2.46) in a triangular scheme (as in Table 2.1). Moreover, we can use the recursion of Algorithm 2 to compute the scheme's entries, where for the case of coincident interpolation points $y_k = y_{k-j}$ (see line 6 in Algorithm 2) we insert the corresponding derivative value $f^{(j)}(y_k)/j!$.

For further illustration, we finally discuss the following example.

Example 2.23. We consider interpolating the *sinc function* $f(x) = \sin(x)/x$. We have

$$f'(x) = \frac{x\cos(x) - \sin(x)}{x^2} \quad \text{and} \quad f''(x) = \frac{2\sin(x) - 2x\cos(x) - x^2\sin(x)}{x^3}.$$

For the interpolation of f, we work with the Hermite data

$$f(0) = 1,\ f'(0) = 0,\ f(\pi) = 0,\ f'(\pi) = -\frac{1}{\pi},\ f''(\pi) = \frac{2}{\pi^2},\ f(2\pi) = 0.$$

This gives the extended set of interpolation points $Y = \{0, 0, \pi, \pi, \pi, 2\pi\}$.

We display the divided differences of the Newton representation (2.46) in a triangular scheme (as in Table 2.1 for $n = 5$) as follows, where we mark the inserted derivative values $f^{(j)}(y_k)/j!$ by a box, respectively.

Y	f_Y					
0	1					
0	1	$\boxed{0}$				
π	0	$-\frac{1}{\pi}$	$-\frac{1}{\pi^2}$			
π	0	$\boxed{-\frac{1}{\pi}}$	0	$\frac{1}{\pi^3}$		
π	0	$\boxed{-\frac{1}{\pi}}$	$\boxed{\frac{1}{\pi^2}}$	$\frac{1}{\pi^3}$	0	
2π	0	0	$\frac{1}{\pi^2}$	0	$-\frac{1}{2\pi^4}$	$-\frac{1}{4\pi^5}$

Given the above divided differences, we see that the polynomial

$$p_5(x) = 1 - \frac{1}{\pi^2}x^2 + \frac{1}{\pi^3}x^2(x-\pi) - \frac{1}{4\pi^5}x^2(x-\pi)^3 \in \mathcal{P}_5$$

is the unique solution to the posed Hermite interpolation problem. ♢

2.5 Error Estimates and Optimal Interpolation Points

In this section, we develop error estimates, i.e., upper bounds on the difference

$$f(x) - p(x) \qquad \text{for } x \in [a,b] \tag{2.47}$$

between f and the interpolation polynomial p. In the following discussion, we regard the problem of Lagrange interpolation, Problem 2.4, as a special case of Hermite interpolation, Problem 2.21. To unify the notations of Problems 2.4 and 2.21 we denote by $Y = \{y_0, \ldots, y_{N-1}\} \subset [a,b]$ the extended set of interpolation points, where for the case of Hermite interpolation we allow coincident interpolation points, according to Problem 2.21 and as in (2.45). We denote the unique solution to the Hermite interpolation problem by p_{N-1}. In particular, the Newton representation (2.46) holds for $p_{N-1} \in \mathcal{P}_{N-1}$.

We can represent the error in (2.47) as follows.

Theorem 2.24. *Let $p_{N-1} \in \mathcal{P}_{N-1}$ be the solution to the Hermite interpolation problem, Problem 2.21. Then we have the pointwise error representation*

$$f(x) - p_{N-1}(x) = [y_0, \ldots, y_{N-1}, x](f) \prod_{k=0}^{N-1} (x - y_k) \qquad \text{for } x \in \mathbb{R}. \tag{2.48}$$

Proof. The error representation in (2.48) is obviously fulfilled for any $x \in Y$. Indeed, in this case, we have $f(x) = p_{N-1}(x)$, and, moreover, the Newtonian knot polynomial

$$\omega_Y(x) := \prod_{k=0}^{N-1} (x - y_k) \tag{2.49}$$

vanishes at every interpolation point from Y.

Now for $x \in \mathbb{R}\backslash Y$, we extend Y by the interpolation point x. Moreover, we let $p_N \in \mathcal{P}_N$ denote the unique polynomial in $\mathcal{P}_N$ which satisfies the Hermite conditions (2.44) and the additional interpolation condition $p_N(x) = f(x)$. In this case, we have the representation

$$p_N(x) = p_{N-1}(x) + [y_0, \ldots, y_{N-1}, x](f) \prod_{k=0}^{N-1} (x - y_k)$$

and so

$$\begin{aligned} f(x) - p_{N-1}(x) &= f(x) - \left(p_N(x) - [y_0, \ldots, y_{N-1}, x](f) \prod_{k=0}^{N-1} (x - y_k) \right) \\ &= [y_0, \ldots, y_{N-1}, x](f) \prod_{k=0}^{N-1} (x - y_k). \end{aligned}$$

■

Theorem 2.24 immediately yields the following upper bound for the interpolation error $f - p$ in (2.47) on the interval $[a, b]$, where we combine the representation in (2.48) with the result of Corollary 2.18 (a).

Corollary 2.25. *Let $p \in \mathcal{P}_{N-1}$ denote the unique solution to the Hermite interpolation problem, Problem 2.21. Then we have for $f \in \mathscr{C}^N$ the pointwise error estimate*

$$|f(x) - p(x)| \leq \frac{1}{N!} \max_{\xi \in [a,b]} |f^{(N)}(\xi)| \cdot \left| \prod_{k=0}^{N-1} (x - y_k) \right| \tag{2.50}$$

in $x \in [a, b]$. ■

As a direct consequence of Corollary 2.25, the *uniform* error estimate

$$\|f - p\|_\infty \leq \frac{\|f^{(N)}\|_\infty}{N!} \cdot \|\omega_Y\|_\infty \qquad \text{for } f \in \mathscr{C}^N[a, b] \tag{2.51}$$

follows from the *pointwise* error estimate in (2.50) for any compact interval $[a, b] \subset \mathbb{R}$ containing the set of interpolation points Y, i.e., $Y \subset [a, b]$.

To reduce the interpolation error in (2.51), we wish to minimize the maximum norm $\|\omega_Y\|_\infty$ of the knot polynomial ω_Y under variation of the interpolation points in $Y \subset [a, b]$. Without loss of generality, we restrict ourselves to the interval $[a, b] = [-1, 1]$. This immediately leads us to the *nonlinear* optimization problem

$$\|\omega_X\|_{\infty,[-1,1]} \longrightarrow \min_{\substack{X \subset [-1,1] \\ |X| = n+1}} ! \tag{2.52}$$

As we show in this section, the minimization problem in (2.52) has a unique solution $X^* \subset [-1,1]$ consisting of $n+1$ *pairwise distinct* interpolation points. This explains our chosen notation $X = Y$ and $n = N-1$, which is in accordance with the Lagrange interpolation problem, Problem 2.4. We formulate the minimization problem in (2.52) more precisely as follows.

Problem 2.26. Determine for $n \in \mathbb{N}_0$ a set $X^* = \{x_0^*, \ldots, x_n^*\} \subset [-1,1]$ of $n+1$ interpolation points, which minimizes the maximum norm of the corresponding knot polynomial ω_{X^*} on $[-1,1]$, so that the upper bound

$$\|\omega_{X^*}\|_{\infty,[-1,1]} \leq \|\omega_X\|_{\infty,[-1,1]} \tag{2.53}$$

holds for all point sets $X = \{x_0, \ldots, x_n\} \subset [-1,1]$ of size $|X| = n+1$. □

For the solution of the minimization problem, Problem 2.26, we work with the `Chebyshev polynomials`

$$T_n(x) = \cos(n \arccos(x)) \qquad \text{for } n \in \mathbb{N}_0, \tag{2.54}$$

where in the subsequent discussion we rely on their following properties.

Theorem 2.27. *The Chebyshev polynomials are generated by the recursion*

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad \text{for } n \in \mathbb{N} \tag{2.55}$$

with the initial values $T_0 \equiv 1$ *and* $T_1(x) = x$.

Proof. The initial values $T_0 \equiv 1$ and $T_1(x) = x$ are obviously consistent with Definition (2.54). For $\phi = \arccos(x)$, we find the representation

$$\cos((n+1)\phi) = 2\cos(\phi)\cos(n\phi) - \cos((n-1)\phi)$$

from standard trigonometric identities, which implies the recursion (2.55). ■

Corollary 2.28. *For* $n \in \mathbb{N}_0$, *the Chebyshev polynomial* T_{n+1} *is an algebraic polynomial of degree* $n+1$ *with leading coefficient* 2^n, *so that an identity of the form*

$$T_{n+1}(x) = 2^n x^{n+1} + q_n(x) \tag{2.56}$$

holds for some $q_n \in \mathcal{P}_n$.

Proof. We prove the identity (2.56) by induction on $n \in \mathbb{N}_0$. For $n = 0$, the statement is trivial. Under the induction hypothesis for $n \in \mathbb{N}_0$, the statement in (2.56) follows, for $n+1$, directly from the recursion in (2.55). ■

Corollary 2.29. *The* $n+1$ *zeros of the Chebyshev polynomial* $T_{n+1} \in \mathcal{P}_{n+1}$ *are, for* $n \in \mathbb{N}_0$, *given by the* `Chebyshev knots`

$$x_k^* = \cos\left(\frac{2k+1}{2n+2}\pi\right) \in [-1,1] \qquad \text{for } 0 \leq k \leq n. \tag{2.57}$$

Moreover, all extrema of T_{n+1} *on* $[-1,1]$ *are attained at the* $n+2$ *points*

$$y_k = \cos\left(\frac{k}{n+1}\pi\right) \in [-1,1] \qquad \text{for } 0 \leq k \leq n+1. \tag{2.58}$$

Proof. For $0 \le k \le n$, we have

$$\begin{aligned} T_{n+1}(x_k^*) &= \cos((n+1)\arccos(x_k^*)) \\ &= \cos\left((n+1)\frac{2k+1}{2(n+1)}\pi\right) = \cos\left((2k+1)\frac{\pi}{2}\right) = 0. \end{aligned}$$

The $n+1$ Chebyshev knots x_k^* in (2.57) are obviously pairwise distinct. Therefore, the Chebyshev polynomial $T_{n+1} \in \mathcal{P}_{n+1} \setminus \{0\}$ has no further zeros. As regards the extrema of T_{n+1}, we have $\|T_{n+1}\|_{\infty,[-1,1]} \le 1$ and, moreover,

$$T_{n+1}(y_k) = \cos\left((n+1)\arccos\left(\cos\left(\frac{k}{n+1}\pi\right)\right)\right) = \cos(k\pi) = (-1)^k,$$

so that all $n+2$ points in $Y = \{y_0, \ldots, y_{n+1}\} \subset [-1,1]$ are extremal points for T_{n+1} on $[-1,1]$. Since T_{n+1} is a polynomial of degree $n+1$, its derivative T'_{n+1} has at most n zeros. Therefore, T_{n+1} has at most n extrema in the open interval $(-1,1)$ and at most $n+2$ extrema in the closed interval $[-1,1]$. But this implies that Y already contains all zeros of T_{n+1} on $[-1,1]$. ∎

Table 2.2. Monomial form of the Chebyshev polynomials $T_n \in \mathcal{P}_n$, $n = 1, \ldots, 12$.

$$\begin{aligned} T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \\ T_6(x) &= 32x^6 - 48x^4 + 18x^2 - 1 \\ T_7(x) &= 64x^7 - 112x^5 + 56x^3 - 7x \\ T_8(x) &= 128x^8 - 256x^6 + 160x^4 - 32x^2 + 1 \\ T_9(x) &= 256x^9 - 576x^7 + 432x^5 - 120x^3 + 9x \\ T_{10}(x) &= 512x^{10} - 1280x^8 + 1120x^6 - 400x^4 + 50x^2 - 1 \\ T_{11}(x) &= 1024x^{11} - 2816x^9 + 2816x^7 - 1232x^5 + 220x^3 - 11x \\ T_{12}(x) &= 2048x^{12} - 6144x^{10} + 6912x^8 - 3584x^6 + 840x^4 - 72x^2 + 1 \end{aligned}$$

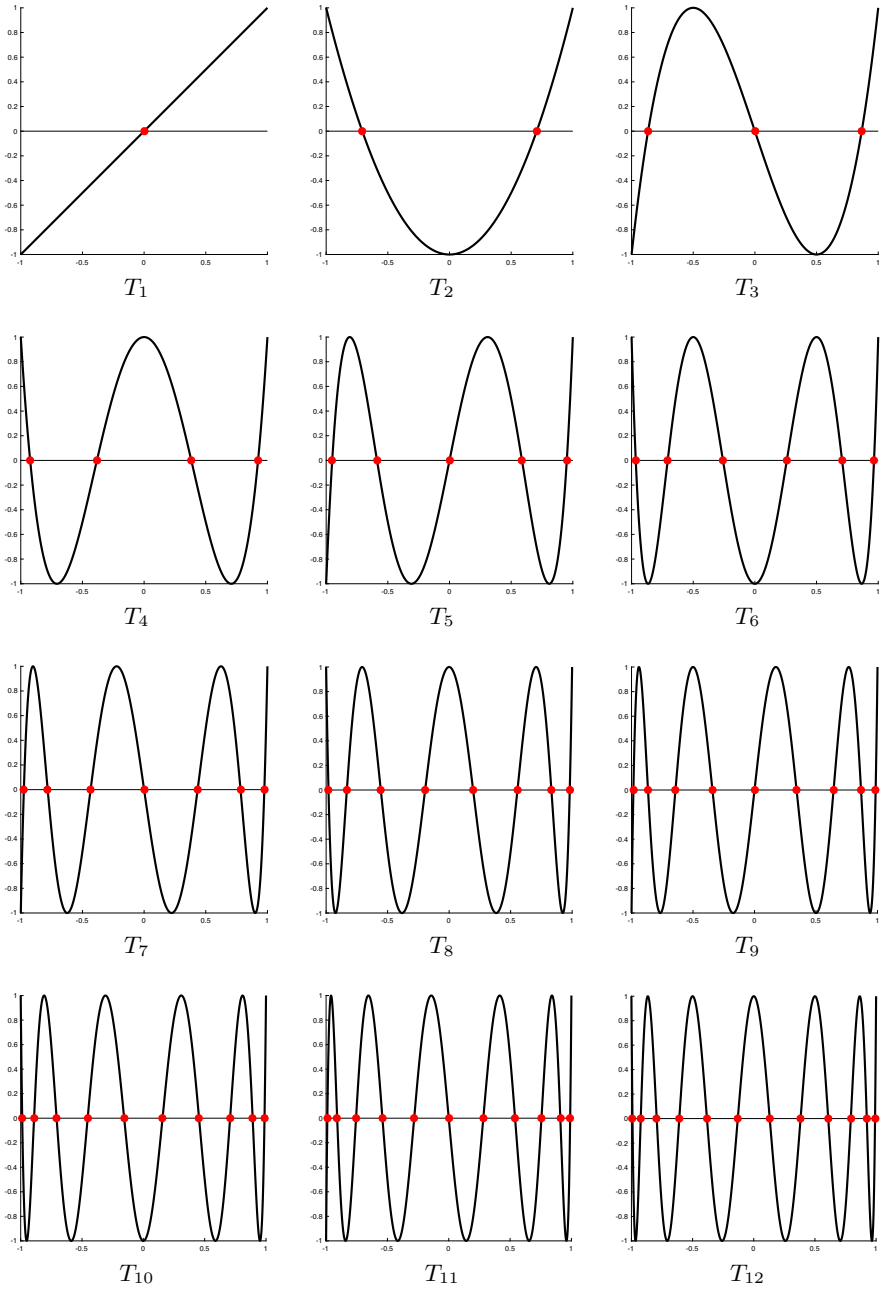


Fig. 2.6. Chebyshev polynomials $T_n \in \mathcal{P}_n$ and their n knots, for $n = 1, \ldots, 12$.

Figure 2.6 shows the graphs of the Chebyshev polynomials $T_n \in \mathcal{P}_n$, for $n = 1, \ldots, 12$, along with their Chebyshev knots (2.57). Moreover, the monomial representations of T_n, for $n = 1, \ldots, 12$, are in Table 2.2.

Corollary 2.30. *For $n \in \mathbb{N}_0$, let $X^* = \{x_0^*, \ldots, x_n^*\} \subset [-1, 1]$ denote the set of Chebyshev knots in (2.57). Then the corresponding knot polynomial ω_{X^*} has the representation*

$$\omega_{X^*} = 2^{-n} T_{n+1}. \tag{2.59}$$

Proof. The knot polynomial ω_X in (2.49) has for any point set X leading coefficient one, in particular for the set X^* of Chebyshev knots. By the representation in (2.56), the polynomial $2^{-n} T_{n+1} \in \mathcal{P}_{n+1}$ has also leading coefficient one. Therefore, the difference

$$q_n = \omega_{X^*} - 2^{-n} T_{n+1}$$

is an algebraic polynomial of degree at most n, i.e., $q_n \in \mathcal{P}_n$. Since $q_n(x^*) = 0$, for $x^* \in X^*$, the polynomial q_n has at least $n + 1$ zeros, and so $q_n \equiv 0$. ∎

Now we can solve the minimization problem in (2.52), Problem 2.26.

Theorem 2.31. *For $n \in \mathbb{N}_0$ the Chebyshev knots $X^* = \{x_0^*, \ldots, x_n^*\}$,*

$$x_k^* = \cos\left(\frac{2k+1}{2n+2}\pi\right) \in [-1, 1] \qquad \text{for } 0 \le k \le n,$$

are the unique solution of the minimization problem (2.52).

Proof. By Corollary 2.30 the knot polynomial $\omega_{X^*} = 2^{-n} T_{n+1} \in \mathcal{P}_{n+1}$ is a multiple of T_{n+1}. Moreover, due to Corollary 2.29 all extrema of ω_{X^*} on $[-1, 1]$ are attained at the $n + 2$ points $Y = \{y_0, \ldots, y_{n+1}\}$ in (2.58), where we have $\|\omega_{X^*}\|_{\infty,[-1,1]} = 2^{-n}$ and

$$\omega_{X^*}(y_k) = 2^{-n}(-1)^k \qquad \text{for } 0 \le k \le n + 1.$$

Now assume that for a point set $X = \{x_0, \ldots, x_n\} \subset [-1, 1]$ its knot polynomial $\omega_X \in \mathcal{P}_{n+1}$ satisfies

$$\|\omega_X\|_{\infty,[-1,1]} < \|\omega_{X^*}\|_{\infty,[-1,1]} = 2^{-n}. \tag{2.60}$$

Then we have $\omega_X(y_k) < \omega_{X^*}(y_k)$, for all *even* indices $k \in \{0, \ldots, n\}$ and $\omega_X(y_k) > \omega_{X^*}(y_k)$, for all *odd* indices $k \in \{1, \ldots, n\}$. Therefore, the difference

$$\omega = \omega_{X^*} - \omega_X$$

has in any of the $n + 1$ intervals $(y_1, y_0), (y_2, y_1), \ldots, (y_{n+1}, y_n)$ at least one sign change, i.e., ω has at least $n + 1$ zeros. Since the knot polynomials $\omega_X, \omega_{X^*} \in \mathcal{P}_{n+1}$ have leading coefficient one, respectively, we see that their

difference $\omega = \omega_{X^*} - \omega_X \in \mathcal{P}_n$ is a polynomial of degree at most n. But this implies $\omega \equiv 0$, i.e.,

$$\omega_X \equiv \omega_{X^*} \in \mathcal{P}_{n+1},$$

in contradiction to (2.60). We can finally conclude that the Chebyshev knots $X^* = \{x_0^*, \ldots, x_n^*\} \subset [-1, 1]$ are the unique solution to the minimization problem in (2.52), Problem 2.26. ■

2.6 Interpolation by Trigonometric Polynomials

In this section, we consider the interpolation of *periodic* functions.

Definition 2.32. *A function $f : \mathbb{R} \longrightarrow \mathbb{R}$ is said to be* `periodic`*, if*

$$f(x) = f(x + T) \qquad \text{for all } x \in \mathbb{R} \tag{2.61}$$

for some $T > 0$. In this case, f is called a T-`periodic` *function. A minimal $T > 0$ satisfying (2.61) is called the* `period` *of f.* ◯

By (2.61) any T-periodic function f is uniquely determined by its function values on the interval $[0, T)$. In the following discussion, we restrict ourselves to 2π-periodic functions. This is without loss of generality since any T-periodic function f can be transformed into a 2π-periodic function by scaling its argument with a scaling factor $T/(2\pi)$, i.e., the function $g : \mathbb{R} \longrightarrow \mathbb{R}$, given as

$$g(x) = f\left(\frac{T}{2\pi} \cdot x\right) \qquad \text{for all } x \in \mathbb{R},$$

is 2π-periodic, if and only if f has period T. We collect all *continuous* and 2π-periodic functions in the linear space

$$\mathscr{C}_{2\pi} = \{f \in \mathscr{C}(\mathbb{R}) \,|\, f(x) = f(x + 2\pi) \text{ for all } x \in \mathbb{R}\}\,.$$

Now let us turn to the interpolation of periodic functions from $\mathscr{C}_{2\pi}$. To this end, we first fix a linear space of interpolants, where it makes sense to choose a finite-dimensional subspace of $\mathscr{C}_{2\pi}$. Obvious examples for functions from $\mathscr{C}_{2\pi}$ are the trigonometric polynomials $\cos(jx)$, for $j \in \mathbb{N}_0$, and $\sin(jx)$, for $j \in \mathbb{N}$. In fact, trigonometric polynomials are suitable choices for the construction of *interpolation spaces* contained in $\mathscr{C}_{2\pi}$. To be more precise on this, we give the following definition.

Definition 2.33. *For $n \in \mathbb{N}_0$, we denote by*

$$\mathcal{T}_n^{\mathbb{R}} = \operatorname{span}_{\mathbb{R}}\{1, \cos(j\,\cdot), \sin(j\cdot) \,|\, 1 \le j \le n\} \subset \mathscr{C}_{2\pi} \tag{2.62}$$

the linear space of `real trigonometric polynomials` *of degree at most n.* ◯

Clearly, $\mathcal{T}_n^{\mathbb{R}}$ is a finite-dimensional linear space, and, moreover, any real-valued trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ can be represented as a linear combination

$$T(x) = \frac{a_0}{2} + \sum_{k=1}^{n} [a_k \cos(kx) + b_k \sin(kx)] \tag{2.63}$$

with coefficients $a_0, \ldots, a_n, b_1, \ldots, b_n \in \mathbb{R}$, the *Fourier*[12] *coefficients* of T. We will provide supporting arguments in favour of the chosen form in (2.63) for the interpolating functions, i.e., as a linear combination of the $2n+1$ (basis) functions in (2.63).

In the following formulation of the interpolation problem we can, due to the 2π-periodicity of the target function $f \in \mathscr{C}_{2\pi}$, restrict ourselves without loss of generality to interpolation points from the interval $[0, 2\pi)$.

Problem 2.34. Compute from a given set $X = \{x_0, x_1, \ldots, x_{2n}\} \subset [0, 2\pi)$ of $2n+1$ pairwise distinct interpolation points and corresponding function values $f_X = (f_0, f_1, \ldots, f_{2n})^T \in \mathbb{R}^{2n+1}$ a real trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ satisfying $T_X = f_X$, i.e.,

$$T(x_j) = f_j \qquad \text{for all } 0 \le j \le 2n. \tag{2.64}$$

□

For the solution to Problem 2.34, our following investigations concerning interpolation of complex-valued functions will be very useful. To this end, we distinguish between *real* (i.e., *real-valued*) trigonometric polynomials, as for $\mathcal{T}_n^{\mathbb{R}}$ in (2.62), and *complex* (i.e., *complex-valued*) trigonometric polynomials. In the following, the symbol i denotes as usual the imaginary unit.

Definition 2.35. *For $N \in \mathbb{N}_0$, the linear space of all* `complex trigonometric polynomials` *of degree at most N is given as*

$$\mathcal{T}_N^{\mathbb{C}} = \operatorname{span}_{\mathbb{C}}\{\exp(\mathrm{i}j\cdot) \,|\, 0 \le j \le N\}. \tag{2.65}$$

◯

Theorem 2.36. *For $N \in \mathbb{N}_0$, the linear space $\mathcal{T}_N^{\mathbb{C}}$ has dimension $N+1$.*

Proof. A complex-valued trigonometric polynomial $p \in \mathcal{T}_N^{\mathbb{C}}$ can be written as a linear combination of the form

$$p(x) = \sum_{k=0}^{N} c_k \mathrm{e}^{\mathrm{i}kx} \tag{2.66}$$

with *complex* coefficients $c_0, \ldots, c_N \in \mathbb{C}$.

[12] Jean Baptiste Joseph Fourier (1768-1830), French mathematician, physicist

We can show the linear independence of the generating function system $\{e^{ik\cdot} \mid 0 \le k \le N\}$ by a simple argument: For $p \equiv 0$ we have

$$0 = \int_0^{2\pi} e^{-imx} \sum_{k=0}^{N} c_k e^{ikx}\, dx = \sum_{k=0}^{N} c_k \int_0^{2\pi} e^{i(k-m)x}\, dx = 2\pi c_m$$

for $m = 0, \ldots, N$, whereby $c_0 = \ldots = c_N = 0$. ■

By the *Euler*[13] *fomula*

$$e^{ix} = \cos(x) + i\sin(x) \tag{2.67}$$

we can represent any real trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ in (2.63) as a complex trigonometric polynomial $p \in \mathcal{T}_N^{\mathbb{C}}$ of the form (2.66). Indeed, by using the Euler formula (2.67) we find the standard trigonometric identities

$$\cos(x) = \frac{1}{2}\left(e^{ix} + e^{-ix}\right) \quad \text{and} \quad \sin(x) = \frac{1}{2i}\left(e^{ix} - e^{-ix}\right) \tag{2.68}$$

and so we obtain for any $T \in \mathcal{T}_n^{\mathbb{R}}$ the representation

$$\begin{aligned} T(x) &= \frac{a_0}{2} + \sum_{k=1}^{n} \left[a_k \cos(kx) + b_k \sin(kx)\right] \\ &= \frac{a_0}{2} + \sum_{k=1}^{n} \left[\frac{a_k}{2}\left(e^{ikx} + e^{-ikx}\right) + \frac{b_k}{2i}\left(e^{ikx} - e^{-ikx}\right)\right] \\ &= \frac{a_0}{2} + \sum_{k=1}^{n} \left[\frac{a_k - ib_k}{2} e^{ikx} + \frac{a_k + ib_k}{2} e^{-ikx}\right] \\ &= \sum_{k=-n}^{n} c_k e^{ikx} = e^{-inx} \sum_{k=0}^{2n} c_{k-n} e^{ikx} \end{aligned}$$

with the *complex* Fourier coefficients

$$c_0 = \frac{1}{2}a_0,\ c_k = \frac{1}{2}(a_k - ib_k),\ c_{-k} = \frac{1}{2}(a_k + ib_k) \quad \text{for } k = 1, \ldots, n. \tag{2.69}$$

Let us now draw an intermediate conclusion.

Proposition 2.37. *Any real trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ in (2.63) can be represented as a product*

$$T(x) = e^{-inx} p(x),$$

where $p \in \mathcal{T}_N^{\mathbb{C}}$ is a complex trigonometric polynomial of the form (2.66), for $N = 2n$. Moreover, the Fourier coefficients of p are uniquely determined by

$$c_{n-k} = \frac{1}{2}(a_k + ib_k),\ c_n = \frac{a_0}{2},\ c_{n+k} = \frac{1}{2}(a_k - ib_k) \quad \text{for } k = 1, \ldots, n, \tag{2.70}$$

where we have applied a periodification to the coefficients c_k in (2.69). ■

[13] Leonhard Euler (1707-1783), Swiss mathematician, physicist, and astronomer

Note that the mapping (2.70) between the *real* Fourier coefficients a_k, b_k of T and the *complex* Fourier coefficients c_k of p is linear,

$$(a_0, \ldots, a_n, b_1, \ldots, b_n)^T \in \mathbb{C}^{2n+1} \longmapsto (c_0, \ldots, c_{2n}) \in \mathbb{C}^{2n+1}.$$

Moreover, this linear mapping is bijective, where its inverse is described as

$$a_0 = 2c_0,\ a_k = c_{n+k} + c_{n-k},\ b_k = \mathrm{i}(c_{n+k} - c_{n-k}) \quad \text{for } k = 1, \ldots, n. \tag{2.71}$$

The Fourier coefficients $a_0, \ldots, a_n, b_1, \ldots, b_n$ are real, if and only if

$$c_{n+k} = \overline{c_{n-k}} \qquad \text{for all } k = 0, \ldots, n.$$

By the bijectivity of the linear mappings in (2.70) and (2.71) between the complex and the real Fourier coefficients, we can determine the dimension of $\mathcal{T}_n^{\mathbb{R}}$. The following result is a direct consequence of Theorem 2.36.

Corollary 2.38. *For $n \in \mathbb{N}_0$, the linear space $\mathcal{T}_n^{\mathbb{R}}$ has dimension $2n+1$.* ■

Now let us return to the interpolation problem, Problem 2.34. For the case of *complex* trigonometric polynomials, we can solve Problem 2.34 as follows.

Theorem 2.39. *Let $X = \{x_0, \ldots, x_N\} \subset [0, 2\pi)$ be a set of $N+1$ pairwise distinct interpolation points and $f_X = (f_0, \ldots, f_N)^T \in \mathbb{C}^{N+1}$ be a data vector of complex function values, for $N \in \mathbb{N}_0$. Then there is a unique complex trigonometric polynomial $p \in \mathcal{T}_N^{\mathbb{C}}$ satisfying $p_X = f_X$, i.e.,*

$$p(x_k) = f_k \qquad \textit{for all } 0 \le k \le N. \tag{2.72}$$

Proof. We regard the linear mapping $L : \mathcal{T}_N^{\mathbb{C}} \longrightarrow \mathbb{C}^{N+1}$, defined as

$$p \in \mathcal{T}_N^{\mathbb{C}} \longmapsto p_X = (p(x_0), \ldots, p(x_N))^T \in \mathbb{C}^{N+1},$$

which assigns every complex trigonometric polynomial $p \in \mathcal{T}_N^{\mathbb{C}}$ in (2.66) to the data vector $p_X \in \mathbb{C}^{N+1}$.

By letting $z_k = \mathrm{e}^{\mathrm{i}x_k} \in \mathbb{C}$, for $0 \le k \le N$, we obtain $N+1$ pairwise distinct interpolation points on the boundary of the unit circle, where we have

$$p(x_k) = \sum_{j=0}^{N} c_j \mathrm{e}^{\mathrm{i}jx_k} = \sum_{j=0}^{N} c_j z_k^j.$$

If $L(p) = 0$, then the complex polynomial p has at least $N+1$ zeros. But in this case, we have $p \equiv 0$, due to the fundamental theorem of algebra. Therefore, the linear mapping L is injective. Due to the dimension formula, L is also surjective, and thus bijective. This already proves the existence and uniqueness of the sought polynomial $p \in \mathcal{T}_N^{\mathbb{C}}$. ■

We finally turn to the solution of the interpolation problem, Problem 2.34, by *real* trigonometric polynomials. The following result is a direct consequence of Theorem 2.39.

Corollary 2.40. *Let $X = \{x_0, \ldots, x_{2n}\} \subset [0, 2\pi)$ be a set of $2n+1$ pairwise distinct interpolation points and $f_X = (f_0, \ldots, f_{2n})^T \in \mathbb{R}^{2n+1}$ be a data vector of real function values, for $n \in \mathbb{N}_0$. Then there is a unique real trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ satisfying $T_X = f_X$.*

Proof. Let $p \in \mathcal{T}_{2n}^{\mathbb{C}}$ be the unique complex trigonometric interpolation polynomial satisfying $p(x_k) = \mathrm{e}^{\mathrm{i}nx_k} f_k$, for $0 \le k \le 2n$, with Fourier coefficients c_j, for $0 \le j \le 2n$. Then we have

$$q(x) := \mathrm{e}^{2\mathrm{i}nx}\overline{p(x)} = \sum_{j=0}^{2n} \overline{c_j}\mathrm{e}^{\mathrm{i}(2n-j)x} = \sum_{j=0}^{2n} \overline{c_{2n-j}}\mathrm{e}^{\mathrm{i}jx} \qquad \text{for } x \in [0, 2\pi)$$

and, moreover, since $f_k \in \mathbb{R}$,

$$q(x_k) = \mathrm{e}^{2\mathrm{i}nx_k}\overline{p(x_k)} = \mathrm{e}^{\mathrm{i}nx_k} f_k = p(x_k) \qquad \text{for all } 0 \le k \le 2n.$$

Therefore, the complex trigonometric polynomial $q \in \mathcal{T}_{2n}^{\mathbb{C}}$ is also a solution to the interpolation problem $q(x_k) = \mathrm{e}^{\mathrm{i}nx_k} f_k$ for all $0 \le k \le 2n$. From the uniqueness of the interpolation by complex trigonometric polynomials we get $q \equiv p$, and so in particular

$$c_j = \overline{c_{2n-j}} \qquad \text{for all } 0 \le j \le 2n. \tag{2.73}$$

The Fourier coefficients of the interpolating real trigonometric polynomial $T \in \mathcal{T}_n^{\mathbb{R}}$ can finally be obtained by the inversion of the complex Fourier coefficients in (2.71). Note that the Fourier coefficients $a_0, \ldots, a_n$, $b_1, \ldots, b_n$ of T are real, due to (2.73). ∎

2.7 The Discrete Fourier Transform

In this section, we explain interpolation by trigonometric polynomials. More specifically, we discuss the special case of $N \in \mathbb{N}$ *equidistant* interpolation points

$$x_k = \frac{2\pi}{N} k \in [0, 2\pi) \qquad \text{for } 0 \le k \le N-1.$$

As we will show, the required Fourier coefficients can be computed efficiently.

In the following discussion, we denote the values of the target function f by $f_k = f(x_k)$, for $0 \le k \le N-1$. Moreover, we use the notation

$$\omega_N = \mathrm{e}^{2\pi\mathrm{i}/N} \qquad \text{for } N \in \mathbb{N}. \tag{2.74}$$

for the *N-th root of unity*.

For further preparation, we make a note of the following observation.

Lemma 2.41. *For $N \in \mathbb{N}$ the N-th root of unity ω_N has the property*

$$\frac{1}{N}\sum_{j=0}^{N-1} \omega_N^{(\ell-k)j} = \delta_{\ell k} \qquad \text{for all } 0 \le \ell, k \le N-1. \tag{2.75}$$

Proof. Let $0 \le \ell, k \le N-1$. Note that for $\ell = k$ the statement in (2.75) is trivial. For $\ell \neq k$ we have $\omega_N^{\ell-k} \neq 1$, so that we can use the standard identity

$$\sum_{j=0}^{N-1} \left(\omega_N^{\ell-k}\right)^j = \frac{\omega_N^{(\ell-k)N} - 1}{\omega_N^{\ell-k} - 1} = \frac{\mathrm{e}^{2\pi \mathrm{i}(\ell-k)} - 1}{\omega_N^{\ell-k} - 1} = 0$$

of geometric series. This already completes our proof for (2.75). ∎

Now we are in a position where we can already give the solution to the posed interpolation problem at equidistant interpolation points.

Theorem 2.42. *For $N \in \mathbb{N}$ equidistant points $x_\ell = 2\pi\ell/N \in [0, 2\pi)$, for $0 \le \ell \le N-1$, and function values $f_X = (f_0, \ldots, f_{N-1})^T \in \mathbb{C}^N$ the Fourier coefficients of the complex trigonometric interpolation polynomial $p \in \mathcal{T}_{N-1}^{\mathbb{C}}$ satisfying $p_X = f_X$ are given as*

$$c_j = \frac{1}{N}\sum_{k=0}^{N-1} f_k \omega_N^{-jk} \qquad \text{for } 0 \le j \le N-1. \tag{2.76}$$

Proof. By using Lemma 2.41, we obtain the identity

$$p(x_\ell) = \sum_{j=0}^{N-1} \frac{1}{N} \sum_{k=0}^{N-1} f_k \omega_N^{-jk} \mathrm{e}^{\mathrm{i}jx_\ell} = \sum_{k=0}^{N-1} f_k \frac{1}{N} \sum_{j=0}^{N-1} \omega_N^{(\ell-k)j} = f_\ell$$

for all $\ell = 0, \ldots, N-1$. ∎

Therefore, the linear mapping in (2.76) yields an automorphism

$$A_N : \mathbb{C}^N \longrightarrow \mathbb{C}^N,$$

which maps the data vector $f_X = (f_0, \ldots, f_{N-1})^T \in \mathbb{C}^N$ on the Fourier coefficients $c = (c_0, \ldots, c_{N-1})^T \in \mathbb{C}^N$ of the complex trigonometric interpolation polynomial $p \in \mathcal{T}_{N-1}^{\mathbb{C}}$ satisfying $p_X = f_X$. The bijective linear mapping A_N, called `discrete Fourier analysis`, is represented by the matrix

$$A_N = \frac{1}{N}\left(\omega_N^{-jk}\right)_{0 \le j,k \le N-1} \in \mathbb{C}^{N \times N}. \tag{2.77}$$

We can characterize the inverse of A_N as follows. The linear mapping

$$A_N^{-1} : \mathbb{C}^N \longrightarrow \mathbb{C}^N,$$

which assigns every vector $c = (c_0, \ldots, c_{N-1})^T \in \mathbb{C}^N$ of Fourier coefficients, for a complex trigonometric polynomial

$$p(x) = \sum_{j=0}^{N-1} c_j \mathrm{e}^{\mathrm{i}jx} \in \mathcal{T}_{N-1}^{\mathbb{C}},$$

to the complex values

$$f_k = p(x_k) = \sum_{j=0}^{N-1} c_j \mathrm{e}^{\mathrm{i}jx_k} = \sum_{j=0}^{N-1} c_j \omega_N^{jk} \quad \text{for } k = 0, \ldots, N-1\,,$$

i.e., $p_X = f_X$, is called `discrete Fourier synthesis`. Therefore, the linear mapping A_N^{-1} is the inverse of A_N, being represented by the matrix

$$A_N^{-1} = \left(\omega_N^{jk}\right)_{0 \le j,k \le N-1} \in \mathbb{C}^{N \times N}. \tag{2.78}$$

The discrete Fourier analysis and the Fourier synthesis are usually referred to as *discrete Fourier transform* and *discrete inverse Fourier transform*. In the following discussion, we derive an efficient method for computing the discrete (inverse) Fourier transform. But we first give a formal introduction for the discrete (inverse) Fourier transform.

Definition 2.43. *The* `discrete Fourier transformation` *(DFT) of*

$$z = (z(0), z(1), \ldots, z(N-1))^T \in \mathbb{C}^N$$

is defined componentwise as

$$\hat{z}(j) = \sum_{k=0}^{N-1} z(k) \omega_N^{-jk} \qquad \textit{for } 0 \le j \le N-1, \tag{2.79}$$

and the `inverse discrete Fourier transform` *(IDFT) of*

$$\hat{z} = (\hat{z}(0), \hat{z}(1), \ldots, \hat{z}(N-1))^T \in \mathbb{C}^N$$

is defined componentwise as

$$z(k) = \frac{1}{N} \sum_{j=0}^{N-1} \hat{z}(j) \omega_N^{jk} \qquad \textit{for } 0 \le k \le N-1.$$

○

The discrete Fourier transform (DFT) and the inverse DFT are represented by the `Fourier matrices` $F_N = NA_N$ and $F_N^{-1} = A_N^{-1}/N$, i.e.,

$$F_N = \left(\omega_N^{-jk}\right)_{0\le j,k\le N-1} \in \mathbb{C}^{N\times N}$$
$$F_N^{-1} = \frac{1}{N}\left(\omega_N^{jk}\right)_{0\le j,k\le N-1} \in \mathbb{C}^{N\times N}.$$

Therefore, with using the notations in Definition 2.43, we have

$$\hat{z} = F_N z \quad \text{and} \quad z = F_N^{-1}\hat{z} \qquad \text{for all } z, \hat{z} \in \mathbb{C}^N.$$

This finally leads us to the *Fourier inversion formula*

$$z = F_N^{-1} F_N z \qquad \text{for all } z \in \mathbb{C}^N.$$

Now let us make one simple example for further illustration.

Example 2.44. We compute the DFT $\hat{z} \in \mathbb{C}^{512}$ of the vector $z \in \mathbb{C}^{512}$ with components $z(k) = 3\sin(2\pi \cdot 7k/512) - 4\cos(2\pi \cdot 8k/512)$. To this end, we regard the Fourier series (from the Fourier inversion formula)

$$z(k) = \frac{1}{512}\sum_{j=0}^{511} \hat{z}(j)\mathrm{e}^{2\pi \mathrm{i} jk/512},$$

whereby we obtain the unique representation of $z \in \mathbb{C}^{512}$ in the *Fourier basis*

$$\left\{\mathrm{e}^{2\pi \mathrm{i} jk/512} \,\middle|\, 0 \le j \le 511\right\}.$$

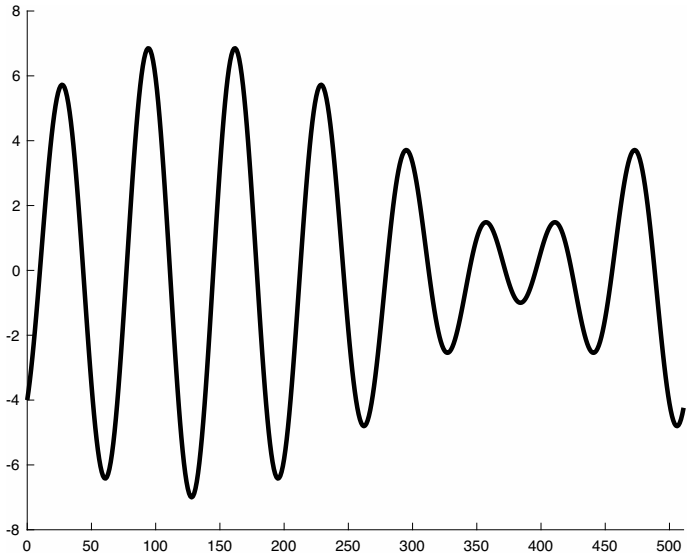
On the other hand, the Euler formula yields the representation

$$\begin{aligned} z(k) &= 3\sin(2\pi 7k/512) - 4\cos(2\pi 8k/512) \\ &= \frac{3}{2\mathrm{i}}\left(\mathrm{e}^{2\pi \mathrm{i}7k/512} - \mathrm{e}^{-2\pi \mathrm{i}7k/512}\right) - \frac{4}{2}\left(\mathrm{e}^{2\pi \mathrm{i}8k/512} + \mathrm{e}^{-2\pi \mathrm{i}8k/512}\right) \\ &= \frac{-3\mathrm{i}}{2}\mathrm{e}^{2\pi \mathrm{i}7k/512} + \frac{3\mathrm{i}}{2}\mathrm{e}^{2\pi \mathrm{i}(-7+512)k/512} - 2\mathrm{e}^{2\pi \mathrm{i}8k/512} - 2\mathrm{e}^{2\pi \mathrm{i}(-8+512)k/512} \\ &= \frac{1}{512}\Big(-3\cdot 256\mathrm{i}\cdot \mathrm{e}^{2\pi \mathrm{i}7k/512} - 1024\cdot \mathrm{e}^{2\pi \mathrm{i}8k/512} \\ &\qquad -1024\cdot \mathrm{e}^{2\pi \mathrm{i}504k/512} + 3\cdot 256i\cdot \mathrm{e}^{2\pi \mathrm{i}505k/512}\Big). \end{aligned}$$

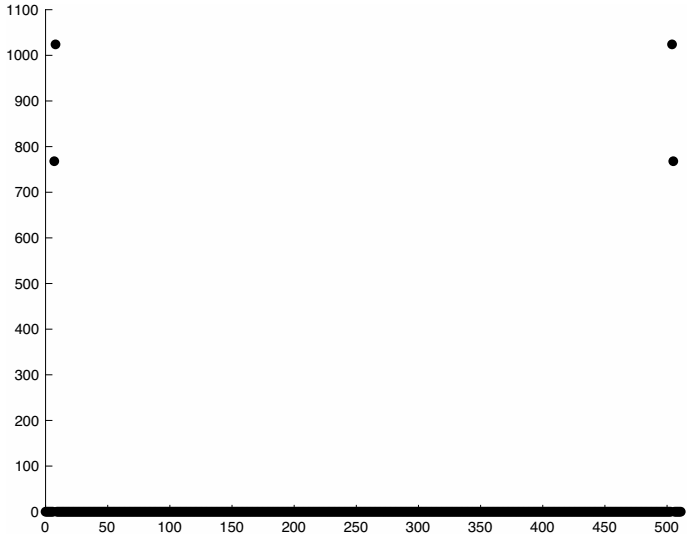
Therefore, we have

$$\hat{z}(7) = -768\mathrm{i}, \quad \hat{z}(8) = -1024, \quad \hat{z}(504) = -1024, \quad \hat{z}(505) = 768\mathrm{i},$$

and, moreover, $\hat{z}(j) = 0$ for all $j \in \{0,\ldots,511\} \setminus \{7, 8, 504, 505\}$. Thereby, the vector $z \in \mathbb{C}^{512}$ has a *sparse* representation by the four non-vanishing Fourier coefficients $\hat{z}(7)$, $\hat{z}(8)$, $\hat{z}(504)$ and $\hat{z}(505)$ (see Figure 2.7). ♢



(a) input vector $z(k)$, $k = 0, \ldots, 511$



(b) amplitude spectrum $|\hat{z}(j)|$

Fig. 2.7. Sparse representation of $z(k) = 3\sin(2\pi \cdot 7k/512) - 4\cos(2\pi \cdot 8k/512)$ with amplitude spectrum $|\hat{z}(j)|$ (see Example 2.44).

Remark 2.45. A componentwise computation of the DFT $\hat{z}$ (or of the IDFT) according to Definition 2.43 requires asymptotically $\mathcal{O}(N^2)$ steps, namely $\mathcal{O}(N)$ steps for each of the N components. □

In the remainder of this section, we explain how to compute the DFT by an *efficient* algorithm, termed the `fast Fourier transform` (FFT), designed by Cooley[14] and Tukey[15] [16]. The Cooley-Tukey algorithm is based on a recursion according to a common (political) principle *divide et impera* (Latin for *divide and conquer*) of Machiavelli[16] dating back to 1513.

The recursion step of the Cooley-Tukey algorithm relies on the identity

$$\omega_{2N}^2 = \omega_N,$$

being applied as follows.

For $N = 2^n$, $n \geq 1$, and $0 \leq j \leq N-1$ we have

$$\begin{aligned}
\hat{z}(j) &= \sum_{k=0}^{N-1} z(k)\omega_N^{-kj} \\
&= \sum_{k \text{ even}} z(k)\omega_N^{-kj} + \sum_{k \text{ odd}} z(k)\omega_N^{-kj} \\
&= \sum_{k=0}^{N/2-1} z(2k)\omega_N^{-2kj} + \sum_{k=0}^{N/2-1} z(2k+1)\omega_N^{-(2k+1)j} \\
&= \sum_{k=0}^{N/2-1} z(2k)\omega_N^{-2kj} + \omega_N^{-j} \sum_{k=0}^{N/2-1} z(2k+1)\omega_N^{-2kj}.
\end{aligned}$$

This already yields for $M = N/2$ the *reduction*

$$\begin{aligned}
\hat{z}(j) &= \sum_{k=0}^{M-1} z(2k)\omega_N^{-2kj} + \omega_N^{-j} \sum_{k=0}^{M-1} z(2k+1)\omega_N^{-2kj} \\
&= \sum_{k=0}^{M-1} u(k)\omega_{N/2}^{-kj} + \omega_N^{-j} \sum_{k=0}^{M-1} v(k)\omega_{N/2}^{-kj} \\
&= \sum_{k=0}^{M-1} u(k)\omega_M^{-kj} + \omega_N^{-j} \sum_{k=0}^{M-1} v(k)\omega_M^{-kj}
\end{aligned}$$

for $j = 0, \ldots, N-1$, where

$$u(k) = z(2k) \quad \text{and} \quad v(k) = z(2k+1) \qquad \text{for } k = 0, 1, \ldots, M-1.$$

[14] James W. Cooley (1926-2016), US American mathematician
[15] John Wilder Tukey (1915-2000), US American mathematician
[16] Niccolò di Bernardo dei Machiavelli (1469-1527), Florentine philosopher

Therefore, we can, for any input vector $z \in \mathbb{C}^N$ of length $N = 2M$, reduce the computation of its DFT $\hat{z}$ to the computation of two DFTs of half length $M = N/2$ each. Indeed, the DFTs of the two vectors $u, v \in \mathbb{C}^M$ yields the DFT of z by

$$\hat{z}(j) = \hat{u}(j) + \omega_N^{-j} \hat{v}(j).$$

From this basic observation, we can already determine the *complexity*, i.e., the asymptotic computational costs, of the fast Fourier transform (FFT).

Theorem 2.46. *For $N = 2^n$, $n \in \mathbb{N}$, the discrete Fourier transform of a vector $z \in \mathbb{C}^N$ is computed by the FFT in asymptotically $\mathcal{O}(N \log(N))$ steps.*

Proof. In the first reduction step the DFT of $z \in \mathbb{C}^N$ with length N is decomposed into two DFTs (for $u, v \in \mathbb{C}^{N/2}$) of length $N/2$ each. By induction, in the m-th reduction step for the current 2^m DFTs of length $N/2^m$, each of these DFT can be decomposed into two DFTs of length $N/2^{m+1}$. After $n = \log_2(N)$ reduction steps we have N atomic DFTs of unit length. But the DFT for a vector z of unit length is trivial: In this case, we have $\hat{z}(0) = z(0)$ for $z = z(0) \in \mathbb{C}^1$, and so the recursion terminates. Altogether, $N \log_2(N)$ steps are performed in the recursion. ∎

We finally discuss one relevant application of the fast Fourier transform. In this application, we consider solving a linear equation system of the form

$$Cx = b \tag{2.80}$$

efficiently, where C is a *cyclic Toeplitz matrix*.

Definition 2.47. *A* `cyclic Toeplitz`[17] `matrix` *has the form*

$$C = \begin{bmatrix} c_0 & c_{N-1} & \cdots & c_2 & c_1 \\ c_1 & c_0 & \ddots & \vdots & c_2 \\ \vdots & c_1 & \ddots & c_{N-1} & \vdots \\ c_{N-2} & \vdots & \ddots & c_0 & c_{N-1} \\ c_{N-1} & c_{N-2} & \cdots & c_1 & c_0 \end{bmatrix} \in \mathbb{C}^{N \times N}$$

where $c = (c_0, \ldots, c_{N-1})^T \in \mathbb{C}^N$ is called the `generating vector` *of C.* ◯

The following observation is quite important for our solution to (2.80).

Proposition 2.48. *Let C be a cyclic Toeplitz matrix with generating vector $c \in \mathbb{C}^N$. Then C is diagonalized by the discrete Fourier transform F_N, so that*

$$F_N C F_N^{-1} = \operatorname{diag}(d),$$

where the eigenvalues $d = (d_0, \ldots, d_{N-1}) \in \mathbb{C}^N$ of C are given by the discrete Fourier transform of c, i.e.,

$$d = F_N c.$$

[17] Otto Toeplitz (1881-1940), German mathematician

Proof. For the entries of the Toeplitz matrix $C = (C_{jk})_{0\le j,k\le N-1}$, we have

$$C_{jk} = c_{(j-k) \bmod N} \qquad \text{for } 0 \le j,k \le N-1.$$

We recall the definition of the Fourier matrices

$$F_N = \left(\omega_N^{-jk}\right)_{0\le j,k\le N-1} \quad \text{and} \quad F_N^{-1} = \frac{1}{N}\left(\omega_N^{jk}\right)_{0\le j,k\le N-1},$$

where $\omega_N = \mathrm{e}^{2\pi \mathrm{i}/N}$. For $0 \le \ell \le N-1$ we let

$$\omega^{(\ell)} = \frac{1}{N}(\omega_N^{j\ell})_{0\le j\le N-1} \in \mathbb{C}^N$$

denote the ℓ-th column of F_N^{-1}. By using the identity

$$\omega_N^{(k-j)\ell} = \left(\omega_N^{(k-j) \bmod N}\right)^\ell$$

we obtain

$$\begin{aligned}(C\omega^{(\ell)})_j &= \frac{1}{N}\sum_{k=0}^{N-1} c_{(j-k) \bmod N} \cdot \omega_N^{k\ell} = \frac{1}{N}\omega_N^{j\ell}\sum_{k=0}^{N-1} c_{(j-k) \bmod N} \cdot \omega_N^{(k-j)\ell} \\ &= \frac{1}{N}\omega_N^{j\ell}\sum_{m=0}^{N-1} c_{m \bmod N} \cdot \omega_N^{-m\ell} = \frac{1}{N}\omega_N^{j\ell}\sum_{m=0}^{N-1} c_m \omega_N^{-m\ell} = \frac{1}{N}\omega_N^{j\ell} d_\ell,\end{aligned}$$

where

$$d_\ell = \sum_{k=0}^{N-1} c_k \omega_N^{-\ell k} \qquad \text{for } 0 \le \ell \le N-1$$

is the ℓ-th component of $d = F_N c$.

Therefore, $\omega^{(\ell)}$ is an eigenvector of C with eigenvalue d_ℓ, i.e.,

$$C\omega^{(\ell)} = d_\ell \omega^{(\ell)} \qquad \text{for } 0 \le \ell \le N-1,$$

whereby

$$CF_N^{-1} = F_N^{-1}\mathrm{diag}(d)$$

or

$$F_N C F_N^{-1} = \mathrm{diag}(d).$$

■

Now we finally regard the linear system (2.80) for a cyclic Toeplitz matrix $C \in \mathbb{C}^{N\times N}$ with generating vector $c \in \mathbb{C}^N$. By application of the discrete Fourier transform F_N to both sides in (2.80) we get the identity

$$F_N C F_N^{-1} F_N x = F_N b.$$

Using Proposition 2.48 leads us to the linear system

$$Dy = r, \tag{2.81}$$

where we let $y = F_N x$ and $r = F_N b$, and where $D = \operatorname{diag}(d)$ for $d = F_N c$. Now the matrix C is non-singular, if and only if none of its eigenvalues in d vanishes. In this case

$$y = \left(\frac{r_0}{d_0}, \ldots, \frac{r_{N-1}}{d_{N-1}} \right)^T \in \mathbb{C}^N$$

is the unique solution of the linear system (2.81). By backward transformation with the inverse discrete Fourier transform F_N^{-1}, we finally obtain the solution of the linear system (2.80) by

$$x = F_N^{-1} y.$$

We summarize the proposed solution for the Toeplitz system (2.80) in Algorithm 3. Note that Algorithm 3 can be implemented *efficiently* by using the fast Fourier transform (FFT): By Theorem 2.46 the performance of the steps in lines 5,6 and 8 of Algorithm 3 by the (inverse) FFT costs only $\mathcal{O}(N \log(N))$ operations each. In this case, a total number of only $\mathcal{O}(N \log(N))$ operations are required for the performance of Algorithm 3. In comparison, the solution of a linear equation system (2.80) via Gauss elimination requiring $\mathcal{O}(N^3)$ operations is far too expensive. But unlike in Algorithm 3, the Toeplitz structure of the matrix C is not used in the Gauss elimination algorithm.

Algorithm 3 Solution of linear Toeplitz systems $Cx = b$ in (2.80)

1: **function** TOEPLITZ-SOLUTION(c,b)
2: **input:** generating vector $c \in \mathbb{C}^N$ of a non-singular
3: cyclic Toeplitz matrix $C \in \mathbb{C}^{N \times N}$;
4: right hand side $b \in \mathbb{C}^N$;
5: **compute** DFT $d = F_N c$;
6: **compute** DFT $r = F_N b$;
7: **let** $y := (r_0/d_0, \ldots, r_{N-1}/d_{N-1})^T$;
8: **compute** IDFT $x = F_N^{-1} y$
9: **output:** solution $x \in \mathbb{C}^N$ of $Cx = b$.
10: **end function**

3 Best Approximations

In this chapter, we analyze fundamental questions of approximation. To this end, let $\mathcal{F}$ be a linear space, equipped with a norm $\|\cdot\|$. Moreover, $\mathcal{S} \subset \mathcal{F}$ be a non-empty subset of $\mathcal{F}$. To approximate one $f \in \mathcal{F} \setminus \mathcal{S}$ by elements from $\mathcal{S}$ we are interested in finding a $s^* \in \mathcal{S}$, whose distance to f is *minimal* among all elements from $\mathcal{S}$. This leads us to the definition of *best approximations*.

Definition 3.1. *Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$. Moreover, let $\mathcal{S} \subset \mathcal{F}$ be a non-empty subset of $\mathcal{F}$. For $f \in \mathcal{F}$, an element $s^* \in \mathcal{S}$ is said to be a* `best approximation` *to f from $\mathcal{S}$ with respect to $(\mathcal{F}, \|\cdot\|)$, or in short: s^* is a* `best approximation` *to f, if*

$$\|s^* - f\| = \inf_{s \in \mathcal{S}} \|s - f\|.$$

Moreover,

$$\eta \equiv \eta(f, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|s - f\|$$

is called the `minimal distance` *between f and $\mathcal{S}$.* ◯

In the following investigations, we will first address questions concerning the existence and uniqueness of best approximations. To this end, we develop sufficient conditions for the linear space $\mathcal{F}$ and the subset $\mathcal{S} \subset \mathcal{F}$, under which we can guarantee for any $f \in \mathcal{F}$ the existence of a best approximation $s^* \in \mathcal{S}$ for f. To guarantee the uniqueness of s^*, we require *strict convexity* for the norm $\|\cdot\|$.

In the following discussion, we develop suitable sufficient and necessary conditions to characterize best approximations. To this end, we first derive *dual* characterizations for best approximations, giving conditions for the elements from the *topological dual space* $\mathcal{F}'$ of linear and continuous functionals on $\mathcal{F}$.

This is followed by *direct* characterizations of best approximations, where we use directional derivatives (*Gâteaux derivatives*) of the norm $\|\cdot\|$. On that occasion, we consider computing directional derivatives of relevant norms explicitly.

To study the material of this chapter (and for the following chapters) we require knowledge of elementary results from optimization and functional analysis. Therefore, we decided to explain a selection of relevant results. But for further reading, we refer to the textbook [33].

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_3

Before we address theoretical questions concerning the existence and uniqueness of best approximations, we first discuss one elementary example, which will illustrate relevant scenarios and phenomena.

Example 3.2. For $\mathcal{F} = \mathbb{R}^2$, let $\mathcal{S} = \{x = (x_1, x_2) \,|\, 2 \leq \|x\|_2 < 3\} \subset \mathbb{R}^2$ be a concentric circle around the origin. Moreover, let $f_\alpha = (\alpha, 0) \in \mathbb{R}^2$, for $\alpha \in \mathbb{R}$. Now we wish to *best-approximate* f_α (according to Definition 3.1) by elements from $\mathcal{S}$. To do so, we first need to fix a norm on $\mathbb{R}^2$. To this end, we work with three different norms on $\mathbb{R}^2$:

- the 1-norm $\|\cdot\|_1$, defined as $\|x\|_1 = |x_1| + |x_2|$ for $x = (x_1, x_2)$;
- the Euclidean norm $\|\cdot\|_2$, defined as $\|x\|_2^2 = |x_1|^2 + |x_2|^2$;
- the maximum norm $\|\cdot\|_\infty$, defined as $\|x\|_\infty = \max(|x_1|, |x_2|)$.

We let $\mathcal{S}_p^* \equiv \mathcal{S}_p^*(f_\alpha)$ denote the set of best approximations to f_α with respect to $\|\cdot\| = \|\cdot\|_p$ at minimal distances $\eta_p \equiv \eta_p(f_\alpha, \mathcal{S})$, for $p = 1, 2, \infty$. For the construction and characterization of best approximations to f_α we distinguish different cases (see Fig. 3.1).

Case (a): Suppose $\alpha \geq 3$. In this case, we have

$$\eta_p = \inf_{s \in \mathcal{S}} \|s - f_\alpha\|_p = \alpha - 3 \qquad \text{for } p = 1, 2, \infty,$$

where

$$\|s - f_\alpha\|_p > \inf_{s \in \mathcal{S}} \|s - f_\alpha\|_p = \alpha - 3 \qquad \text{for all } s \in \mathcal{S},$$

i.e., there is *no* best approximation to f_α from $\mathcal{S}$, and so $\mathcal{S}_p^* = \emptyset$.

Case (b): Suppose $\alpha \in (0, 2)$. In this case, we have

$$\eta_1 = \eta_2 = 2 - \alpha \qquad \text{and} \qquad \eta_\infty = \frac{\sqrt{8 - \alpha^2} - \alpha}{2}$$

and, moreover, $\mathcal{S}_p^* = \{(2, 0)\}$ for $p = 1, 2$ and $\mathcal{S}_\infty^* = \left\{ \left(\frac{\sqrt{8-\alpha^2}+\alpha}{2}, \pm \frac{\sqrt{8-\alpha^2}-\alpha}{2} \right) \right\}$.

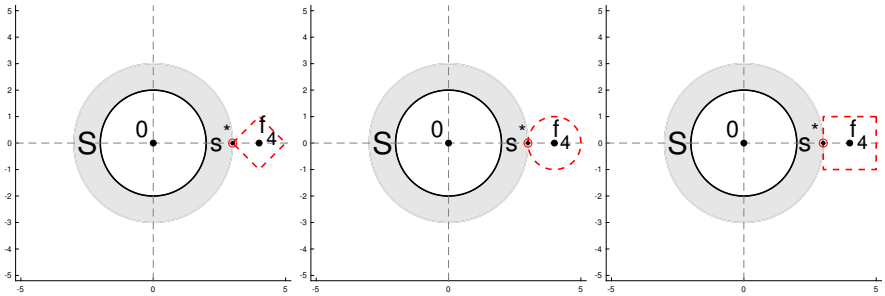
Case (c): Suppose $\alpha = 0$. In this case, we have

$$\eta_1 = \eta_2 = 2 \qquad \text{and} \qquad \eta_\infty = \sqrt{2}$$

where $\mathcal{S}_1^* = \{(\pm 2, 0), (0, \pm 2)\}$, $\mathcal{S}_2^* = \{x \in \mathcal{S} \,|\, \|x\|_2 = 2\}$, $\mathcal{S}_\infty^* = \{(\pm\sqrt{2}, \pm\sqrt{2})\}$. Therefore, there exists, for any of the three norms $\|\cdot\|_p$, $p = 1, 2, \infty$, a best approximation to f_0. In either case, however, the best approximations are not unique. For $\|\cdot\|_2$ there are even uncountably many best approximations to f_0.

Case (d): For $\alpha \in [2, 3)$ we have $f_\alpha \in \mathcal{S}$ and so $\mathcal{S}_p^* = \{f_\alpha\}$ with $\eta_p = 0$.

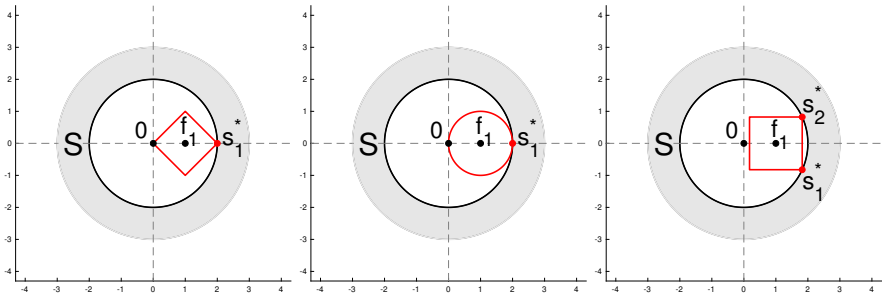
For any other case, i.e., for $\alpha < 0$, we can analyze the set of best approximations by using one of the (symmetric) cases (a)-(d). ♢



$\alpha = 4,\ \|\cdot\| = \|\cdot\|_1$
$\mathcal{S}_1^* = \emptyset,\ \eta_1 = 1$

$\alpha = 4,\ \|\cdot\| = \|\cdot\|_2$
$\mathcal{S}_2^* = \emptyset,\ \eta_2 = 1$

$\alpha = 4,\ \|\cdot\| = \|\cdot\|_\infty$
$\mathcal{S}_\infty^* = \emptyset,\ \eta_\infty = 1$



$\alpha = 1,\ \|\cdot\| = \|\cdot\|_1$

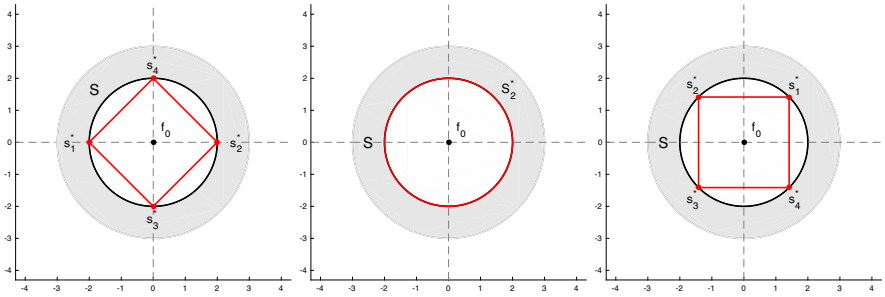
$\mathcal{S}_1^* = \{(2,0)\}$
$\eta_1 = 1$

$\alpha = 1,\ \|\cdot\| = \|\cdot\|_2$

$\mathcal{S}_2^* = \{(2,0)\}$
$\eta_2 = 1$

$\alpha = 1,\ \|\cdot\| = \|\cdot\|_\infty$

$\mathcal{S}_\infty^* = \left\{\left(\frac{\sqrt{7}+1}{2}, \pm\frac{\sqrt{7}-1}{2}\right)\right\}$
$\eta_\infty = \frac{\sqrt{7}-1}{2}$



$\alpha = 0,\ \|\cdot\| = \|\cdot\|_1$

$\mathcal{S}_1^* = \{(\pm 2, 0), (0, \pm 2)\}$
$\eta_1 = 2$

$\alpha = 0,\ \|\cdot\| = \|\cdot\|_2$

$\mathcal{S}_2^* = \{x \in \mathbb{R}^2 \mid \|x\|_2 = 2\}$
$\eta_2 = 2$

$\alpha = 0,\ \|\cdot\| = \|\cdot\|_\infty$

$\mathcal{S}_\infty^* = \{(\pm\sqrt{2}, \pm\sqrt{2})\}$
$\eta_\infty = \sqrt{2}$

Fig. 3.1. Approximation of $f_\alpha = (\alpha, 0) \in \mathbb{R}^2$, for $\alpha = 4, 1, 0$, by elements from the approximation set $\mathcal{S} = \{x = (x_1, x_2) \,|\, 2 \le \|x\|_2 < 3\} \subset \mathbb{R}^2$ and with respect to the norms $\|\cdot\|_p$ for $p = 1, 2, \infty$ (see Example 3.2).

3.1 Existence

In the following discussion, the notions of *compactness*, *completeness*, and *continuity* play an important role. We assume that their definitions and further properties are familiar from analysis. Nevertheless, let us recall the continuity of functionals. Throughout this chapter, $\mathcal{F}$ denotes a linear space with norm $\|\cdot\|$.

Definition 3.3. *A functional $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ is said to be* `continuous` *at $u \in \mathcal{F}$, if for any convergent sequence $(u_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ with limit $u \in \mathcal{F}$, i.e.,*

$$\|u_n - u\| \longrightarrow 0 \quad \text{for } n \to \infty,$$

we have

$$\varphi(u_n) \longrightarrow \varphi(u) \quad \text{for } n \to \infty.$$

Moreover, φ is called `continuous on` *$\mathcal{F}$, if φ is continuous at every $u \in \mathcal{F}$.* ◯

Now recall that any continuous functional attains its minimum (and its maximum) on compact sets. Any compact set is closed and bounded. The converse, however, is only true in *finite-dimensional* spaces.

For the discussion in this section, we need the continuity of norms. This requirement is already covered by the following result.

Theorem 3.4. *Every norm is continuous.*

Proof. Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$. Moreover let $v \in \mathcal{F}$ and $(v_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ be a convergent sequence in $\mathcal{F}$ with limit v, i.e.,

$$\|v_n - v\| \longrightarrow 0 \quad \text{for } n \to \infty.$$

Now, by the triangle inequality for the norm $\|\cdot\|$, this implies

$$|\|v_n\| - \|v\|| \leq \|v_n - v\| \longrightarrow 0 \quad \text{for } n \to \infty$$

and therefore

$$\|v_n\| \longrightarrow \|v\| \quad \text{for } n \to \infty,$$

i.e., $\|\cdot\|$ is continuous at $v \in \mathcal{F}$. Since we did not pose any further conditions on $v \in \mathcal{F}$, the norm $\|\cdot\|$ is continuous on $\mathcal{F}$. ■

The above result allows us to prove a first elementary result concerning the existence of best approximations.

Theorem 3.5. *Let $\mathcal{S} \subset \mathcal{F}$ be compact. Then there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.*

Proof. For $f \in \mathcal{F}$ the functional $\varphi : \mathcal{F} \longrightarrow [0, \infty)$, defined as

$$\varphi(v) = \|v - f\| \qquad \text{for } v \in \mathcal{F},$$

is continuous on $\mathcal{F}$. In particular, φ attains its minimum on the compact set $\mathcal{S}$, i.e., there is one $s^* \in \mathcal{S}$ satisfying

$$\varphi(s^*) = \|s^* - f\| \leq \|s - f\| = \varphi(s) \qquad \text{for all } s \in \mathcal{S}.$$

■

From this result, we can further conclude as follows.

Corollary 3.6. *Let $\mathcal{F}$ be finite-dimensional, and $\mathcal{S} \subset \mathcal{F}$ be closed in $\mathcal{F}$. Then there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.*

Proof. For $s_0 \in \mathcal{S}$ and $f \in \mathcal{F}$ the non-empty set

$$\mathcal{S}_0 = \mathcal{S} \cap \{v \in \mathcal{F} \,|\, \|v - f\| \leq \|s_0 - f\|\} \subset \mathcal{S}$$

is closed and bounded, i.e., $\mathcal{S}_0 \subset \mathcal{F}$ is compact. By Theorem 3.5 there is one best approximation $s^* \in \mathcal{S}_0$ to f from $\mathcal{S}_0$, so that

$$\|s^* - f\| \leq \|s - f\| \qquad \text{for all } s \in \mathcal{S}_0.$$

Moreover, for any $s \in \mathcal{S} \setminus \mathcal{S}_0$ we have the inequality

$$\|s - f\| > \|s_0 - f\| \geq \|s^* - f\|,$$

so that altogether,

$$\|s^* - f\| \leq \|s - f\| \qquad \text{for all } s \in \mathcal{S},$$

i.e., $s^* \in \mathcal{S}_0 \subset \mathcal{S}$ is a best approximation to f from $\mathcal{S}$. ■

Corollary 3.7. *Let $\mathcal{S} \subset \mathcal{F}$ be a closed subset of $\mathcal{F}$. If $\mathcal{S}$ is contained in a finite-dimensional linear subspace $\mathcal{R} \subset \mathcal{F}$ of $\mathcal{F}$, i.e., $\mathcal{S} \subset \mathcal{R}$, then there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.*

Proof. Regard, for $f \in \mathcal{F}$, the finite-dimensional linear space

$$\mathcal{R}_f = \operatorname{span}\{f, r_1, \ldots, r_n\} \subset \mathcal{F},$$

where $\{r_1, \ldots, r_n\}$ be a basis of $\mathcal{R}$. Then there exists, by Corollary 3.6 a best approximation $s^* \in \mathcal{S}$ to $f \in \mathcal{R}_f$, where in particular,

$$\|s^* - f\| \leq \|s - f\| \qquad \text{for all } s \in \mathcal{S}.$$

■

The result of Corollary 3.7 holds in particular for the case $\mathcal{R} = \mathcal{S}$.

Corollary 3.8. *Let $\mathcal{S} \subset \mathcal{F}$ be a finite-dimensional subspace of $\mathcal{F}$. Then there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.* ■

In the above results concerning the existence of best approximations, we require that $\mathcal{S} \subset \mathcal{F}$ is contained in a finite-dimensional linear space. For the approximation in Euclidean spaces $\mathcal{F}$, we can refrain from using this restriction. To this end, the following geometric identities are of fundamental importance.

Theorem 3.9. *Let $\mathcal{F}$ be a Euclidean space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$. Then the* `parallelogram identity`

$$\|v+w\|^2 + \|v-w\|^2 = 2\|v\|^2 + 2\|w\|^2 \qquad \text{for all } v, w \in \mathcal{F} \tag{3.1}$$

holds. If $\mathcal{F}$ is a Euclidean space over the real numbers $\mathbb{R}$, then the `polarization identity`

$$(v,w) = \frac{1}{4}\left(\|v+w\|^2 - \|v-w\|^2\right) \qquad \text{for all } v, w \in \mathcal{F}. \tag{3.2}$$

holds. If $\mathcal{F}$ is a Euclidean space over the complex numbers $\mathbb{C}$, then the `polarization identity` *holds as*

$$(v,w) = \frac{1}{4}\left(\|v+w\|^2 - \|v-w\|^2 + \mathrm{i}\|v+\mathrm{i}w\|^2 - \mathrm{i}\|v-\mathrm{i}w\|^2\right) \tag{3.3}$$

for all $v, w \in \mathcal{F}$.

Proof. Equations (3.1),(3.2) follow directly from the identities

$$\|v \pm w\|^2 = (v \pm w, v \pm w) = (v,v) \pm 2(v,w) + (w,w) = \|v\|^2 \pm 2(v,w) + \|w\|^2.$$

Likewise, the polarization identity (3.3) can be verified by elementary calculations. ■

For the geometric interpretation of the parallelogram identity, we make the following remark.

For any parallelogram the sum of square lengths of the four edges coincides with the sum of the square lengths of the two diagonals (see Fig. 3.2).

For the statement in Theorem 3.9, the converse is true, according to the theorem of *Jordan*[1] *and von Neumann*[2] [40].

Theorem 3.10. (Jordan-von Neumann theorem, 1935).
Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$, for which the parallelogram identity (3.1) holds. Then there is an inner product $(\cdot,\cdot)$ on $\mathcal{F}$, so that

$$(v,v) = \|v\|^2 \qquad \text{for all } v \in \mathcal{F}, \tag{3.4}$$

i.e., $\mathcal{F}$ is a Euclidean space.

[1] Pascual Jordan (1902-1980), German mathematician and physicist
[2] John von Neumann (1903-1957), Hungarian-US American mathematician

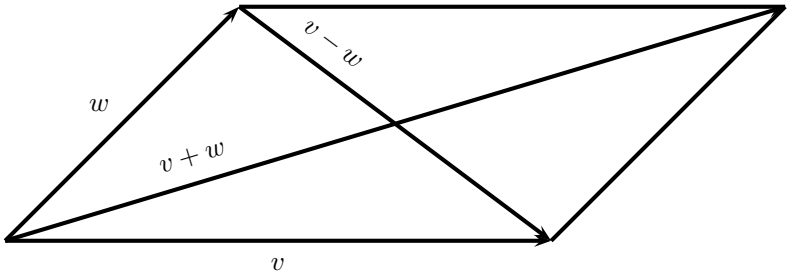


Fig. 3.2. On the geometry of the parallelogram identity (see Theorem 3.9).

Proof. Let $\mathcal{F}$ be a linear space over $\mathbb{R}$. By using the norm $\|\cdot\|$ of $\mathcal{F}$ we define a mapping $(\cdot,\cdot) : \mathcal{F}\times\mathcal{F} \longrightarrow \mathbb{R}$ through the polarization identity (3.2), i.e., we let

$$(v,w) := \frac{1}{4}\left(\|v+w\|^2 - \|v-w\|^2\right) \qquad \text{for } v,w \in \mathcal{F}.$$

Obviously, we have (3.4) and so $(\cdot,\cdot)$ is positive definite. Moreover, $(\cdot,\cdot)$ is obviously symmetric, so that $(v,w) = (w,v)$ for all $v,w \in \mathcal{F}$.

It remains to verify the linearity

$$(\alpha u + \beta v, w) = \alpha(u,w) + \beta(v,w) \qquad \text{for all } \alpha,\beta \in \mathbb{R}, u,v,w \in \mathcal{F}. \tag{3.5}$$

To this end, we note the property

$$(-v,w) = -(v,w) \qquad \text{for all } v,w \in \mathcal{F}, \tag{3.6}$$

which immediately follows from the definition of $(\cdot,\cdot)$. In particular, we have

$$(0,w) = 0 \qquad \text{for all } w \in \mathcal{F}.$$

Moreover, by the parallelogram identity (3.1) we obtain

$$\begin{aligned}(u,w) + (v,w) &= \frac{1}{4}\left(\|u+w\|^2 - \|u-w\|^2 + \|v+w\|^2 - \|v-w\|^2\right)\\ &= \frac{1}{2}\left(\left\|\frac{1}{2}(u+v)+w\right\|^2 - \left\|\frac{1}{2}(u+v)-w\right\|^2\right)\\ &= 2\left(\frac{1}{2}(u+v),w\right),\end{aligned}$$

which, for $v = 0$, implies

$$(u, w) = 2\left(\frac{1}{2}u, w\right) \qquad \text{for all } u, w \in \mathcal{F} \tag{3.7}$$

and thereby the *additivity*

$$(u, w) + (v, w) = (u + v, w) \qquad \text{for all } u, v, w \in \mathcal{F}. \tag{3.8}$$

From (3.7),(3.8) we obtain for $m, n \in \mathbb{N}$ the identities

$$m(u, w) = (mu, w) \qquad \text{for all } u, w \in \mathcal{F}$$
$$\frac{1}{2^n}(u, w) = \left(\frac{1}{2^n}u, w\right) \qquad \text{for all } u, w \in \mathcal{F}$$

by induction on $m \in \mathbb{N}$ and by induction on $n \in \mathbb{N}$, respectively.

In combination with (3.6) and (3.8) this implies the *homogeneity*

$$(\alpha u, w) = \alpha(u, w) \qquad \text{for all } u, w \in \mathcal{F} \tag{3.9}$$

for all *dyadic* numbers $\alpha \in \mathbb{Q}$ of the form

$$\alpha = m + \sum_{k=1}^{n} \frac{\alpha_k}{2^k} \qquad \text{for } m \in \mathbb{Z}, n \in \mathbb{N}, \alpha_k \in \{0, 1\}, 1 \leq k \leq n.$$

Since any real number $\alpha \in \mathbb{R}$ can be approximated arbitrarily well by a dyadic number, the continuity of the norm $\|\cdot\|$ implies the homogeneity (3.9) even for all $\alpha \in \mathbb{R}$. Together with the additivity (3.8) this implies the linearity (3.5). Therefore, $(\cdot, \cdot)$ is an inner product over $\mathbb{R}$.

If $\mathcal{F}$ is a linear space over $\mathbb{C}$, then we define $(\cdot, \cdot) : \mathcal{F} \times \mathcal{F} \longrightarrow \mathbb{C}$ through the polarization identity (3.3), for which we then verify the properties of an inner product for $(\cdot, \cdot)$ in (3.2), by analogy. ■

Given the above characterization of Euclidean norms by the parallelogram identity (3.1) and the polarization identities (3.2),(3.3), approximation in Euclidean spaces is of particular importance. From the equivalence relation in Theorems 3.9 and 3.10, we can immediately draw the following conclusion.

Corollary 3.11. *Every inner product is continuous.*

Proof. Let $\mathcal{F}$ be a Euclidean space over $\mathbb{R}$ with inner product $(\cdot, \cdot)$. Moreover, let $(v_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ and $(w_n)_{n\in\mathbb{N}} \subset \mathcal{F}$ be convergent sequences in $\mathcal{F}$ with limit elements $v \in \mathcal{F}$ and $w \in \mathcal{F}$. From the polarization identity (3.2) and by the continuity of the norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$, from Theorem 3.4, we have

$$(v_n, w_m) = \frac{1}{4}\left(\|v_n + w_m\|^2 - \|v_n - w_m\|^2\right)$$
$$\longrightarrow \frac{1}{4}\left(\|v + w\|^2 - \|v - w\|^2\right) = (v, w) \qquad \text{for } n, m \to \infty.$$

For the case where $\mathcal{F}$ is a Euclidean space over $\mathbb{C}$, we show the continuity of $(\cdot, \cdot)$ from the polarization identity (3.3), by analogy. ■

Now we return to the question for the existence of best approximations. In Euclidean spaces $\mathcal{F}$ we can rely on the parallelogram identity (3.1). Moreover, we need the *completeness* of $\mathcal{F}$. On this occasion, we recall the following definition.

Definition 3.12. *A complete Euclidean space is called* `Hilbert`[3] `space`. ◯

Moreover, let us recall the notion of strictly convex sets.

Definition 3.13. *A non-empty subset $\mathcal{K} \subset \mathcal{F}$ is called* `convex`*, if for any $u, v \in \mathcal{K}$ the straight line*

$$[u, v] = \{\lambda u + (1 - \lambda)v \,|\, \lambda \in [0, 1]\}$$

between u and v lies in $\mathcal{K}$, i.e., if $[u, v] \subset \mathcal{K}$ for all $u, v \in \mathcal{K}$.

If for any $u, v \in \mathcal{K}$, $u \neq v$, the open straight line

$$(u, v) = \{\lambda u + (1 - \lambda)v \,|\, \lambda \in (0, 1)\}$$

is contained in the interior of $\mathcal{K}$, then $\mathcal{K}$ is called `strictly convex`. ◯

Now we prove an important result concerning the existence of best approximations in Hilbert spaces.

Theorem 3.14. *Let $\mathcal{F}$ be a Hilbert space with inner product $(\cdot, \cdot)$ and norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$. Moreover, let $\mathcal{S} \subset \mathcal{F}$ be a closed and convex subset of $\mathcal{F}$. Then there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.*

Proof. Let $(s_n)_{n\in\mathbb{N}} \subset \mathcal{S}$ be a `minimal sequence` in $\mathcal{S}$, i.e.,

$$\|s_n - f\| \longrightarrow \eta(f, \mathcal{S}) \quad \text{for } n \to \infty$$

with minimal distance $\eta \equiv \eta(f, \mathcal{S}) = \inf_{s\in\mathcal{S}} \|s - f\|$.

From the parallelogram identity (3.1) we obtain the estimate

$$\begin{aligned} \|s_n - s_m\|^2 &= 2\|s_n - f\|^2 + 2\|s_m - f\|^2 - 4\left\|\frac{s_n + s_m}{2} - f\right\|^2 \\ &\leq 2\|s_n - f\|^2 + 2\|s_m - f\|^2 - 4\eta^2. \end{aligned}$$

Therefore, for any $\varepsilon > 0$ there is one $N \equiv N(\varepsilon) \in \mathbb{N}$ satisfying

$$\|s_n - s_m\| < \varepsilon \qquad \text{for all } n, m \geq N,$$

i.e., $(s_n)_{n\in\mathbb{N}}$ is a `Cauchy`[4] `sequence` in the Hilbert space $\mathcal{F}$, and therefore convergent in $\mathcal{F}$. Since $\mathcal{S}$ is a closed set, the limit element s^* lies in $\mathcal{S}$, and we have

$$\eta = \lim_{n\to\infty} \|s_n - f\| = \|s^* - f\|,$$

i.e., $s^* \in \mathcal{S}$ is a best approximation to f. ■

[3] David Hilbert (1862-1943), German mathematician

[4] Augustin-Louis Cauchy (1789-1857), French mathematician

Remark 3.15. The required convexity for $\mathcal{S}$ is necessary for the result of Theorem 3.14. In order to see this, we regard the sequence space

$$\ell^2 \equiv \ell^2(\mathbb{R}) = \left\{ x = (x_k)_{k\in\mathbb{N}} \subset \mathbb{R} \,\middle|\, \sum_{k=1}^{\infty} |x_k|^2 < \infty \right\} \tag{3.10}$$

consisting of all square summable sequences of real numbers. The sequence space ℓ^2, being equipped with the inner product

$$(x, y) = \sum_{k=1}^{\infty} x_k y_k \qquad \text{for } x = (x_k)_{k\in\mathbb{N}}, y = (y_k)_{k\in\mathbb{N}} \in \ell^2$$

is a Hilbert space with the ℓ^2-norm

$$\|x\|_2 := \sqrt{\sum_{k=1}^{\infty} |x_k|^2} \qquad \text{for } x = (x_k)_{k\in\mathbb{N}} \in \ell^2.$$

Now we regard the subset

$$\mathcal{S} = \left\{ x^{(k)} = \left(1 + \frac{1}{k}\right) e_k \,\middle|\, k \in \mathbb{N} \right\} \subset \ell^2,$$

where $e_k \in \ell^2$ is the sequence with $(e_k)_j = \delta_{jk}$, for $j, k \in \mathbb{N}$. Note that the elements $x^{(k)} \in \mathcal{S}$ are isolated in ℓ^2, and so $\mathcal{S}$ is closed. But $\mathcal{S}$ is *not* convex. Now we have $\eta(0, \mathcal{S}) = 1$ for the minimal distance between $0 \in \ell^2$ and $\mathcal{S}$, and, moreover,

$$\|x^{(k)} - 0\|_2 > 1 \qquad \text{for all } x^{(k)} \in \mathcal{S}.$$

Hence there exists no $x^{(k)} \in \mathcal{S}$ with unit distance to the origin.

Finally, we remark that the result of Theorem 3.14 does *not* generalize to Banach spaces. To see this, a counterexample can for instance be found in [42, Section 5.2]. □

3.2 Uniqueness

In the following discussion, the notion of (*strict*) convexity for point sets, functions, functionals and norms plays an important role. Recall the relevant definitions for sets (see Definition 3.13) and for functions (see Definition 3.20), as these should be familiar from analysis.

Now we note some fundamental results, where $\mathcal{F}$ denotes, throughout this section, a linear space with norm $\|\cdot\|$. We start with a relevant example for a convex set.

Theorem 3.16. *Let $\mathcal{S} \subset \mathcal{F}$ be convex and $f \in \mathcal{F}$. Then the set*

$$\mathcal{S}^* \equiv \mathcal{S}^*(f, \mathcal{S}) = \{s^* \in \mathcal{S} \mid \|s^* - f\| = \inf_{s \in \mathcal{S}} \|s - f\|\} \subset \mathcal{S}$$

of best approximations $s^* \in \mathcal{S}$ to f is convex.

Proof. Let $s_1^*, s_2^* \in \mathcal{S}^*$ be two best approximations to $f \in \mathcal{F}$. Then, for any element

$$s_\lambda^* = \lambda s_1^* + (1-\lambda) s_2^* \in [s_1^*, s_2^*] \subset \mathcal{S} \qquad \text{for } \lambda \in [0,1] \tag{3.11}$$

we have

$$\begin{aligned} \|s_\lambda^* - f\| &= \|(\lambda s_1^* + (1-\lambda)s_2^*) - (\lambda + (1-\lambda))f\| \\ &= \|\lambda(s_1^* - f) + (1-\lambda)(s_2^* - f)\| \\ &\le \lambda \|s_1^* - f\| + (1-\lambda)\|s_2^* - f\| \\ &= \lambda \inf_{s \in \mathcal{S}} \|s - f\| + (1-\lambda) \inf_{s \in \mathcal{S}} \|s - f\| \\ &= \inf_{s \in \mathcal{S}} \|s - f\|, \end{aligned}$$

i.e., $s_\lambda^* = \lambda s_1^* + (1-\lambda)s_2^* \in [s_1^*, s_2^*]$, for $\lambda \in [0,1]$, lies in $\mathcal{S}^*$. ∎

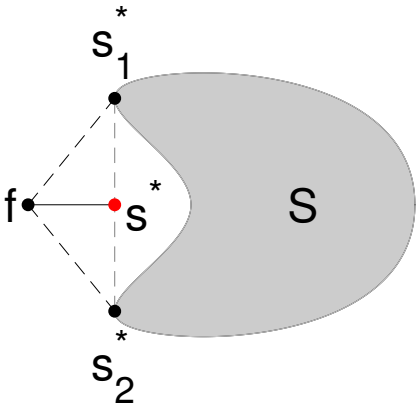


Fig. 3.3. $\mathcal{S}$ is not convex and for $s^* \notin \mathcal{S}$ we have $\|s^* - f\| < \eta(f, \mathcal{S})$, cf. Remark 3.17.

We continue with the following remarks concerning Theorem 3.16.

Remark 3.17. If $\mathcal{S} \subset \mathcal{F}$ is in the situation of Theorem 3.16 *not* convex, then any element $s^* \in [s_1^*, s_2^*]$ is at least as close to $f \in \mathcal{F}$ as s_1^* and s_2^* are, i.e.,

$$\|s^* - f\| \le \eta \equiv \eta(f, \mathcal{S}) \qquad \text{for all } s^* \in [s_1^*, s_2^*].$$

Eventually, one $s^* \in [s_1^*, s_2^*]$ could even lie closer to f than s_1^*, s_2^*, so that $\|s^* - f\| < \eta$, as the example in Figure 3.3 shows. □

Remark 3.18. For a convex subset $\mathcal{S} \subset \mathcal{F}$, and in the case of non-unique best approximations, we can explain the situation as follows. If there are at least two best approximations $s_1^* \neq s_2^*$ to f, then *all* $s^* \in [s_1^*, s_2^*]$ are contained in the set of best approximations $\mathcal{S}^*$, and so the distance between f and elements in $[s_1^*, s_2^*]$ is constant, i.e.,

$$\|s^* - f\| = \eta(f, \mathcal{S}) \qquad \text{for all } s^* \in [s_1^*, s_2^*].$$

□

To further illustrate this, let us make one simple example.

Example 3.19. For $\mathcal{S} = \{x \in \mathbb{R}^2 \mid \|x\|_\infty \leq 1\}$ and $f = (2, 0)$, the set $\mathcal{S}^*$ of best approximations to f from $\mathcal{S}$ with respect to the maximum norm $\|\cdot\|_\infty$ is given by

$$\mathcal{S}^* = \big\{(1, \alpha) \in \mathbb{R}^2 \,|\, \alpha \in [-1, 1]\big\} \subset \mathcal{S}$$

with the minimal distance

$$\eta(f, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|s - f\|_\infty = 1.$$

For $s_1^*, s_2^* \in \mathcal{S}^*$ every element $s^* \in [s_1^*, s_2^*]$ lies in $\mathcal{S}^*$ (see Figure 3.4). ♢

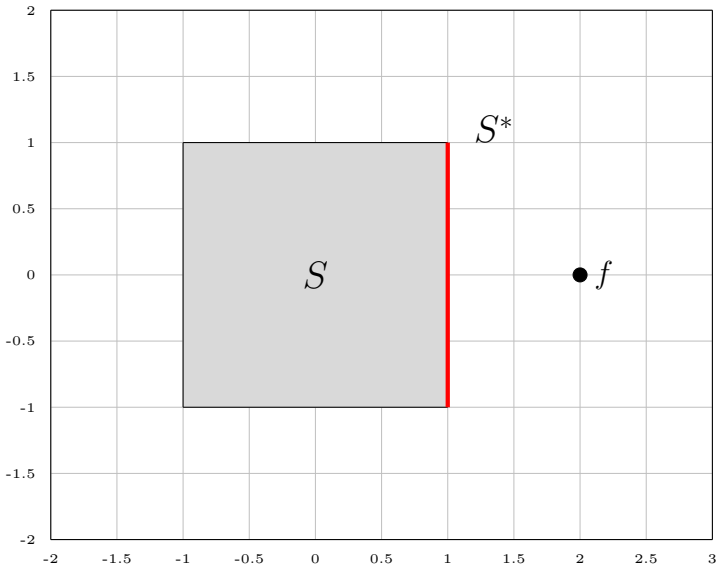


Fig. 3.4. $\mathcal{S}^* = \big\{(1, \alpha) \in \mathbb{R}^2 \,|\, \alpha \in [-1, 1]\big\}$ is the set of best approximations to $f = (2, 0)$ from $\mathcal{S} = \{x \in \mathbb{R}^2 \mid \|x\|_\infty \leq 1\}$ with respect to $\|\cdot\|_\infty$ (see Example 3.19).

Next, we recall the definition for (strictly) convex functions.

Definition 3.20. *A function* $f : [a,b] \longrightarrow \mathbb{R}$ *is called* `convex` *on an interval* $[a,b] \subset \mathbb{R}$*, if for all* $x, y \in [a,b]$ *the inequality*

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y) \qquad \text{for all } \lambda \in [0,1]$$

holds; f *is said to be* `strictly convex` *on* $[a,b]$*, if for all* $x, y \in [a,b]$*,* $x \neq y$*, we have*

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda) f(y) \qquad \text{for all } \lambda \in (0,1).$$

◯

An important property of convex functions is described by the *Jensen*[5] *inequality* [39], whereby the value of a convex function, when evaluated at a finite convex combination of arguments, is bounded above by the corresponding convex combination of functions values at these arguments.

Theorem 3.21. (Jensen inequality, 1906).
Let $f : [a,b] \longrightarrow \mathbb{R}$ *be a convex function, and* $\{x_1, \ldots, x_n\} \subset [a,b]$ *a set of* $n \geq 2$ *points. Then, the* `Jensen inequality`

$$f\left(\sum_{j=1}^{n} \lambda_j x_j\right) \leq \sum_{j=1}^{n} \lambda_j f(x_j) \qquad \text{for all } \lambda_j \in (0,1) \text{ with } \sum_{j=1}^{n} \lambda_j = 1$$

holds. If f *is strictly convex, then equality holds, if and only if all points coincide, i.e.,* $x_1 = \ldots = x_n$*.*

Proof. We prove the statement of Jensen's inequality by induction on n.
Initial step: For $n = 2$, the statement of Jensen's inequality is obviously true.
Induction hypothesis: Assume the statement holds for n points $\{x_1, \ldots, x_n\}$.
Induction step $(n \longrightarrow n+1)$*:* For $n+1$ points $\{x_1, \ldots, x_n, x_{n+1}\} \subset [a,b]$ and

$$\lambda_1, \ldots, \lambda_n, \lambda_{n+1} \in (0,1) \qquad \text{with } \sum_{j=1}^{n} \lambda_j = 1 - \lambda_{n+1}$$

we have

$$\begin{aligned} f\left(\sum_{j=1}^{n+1} \lambda_j x_j\right) &= f\left((1-\lambda_{n+1}) \sum_{j=1}^{n} \frac{\lambda_j}{1-\lambda_{n+1}} x_j + \lambda_{n+1} x_{n+1}\right) \\ &\leq (1-\lambda_{n+1}) f\left(\sum_{j=1}^{n} \frac{\lambda_j}{1-\lambda_{n+1}} x_j\right) + \lambda_{n+1} f(x_{n+1}) \end{aligned}$$

[5] Johan Ludwig Jensen (1859-1925), Danish mathematician

by the convexity of f. By the induction hypothesis, we can further conclude

$$f\left(\sum_{j=1}^{n}\frac{\lambda_j}{1-\lambda_{n+1}}x_j\right)\leq\sum_{j=1}^{n}\frac{\lambda_j}{1-\lambda_{n+1}}f(x_j)=\frac{1}{1-\lambda_{n+1}}\sum_{j=1}^{n}\lambda_j f(x_j) \quad (3.12)$$

and thus, altogether,

$$f\left(\sum_{j=1}^{n+1}\lambda_j x_j\right)\leq\sum_{j=1}^{n+1}\lambda_j f(x_j). \quad (3.13)$$

If f is strictly convex, then equality holds in (3.12) only for $x_1=\ldots=x_n$ (by induction hypothesis), and, moreover, equality in (3.13) holds only for

$$x_{n+1}=\sum_{j=1}^{n}\frac{\lambda_j}{1-\lambda_{n+1}}x_j,$$

thus altogether only for $x_1=\ldots=x_n=x_{n+1}$. ■

Next, we introduce the convexity for functionals.

Definition 3.22. *A functional $\varphi:\mathcal{F}\longrightarrow\mathbb{R}$ is said to be* `convex` *on $\mathcal{F}$, if for all $u,v\in\mathcal{F}$ the inequality*

$$\varphi(\lambda u+(1-\lambda)v)\leq\lambda\varphi(u)+(1-\lambda)\varphi(v)\qquad \text{for all } \lambda\in[0,1] \quad (3.14)$$

holds. ◯

Remark 3.23. Every norm $\|\cdot\|:\mathcal{F}\longrightarrow[0,\infty)$ is a convex functional on $\mathcal{F}$. Indeed, for any $u,v\in\mathcal{F}$ we find the inequality

$$\|\lambda u+(1-\lambda)v\|\leq\lambda\|u\|+(1-\lambda)\|v\|\qquad \text{for all } \lambda\in[0,1] \quad (3.15)$$

due to the triangle inequality and the homogeneity of $\|\cdot\|$. Moreover, equality in (3.15) holds for all pairs of linearly dependent elements $u,v\in\mathcal{F}$ with $u=\alpha v$ for positive scalar $\alpha>0$, i.e., we have

$$\|\lambda\alpha v+(1-\lambda)v\|=\lambda\|\alpha v\|+(1-\lambda)\|v\|\qquad \text{for all } \lambda\in[0,1] \quad (3.16)$$

by the homogeneity of $\|\cdot\|$. □

We introduce the notion of a *strictly* convex norm classically as follows.

Definition 3.24. *A norm $\|\cdot\|$ is called* `strictly convex` *on $\mathcal{F}$, if the unit ball $\mathcal{B}=\{u\in\mathcal{F}\,|\,\|u\|\leq 1\}\subset\mathcal{F}$ is strictly convex.* ◯

As we will show, not every norm is strictly convex. But before we do so, our "classical" introduction for strictly convex norms in Definition 3.24 deserves a comment.

Remark 3.25. Had we introduced the strict convexity of $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ in Definition 3.22 in a straightforward manner through the inequality

$$\varphi(\lambda u + (1-\lambda)v) < \lambda\varphi(u) + (1-\lambda)\varphi(v) \qquad \text{for all } \lambda \in (0,1), \tag{3.17}$$

then *no* norm would be strictly convex in this particular sense! This important observation is verified by the counterexample in (3.16). □

When working with strictly convex norms $\|\cdot\|$ (according to Definition 3.24), we can exclude non-uniqueness of best approximations, if $\mathcal{S} \subset \mathcal{F}$ is convex. To explain this, we need to further analyze strictly convex norms. To this end, we first prove the following useful characterization.

Theorem 3.26. *Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$. Then the following statements are equivalent.*

(a) The norm $\|\cdot\|$ is strictly convex.
(b) The unit ball $\mathcal{B} = \{u \in \mathcal{F} \,|\, \|u\| \le 1\} \subset \mathcal{F}$ is strictly convex.
(c) The inequality $\|u+v\| < 2$ holds for all $u \neq v$, with $\|u\| = \|v\| = 1$.
(d) The equality $\|u+v\| = \|u\| + \|v\|$, $v \neq 0$, implies $u = \alpha v$ for some $\alpha \ge 0$.

Proof. Note that the equivalence $(a) \Leftrightarrow (b)$ holds by Definition 3.24.

$(b) \Rightarrow (c)$: The strict convexity of $\mathcal{B}$ implies $\|(u+v)/2\| < 1$ for $u \neq v$ with $\|u\| = \|v\| = 1$, and so in this case we have $\|u+v\| < 2$.

$(c) \Rightarrow (d)$: For $u = 0$ statement (d) holds with $\alpha = 0$. Now suppose $u, v \in \mathcal{F} \setminus \{0\}$ satisfy $\|u+v\| = \|u\| + \|v\|$. Without loss of generality, we may assume $\|u\| \le \|v\|$ (otherwise we swap u and v). In this case, in the sequence of inequalities

$$\begin{aligned} 2 &\ge \left\| \frac{u}{\|u\|} + \frac{v}{\|v\|} \right\| = \left\| \left(\frac{u}{\|u\|} + \frac{v}{\|u\|} \right) - \left(\frac{v}{\|u\|} - \frac{v}{\|v\|} \right) \right\| \\ &\ge \left\| \frac{u}{\|u\|} + \frac{v}{\|u\|} \right\| - \left\| \frac{v}{\|u\|} - \frac{v}{\|v\|} \right\| = \frac{\|u+v\|}{\|u\|} - \left| \frac{1}{\|u\|} - \frac{1}{\|v\|} \right| \|v\| \\ &= \frac{\|u\| + \|v\|}{\|u\|} - \left(\frac{1}{\|u\|} - \frac{1}{\|v\|} \right) \|v\| = 2 \end{aligned}$$

equality holds everywhere, in particular

$$\left\| \frac{u}{\|u\|} + \frac{v}{\|v\|} \right\| = 2.$$

From (c) we can conclude $u/\|u\| = v/\|v\|$ and therefore

$$u = \alpha v \qquad \text{for } \alpha = \frac{\|u\|}{\|v\|} > 0.$$

$(d) \Rightarrow (b)$: Suppose $u, v \in \mathcal{B}$, $u \neq v$, i.e., $\|u\| \le 1$ and $\|v\| \le 1$. Then we find for any $\lambda \in (0,1)$ the inequality

$$\|\lambda u + (1-\lambda)v\| \le \lambda\|u\| + (1-\lambda)\|v\| < 1,$$

provided that $\|u\| < 1$ or $\|v\| < 1$. Otherwise, i.e., if $\|u\| = \|v\| = 1$, we have

$$\|\lambda u\| + \|(1-\lambda)v\| = \lambda\|u\| + (1-\lambda)\|v\| = 1 \qquad \text{for } \lambda \in (0,1).$$

If $\|\lambda u + (1-\lambda)v\| = 1$, then we have $\lambda u = \alpha(1-\lambda)v$ for one $\alpha > 0$ from (d). Therefore, we have $u = v$, since $\|u\| = \|v\|$. This, however, is in contradiction to the assumption $u \neq v$. Therefore, we have, also for this case,

$$\|\lambda u + (1-\lambda)v\| < 1 \qquad \text{for all } \lambda \in (0,1).$$

■

Next, we make explicit examples of strictly convex norms. A first simple example is the absolute value $|\cdot|$, taken as a norm on $\mathbb{R}$.

Remark 3.27. The absolute value $|\cdot|$ is a strictly convex norm on $\mathbb{R}$. Indeed, in the equivalence (c) of Theorem 3.26 we can only use the two points $u = -1$ and $v = 1$, where we have $|u+v| = 0 < 2$. But note that the absolute value, when regarded as a function $|\cdot| : \mathbb{R} \longrightarrow \mathbb{R}$ is *not* strictly convex on $\mathbb{R}$. □

Further examples are Euclidean norms.

Theorem 3.28. *Every Euclidean norm is strictly convex.*

Proof. Let $\mathcal{F}$ be a linear space with Euclidean norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$. By Theorem 3.9, the parallelogram inequality (3.1) holds in $\mathcal{F}$, and so

$$\left\|\frac{u+v}{2}\right\|^2 + \left\|\frac{u-v}{2}\right\|^2 = \frac{\|u\|^2}{2} + \frac{\|v\|^2}{2} \qquad \text{for all } u, v \in \mathcal{F}.$$

For $u, v \in \mathcal{F}$, $u \neq v$, with $\|u\| = \|v\|$ we thus have

$$\left\|\frac{u+v}{2}\right\|^2 < \|u\|^2 = \|v\|^2,$$

or, $\|u+v\| < 2$ for $\|u\| = \|v\| = 1$.

By statement (c) in Theorem 3.26, we see that $\|\cdot\|$ is strictly convex. ■

Next, we regard the linear space of all bounded sequences,

$$\ell^\infty \equiv \ell^\infty(\mathbb{R}) = \left\{ x = (x_k)_{k\in\mathbb{N}} \subset \mathbb{R} \;\middle|\; \sup_{k\in\mathbb{N}} |x_k| < \infty \right\},$$

equipped with the ℓ^∞-norm

$$\|x\|_\infty := \sup_{k\in\mathbb{N}} |x_k| \qquad \text{for } x = (x_k)_{k\in\mathbb{N}} \in \ell^\infty.$$

Moreover, we regard for $1 \le p < \infty$ the linear subspaces

$$\ell^p \equiv \ell^p(\mathbb{R}) = \left\{ x = (x_k)_{k\in\mathbb{N}} \subset \mathbb{R} \;\middle|\; \sum_{k=1}^{\infty} |x_k|^p < \infty \right\} \subset \ell^\infty, \tag{3.18}$$

equipped with the ℓ^p-norm

$$\|x\|_p := \left(\sum_{k=1}^{\infty} |x_k|^p \right)^{1/p} \qquad \text{for } x = (x_k)_{k\in\mathbb{N}} \in \ell^p.$$

To further analyze the ℓ^p-norms we prove the *Hölder*[6] *inequality.*

Theorem 3.29. (Hölder inequality, 1889).
Let $1 < p, q < \infty$ *satisfy* $1/p + 1/q = 1$. *Then, the* `Hölder inequality`

$$\|xy\|_1 \le \|x\|_p \|y\|_q \qquad \textit{for all } x \in \ell^p, y \in \ell^q, \tag{3.19}$$

holds with equality in (3.19), if and only if either $x = 0$ *or* $y = 0$ *or*

$$|x_k|^{p-1} = \alpha |y_k| \quad \textit{with } \alpha = \frac{\|x\|_p^{p-1}}{\|y\|_q} > 0 \qquad \textit{for } y \neq 0. \tag{3.20}$$

Proof. For $1 < p, q < \infty$ with $1/p + 1/q = 1$ let

$$x = (x_k)_{k\in\mathbb{N}} \in \ell^p \quad \text{and} \quad y = (y_k)_{k\in\mathbb{N}} \in \ell^q.$$

For $x = 0$ or $y = 0$ the Hölder inequality (3.19) is trivial. Now suppose $x, y \neq 0$. Then, we find for $k \in \mathbb{N}$ the estimate

$$-\log\left(\frac{1}{p} \frac{|x_k|^p}{\|x\|_p^p} + \frac{1}{q} \frac{|y_k|^q}{\|y\|_q^q} \right) \le -\frac{1}{p} \log\left(\frac{|x_k|^p}{\|x\|_p^p} \right) - \frac{1}{q} \log\left(\frac{|y_k|^q}{\|y\|_q^q} \right) \tag{3.21}$$

by the Jensen inequality, Theorem 3.21, here applied to the strictly convex function $-\log : (0, \infty) \longrightarrow \mathbb{R}$. This yields the *Young*[7] *inequality*

$$\frac{|x_k y_k|}{\|x\|_p \|y\|_q} = \left(\frac{|x_k|^p}{\|x\|_p^p} \right)^{1/p} \left(\frac{|y_k|^q}{\|y\|_q^q} \right)^{1/q} \le \frac{1}{p} \frac{|x_k|^p}{\|x\|_p^p} + \frac{1}{q} \frac{|y_k|^q}{\|y\|_q^q}. \tag{3.22}$$

Moreover, by Theorem 3.21, we have equality in (3.21), and therefore equality in (3.22), if and only if

$$\frac{|x_k|^p}{\|x\|_p^p} = \frac{|y_k|^q}{\|y\|_q^q}. \tag{3.23}$$

By $q = p/(p-1)$ we see that (3.23) is equivalent to

[6] Hölder, Otto (1859-1937), German mathematician
[7] William Henry Young (1863-1942), English mathematician

$$\frac{|x_k|}{\|x\|_p} = \left(\frac{|y_k|}{\|y\|_q}\right)^{1/(p-1)}. \tag{3.24}$$

Therefore, we have equality in (3.22), if and only if (3.20) holds. Summing up both sides in the Young inequality (3.22) over k, we find

$$\sum_{k=1}^{\infty} \frac{|x_k y_k|}{\|x\|_p \|y\|_q} \le \sum_{k=1}^{\infty} \frac{1}{p} \frac{|x_k|^p}{\|x\|_p^p} + \sum_{k=1}^{\infty} \frac{1}{q} \frac{|y_k|^q}{\|y\|_q^q} = \frac{1}{p} + \frac{1}{q} = 1,$$

and this already proves the Hölder inequality (3.19), with equality, if and only if (3.20) holds for all $k \in \mathbb{N}$. ■

Now we can show the strict convexity of the ℓ^p-norms, for $1 < p < \infty$.

Theorem 3.30. *For $1 < p < \infty$, the ℓ^p-norm $\|\cdot\|_p$ on ℓ^p is strictly convex.*

Proof. For $1 < p < \infty$, let $1 < q < \infty$ be the *conjugate Hölder exponent* of p satisfying $1/p + 1/q = 1$.

For

$$x = (x_k)_{k\in\mathbb{N}} \quad \text{and} \quad y = (y_k)_{k\in\mathbb{N}} \in \ell^p,$$

where $x \neq y$ and $\|x\|_p = \|y\|_p = 1$, we wish to prove the inequality

$$\|x + y\|_p < 2, \tag{3.25}$$

in which case the norm $\|\cdot\|_p$ would be strictly convex by the equivalence statement (c) in Theorem 3.26.

For $s_k := |x_k + y_k|^{p-1}$ and $s := (s_k)_{k\in\mathbb{N}} \in \ell^q$ we have

$$\begin{aligned} \|x + y\|_p^p &= \sum_{k=1}^{\infty} |x_k + y_k||s_k| \\ &\le \sum_{k=1}^{\infty} (|x_k||s_k| + |y_k||s_k|) \qquad (3.26) \\ &\le \|x\|_p \|s\|_q + \|y\|_p \|s\|_q, \qquad (3.27) \end{aligned}$$

where we applied the Hölder inequality (3.19) in (3.27) twice.

By $p = (p-1)q$, we have

$$\|s\|_q = \left(\sum_{k=1}^{\infty} |x_k + y_k|^{(p-1)q}\right)^{1/q} = \left(\sum_{k=1}^{\infty} |x_k + y_k|^p\right)^{\frac{1}{p}\cdot\frac{p}{q}} = \|x + y\|_p^{p-1}$$

and this implies, in combination with (3.27), the *Minkowski*[8] *inequality*

[8] Hermann Minkowski (1864-1909), German mathematician and physicist

$$\|x+y\|_p \le \|x\|_p + \|y\|_p \qquad \text{for all } x, y \in \ell^p,$$

in particular,

$$\|x+y\|_p \le 2 \qquad \text{for } \|x\|_p = \|y\|_p = 1.$$

If $\|x+y\|_p = 2$ for $\|x\|_p = \|y\|_p = 1$, then we have equality in both (3.26) and (3.27). But equality in (3.27) is by (3.20) equivalent to the two conditions

$$|x_k|^{p-1} = \alpha |s_k| \quad \text{and} \quad |y_k|^{p-1} = \alpha |s_k| \qquad \text{with } \alpha = \frac{1}{\|s\|_q},$$

which implies

$$|x_k| = |y_k| \qquad \text{for all } k \in \mathbb{N}.$$

In this case, we have equality in (3.26), if and only if $\operatorname{sgn}(x_k) = \operatorname{sgn}(y_k)$, for all $k \in \mathbb{N}$, i.e., equality in (3.26) and (3.27) implies $x = y$.

Therefore, the inequality (3.25) holds for all $x \neq y$ with $\|x\|_p = \|y\|_p = 1$. ■

Remark 3.31. Theorem 3.30 can be generalized to L^p-norms,

$$\|u\|_p := \left(\int_{\mathbb{R}^d} |u(x)|^p \, \mathrm{d}x \right)^{1/p} \qquad \text{for } u \in \mathrm{L}^p,$$

for $1 < p < \infty$, where $\mathrm{L}^p \equiv \mathrm{L}^p(\mathbb{R}^d)$ is the linear space of all functions whose p-th power is *Lebesgue*[9] *integrable*. Indeed, in this case (in analogy to Theorem 3.29) the *Hölder inequality*

$$\|uv\|_1 \le \|u\|_p \|v\|_q \qquad \text{for all } u \in \mathrm{L}^p, v \in \mathrm{L}^q$$

holds for $1 < p, q < \infty$ satisfying $1/p + 1/q = 1$. This implies (as in the proof of Theorem 3.30) the *Minkowski inequality*

$$\|u+v\|_p \le \|u\|_p + \|v\|_p \qquad \text{for all } u, v \in \mathrm{L}^p,$$

where for $1 < p < \infty$ we have equality, if and only if $u = \alpha v$ for some $\alpha \ge 0$ (see [35, Theorem 12.6]). Therefore, the L^p-norm $\|\cdot\|_p$, forr $1 < p < \infty$, is by equivalence statement (d) in Theorem 3.26 strictly convex. □

We can conclude our statement from Remark 3.31 as follows.

Theorem 3.32. *For $1 < p < \infty$, the L^p-norm $\|\cdot\|_p$ on L^p is strictly convex.* □

But there are norms that are *not* strictly convex. Here are two examples.

[9] Henri Léon Lebesgue (1875-1941), French mathematician

Example 3.33. The ℓ^1-norm $\|\cdot\|_1$ on ℓ^1 in (3.18), defined as

$$\|x\|_1 = \sum_{k=1}^{\infty} |x_k| \qquad \text{for } x = (x_k)_{k\in\mathbb{N}} \in \ell^1,$$

is not strictly convex, since the unit ball $\mathcal{B}_1 = \{x \in \ell^1 \,|\, \|x\|_1 \le 1\} \subset \ell^1$ is not strictly convex. Indeed, for any pair of two unit vectors $e_j, e_k \in \ell^1$, $j \ne k$, we have $\|e_j\|_1 = \|e_k\|_1 = 1$ and

$$\|\lambda e_j + (1-\lambda)e_k\|_1 = \lambda + (1-\lambda) = 1 \qquad \text{for all } \lambda \in (0,1).$$

Thus, by Theorem 3.26, statement (b), the ℓ^1-norm $\|\cdot\|_1$ is not strictly convex.

Likewise, we show that for the linear space ℓ^∞ of all bounded sequences the ℓ^∞-norm $\|\cdot\|_\infty$, defined as

$$\|x\|_\infty = \sup_{k\in\mathbb{N}} |x_k| \qquad \text{for } x = (x_k)_{k\in\mathbb{N}} \in \ell^\infty,$$

is not strictly convex. This is because for any $e_k \in \ell^\infty$, $k \in \mathbb{N}$, and the constant sequence $\mathbf{1} = (1)_{k\in\mathbb{N}} \in \ell^\infty$, we have $\|e_k\|_\infty = \|\mathbf{1}\|_\infty = 1$ and

$$\|\lambda e_k + (1-\lambda)\mathbf{1}\|_\infty = 1 \qquad \text{for all } \lambda \in (0,1).$$

♢

Example 3.34. For the linear space $\mathscr{C}([0,1]^d)$ of all continuous functions on the unit cube $[0,1]^d \subset \mathbb{R}^d$, the maximum norm $\|\cdot\|_\infty$, defined as

$$\|u\|_\infty = \max_{x\in[0,1]^d} |u(x)| \qquad \text{for } u \in \mathscr{C}([0,1]^d),$$

is not strictly convex. To see this we take a continuous function $u_1 \in \mathscr{C}([0,1]^d)$ satisfying $\|u_1\|_\infty = 1$ and another continuous function $u_2 \in \mathscr{C}([0,1]^d)$ satisfying $\|u_2\|_\infty = 1$, so that $|u_1|$ and $|u_2|$ attain their maximum on $[0,1]^d$ at one point $x^* \in [0,1]^d$, respectively, i.e.,

$$\|u_1\|_\infty = \max_{x\in[0,1]^d} |u_1(x)| = |u_1(x^*)| = |u_2(x^*)| = \max_{x\in[0,1]^d} |u_2(x)| = \|u_2\|_\infty = 1.$$

This then implies for $u_\lambda = \lambda u_1 + (1-\lambda)u_2 \in (u_1, u_2)$, with $\lambda \in (0,1)$,

$$|u_\lambda(x)| \le \lambda|u_1(x)| + (1-\lambda)|u_2(x)| \le 1 \qquad \text{for all } x \in [0,1]^d$$

where equality holds for $x = x^*$, whereby $\|u_\lambda\|_\infty = 1$ for all $\lambda \in (0,1)$.

In this case, the unit ball $\mathcal{B} = \{u \in \mathscr{C}([0,1]^d) \,|\, \|u\|_\infty \le 1\}$ is not strictly convex, i.e., $\|\cdot\|_\infty$ is not strictly convex by statement (b) in Theorem 3.26.

To make an explicit example for the required functions u_1 and u_2, we take the *geometric mean* $u_g \in \mathscr{C}([0,1]^d)$ and the *arithmetic mean* $u_a \in \mathscr{C}([0,1]^d)$,

$$u_g(x) = \sqrt[d]{x_1 \cdot \ldots \cdot x_d} \leq \frac{x_1 + \ldots + x_d}{d} = u_a(x),$$

for $x = (x_1, \ldots, x_d) \in [0,1]^d$. Obviously, we have $\|u_g\|_\infty = \|u_a\|_\infty = 1$, where u_g and u_a attain their unique maximum on $[0,1]^d$ at $\mathbf{1} = (1, \ldots, 1) \in [0,1]^d$. ♢

Now we consider the Euclidean space $\mathbb{R}^d$, for $d \in \mathbb{N}$, as a linear subspace of the sequence space ℓ^p, $1 \leq p \leq \infty$, via the canonical embedding $i : \mathbb{R}^d \hookrightarrow \ell^p$,

$$x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d \longmapsto i(x) = (x_1, \ldots, x_d, 0, \ldots, 0, \ldots) \in \ell^p.$$

This allows us to formulate the following statements concerning the strict convexity of the ℓ^p-norms $\|\cdot\|_p$ on $\mathbb{R}^d$.

Corollary 3.35. *For the ℓ^p-norms $\|\cdot\|_p$ on $\mathbb{R}^d$, defined as*

$$\|x\|_p^p = \sum_{k=1}^{d} |x_k|^p \text{ for } 1 \leq p < \infty \quad \text{and} \quad \|x\|_\infty = \max_{1 \leq k \leq d} |x_k|$$

the following statements are true.

(a) For $1 < p < \infty$, the ℓ^p-norms $\|\cdot\|_p$ are strictly convex on $\mathbb{R}^d$.
(b) For $d > 1$, the ℓ^1-norm $\|\cdot\|_1$ is not strictly convex on $\mathbb{R}^d$.
(c) For $d > 1$, the ℓ^∞-norm $\|\cdot\|_\infty$ is not strictly convex on $\mathbb{R}^d$. ■

Remark 3.36. In statements (b), (c) of Corollary 3.35, we excluded the case $d = 1$, since in this univariate setting the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$ coincide with the strictly convex norm $|\cdot|$ on $\mathbb{R}$ (see Remark 3.27). □

Now we formulate the main result of this section.

Theorem 3.37. *Let $\mathcal{F}$ be a linear space, equipped with a strictly convex norm $\|\cdot\|$. Moreover, assume $\mathcal{S} \subset \mathcal{F}$ is convex and $f \in \mathcal{F}$. If there exists a best approximation $s^* \in \mathcal{S}$ to f, then s^* is unique.*

Proof. Suppose $s_1^*, s_2^* \in \mathcal{S}$ are two different best approximations to f from $\mathcal{S}$, i.e., $s_1^* \neq s_2^*$. Then we have

$$\|s_1^* - f\| = \|s_2^* - f\| = \inf_{s \in \mathcal{S}} \|s - f\|,$$

which, in combination with the strict convexity of the norm $\|\cdot\|$, implies

$$\left\| \frac{s_1^* + s_2^*}{2} - f \right\| = \left\| \frac{(s_1^* - f) + (s_2^* - f)}{2} \right\| < \|s_1^* - f\| = \|s_2^* - f\|. \tag{3.28}$$

Due to the assumed convexity for $\mathcal{S}$, the element $s^* = (s_1^* + s_2^*)/2$ lies in $\mathcal{S}$. Moreover, s^* is closer to f than s_1^* and s_2^*, by (3.28). But this is in contradiction to the optimality of s_1^* and s_2^*. ■

We remark that the strict convexity of the norm $\|\cdot\|$ gives, in combination with the convexity of $\mathcal{S} \subset \mathcal{F}$, only a *sufficient* condition for the uniqueness of the best approximation. Now we show that this condition is not necessary. To this end, we make a simple example.

Example 3.38. We regard the maximum norm $\|\cdot\|_\infty$ on $\mathcal{F} = \mathbb{R}^2$. Moreover, we let $f = (0,1) \in \mathbb{R}^2$ and $\mathcal{S} = \{(\alpha,\alpha) \,|\, \alpha \in \mathbb{R}\} \subset \mathbb{R}^2$. Then, $s^* = (\frac{1}{2}, \frac{1}{2}) \in \mathcal{S}$ is the *unique* best approximation to f from $\mathcal{S}$ with respect to $\|\cdot\|_\infty$, although $\|\cdot\|_\infty$ is *not* stricty convex (see Figure 3.5). ♢

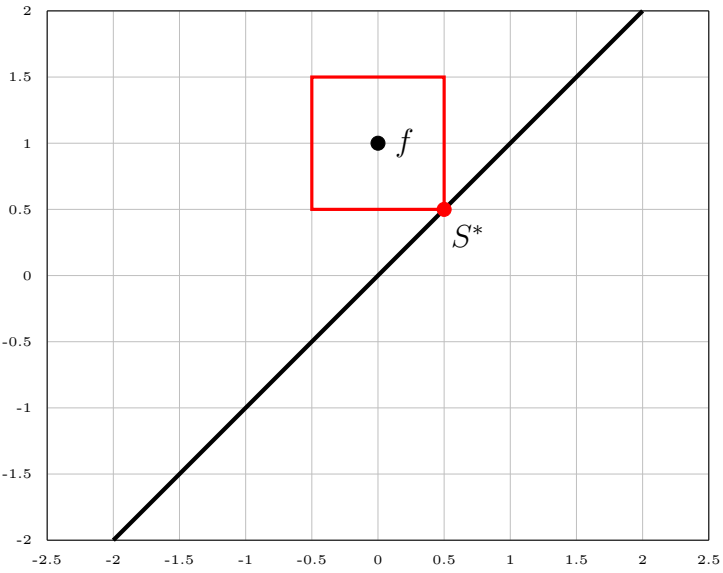


Fig. 3.5. $s^* = (\frac{1}{2}, \frac{1}{2}) \in \mathcal{S} = \{(\alpha,\alpha) \,|\, \alpha \in \mathbb{R}\}$ is the *unique* best approximation to $f = (0,1)$ w.r.t. $\|\cdot\|_\infty$, although $\|\cdot\|_\infty$ is *not* strictly convex (see Example 3.38).

We finally summarize our discussion concerning the uniqueness of best approximations, where we note three immediate conclusions from Theorem 3.37.

Corollary 3.39. *Let $\mathcal{F}$ be a Euclidean space and $\mathcal{S} \subset \mathcal{F}$ be convex. Then there is for any $f \in \mathcal{F}$ at most one best approximation $s^* \in \mathcal{S}$ to f.* ■

Corollary 3.40. *Let $\mathcal{S} \subset \mathrm{L}^p$ be convex for $1 < p < \infty$. Then there is for any $f \in \mathrm{L}^p$ at most one best approximation $s^* \in \mathcal{S}$ to f w.r.t. $\|\cdot\|_p$.* ■

Corollary 3.41. *Let $\mathcal{S} \subset \ell^p$ be convex for $1 < p < \infty$. Then there is for any $f \in \ell^p$ at most one best approximation $s^* \in \mathcal{S}$ to f w.r.t. $\|\cdot\|_p$.* ■

We finally formulate an important result concerning the approximation of continuous functions from $\mathscr{C}[-1,1]$ by approximation spaces $\mathcal{S} \subset \mathscr{C}[-1,1]$ that are *invariant under reflections of the argument* (in short: *reflection-invariant*), i.e., for any $s(x) \in \mathcal{S}$ the function $s(-x)$ lies also in $\mathcal{S}$. For example, the linear space $\mathcal{P}_n$ of all algebraic polynomials of degree at most $n \in \mathbb{N}_0$ is reflection-invariant. For the following observation, the uniqueness of the best approximation plays an important role.

Proposition 3.42. *Let $f \in \mathscr{C}[-1,1]$ be an even function and, moreover, let $\mathcal{S} \subset \mathscr{C}[-1,1]$ be a reflection-invariant subset of $\mathscr{C}[-1,1]$. If there exists a unique best approximation $s_p^* \in \mathcal{S}$ to f with respect to the L^p-norm $\|\cdot\|_p$, for $1 \le p \le \infty$, then s_p^* to f is an even function.*

Proof. Let $f \in \mathscr{C}[-1,1]$ be an even function, i.e., $f(x) = f(-x)$ for all $x \in [-1,1]$. Moreover, let $s_p^* \in \mathcal{S}$ be the unique best approximation to f with respect to $\|\cdot\|_p$, for $1 \le p \le \infty$. We regard the *reflected* function r_p^* for s_p^*, defined as

$$r_p^*(x) = s_p^*(-x) \qquad \text{for } x \in [-1,1].$$

By our assumption we have r_p^* in $\mathcal{S}$.

Case $p = \infty$: For the distance between r_∞^* and f with respect to $\|\cdot\|_\infty$,

$$\begin{aligned}\|r_\infty^* - f\|_\infty &= \max_{x\in[-1,1]} |r_\infty^*(x) - f(x)| = \max_{x\in[-1,1]} |r_\infty^*(-x) - f(-x)| \\ &= \max_{x\in[-1,1]} |s_\infty^*(x) - f(x)| = \|s_\infty^* - f\|_\infty = \eta_\infty(f,\mathcal{S}),\end{aligned}$$

we obtain the minimal distance between f and $\mathcal{S}$ with respect to $\|\cdot\|_\infty$, i.e., $r_\infty^* \in \mathcal{S}$ is the best approximation to f. Now the uniqueness of the best approximation implies our statement $s_\infty^*(x) = r_\infty^*(x)$, or, $s_\infty^*(x) = s_\infty^*(-x)$ for all $x \in [-1,1]$, i.e., s_∞^* is an even function on $[-1,1]$.

Case $1 \le p < \infty$: In this case, we regard the distance between r_p^* and f in the L^p-Norm,

$$\begin{aligned}\|r_p^* - f\|_p^p &= \int_{-1}^{1} |r_p^*(x) - f(x)|^p \mathrm{d}x = \int_{-1}^{1} |r_p^*(-x) - f(-x)|^p \mathrm{d}x \\ &= \int_{-1}^{1} |s_p^*(x) - f(x)|^p \mathrm{d}x = \|s_p^* - f\|_p^p = \eta_p^p(f,\mathcal{S}),\end{aligned}$$

whereby we get the minimal distance $\eta_p(f,\mathcal{S})$ between f and $\mathcal{S}$ with respect to $\|\cdot\|_p$. Again, by the uniqueness of the best approximation we obtain the stated result by

$$s_p^*(x) = r_p^*(x) = s_p^*(-x) \qquad \text{for all } x \in [-1,1].$$

■

For an alternative proof of Proposition 3.42, we refer to Exercise 3.74.

3.3 Dual Characterization

In this section and in the following section, we develop necessary and sufficient conditions to characterize best approximations. We begin with *dual* characterizations. To this end, let $\mathcal{F}$ be a normed linear space. We introduce the `topological dual`, or, the `dual space` of $\mathcal{F}$ as usual by

$$\mathcal{F}' = \left\{\varphi : \mathcal{F} \longrightarrow \mathbb{R} \,\middle|\, \varphi \text{ linear and continuous}\right\}.$$

The elements from the linear space $\mathcal{F}'$ are called `dual functionals`. On this occasion, we recall the notions of *linearity*, *continuity* and *boundedness* of functionals. We start with linearity.

Definition 3.43. *A functional $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ is called* `linear` *on $\mathcal{F}$, if*

$$\varphi(\alpha u + \beta v) = \alpha\varphi(u) + \beta\varphi(v) \qquad \text{for all } u, v \in \mathcal{F} \text{ and all } \alpha, \beta \in \mathbb{R}.$$

◯

We had introduced continuity already in Definition 3.3. Next, we turn to the boundedness of functionals. In the following discussion, $\mathcal{F}$ denotes a linear space with norm $\|\cdot\|$.

Definition 3.44. *A functional $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ is said to be* `bounded` *on $\mathcal{F}$, if there exists a constant $C \equiv C_\varphi > 0$ satisfying*

$$|\varphi(u)| \le C\|u\| \qquad \text{for all } u \in \mathcal{F}. \tag{3.29}$$

We call such a constant C `upper bound` *for φ.* ◯

Now we can introduce a norm for the dual space $\mathcal{F}'$, by using the norm $\|\cdot\|$ of $\mathcal{F}$. To this end, we take for any functional $\varphi \in \mathcal{F}'$ the *smallest* upper bound $C \equiv C_\varphi$ in (3.29). To be more precise on this, we define by

$$\|\varphi\| = \sup_{\substack{u \in \mathcal{F} \\ u \neq 0}} \frac{|\varphi(u)|}{\|u\|} = \sup_{\substack{u \in \mathcal{F} \\ \|u\| = 1}} |\varphi(u)|$$

a mapping $\|\cdot\| : \mathcal{F}' \longrightarrow \mathbb{R}$. As it can be verified by elementary calculations, $\|\cdot\|$ is a norm on $\mathcal{F}'$, according to Definition 1.1. In other words, the dual space $\mathcal{F}'$ is a linear space with norm $\|\cdot\|$.

The following result for linear functionals is quite important.

Theorem 3.45. *For a linear functional $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$, the following statements are equivalent.*

(a) φ is continuous at one $u_0 \in \mathcal{F}$.
(b) φ is continuous on $\mathcal{F}$.
(c) φ is bounded on $\mathcal{F}$.

Proof. $(a) \Rightarrow (b)$: Let φ be continuous at $u_0 \in \mathcal{F}$, and, moreover, let $(u_n)_{n\in\mathbb{N}}$ be a convergent sequence in $\mathcal{F}$ with limit $u \in \mathcal{F}$. Then we have

$$\varphi(u_n) = \varphi(u_n - u + u_0) + \varphi(u - u_0) \longrightarrow \varphi(u_0) + \varphi(u - u_0) = \varphi(u) \quad \text{for } n \to \infty.$$

Therefore, φ is continuous at every $u \in \mathcal{F}$, i.e., φ is continuous on $\mathcal{F}$.

The implication $(b) \Rightarrow (a)$ is trivial, so the equivalence $(a) \Leftrightarrow (b)$ is shown.

$(c) \Rightarrow (b)$: Let φ be bounded on $\mathcal{F}$, i.e., we have (3.29) for some $C > 0$. This implies $\varphi(u_n) \longrightarrow 0$, $n \to \infty$, for every zero sequence $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$, and so φ is continuous at zero. By the equivalence $(a) \Leftrightarrow (b)$ is φ continuous on $\mathcal{F}$.

$(b) \Rightarrow (c)$: Let φ be continuous on $\mathcal{F}$. Suppose φ is not bounded on $\mathcal{F}$. Then, there is a sequence $(u_n)_{n\in\mathbb{N}}$ in $\mathcal{F}$ satisfying

$$\|u_n\| = 1 \quad \text{and} \quad |\varphi(u_n)| > n \qquad \text{for all } n \in \mathbb{N},$$

since otherwise there would exist an upper bound $N \in \mathbb{N}$ for φ (i.e., φ would be bounded). In this case, the sequence $(v_n)_{n\in\mathbb{N}}$, defined as

$$v_n = \frac{u_n}{|\varphi(u_n)|} \qquad \text{for } n \in \mathbb{N},$$

is a zero sequence in $\mathcal{F}$ by

$$\|v_n\| = \frac{1}{|\varphi(u_n)|} \longrightarrow 0 \quad \text{for } n \to \infty,$$

and so, by continuity of φ, we have

$$\varphi(v_n) \longrightarrow \varphi(0) = 0 \quad \text{for } n \to \infty.$$

But this is in contradiction to $|\varphi(v_n)| = 1$ for all $n \in \mathbb{N}$. ■

Now we are in a position where we can formulate a sufficient condition for the dual characterization of best approximations.

Theorem 3.46. *Let $\mathcal{S} \subset \mathcal{F}$ be a non-empty subset of $\mathcal{F}$. Moreover, let $f \in \mathcal{F}$ and $s^* \in \mathcal{S}$. Suppose that $\varphi \in \mathcal{F}'$ is a dual functional satisfying the following properties.*

(a) $\|\varphi\| = 1$.
(b) $\varphi(s^* - f) = \|s^* - f\|$.
(c) $\varphi(s - s^*) \geq 0$ *for all* $s \in \mathcal{S}$.

Then s^* is a best approximation to f.

Proof. For $s \in \mathcal{S}$, we have $\varphi(s - f) \leq \|s - f\|$, due to (a). Moreover, we have

$$\|s - f\| \geq \varphi(s - f) = \varphi(s - s^*) + \varphi(s^* - f) \geq \|s^* - f\|$$

by (b) and (c). Therefore, s^* is a best approximation to f. ■

Note that the above characterization in Theorem 3.46 only requires $\mathcal{S}$ to be non-empty. However, if we assume $\mathcal{S} \subset \mathcal{F}$ to be convex, then we can show that the sufficient condition in Theorem 3.46 is also necessary. To this end, we need the following *separation theorem for convex sets*, which can be viewed as a geometric implication from the well-known *Hahn*[10]*-Banach*[11] *theorem* (see [33, Section 16.1]) which was proven by Mazur[12] in [3].

Theorem 3.47. (Banach-Mazur separation theorem, 1933). *Let $\mathcal{K}_1, \mathcal{K}_2 \subset \mathcal{F}$ be two non-empty, disjoint and convex subsets in a normed linear space $\mathcal{F}$. Moreover, suppose $\mathcal{K}_1$ is an open set. Then there exists a separating functional $\varphi \in \mathcal{F}'$ for $\mathcal{K}_1$ and $\mathcal{K}_2$, i.e., we have*

$$\varphi(u_1) < \varphi(u_2) \qquad \text{for all } u_1 \in \mathcal{K}_1, u_2 \in \mathcal{K}_2.$$

□

On the Banach-Mazur separation theorem, we can formulate a sufficient and necessary condition for the dual characterization of best approximations.

Theorem 3.48. *Let $\mathcal{S} \subset \mathcal{F}$ be a convex set in $\mathcal{F}$. Moreover, suppose $f \in \mathcal{F} \setminus \mathcal{S}$. Then, $s^* \in \mathcal{S}$ is a best approximation to f, if and only if there exists a dual functional $\varphi \in \mathcal{F}'$ satisfying the following properties.*

(a) $\|\varphi\| = 1$.
(b) $\varphi(s^* - f) = \|s^* - f\|$.
(c) $\varphi(s - s^*) \geq 0$ *for all* $s \in \mathcal{S}$.

Proof. Note that the sufficiency of the statement is covered by Theorem 3.46.

To prove the necessity, suppose that $s^* \in \mathcal{S}$ is a best approximation to f. Regard the open ball

$$\mathcal{B}_\eta(f) = \{u \in \mathcal{F} \,|\, \|u - f\| < \|s^* - f\|\} \subset \mathcal{F}$$

around f with radius $\eta = \|s^* - f\|$. Note that for $\mathcal{K}_1 = \mathcal{B}_\eta(f)$ and $\mathcal{K}_2 = \mathcal{S}$ the assumptions of the Banach-Mazur separation theorem, Theorem 3.47, are satisfied. Therefore, there is a separating functional $\varphi \in \mathcal{F}'$ with

$$\varphi(u) < \varphi(s) \qquad \text{for all } u \in \mathcal{B}_\eta(f) \text{ and } s \in \mathcal{S}. \tag{3.30}$$

Now let $(u_n)_{n\in\mathbb{N}} \subset \mathcal{B}_\eta(f)$ be a convergent sequence with limit element s^*, i.e., $u_n \longrightarrow s^*$ for $n \to \infty$. By the continuity of φ, this implies

$$\varphi(u_n) \longrightarrow \varphi(s^*) = \inf_{s\in\mathcal{S}} \varphi(s),$$

[10] Hans Hahn (1879-1934), Austrian mathematician and philosopher
[11] Stefan Banach (1892-1945), Polish mathematician
[12] Stanisław Mazur (1905-1981), Polish mathematician

i.e., $\varphi(s^*) \leq \varphi(s)$ for all $s \in \mathcal{S}$, and so φ has property (c).

To show properties (a) and (b), let $v \in \mathcal{F}$ with $\|v\| < 1$. Then, $u = \eta v + f$ lies in $\mathcal{B}_\eta(f)$. With (3.30) and by the linearity of φ, we have

$$\varphi(v) = \varphi\left(\frac{u-f}{\|s^*-f\|}\right) < \varphi\left(\frac{s^*-f}{\|s^*-f\|}\right).$$

This implies

$$\|\varphi\| = \sup_{\|v\|\leq 1} |\varphi(v)| \leq \varphi\left(\frac{s^*-f}{\|s^*-f\|}\right)$$

and, moreover, by using the continuity of φ once more, we have

$$\|\varphi\| = \frac{\varphi(s^*-f)}{\|s^*-f\|} \qquad \Longleftrightarrow \qquad \varphi(s^*-f) = \|\varphi\| \cdot \|s^*-f\|.$$

If we finally normalize the length of $\varphi \neq 0$, by scaling φ to unit norm, i.e., $\|\varphi\| = 1$, then φ satisfies properties (a) and (b). ■

3.4 Direct Characterization

In this section, we develop necessary and sufficient conditions for the minimization of convex functionals. We then apply these conditions to norms to obtain useful characterizations for best approximations. To this end, we work with *Gâteaux*[13] *derivatives* to compute directional derivatives for relevant norms. In the following discussion, $\mathcal{F}$ denotes a linear space.

Definition 3.49. *For a functional* $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$,

$$\varphi'_+(u,v) := \lim_{h \searrow 0} \frac{1}{h}\left(\varphi(u+hv) - \varphi(u)\right) \qquad \text{for } u, v \in \mathcal{F} \tag{3.31}$$

is said to be the `Gâteaux derivative` *of* φ *at* u *in direction* v*, provided that the limit on the right hand side in (3.31) exists.* ◯

For *convex* functionals $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ we can show that the limit on the right hand side in (3.31) exists.

Theorem 3.50. *Let* $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ *be a convex functional. Then, the Gâteaux derivative* $\varphi'_+(u,v)$ *exists for all* $u, v \in \mathcal{F}$*. Moreover, the inequality*

$$-\varphi'_+(u,-v) \leq \varphi'_+(u,v) \qquad \text{for all } u, v \in \mathcal{F}$$

holds.

[13] René Gâteaux (1889-1914), French mathematician

Proof. Let $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ be a convex functional. We show that for any $u, v \in \mathcal{F}$ the difference quotient $D_{u,v} : (0, \infty) \longrightarrow \mathbb{R}$, defined as

$$D_{u,v}(h) = \frac{1}{h}\left(\varphi(u + hv) - \varphi(u)\right) \qquad \text{for } h > 0, \tag{3.32}$$

is a monotonically increasing function in $h > 0$, which, moreover, is bounded below. To verify the monotonicity, we regard the convex combination

$$u + h_1 v = \frac{h_2 - h_1}{h_2} u + \frac{h_1}{h_2}(u + h_2 v) \qquad \text{for } h_2 > h_1 > 0.$$

The convexity of φ then implies the inequality

$$\varphi(u + h_1 v) \le \frac{h_2 - h_1}{h_2}\varphi(u) + \frac{h_1}{h_2}\varphi(u + h_2 v)$$

and, after elementary calculations, the monotonicity

$$D_{u,v}(h_1) = \frac{1}{h_1}(\varphi(u + h_1 v) - \varphi(u)) \le \frac{1}{h_2}(\varphi(u + h_2 v) - \varphi(u)) = D_{u,v}(h_2).$$

If we now form the convex combination

$$u = \frac{h_2}{h_1 + h_2}(u - h_1 v) + \frac{h_1}{h_1 + h_2}(u + h_2 v) \qquad \text{for } h_1, h_2 > 0,$$

we obtain, by using the convexity of φ, the inequality

$$\varphi(u) \le \frac{h_2}{h_1 + h_2}\varphi(u - h_1 v) + \frac{h_1}{h_1 + h_2}\varphi(u + h_2 v)$$

and, after elementary calculations, we obtain the estimate

$$\begin{aligned} -D_{u,-v}(h_1) &= -\frac{1}{h_1}(\varphi(u - h_1 v) - \varphi(u)) \\ &\le \frac{1}{h_2}(\varphi(u + h_2 v) - \varphi(u)) = D_{u,v}(h_2). \end{aligned} \tag{3.33}$$

This implies that the monotonically increasing difference quotient $D_{u,v}$ is bounded from below for all $u, v \in \mathcal{F}$. In particular, $D_{u,-v}$ is a monotonically increasing function that is bounded from below. Therefore, the Gâteaux derivatives $\varphi'_+(u, v)$ and $\varphi'_+(u, -v)$ exist. By (3.33), we finally have

$$-\frac{1}{h}(\varphi(u - hv) - \varphi(u)) \le -\varphi'_+(u, -v) \le \varphi'_+(u, v) \le \frac{1}{h}(\varphi(u + hv) - \varphi(u))$$

for all $h > 0$, as stated. ∎

Now we note a few elementary properties of the Gâteaux derivative.

Theorem 3.51. *Let $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ be a convex functional. Then the Gâteaux derivative φ'_+ of φ has for all $u, v, w \in \mathcal{F}$ the following properties.*

(a) $\varphi'_+(u, \alpha v) = \alpha \varphi'_+(u, v)$ *for all* $\alpha \geq 0$.
(b) $\varphi'_+(u, v + w) \leq \varphi'_+(u, v) + \varphi'_+(u, w)$.
(c) $\varphi'_+(u, \cdot) : \mathcal{F} \longrightarrow \mathbb{R}$ *is a convex functional.*

Proof. (a): The case $\alpha = 0$ is trivial. For $\alpha > 0$ we have

$$\begin{aligned}\varphi'_+(u, \alpha v) &= \lim_{h \searrow 0} \frac{1}{h}(\varphi(u + h\alpha v) - \varphi(u)) \\ &= \alpha \lim_{h \searrow 0} \frac{1}{h\alpha}(\varphi(u + h\alpha v) - \varphi(u)) = \alpha \varphi'_+(u, v).\end{aligned}$$

(b): The representation

$$u + h(v + w) = \frac{1}{2}(u + 2hv) + \frac{1}{2}(u + 2hw),$$

in combination with the convexity of φ, implies

$$\begin{aligned}\varphi'_+(u, v + w) &= \lim_{h \searrow 0} \frac{1}{h}(\varphi(u + h(v + w)) - \varphi(u)) \\ &\leq \lim_{h \searrow 0} \frac{1}{h}\left(\frac{1}{2}\varphi(u + 2hv) + \frac{1}{2}\varphi(u + 2hw) - \varphi(u)\right) \\ &= \lim_{h \searrow 0} \frac{1}{2h}(\varphi(u + 2hv) - \varphi(u)) + \lim_{h \searrow 0} \frac{1}{2h}(\varphi(u + 2hw) - \varphi(u)) \\ &= \varphi'_+(u, v) + \varphi'_+(u, w).\end{aligned}$$

(c): For $u \in \mathcal{F}$, the Gâteaux derivative $\varphi'_+(u, \cdot) : \mathcal{F} \longrightarrow \mathbb{R}$ is convex, since

$$\begin{aligned}\varphi'_+(u, \lambda v + (1 - \lambda)w) &\leq \varphi'_+(u, \lambda v) + \varphi'_+(u, (1 - \lambda)w) \\ &= \lambda \varphi'_+(u, v) + (1 - \lambda)\varphi'_+(u, w)\end{aligned}$$

holds for all $\lambda \in [0, 1]$, by using properties (a) and (b). ■

Remark 3.52. By the properties (a) and (b) in Theorem 3.51, we call the functional $\varphi'_+(u, \cdot) : \mathcal{F} \longrightarrow \mathbb{R}$ *sublinear*. We can show that the sublinearity of $\varphi'_+(u, \cdot)$, for all $u \in \mathcal{F}$, in combination with the inequality

$$\varphi'_+(u, v - u) \leq \varphi(v) - \varphi(u) \qquad \text{for all } u, v \in \mathcal{F},$$

implies the convexity of φ. To see this, we refer to Exercise 3.80. □

Now we show further elementary properties of the Gâteaux derivative.

Theorem 3.53. *Let $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ be a continuous functional. Suppose that for $u, v \in \mathcal{F}$ the Gâteaux derivative $\varphi'_+(u, v)$ exists. Moreover, suppose that $F : \mathbb{R} \longrightarrow \mathbb{R}$ has a continuous derivative, i.e., $F \in \mathscr{C}^1(\mathbb{R})$. Then the Gâteaux derivative $(F \circ \varphi)'_+(u, v)$ of the composition $F \circ \varphi : \mathcal{F} \longrightarrow \mathbb{R}$ exists at u in direction v, and, moreover, the* `chain rule`

$$(F \circ \varphi)'_+(u, v) = F'(\varphi(u)) \cdot \varphi'_+(u, v) \tag{3.34}$$

holds.

Proof. For $x := \varphi(u)$ and $x_h := \varphi(u + hv)$, for $h > 0$, we let

$$G(x_h) := \begin{cases} \dfrac{F(x_h) - F(x)}{x_h - x} & \text{for } x_h \neq x, \\ F'(x) & \text{for } x_h = x. \end{cases}$$

By the continuity of φ we have $x_h \longrightarrow x$ for $h \searrow 0$. Since $F \in \mathscr{C}^1(\mathbb{R})$ this implies

$$F'(x) = \lim_{x_h \to x} G(x_h) = \lim_{h \searrow 0} G(\varphi(u + hv)) = F'(\varphi(u)).$$

Moreover, we have

$$F(x_h) - F(x) = G(x_h)(x_h - x) \qquad \text{for all } h > 0.$$

This finally implies

$$\begin{aligned} (F \circ \varphi)'_+(u, v) &= \lim_{h \searrow 0} \frac{1}{h} \left(F(\varphi(u + hv)) - F(\varphi(u))\right) \\ &= \lim_{h \searrow 0} \frac{1}{h} \left(F(x_h) - F(x)\right) \\ &= \lim_{h \searrow 0} G(x_h) \cdot \lim_{h \searrow 0} \frac{1}{h} \left(x_h - x\right) \\ &= \lim_{h \searrow 0} G(\varphi(u + hv)) \cdot \lim_{h \searrow 0} \frac{1}{h} \left(\varphi(u + hv) - \varphi(u)\right) \\ &= F'(\varphi(u)) \cdot \varphi'_+(u, v), \end{aligned}$$

proving both the existence of $(F \circ \varphi)'_+(u, v)$ and the chain rule in (3.34). ■

We now formulate a fundamental sufficient and necessary condition for the characterization of minima for convex functionals.

Theorem 3.54. *Let $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ be a convex functional. Moreover, let $\mathcal{K} \subset \mathcal{F}$ be convex and $u_0 \in \mathcal{K}$. Then the following statements are equivalent.*

(a) $\varphi(u_0) = \inf_{u \in \mathcal{K}} \varphi(u)$.
(b) $\varphi'_+(u_0, u - u_0) \geq 0$ *for all* $u \in \mathcal{K}$.

Proof. $(b) \Rightarrow (a)$: Suppose $\varphi'_+(u_0, u-u_0) \geq 0$ for $u \in \mathcal{K}$. Then we have, due to the monotonicity of the difference quotient $D_{u_0,u-u_0}$ in (3.32), in particular for $h = 1$,

$$0 \leq \varphi'_+(u_0, u - u_0) \leq \varphi(u_0 + (u - u_0)) - \varphi(u_0) = \varphi(u) - \varphi(u_0)$$

and thus $\varphi(u) \geq \varphi(u_0)$.

$(a) \Rightarrow (b)$: Suppose $\varphi(u_0) = \inf_{u\in\mathcal{K}} \varphi(u)$.

Then for $u \in \mathcal{K}$ and small enough $h > 0$ the inequality

$$\frac{1}{h}(\varphi(u_0 + h(u - u_0)) - \varphi(u_0)) \geq 0$$

holds, since, due to convexity of $\mathcal{K}$, we have

$$u_0 + h(u - u_0) = hu + (1 - h)u_0 \in \mathcal{K} \qquad \text{for all } h \in (0, 1).$$

This finally implies

$$\varphi'_+(u_0, u - u_0) = \lim_{h\searrow 0} \frac{1}{h}(\varphi(u_0 + h(u - u_0)) - \varphi(u_0)) \geq 0.$$

■

Now we wish to use the conditions from Theorem 3.54 to derive direct characterizations of best approximations to $f \in \mathcal{F}$. To this end, we regard the distance functional $\varphi_f : \mathcal{F} \longrightarrow \mathbb{R}$, defined as

$$\varphi_f(v) = \|v - f\| \qquad \text{for } v \in \mathcal{F}.$$

Note that the distance functional φ_f is, as a composition between two continuous functionals, namely the translation about f and the norm $\|\cdot\|$, a continuous functional. Moreover, $\varphi_f : \mathcal{F} \longrightarrow \mathbb{R}$ is convex. Indeed, for $\lambda \in [0, 1]$ and $v_1, v_2 \in \mathcal{F}$ we have

$$\begin{aligned}
&\varphi_f(\lambda v_1 + (1 - \lambda)v_2) \\
&= \|\lambda v_1 + (1 - \lambda)v_2 - f\| = \|\lambda(v_1 - f) + (1 - \lambda)(v_2 - f)\| \\
&\leq \lambda\|v_1 - f\| + (1 - \lambda)\|v_2 - f\| = \lambda\varphi_f(v_1) + (1 - \lambda)\varphi_f(v_2).
\end{aligned}$$

Therefore, φ_f has a Gâteaux derivative, for which the chain rule (3.34) holds.

Now the direct characterization from Theorem 3.54 can be applied to the distance functional φ_f. This leads us to a corresponding equivalence, which is referred to as *Kolmogorov*[14] *criterion*.

For the Gâteaux derivative of the norm $\varphi = \|\cdot\| : \mathcal{F} \longrightarrow \mathbb{R}$ we will henceforth use the notation

$$\|'_+(u, v) := \varphi'_+(u, v) \qquad \text{for } u, v \in \mathcal{F}.$$

[14] Andrey Nikolaevich Kolmogorov (1903-1987), Russian mathematician

Corollary 3.55. (Kolmogorov criterion).
For $f \in \mathcal{F}$, $\mathcal{S} \subset \mathcal{F}$ convex, and $s^* \in \mathcal{S}$ the following statements are equivalent.

(a) s^* is a best approximation to f.
(b) $\|'_+(s^* - f, s - s^*) \geq 0$ for all $s \in \mathcal{S}$.

Proof. Using $\varphi(u) = \|u - f\|$ in Theorem 3.54, we see that $s^* \in \mathcal{S}$ is a best approximation to f, if and only if

$$\begin{aligned}
\varphi'_+(s^*, s - s^*) &= \lim_{h \searrow 0} \frac{1}{h}(\varphi(s^* + h(s - s^*)) - \varphi(s^*)) \\
&= \lim_{h \searrow 0} \frac{1}{h}(\|s^* + h(s - s^*) - f\| - \|s^* - f\|) \\
&= \lim_{h \searrow 0} \frac{1}{h}(\|s^* - f + h(s - s^*)\| - \|s^* - f\|) \\
&= \|'_+(s^* - f, s - s^*) \geq 0 \qquad \text{for all } s \in \mathcal{S}.
\end{aligned}$$

■

Remark 3.56. For proving the implication $(b) \Rightarrow (a)$ in Theorem 3.54 we did not use the convexity of $\mathcal{K}$. Therefore, we can specialize the equivalence in Corollary 3.55 to establish the implication

$$\|'_+(s^* - f, s - s^*) \geq 0 \text{ for all } s \in \mathcal{S} \quad \Longrightarrow \quad s^* \text{ is best approximation to } f$$

for subsets $\mathcal{S} \subset \mathcal{F}$ that are not necessarily convex. □

Now we use the Gâteaux derivative to prove a characterization concerning the uniqueness of best approximations. To this end, we introduce a sufficient condition, which will be more restrictive than that of (plain) uniqueness.

Definition 3.57. *Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$, $\mathcal{S} \subset \mathcal{F}$ be a subset of $\mathcal{F}$, and $f \in \mathcal{F}$. Then $s^* \in \mathcal{S}$ is said to be the* `strongly unique best approximation` *to f, if there is a constant $\alpha > 0$ satisfying*

$$\|s - f\| - \|s^* - f\| \geq \alpha \|s - s^*\| \qquad \textit{for all } s \in \mathcal{S}.$$

○

Relying on our previous investigations (especially on Theorem 3.54 and Corollary 3.55), we can characterize the strong uniqueness of best approximations for convex subsets $\mathcal{S} \subset \mathcal{F}$ directly as follows.

Theorem 3.58. *Let $\mathcal{F}$ be a linear space with norm $\|\cdot\|$ and $\mathcal{S} \subset \mathcal{F}$ be convex. Moreover, suppose $f \in \mathcal{F}$. Then the following statements are equivalent.*

(a) $s^* \in \mathcal{S}$ is the strongly unique best approximation to f.
(b) There is one $\alpha > 0$ satisfying $\|'_+(s^* - f, s - s^*) \geq \alpha \|s - s^*\|$ for all $s \in \mathcal{S}$.

Proof. Suppose $f \in \mathcal{F}$ and $s^* \in \mathcal{S}$. For $f \in \mathcal{F}$ we regard the convex distance functional $\varphi : \mathcal{S} \longrightarrow [0, \infty)$, defined as $\varphi(s) = \|s - f\|$ for $s \in \mathcal{S}$. Now any element in $\mathcal{S} \setminus \{s^*\}$ can be written as a convex combination of the form

$$s^* + h(s - s^*) = hs + (1 - h)s^* \in \mathcal{S} \qquad \text{with } s \in \mathcal{S} \setminus \{s^*\} \text{ and } 1 \geq h > 0.$$

Thereby, we can formulate the strong uniqueness of s^*, for some $\alpha > 0$, as

$$\frac{1}{\|s - s^*\|} \frac{1}{h} (\varphi(s^* + h(s - s^*)) - \varphi(s^*)) \geq \alpha \quad \text{for all } s \in \mathcal{S} \setminus \{s^*\} \text{ and } h > 0.$$

By the monotonicity of the Gâteaux derivative, this is equivalent to

$$\varphi'_+(s^*, s - s^*) = \lim_{h \searrow 0} \frac{1}{h} (\varphi(s^* + h(s - s^*)) - \varphi(s^*)) \geq \alpha \|s - s^*\|$$

for all perturbation directions $s - s^* \in \mathcal{S}$, or, in other words, we have

$$\|'_+(s^* - f, s - s^*) \geq \alpha \|s - s^*\| \qquad \text{for all } s \in \mathcal{S}.$$

■

Next, we add an important stability estimate to our discussion on strongly unique best approximations. The following result is due to Freud[15].

Theorem 3.59. (Freud).
For a linear space $\mathcal{F}$ with norm $\|\cdot\|$ and a subset $\mathcal{S} \subset \mathcal{F}$, let $s_f^* \in \mathcal{S}$ be the strongly unique best approximation to $f \in \mathcal{F}$ with constant $\alpha > 0$. Moreover, let $s_g^* \in \mathcal{S}$ be a best approximation to $g \in \mathcal{F}$. Then the stability estimate

$$\|s_g^* - s_f^*\| \leq \frac{2}{\alpha} \|g - f\|$$

holds.

Proof. By the strong uniqueness of s_f^*, we have the estimate

$$\|s_g^* - f\| - \|s_f^* - f\| \geq \alpha \|s_g^* - s_f^*\|,$$

which further implies the inequalities

$$\begin{aligned}
\|s_g^* - s_f^*\| &\leq \frac{1}{\alpha} (\|s_g^* - f\| - \|s_f^* - f\|) \\
&\leq \frac{1}{\alpha} (\|s_g^* - g\| + \|g - f\| - \|s_f^* - f\|) \\
&\leq \frac{1}{\alpha} (\|s_f^* - g\| + \|g - f\| - \|s_f^* - f\|) \\
&\leq \frac{1}{\alpha} (\|s_f^* - f\| + \|f - g\| + \|g - f\| - \|s_f^* - f\|) \\
&= \frac{2}{\alpha} \|g - f\|,
\end{aligned}$$

which already proves the stated stability estimate. ■

[15] Géza Freud (1922-1979), Hungarian mathematician

Remark 3.60. If the best approximation $s_g^* \in \mathcal{S}$ is for any $g \in \mathcal{F}$ unique, then the mapping $g \longmapsto s_g^*$ is well-defined. Further, due to the Freud theorem, Theorem 3.59, this mapping is continuous at all elements $f \in \mathcal{F}$ which have a strongly unique best approximation $s_f^* \in \mathcal{S}$.

If every $f \in \mathcal{F}$ has a strongly unique best approximation $s_f^* \in \mathcal{S}$, such that their corresponding constants $\alpha_f > 0$ are on $\mathcal{F}$ uniformly bounded away from zero, i.e., if there is some $\alpha_0 > 0$ satisfying $\alpha_f \geq \alpha_0 > 0$ for all $f \in \mathcal{F}$, then the mapping $f \longmapsto s_f^*$ is, due to the Freud theorem, Theorem 3.59, by

$$\|s_g^* - s_f^*\| \leq \frac{2}{\alpha_0}\|g - f\| \qquad \text{for all } f, g \in \mathcal{F},$$

Lipschitz continuous on $\mathcal{F}$ with *Lipschitz constant* $2/\alpha_0$ (see Definition 6.64). □

Next, we will explicitly compute Gâteaux derivatives for relevant norms $\|\cdot\|$. But first we make the following simple observation.

Remark 3.61. For the Gâteaux derivative of $\varphi(u) = \|u\|$ at $u = 0$, we have

$$\|'_+(0, v) = \lim_{h \searrow 0} \frac{1}{h} \left(\|0 + hv\| - \|0\|\right) = \|v\|$$

for any direction $v \in \mathcal{F}$. □

Now we compute Gâteaux derivatives for Euclidean norms.

Theorem 3.62. *The Gâteaux derivative of any Euclidean norm $\|\cdot\|$ is given as*

$$\|'_+(u, v) = \left(\frac{u}{\|u\|}, v\right) \qquad \textit{for all } u \in \mathcal{F} \setminus \{0\} \textit{ and } v \in \mathcal{F}.$$

Proof. Let $\mathcal{F}$ be a Euclidean space with norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$.

For $\varphi(u) = \|u\|$ the chain rule in (3.34) holds in particular for $F(x) = x^2$, i.e.,

$$\left(\|^2\right)'_+ (u, v) = 2\|u\| \cdot \|'_+(u, v) \qquad \text{for all } u, v \in \mathcal{F}. \tag{3.35}$$

Moreover, we have

$$\begin{aligned}
\left(\|^2\right)'_+ (u, v) &= \lim_{h \searrow 0} \frac{1}{h} \left(\|u + hv\|^2 - \|u\|^2\right) \\
&= \lim_{h \searrow 0} \frac{1}{h} \left(\|u\|^2 + 2h(u, v) + h^2\|v\|^2 - \|u\|^2\right) \\
&= \lim_{h \searrow 0} \frac{1}{h} \left(2h(u, v) + h^2\|v\|^2\right) \\
&= 2(u, v).
\end{aligned}$$

This implies, for $u \neq 0$ with (3.35),

$$\|'_+(u,v) = \left(\frac{u}{\|u\|}, v\right) \qquad \text{for all } v \in \mathcal{F}.$$

■

The Gâteaux derivative of the absolute value $|\cdot|$ is rather elementary.

Lemma 3.63. *For the absolute-value function* $|\cdot| : \mathbb{R} \longrightarrow [0,\infty)$*, we have*

$$|'_+(x,y) = y\,\mathrm{sgn}(x) \qquad \textit{for all } x \neq 0 \textit{ and } y \in \mathbb{R}.$$

Proof. First, note that for $x \neq 0$ we have

$$|x + hy| = |x| + hy\,\mathrm{sgn}(x) \qquad \text{for } h|y| < |x|. \tag{3.36}$$

This implies

$$|'_+(x,y) = \lim_{h \searrow 0} \frac{1}{h}(|x+hy| - |x|) = \lim_{h \searrow 0} \frac{1}{h}(|x| + hy\,\mathrm{sgn}(x) - |x|) = y\,\mathrm{sgn}(x).$$

■

By using the observation (3.36), we can compute the Gâteaux derivatives for all L^p-norms $\|\cdot\|_p$, for $1 \leq p \leq \infty$, on the linear space

$$\mathscr{C}(\Omega) = \{u : \Omega \longrightarrow \mathbb{R} \,|\, u \text{ continuous on } \Omega\}$$

of all continuous functions on a compact domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$. We begin with the maximum norm $\|\cdot\|_\infty$ on $\mathscr{C}(\Omega)$, defined as

$$\|u\|_\infty = \max_{x \in \Omega} |u(x)| \qquad \text{for } u \in \mathscr{C}(\Omega).$$

Theorem 3.64. *Let* $\Omega \subset \mathbb{R}^d$ *be compact. Then, for the Gâteaux derivative of the maximum norm* $\|\cdot\| = \|\cdot\|_\infty$ *on* $\mathscr{C}(\Omega)$*, we have*

$$\|'_+(u,v) = \max_{\substack{x \in \Omega \\ |u(x)| = \|u\|_\infty}} v(x)\,\mathrm{sgn}(u(x))$$

for any $u, v \in \mathscr{C}(\Omega), u \not\equiv 0$.

Proof. Suppose $u \in \mathscr{C}(\Omega)$, $u \not\equiv 0$, and $v \in \mathscr{C}(\Omega)$.
"$\geq$": We first show the inequality

$$\|'_+(u,v) \geq \max_{\substack{x \in \Omega \\ |u(x)| = \|u\|_\infty}} v(x)\,\mathrm{sgn}(u(x)).$$

Suppose $x \in \Omega$ with $|u(x)| = \|u\|_\infty$. Then, by (3.36) we have the inequality

$$\begin{aligned}\frac{1}{h}(\|u+hv\|_\infty - \|u\|_\infty) &\geq \frac{1}{h}(|u(x)+hv(x)| - |u(x)|)\\ &= \frac{1}{h}(|u(x)| + hv(x)\operatorname{sgn}(u(x)) - |u(x)|)\\ &= v(x)\operatorname{sgn}(u(x))\end{aligned}$$

for $h|v(x)| < |u(x)|$, which by $h \searrow 0$ already implies the stated inequality.

"$\leq$": To verify the inequality

$$\|'_+(u,v) \leq \max_{\substack{x\in\Omega\\ |u(x)|=\|u\|_\infty}} v(x)\operatorname{sgn}(u(x))$$

we regard a strictly monotonically decreasing zero sequence $(h_k)_{k\in\mathbb{N}}$ of positive real numbers, so that $\lim_{k\to\infty} h_k = 0$. Note that for any element $h_k > 0$ there exists some $x_{h_k} \in \Omega$ satisfying

$$\|u + h_k v\|_\infty = |u(x_{h_k}) + h_k v(x_{h_k})| \qquad \text{for } k \in \mathbb{N}.$$

Since Ω is compact, the sequence $(x_{h_k})_{k\in\mathbb{N}}$ has a convergent subsequence $(x_{h_{k_\ell}})_{\ell\in\mathbb{N}} \subset \Omega$ with limit element $\lim_{\ell\to\infty} x_{h_{k_\ell}} = x \in \Omega$. For $\ell \to \infty$ we get

$$\|u + h_{k_\ell} v\|_\infty = |u(x_{h_{k_\ell}}) + h_{k_\ell} v(x_{h_{k_\ell}})| \longrightarrow \|u\|_\infty = |u(x)| \quad \text{with } h_{k_\ell} \searrow 0,$$

i.e., *every* accumulation point of $(x_{h_k})_{k\in\mathbb{N}}$ is an extremum of u in Ω.

Moreover, by (3.36) we obtain the inequality

$$\begin{aligned}&\frac{1}{h_k}(\|u+h_k v\|_\infty - \|u\|_\infty)\\ &\leq \frac{1}{h_k}(|u(x_{h_k}) + h_k v(x_{h_k})| - |u(x_{h_k})|)\\ &= \frac{1}{h_k}(|u(x_{h_k})| + h_k v(x_{h_k})\operatorname{sgn}(u(x_{h_k})) - |u(x_{h_k})|)\\ &= v(x_{h_k})\operatorname{sgn}(u(x_{h_k}))\end{aligned}$$

for $h_k|v(x_{h_k})| < |u(x_{h_k})|$, whereby

$$\lim_{\ell\to\infty} \frac{1}{h_{k_\ell}}(\|u+h_{k_\ell}v\|_\infty - \|u\|_\infty) = \|'_+(u,v) \leq v(x)\operatorname{sgn}(u(x)),$$

and where $x \in \Omega$ is an extremum of u in Ω, i.e., $|u(x)| = \|u\|_\infty$. ■

Theorem 3.65. *Let $\Omega \subset \mathbb{R}^d$ be compact and suppose $u, v \in \mathscr{C}(\Omega)$. Then, we have*

$$\|'_+(u,v) = \int_{\Omega_+} v(x)\operatorname{sgn}(u(x))\,\mathrm{d}x + \int_{\Omega_0} |v(x)|\,\mathrm{d}x \tag{3.37}$$

for the Gâteaux derivative of the L^1-norm $\|\cdot\| = \|\cdot\|_1$ on $\mathscr{C}(\Omega)$,

$$\|u\|_1 = \int_\Omega |u(x)|\,\mathrm{d}x \qquad \text{for } u \in \mathscr{C}(\Omega),$$

where $\Omega_0 := \{x \in \Omega \,|\, u(x) = 0\} \subset \Omega$ and $\Omega_+ := \Omega \setminus \Omega_0$.

Proof. For $u, v \in \mathscr{C}(\Omega)$, we have

$$\begin{aligned}
&\frac{1}{h}(\|u + hv\|_1 - \|u\|_1) \\
&= \frac{1}{h}\left(\int_\Omega |u(x) + hv(x)|\,\mathrm{d}x - \int_\Omega |u(x)|\,\mathrm{d}x\right) \\
&= \frac{1}{h}\int_{\Omega_+} (|u(x) + hv(x)| - |u(x)|)\,\mathrm{d}x + \int_{\Omega_0} |v(x)|\,\mathrm{d}x. \qquad (3.38)
\end{aligned}$$

By $\Omega_h := \{x \in \Omega_+ \,|\, h \cdot |v(x)| < |u(x)|\} \subset \Omega_+$, for $h > 0$, and with (3.36) we represent the first integral in (3.38) as

$$\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}^d} \chi_{\Omega_+}(x)(|u(x) + hv(x)| - |u(x)|)\,\mathrm{d}x \\
&= \int_{\mathbb{R}^d} \chi_{\Omega_h}(x) v(x)\,\mathrm{sgn}(u(x))\,\mathrm{d}x \\
&\quad + \frac{1}{h}\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\,(|u(x) + hv(x)| - |u(x)|)\,\mathrm{d}x, \qquad (3.39)
\end{aligned}$$

where for $\Omega \subset \mathbb{R}^d$

$$\chi_\Omega(x) = \begin{cases} 1 & \text{for } x \in \Omega, \\ 0 & \text{for } x \notin \Omega, \end{cases}$$

denotes the *indicator function* (i.e., the *characteristic function*) for Ω.

Now we estimate the integral in (3.39) from above by

$$\begin{aligned}
&\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\,(|u(x) + hv(x)| - |u(x)|)\,\mathrm{d}x \\
&\le \int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\,(|u(x)| + h|v(x)| - |u(x)|)\,\mathrm{d}x \\
&= h \cdot \int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)|v(x)|\,\mathrm{d}x. \qquad (3.40)
\end{aligned}$$

Since $\chi_{\Omega_h} \longrightarrow \chi_{\Omega_+}$, or, $\chi_{\Omega_+\setminus\Omega_h} \longrightarrow 0$, for $h \searrow 0$, the statement in (3.37) follows from the representations (3.38), (3.39) and (3.40). ∎

To compute the Gâteaux derivatives for the remaining L^p-norms $\|\cdot\|_p$,

$$\|u\|_p = \left(\int_\Omega |u(x)|^p\,\mathrm{d}x\right)^{1/p} \qquad \text{for } u \in \mathscr{C}(\Omega),$$

for $1 < p < \infty$, we need the following lemma.

Lemma 3.66. *For $1 < p < \infty$, let $\varphi(u) = \|u\|_p$ for $u \in \mathscr{C}(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is assumed to be compact. Then, we have*

$$(\varphi^p)'_+(u, v) = p\int_\Omega |u(x)|^{p-1} v(x)\,\mathrm{sgn}(u(x))\,\mathrm{d}x \qquad (3.41)$$

for all $u, v \in \mathscr{C}(\Omega), u \not\equiv 0$.

Proof. For $u, v \in \mathscr{C}(\Omega)$, we have

$$\begin{aligned}
&\frac{1}{h}(\varphi^p(u+hv) - \varphi^p(u)) \\
&= \frac{1}{h}\left(\|u+hv\|_p^p - \|u\|_p^p\right) = \frac{1}{h}\left(\int_\Omega |u(x)+hv(x)|^p \,\mathrm{d}x - \int_\Omega |u(x)|^p \,\mathrm{d}x\right) \\
&= \frac{1}{h}\left(\int_{\Omega_+} \left(|u(x)+hv(x)|^p - |u(x)|^p\right) \,\mathrm{d}x\right) + h^{p-1}\int_{\Omega_0} |v(x)|^p \,\mathrm{d}x, \qquad (3.42)
\end{aligned}$$

where $\Omega_0 = \{x \in \Omega \,|\, u(x) = 0\}$ and $\Omega_+ = \Omega \setminus \Omega_0$.

For $x \in \Omega_h = \{x \in \Omega \,|\, h \cdot |v(x)| < |u(x)|\} \subset \Omega_+$, where $h > 0$, we have

$$\begin{aligned}
|u(x)+hv(x)|^p &= \left(|u(x)| + hv(x)\operatorname{sgn}(u(x))\right)^p \\
&= |u(x)|^p + p \cdot |u(x)|^{p-1} \cdot hv(x)\operatorname{sgn}(u(x)) + o(h) \text{ for } h \searrow 0
\end{aligned}$$

by (3.36) and by Taylor[16] expansion of $F(u) = u^p$ at $|u|$.

Thereby, we can split the first integral in (3.42) into the sum

$$\begin{aligned}
&\frac{1}{h}\left(\int_{\mathbb{R}^d} \chi_{\Omega_+}(x)\left(|u(x)+hv(x)|^p - |u(x)|^p\right) \,\mathrm{d}x\right) \\
&= p\int_{\mathbb{R}^d} \chi_{\Omega_h}(x)|u(x)|^{p-1}v(x)\operatorname{sgn}(u(x)) \,\mathrm{d}x + o(1) \qquad (3.43) \\
&\quad + \frac{1}{h}\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\left(|u(x)+hv(x)|^p - |u(x)|^p\right) \,\mathrm{d}x. \qquad (3.44)
\end{aligned}$$

Now we estimate the expression in (3.44) from above by

$$\begin{aligned}
&\frac{1}{h}\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\left(|u(x)+hv(x)|^p - |u(x)|^p\right) \,\mathrm{d}x \\
&\le \frac{1}{h}\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)\left((|u(x)| + h|v(x)|)^p - |u(x)|^p\right) \,\mathrm{d}x \\
&= p\int_{\mathbb{R}^d} \chi_{\Omega_+\setminus\Omega_h}(x)|u(x)|^{p-1}|v(x)| \,\mathrm{d}x + o(1) \quad \text{for } h \searrow 0. \qquad (3.45)
\end{aligned}$$

Since $\chi_{\Omega_h} \longrightarrow \chi_{\Omega_+}$, or, $\chi_{\Omega_+\setminus\Omega_h} \longrightarrow 0$, for $h \searrow 0$, the stated representation in (3.41) follows from (3.42), (3.43), (3.44), and (3.45). ∎

Now we can finally provide the Gâteaux derivatives for the remaining L^p-norms $\|\cdot\|_p$, for $1 < p < \infty$.

Theorem 3.67. *Let $\Omega \subset \mathbb{R}^d$ be compact. Moreover, suppose $1 < p < \infty$. Then, for the Gâteaux derivative of the L^p-norm $\|\cdot\| = \|\cdot\|_p$ on $\mathscr{C}(\Omega)$, we have*

$$\|'_+(u,v) = \frac{1}{\|u\|_p^{p-1}}\int_\Omega |u(x)|^{p-1}v(x)\operatorname{sgn}(u(x)) \,\mathrm{d}x$$

for all $u, v \in \mathscr{C}(\Omega)$, $u \not\equiv 0$.

[16] Brook Taylor (1685-1731), English mathematician

Proof. The statement follows from the chain rule (3.34) in Theorem 3.53 with $F(x) = x^p$ in combination with the representation of the Gâteaux derivative $(\varphi^p)'_+$ in Lemma 3.66, whereby

$$\varphi'_+(u,v) = \frac{(\varphi^p)'_+(u,v)}{p\varphi^{p-1}(u)} = \frac{p}{p\|u\|_p^{p-1}} \int_\Omega |u(x)|^{p-1} v(x)\,\mathrm{sgn}(u(x))\,\mathrm{d}x,$$

for $\varphi(u) = \|u\|_p$. ∎

3.5 Exercises

Exercise 3.68. Consider approximating the parabola $f(x) = x^2$ on the unit interval $[0,1]$ by linear functions of the form

$$g_\xi(x) = \xi \cdot x \qquad \text{for } \xi \in \mathbb{R}$$

with respect to the p-norms $\|\cdot\|_p$, for $p = 1, 2, \infty$, respectively. To this end, first compute the distance function

$$\eta_p(\xi) = \|g_\xi - f\|_p.$$

Then, determine the best approximation g_{ξ^*} to f satisfying

$$\|g_{\xi^*} - f\|_p = \inf_{\xi\in\mathbb{R}} \|g_\xi - f\|_p$$

along with the minimal distance $\eta_p(\xi^*)$, for each of the three cases $p = 1, 2, \infty$.

Exercise 3.69. Suppose we wish to approximate the identity $f(x) = x$ on the unit interval $[0,1]$, by an exponential sum of the form

$$p_\xi(x) = \xi_1 \mathrm{e}^{\xi_2 x} + \xi_3 \qquad \text{for } \xi = (\xi_1, \xi_2, \xi_3)^T \in \mathbb{R}^3,$$

and with respect to the maximum norm $\|\cdot\|_\infty$.

Show that there is *no* best approximation to f from $\mathcal{S} = \{p_\xi \,|\, \xi \in \mathbb{R}^3\}$.

Hint: Use the parameter sequence $\xi^{(k)} = (k, 1/k, -k)^T$, for $k \in \mathbb{N}$.

Exercise 3.70. Regard the linear space $\mathscr{C}[-\pi,\pi]$, equipped with the norm

$$\|g\| := \|g\|_1 + \|g\|_\infty \qquad \text{for } g \in \mathscr{C}[-\pi,\pi].$$

Moreover, let $f(x) = x$, for $-\pi \le x \le \pi$, and

$$\mathcal{S} = \left\{\alpha \sin^2(\cdot) \,|\, \alpha \in \mathbb{R}\right\} \subset \mathscr{C}[-\pi,\pi].$$

Analyze the existence and uniqueness of the approximation problem

$$\min_{s\in\mathcal{S}} \|s - f\|.$$

Exercise 3.71. Let $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ be a convex functional on a linear space $\mathcal{F}$. Prove the following statements for φ.

(a) If φ has a (global) maximum on $\mathcal{F}$, then φ is constant.
(b) A *local* minimum of φ is also a *global* minimum of φ.

Exercise 3.72. Let $(\mathcal{F}, \|\cdot\|)$ be a normed linear space, whose norm $\|\cdot\|$ is *not* strictly convex. Show that there exists an element $f \in \mathcal{F}$, a linear subspace $\mathcal{S} \subset \mathcal{F}$, and *distinct* best approximations $s_1^*, s_2^* \in \mathcal{S}$ to f, $s_1^* \neq s_2^*$, satisfying

$$\eta(f, \mathcal{S}) = \|s_1^* - f\| = \|s_2^* - f\|.$$

Hint: Take for *suitable* $f_1, f_2 \in \mathcal{F}$, $f_1 \neq f_2$, satisfying $\|f_1\| = \|f_2\| = 1$ and $\|f_1 + f_2\| = 2$, the element $f = \frac{1}{2}(f_1 + f_2) \in \mathcal{F}$ and the linear subspace $\mathcal{S} = \{\alpha(f_1 - f_2) \,|\, \alpha \in \mathbb{R}\} \subset \mathcal{F}$.

Exercise 3.73. Transfer the result of Proposition 3.42 to the case of *odd* functions $f \in \mathscr{C}[-1, 1]$. To this end, formulate and prove a corresponding result for subsets $\mathcal{S} \subset \mathscr{C}[-1, 1]$ that are *invariant under point reflections*, i.e., for any $s(x) \in \mathcal{S}$, we have $-s(-x) \in \mathcal{S}$.

Exercise 3.74. Let $(\mathcal{F}, \|\cdot\|)$ be a normed linear space and $T : \mathcal{F} \longrightarrow \mathcal{F}$ be a linear operator which is *isometric* on $\mathcal{F}$, i.e., $\|Tv\| = \|v\|$ for all $v \in \mathcal{F}$. Moreover, let $\mathcal{S} \subset \mathcal{F}$ be a non-empty subset of $\mathcal{F}$ satisfying $T(\mathcal{S}) \subset \mathcal{S}$.

First prove statements (a) and (b), before you analyze question (c).

(a) If $s^* \in \mathcal{S}$ is a best approximation to $f \in \mathcal{F}$ and $T(\mathcal{S}) = \mathcal{S}$, then $Ts^* \in \mathcal{S}$ is a best approximation to $Tf \in \mathcal{F}$.
(b) If $f \in \mathcal{F}$ is a fixed point of T in $\mathcal{F}$, i.e., $Tf = f$, and $s^* \in \mathcal{S}$ is a unique best approximation to f, then s^* is a fixed point of T in $\mathcal{S}$.
(c) Suppose $f \in \mathcal{F}$ is a fixed point of T in $\mathcal{F}$. Moreover, suppose there is *no* fixed point of T in $\mathcal{S}$, which is also a best approximation to f. Can you draw conclusions concerning the uniqueness of best approximation to f?

Use the results from this exercise to prove Proposition 3.42.

Exercise 3.75. In this exercise, we analyze the existence of *discontinuous* linear functionals φ on $(\mathscr{C}[0, 1], \|\cdot\|_2)$ and on $(\mathscr{C}[0, 1], \|\cdot\|_\infty)$. Give examples, if possible.

(a) Are there *discontinuous* linear functionals on $(\mathscr{C}[0, 1], \|\cdot\|_2)$?
(b) Are there *discontinuous* linear functionals on $(\mathscr{C}[0, 1], \|\cdot\|_\infty)$?

Exercise 3.76. Let $a \leq x_0 < \ldots < x_n \leq b$ be a sequence of pairwise distinct points in $[a,b] \subset \mathbb{R}$ and $\lambda_0, \ldots, \lambda_n \in \mathbb{R}$. Show that the mapping $\varphi : \mathscr{C}[a,b] \longrightarrow \mathbb{R}$, defined as

$$\varphi(f) = \sum_{k=0}^{n} \lambda_k \, f(x_k) \qquad \text{for } f \in \mathscr{C}[a,b],$$

is a continuous linear functional on $(\mathscr{C}[a,b], \|\cdot\|_\infty)$ with operator norm

$$\|\varphi\|_\infty = \sum_{k=0}^{n} |\lambda_k|.$$

Exercise 3.77. Let $(\mathcal{F}, \|\cdot\|)$ be a normed linear space and $\mathcal{S} \subset \mathcal{F}$ be a finite-dimensional linear subspace of $\mathcal{F}$. Moreover, suppose $f \in \mathcal{F}$.

Prove the following statements on linear functionals from the dual space $\mathcal{F}'$.

(a) If $\varphi \in \mathcal{F}'$ satisfies $\|\varphi\| \leq 1$ and $\varphi(\mathcal{S}) = 0$, i.e., $\varphi(s) = 0$ for all $s \in \mathcal{S}$, then we have

$$\eta(f, \mathcal{S}) = \inf_{s \in \mathcal{S}} \|s - f\| \geq |\varphi(f)|$$

for the minimal distance $\eta(f, \mathcal{S})$ between f and $\mathcal{S}$.

(b) There exists one $\varphi \in \mathcal{F}'$ satisfying $\|\varphi\| \leq 1$ and $\varphi(\mathcal{S}) = 0$, such that

$$|\varphi(f)| = \eta(f, \mathcal{S}).$$

If $\eta(f, \mathcal{S}) > 0$, then $\|\varphi\| = 1$.

Exercise 3.78. Consider the linear space $\mathcal{F} = \mathscr{C}([0,1]^2)$, equipped with the maximum norm $\|\cdot\|_\infty$. Approximate the function

$$f(x,y) = x \cdot y \qquad \text{for } (x,y)^T \in [0,1]^2$$

by a function from the linear approximation space

$$\mathcal{S} = \left\{ s \in \mathcal{F} \,|\, s(x,y) = s_1(x) + s_2(y) \text{ for } (x,y)^T \in [0,1]^2 \text{ with } s_1, s_2 \in \mathscr{C}[0,1] \right\}.$$

(a) Construct a linear functional of the form

$$\varphi(g) = \sum_{j=1}^{4} \lambda_j \, g(x_j, y_j) \qquad \text{for } g \in \mathcal{F}$$

to estimate the minimal distance $\eta(f, \mathcal{S})$ between f and $\mathcal{S}$, where

$$\eta(f, \mathcal{S}) \geq \frac{1}{4}.$$

(b) Show that

$$s^*(x,y) = \frac{x}{2} + \frac{y}{2} - \frac{1}{4} \qquad \text{for } (x,y)^T \in [0,1]^2$$

is a best approximation to f from $\mathcal{S}$ with respect to $\|\cdot\|_\infty$.

Exercise 3.79. Show that the function $\varphi : \mathbb{R}^2 \longrightarrow \mathbb{R}$, defined as

$$\varphi(x,y) = \begin{cases} \frac{xy^2}{x^2+y^4} & \text{for } (x,y) \neq 0 \\ 0 & \text{for } (x,y) = 0 \end{cases} \qquad \text{for } (x,y)^T \in \mathbb{R}^2,$$

has a Gâteaux derivative at zero, although φ is *not* continuous at zero.

Exercise 3.80. Let $\mathcal{F}$ be a linear space and $\varphi : \mathcal{F} \longrightarrow \mathbb{R}$ a functional on $\mathcal{F}$. Prove the following statements (related to Remark 3.52).

(a) If φ is convex on $\mathcal{F}$, then the Gâteaux derivative φ'_+ is monotone, i.e.,

$$\varphi'_+(u_1, u_1 - u_2) - \varphi'_+(u_2, u_1 - u_2) \geq 0 \qquad \text{for all } u_1, u_2 \in \mathcal{F}.$$

(b) Assume that the Gâteaux derivative $\varphi'_+(u,v)$ exists for all $u, v \in \mathcal{F}$. Moreover, assume that $\varphi'_+(u,\cdot) : \mathcal{F} \longrightarrow \mathbb{R}$ is sublinear for all $u \in \mathcal{F}$. If the inequality

$$\varphi'_+(u, v-u) \leq \varphi(v) - \varphi(u) \qquad \text{for all } u, v \in \mathcal{F},$$

holds, then φ is convex on $\mathcal{F}$.

4 Euclidean Approximation

In this chapter, we study approximation in *Euclidean* spaces. Therefore, $\mathcal{F}$ denotes a linear space, equipped with a *Euclidean norm* $\|\cdot\|$, i.e., $\|\cdot\|$ is defined by an inner product,

$$\|f\| = (f, f)^{1/2} \qquad \text{for } f \in \mathcal{F}.$$

From the preceding chapter, we understand that Euclidean approximation has fundamental advantages, in particular for the existence and uniqueness of best approximations. We briefly summarize our previous results as follows.

Existence of best approximations: For a *Hilbert space* $\mathcal{F}$, i.e., $\mathcal{F}$ is *complete* with respect to $\|\cdot\|$, and a *closed* and *convex* subset $\mathcal{S} \subset \mathcal{F}$, there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f.

Uniqueness best approximations: For *convex* $\mathcal{S} \subset \mathcal{F}$, a best approximation $s^* \in \mathcal{S}$ to $f \in \mathcal{F}$ is unique, due to the *strict convexity* of $\|\cdot\|$.

The above statements are based on Theorems 3.14, 3.28, and Corollary 3.39 from Chapter 3. Recall that for the existence and uniqueness of s^*, the *parallelogram identity* (3.1) plays a central role. Moreover, according to the Jordan-von Neumann theorem, Theorem 3.10, the parallelogram identity holds only in Euclidean spaces. Therefore, the problem of Euclidean approximation is fundamentally different from the problem of approximation in non-Euclidean spaces.

In this chapter, we explain further advantages of Euclidean approximation. To this end, we rely on the characterizations for best approximations. In particular, we make use of the Kolmogorov criterion, Corollary 3.55, in combination with the representation of Gâteaux derivatives for Euclidean norms from Theorem 3.62. This yields for *finite-dimensional* approximation spaces $\mathcal{S} \subset \mathcal{F}$ constructive methods to compute best approximations by *orthogonal projection* $\Pi : \mathcal{F} \longrightarrow \mathcal{S}$ of $f \in \mathcal{F}$ on $\mathcal{S}$.

We treat two important special cases of Euclidean approximation: Firstly, the approximation of 2π-periodic continuous functions by trigonometric polynomials, where $\mathcal{F} = \mathscr{C}_{2\pi}$ and $\mathcal{S} = \mathcal{T}_n$. Secondly, the approximation of continuous functions by algebraic polynomials, in which case $\mathcal{F} = \mathscr{C}[a, b]$, for a compact interval $[a, b] \subset \mathbb{R}$, and $\mathcal{S} = \mathcal{P}_n$.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_4

4.1 Construction of Best Approximations

In this section, we apply the characterizations for best approximations from the previous chapter to Euclidean spaces. To this end, we assume that the approximation space $\mathcal{S} \subset \mathcal{F}$ is a linear subspace of the Euclidean space $\mathcal{F}$. The application of the Kolmogorov criterion immediately provides the following fundamental result.

Theorem 4.1. *Let $\mathcal{F}$ be a Euclidean space with inner product $(\cdot,\cdot)$. Moreover, suppose $\mathcal{S} \subset \mathcal{F}$ is a convex subset of $\mathcal{F}$. Then the following statements are equivalent.*

(a) $s^* \in \mathcal{S}$ is a best approximation to $f \in \mathcal{F} \setminus \mathcal{S}$.
(b) We have $(s^* - f, s - s^*) \geq 0$ for all $s \in \mathcal{S}$.

Proof. Under the stated assumptions, the equivalence of the Kolmogorov criterion, Corollary 3.55, holds. Thereby, a best approximation $s^* \in \mathcal{S}$ to f is characterized by the necessary and sufficient condition

$$\|'_+(s^* - f, s - s^*) = \left(\frac{s^* - f}{\|s^* - f\|}, s - s^*\right) \geq 0 \qquad \text{for all } s \in \mathcal{S},$$

using the representation of the Gâteaux derivative from Theorem 3.62. ■

Remark 4.2. If $\mathcal{S} \subset \mathcal{F}$ is a *linear subspace* of $\mathcal{F}$, then the variational inequality in statement (b) of Theorem 4.1 immediately leads us to the necessary and sufficient condition

$$(s^* - f, s) = 0 \qquad \text{for all } s \in \mathcal{S}, \tag{4.1}$$

i.e., in this case, $s^* \in \mathcal{S}$ is a best approximation to $f \in \mathcal{F} \setminus \mathcal{S}$, if and only if the orthogonality relation $s^* - f \perp \mathcal{S}$ holds. □

Note that the equivalence statement in Remark 4.2 identifies a best approximation $s^* \in \mathcal{S}$ to $f \in \mathcal{F}$ as the unique *orthogonal projection* of f onto $\mathcal{S}$. In Section 4.2, we will study the projection operator $\Pi : \mathcal{F} \longrightarrow \mathcal{S}$, which assigns every $f \in \mathcal{F}$ to its unique best approximation $s^* \in \mathcal{S}$ in more detail.

Before doing so, we first use the orthogonality in (4.1) to characterize best approximations $s^* \in \mathcal{S}$ for convex subsets $\mathcal{S} \subset \mathcal{F}$. To this end, we work with the dual characterization of Theorem 3.46.

Theorem 4.3. *Let $\mathcal{F}$ be a Euclidean space with inner product $(\cdot,\cdot)$ and let $\mathcal{S} \subset \mathcal{F}$ be a convex subset of $\mathcal{F}$. Moreover, suppose that $s^* \in \mathcal{S}$ satisfies $s^* - f \perp \mathcal{S}$. Then, s^* is the unique best approximation to f.*

Proof. The linear functional $\varphi \in \mathcal{F}'$, defined as

$$\varphi(u) = \left(\frac{s^* - f}{\|s^* - f\|}, u\right) \qquad \text{for } u \in \mathcal{F},$$

satisfies all three conditions from the dual characterization of Theorem 3.46: Indeed, the first condition, $\|\varphi\| = 1$, follows from the Cauchy[1]-Schwarz[2] inequality,

$$|\varphi(u)| = \left| \left(\frac{s^* - f}{\|s^* - f\|}, u \right) \right| \leq \frac{\|s^* - f\|}{\|s^* - f\|} \cdot \|u\| = \|u\| \qquad \text{for all } u \in \mathcal{F},$$

where equality holds for $u = s^* - f \in \mathcal{F}$, since

$$\varphi(s^* - f) = \left(\frac{s^* - f}{\|s^* - f\|}, s^* - f \right) = \frac{\|s^* - f\|^2}{\|s^* - f\|} = \|s^* - f\|.$$

Therefore, φ also satisfies the second condition in Theorem 3.46. By $s^* - f \perp \mathcal{S}$ we have

$$\varphi(s) = \left(\frac{s^* - f}{\|s^* - f\|}, s \right) = 0 \qquad \text{for all } s \in \mathcal{S},$$

and so φ finally satisfies the third condition in Theorem 3.46.

In conclusion, s^* is a best approximation to f. The uniqueness of s^* follows from the strict convexity of the Euclidean norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$. ■

Now we consider the special case of Euclidean approximation by *finite-dimensional* approximation spaces $\mathcal{S}$. Therefore, suppose that $\mathcal{S} \subset \mathcal{F}$ is a linear subspace with $\dim(\mathcal{S}) < \infty$. According to Corollary 3.8, there exists for any $f \in \mathcal{F}$ a best approximation $s^* \in \mathcal{S}$ to f and, moreover, s^* is unique, due to Theorem 3.37.

Suppose that $\mathcal{S}$ is spanned by $n \in \mathbb{N}$ basis elements $\{s_1, \ldots, s_n\} \subset \mathcal{F}$, i.e.,

$$\mathcal{S} = \operatorname{span}\{s_1, \ldots, s_n\} \subset \mathcal{F}$$

so that $\dim(\mathcal{S}) = n < \infty$. To compute the unique best approximation $s^* \in \mathcal{S}$ to some $f \in \mathcal{F}$ we utilize the representation

$$s^* = \sum_{j=1}^{n} c_j^* s_j \in \mathcal{S}. \tag{4.2}$$

Now the (necessary and sufficient) orthogonality condition $s^* - f \perp \mathcal{S}$ from Remark 4.2, or, in (4.1) is equivalent to the requirement

$$(s^*, s_k) = (f, s_k) \qquad \text{for all } 1 \leq k \leq n.$$

Therefore, the representation for s^* in (4.2) leads us to the n linear conditions

$$\sum_{j=1}^{n} c_j^* (s_j, s_k) = (f, s_k) \qquad \text{for } 1 \leq k \leq n$$

[1] Augustin-Louis Cauchy (1789-1857), French mathematician
[2] Hermann Amandus Schwarz (1843-1921), German mathematician

and so to the linear equation system

$$\begin{bmatrix} (s_1,s_1) & (s_2,s_1) & \cdots & (s_n,s_1) \\ (s_1,s_2) & (s_2,s_2) & \cdots & (s_n,s_2) \\ \vdots & \vdots & \ddots & \vdots \\ (s_1,s_n) & (s_2,s_n) & \cdots & (s_n,s_n) \end{bmatrix} \cdot \begin{bmatrix} c_1^* \\ c_2^* \\ \vdots \\ c_n^* \end{bmatrix} = \begin{bmatrix} (f,s_1) \\ (f,s_2) \\ \vdots \\ (f,s_n) \end{bmatrix},$$

or, in short,

$$Gc^* = b \tag{4.3}$$

with the **Gram**[3] **matrix** $G = ((s_j, s_k))_{1 \le k,j \le n} \in \mathbb{R}^{n \times n}$, the unknown coefficient vector $c^* = (c_1^*, \ldots, c_n^*)^T \in \mathbb{R}^n$ of s^* in (4.2) and the right hand side $b = ((f,s_1), \ldots, (f,s_n))^T \in \mathbb{R}^n$. Therefore, the solution $c^* \in \mathbb{R}^n$ of the linear system (4.3) yields the unknown coefficients of s^* in (4.2).

Due to the existence and uniqueness of the best approximation s^*, the Gram matrix G must be regular. We specialize this statement on G as follows.

Theorem 4.4. *The Gram matrix G in (4.3) is symmetric positive definite.*

Proof. The symmetry of G follows from the symmetry of the inner product, whereby $(s_j, s_k) = (s_k, s_j)$ for all $1 \le j,k \le n$. Moreover, G is positive definite, which also immediately follows from the properties of the inner product,

$$c^T G c = \sum_{j,k=1}^{n} c_j c_k (s_j, s_k) = \left(\sum_{j=1}^{n} c_j s_j, \sum_{k=1}^{n} c_k s_k \right) = \left\| \sum_{j=1}^{n} c_j s_j \right\|^2 > 0$$

for all $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n \setminus \{0\}$. ∎

Given our investigations in this section, the problem of Euclidean approximation by finite-dimensional approximation spaces $\mathcal{S}$ seems to be solved. The unique best approximation s^* can be determined by the unique solution of the linear system (4.3), i.e., computing s^* is equivalent to solving (4.3).

But note that we have not posed any conditions on the basis of $\mathcal{S}$, yet.

Next, we show that by suitable choices for bases of $\mathcal{S}$, we can avoid the linear system (4.3). Indeed, for an **orthogonal basis** $\{s_1, \ldots, s_n\}$ of $\mathcal{S}$, i.e.,

$$(s_j, s_k) = \begin{cases} 0 & \text{for } j \ne k, \\ \|s_j\|^2 > 0 & \text{for } j = k, \end{cases}$$

the Gram matrix G is a diagonal matrix,

$$G = \operatorname{diag}(\|s_1\|^2, \ldots, \|s_n\|^2) = \begin{bmatrix} \|s_1\|^2 & & & \\ & \|s_2\|^2 & & \\ & & \ddots & \\ & & & \|s_n\|^2 \end{bmatrix},$$

[3] Jørgen Pedersen Gram (1850-1916), Danish mathematician

in which case the solution c^* of (4.3) is given by

$$c^* = \left(\frac{(f, s_1)}{\|s_1\|^2}, \ldots, \frac{(f, s_n)}{\|s_n\|^2} \right)^T \in \mathbb{R}^n.$$

For an `orthonormal basis` $\{s_1, \ldots, s_n\}$ of $\mathcal{S}$, i.e., $(s_j, s_k) = \delta_{jk}$, the Gram matrix G is the identity matrix, $G = I_n \in \mathbb{R}^{n \times n}$, in which case

$$c^* = ((f, s_1), \ldots, (f, s_n))^T \in \mathbb{R}^n.$$

In the following of this chapter, we develop suitable constructions and characterizations for orthogonal bases in relevant applications. Before doing so, we summarize the discussion of this section, and, moreover, we derive a few elementary properties of orthogonal bases.

4.2 Orthogonal Bases and Orthogonal Projections

From our discussion in the previous section, we can explicitly represent, for a fixed orthogonal basis (orthonormal basis), $\{s_1, \ldots, s_n\}$ of $\mathcal{S}$, the unique best approximation $s^* \in \mathcal{S}$ to f, for any $f \in \mathcal{F}$.

Theorem 4.5. *Let $\mathcal{F}$ be a Euclidean space with inner product $(\cdot, \cdot)$. Moreover, let $\mathcal{S} \subset \mathcal{F}$ be a finite-dimensional linear subspace with orthogonal basis $\{s_1, \ldots, s_n\}$. Then, for any $f \in \mathcal{F}$,*

$$s^* = \sum_{j=1}^{n} \frac{(f, s_j)}{\|s_j\|^2} s_j \in \mathcal{S} \tag{4.4}$$

is the unique best approximation to f. For the special case of an orthonormal basis $\{s_1, \ldots, s_n\}$ for $\mathcal{S}$ we have the representation

$$s^* = \sum_{j=1}^{n} (f, s_j) s_j \in \mathcal{S}.$$

■

Now we study the linear and surjective operator $\Pi : \mathcal{F} \longrightarrow \mathcal{S}$, which maps any element $f \in \mathcal{F}$ to its unique best approximation $s^* \in \mathcal{S}$. But first we note that the optimality of the best approximation $s^* = \Pi(f)$ immediately implies the *stability estimate*

$$\|(I - \Pi)(f)\| \leq \|f - s\| \qquad \text{for all } f \in \mathcal{F}, s \in \mathcal{S} \tag{4.5}$$

where I denotes the identity on $\mathcal{F}$. Moreover, for $f = s$ in (4.5) we get

$$\Pi(s) = s \quad \text{for all } s \in \mathcal{S}$$

and so Π is a `projection operator`, i.e., $\Pi \circ \Pi = \Pi$. By the characterization of the best approximation $s^* = \Pi(f) \in \mathcal{S}$ in (4.1), the operator Π is an `orthogonal projection`, since

$$f - \Pi(f) = (I - \Pi)(f) \perp \mathcal{S} \qquad \text{for all } f \in \mathcal{F},$$

i.e., the linear operator $I - \Pi : \mathcal{F} \longrightarrow \mathcal{S}^\perp$ maps onto the *orthogonal complement* $\mathcal{S}^\perp \subset \mathcal{F}$ of $\mathcal{S}$ in $\mathcal{F}$. Moreover, $I - \Pi$ is also a projection operator, since for any $f \in \mathcal{F}$, we have

$$\begin{aligned}((I - \Pi) \circ (I - \Pi))(f) &= (I - \Pi)(f - \Pi(f)) \\ &= f - \Pi(f) - \Pi(f) + (\Pi \circ \Pi)(f) \\ &= f - \Pi(f) = (I - \Pi)(f).\end{aligned}$$

The orthogonality of Π immediately implies another well-known result.

Theorem 4.6. *The* `Pythagoras`[4] `theorem`

$$\|f - \Pi(f)\|^2 + \|\Pi(f)\|^2 = \|f\|^2 \qquad \textit{for all } f \in \mathcal{F} \tag{4.6}$$

holds.

Proof. For $f \in \mathcal{F}$ we have

$$\begin{aligned}\|f\|^2 &= \|f - \Pi(f) + \Pi(f)\|^2 \\ &= \|f - \Pi(f)\|^2 + 2(f - \Pi(f), \Pi(f)) + \|\Pi(f)\|^2 \\ &= \|f - \Pi(f)\|^2 + \|\Pi(f)\|^2.\end{aligned}$$

■

The Pythagoras theorem implies two further *stability results.*

Corollary 4.7. *For $I \neq \Pi$ the stability estimates*

$$\|(I - \Pi)(f)\| \leq \|f\| \quad \textit{and} \quad \|\Pi(f)\| \leq \|f\| \qquad \textit{for all } f \in \mathcal{F} \tag{4.7}$$

hold. In particular, we have

$$\|I - \Pi\| = 1 \quad \textit{and} \quad \|\Pi\| = 1$$

for the operator norms of $I - \Pi$ and Π.

Proof. The stability estimates in (4.7) follow directly from the Pythagoras theorem, Theorem 4.6. In the first inequality in (4.7), we have equality for every element $f - \Pi(f) \in \mathcal{S}^\perp$, whereas in the second inequality, we have equality for every $s \in \mathcal{S}$. Thereby, the operator norms of $I - \Pi$ and Π are already determined by

$$\|I - \Pi\| = \sup_{f \neq 0} \frac{\|(I - \Pi)(f)\|}{\|f\|} = 1 \quad \text{and} \quad \|\Pi\| = \sup_{f \neq 0} \frac{\|\Pi(f)\|}{\|f\|} = 1.$$

■

[4] Pythagoras of Samos (around 570-510 BC), ancient Greek philosopher

Next, we compute for $f \in \mathcal{F}$ the norm $\|\Pi(f)\|$ of $\Pi(f) = s^*$. To this end, we utilize, for a fixed orthogonal basis $\{s_1, \ldots, s_n\}$ of $\mathcal{S}$ the representation in (4.4), whereby

$$\Pi(f) = \sum_{j=1}^{n} \frac{(f, s_j)}{\|s_j\|^2} s_j \in \mathcal{S} \qquad \text{for } f \in \mathcal{F}. \tag{4.8}$$

In particular, for $s \in \mathcal{S}$, we have the representation

$$\Pi(s) = s = \sum_{j=1}^{n} \frac{(s, s_j)}{\|s_j\|^2} s_j \in \mathcal{S} \qquad \text{for all } s \in \mathcal{S}. \tag{4.9}$$

Theorem 4.8. *Let $\{s_1, \ldots, s_n\} \subset \mathcal{S}$ be an orthogonal basis of $\mathcal{S}$. Then, the* `Parseval`[5] `identity`

$$(\Pi(f), \Pi(g)) = \sum_{j=1}^{n} \frac{(f, s_j)(g, s_j)}{\|s_j\|^2} \qquad \textit{for all } f, g \in \mathcal{F} \tag{4.10}$$

holds, where in particular

$$\|\Pi(f)\|^2 = \sum_{j=1}^{n} \frac{|(f, s_j)|^2}{\|s_j\|^2} \qquad \textit{for all } f \in \mathcal{F}. \tag{4.11}$$

Proof. By the representation of Π in (4.8), we have

$$\begin{aligned}(\Pi(f), \Pi(g)) &= \left(\sum_{j=1}^{n} \frac{(f, s_j)}{\|s_j\|^2} s_j, \sum_{k=1}^{n} \frac{(g, s_k)}{\|s_k\|^2} s_k \right) \\ &= \sum_{j,k=1}^{n} \frac{(f, s_j)}{\|s_j\|^2} \frac{(g, s_k)}{\|s_k\|^2} (s_j, s_k) = \sum_{j=1}^{n} \frac{(f, s_j)(g, s_j)}{\|s_j\|^2}\end{aligned}$$

for all $f, g \in \mathcal{F}$. For $f = g$ we obtain the stated representation in (4.11). ∎

We finally add another important result.

Theorem 4.9. *Let $\{s_1, \ldots, s_n\} \subset \mathcal{S}$ be an orthogonal basis of $\mathcal{S}$. Then the* `Bessel`[6] `inequality`

$$\|\Pi(f)\|^2 = \sum_{j=1}^{n} \frac{|(f, s_j)|^2}{\|s_j\|^2} \le \|f\|^2 \qquad \textit{for all } f \in \mathcal{F} \tag{4.12}$$

holds. Moreover, we have the identity

$$\|f - \Pi(f)\|^2 = \|f\|^2 - \sum_{j=1}^{n} \frac{|(f, s_j)|^2}{\|s_j\|^2} \le \|f\|^2 \qquad \textit{for all } f \in \mathcal{F}.$$

[5] Marc-Antoine Parseval des Chênes (1755-1836), French mathematician
[6] Friedrich Wilhelm Bessel (1784-1846), German astronomer, mathematician

Proof. The Bessel inequality follows from the second stability estimate in (4.7) in combination with the representation in (4.11). The second statement follows from the Pythagoras theorem (4.6) and the representation (4.11). ■

4.3 Fourier Partial Sums

In this section, we study one concrete example for Euclidean approximation. In this particular case, we wish to approximate a continuous 2π-periodic function by real-valued trigonometric polynomials. To this end, we equip the linear space of all *real-valued* continuous 2π-periodic functions

$$\mathscr{C}_{2\pi} \equiv \mathscr{C}_{2\pi}^{\mathbb{R}} = \{f : \mathbb{R} \longrightarrow \mathbb{R} \,|\, f \in \mathscr{C}(\mathbb{R}) \text{ and } f(x) = f(x + 2\pi) \text{ for all } x \in \mathbb{R}\}$$

with the inner product

$$(f, g)_{\mathbb{R}} = \frac{1}{\pi} \int_0^{2\pi} f(x) g(x) \, \mathrm{d}x \qquad \text{for } f, g \in \mathscr{C}_{2\pi}, \tag{4.13}$$

which by $\| \cdot \|_{\mathbb{R}} = (\cdot, \cdot)_{\mathbb{R}}^{1/2}$ defines a Euclidean norm on $\mathscr{C}_{2\pi}$, so that

$$\|f\|_{\mathbb{R}}^2 = \frac{1}{\pi} \int_0^{2\pi} |f(x)|^2 \, \mathrm{d}x \qquad \text{for } f \in \mathscr{C}_{2\pi}.$$

Therefore, $\mathscr{C}_{2\pi}$ with $\| \cdot \|_{\mathbb{R}} = (\cdot, \cdot)_{\mathbb{R}}^{1/2}$ is a Euclidean space.

For the approximation space, we consider choosing the linear space of all real-valued trigonometric polynomials of degree at most $n \in \mathbb{N}_0$,

$$\mathcal{T}_n \equiv \mathcal{T}_n^{\mathbb{R}} = \operatorname{span}\left\{ \frac{1}{\sqrt{2}}, \cos(j \cdot), \sin(j \cdot) \,\middle|\, 1 \le j \le n \right\} \qquad \text{for } n \in \mathbb{N}_0.$$

Therefore, by using the notations introduced at the outset of this chapter, we consider the special case of the Euclidean space $\mathcal{F} = \mathscr{C}_{2\pi}$, equipped with the norm $\| \cdot \|_{\mathbb{R}} = (\cdot, \cdot)_{\mathbb{R}}^{1/2}$, and the linear approximation space $\mathcal{S} = \mathcal{T}_n \subset \mathscr{C}_{2\pi}$ of finite dimension $\dim(\mathcal{T}_n) = 2n + 1$, for $n \in \mathbb{N}_0$.

Remark 4.10. In the following chapters, we also use *complex-valued* trigonometric polynomials from $\mathcal{T}_n^{\mathbb{C}}$ for the approximation of *complex-valued* continuous 2π-periodic functions from

$$\mathscr{C}_{2\pi}^{\mathbb{C}} = \{f : \mathbb{R} \longrightarrow \mathbb{C} \,|\, f \in \mathscr{C}(\mathbb{R}) \text{ and } f(x) = f(x + 2\pi) \text{ for all } x \in \mathbb{R}\}.$$

In that case, we equip $\mathscr{C}_{2\pi}^{\mathbb{C}}$ with the inner product

$$(f, g)_{\mathbb{C}} = \frac{1}{2\pi} \int_0^{2\pi} f(x) \overline{g(x)} \, \mathrm{d}x \qquad \text{for } f, g \in \mathscr{C}_{2\pi}^{\mathbb{C}}, \tag{4.14}$$

and thereby obtain the Euclidean norm $\| \cdot \|_{\mathbb{C}} = (\cdot, \cdot)_{\mathbb{C}}^{1/2}$ on $\mathscr{C}_{2\pi}^{\mathbb{C}}$. The different scalar factors, $1/\pi$ for $(\cdot, \cdot)_{\mathbb{R}}$ in (4.13) and $1/(2\pi)$ for $(\cdot, \cdot)_{\mathbb{C}}$ in (4.14), will be useful later. To keep notations simple, we will from now use $(\cdot, \cdot) = (\cdot, \cdot)_{\mathbb{R}}$ and $\| \cdot \| = \| \cdot \|_{\mathbb{R}}$ for the inner product (4.13) and the norm on $\mathscr{C}_{2\pi} \equiv \mathscr{C}_{2\pi}^{\mathbb{R}}$. □

For the approximation of $f \in \mathscr{C}_{2\pi}$, we use fundamental results, as developed in the previous section. In particular, we make use of *orthonormal systems* to construct best approximations to f. To this end, we take note of the following important result.

Theorem 4.11. *For $n \in \mathbb{N}_0$, the real-valued trigonometric polynomials*

$$\left\{ \frac{1}{\sqrt{2}}, \cos(j\,\cdot), \sin(j\,\cdot) \,\middle|\, 1 \le j \le n \right\} \tag{4.15}$$

form an orthonormal system in $\mathscr{C}_{2\pi}$.

Proof. From the usual addition theorems for trigonometric polynomials we get the identities

$$2\cos(jx)\cos(kx) = \cos((j-k)x) + \cos((j+k)x) \tag{4.16}$$
$$2\sin(jx)\sin(kx) = \cos((j-k)x) - \cos((j+k)x) \tag{4.17}$$
$$2\sin(jx)\cos(kx) = \sin((j-k)x) + \sin((j+k)x). \tag{4.18}$$

The 2π-periodicity of $\cos((j \pm k)x)$ and $\sin((j \pm k)x)$ implies

$$(\cos(j\,\cdot), \cos(k\,\cdot)) = \frac{1}{2\pi} \int_0^{2\pi} [\cos((j-k)x) + \cos((j+k)x)]\,\mathrm{d}x = 0$$
$$(\sin(j\,\cdot), \sin(k\,\cdot)) = \frac{1}{2\pi} \int_0^{2\pi} [\cos((j-k)x) - \cos((j+k)x)]\,\mathrm{d}x = 0$$

for $j \neq k$ and

$$(\sin(j\,\cdot), \cos(k\,\cdot)) = \frac{1}{2\pi} \int_0^{2\pi} [\sin((j-k)x) + \sin((j+k)x)]\,\mathrm{d}x = 0$$

for all $j, k \in \{1, \ldots, n\}$. Moreover, we have

$$\left(\frac{1}{\sqrt{2}}, \cos(j\cdot)\right) = \frac{1}{\sqrt{2}\pi} \int_0^{2\pi} \cos(jx)\,\mathrm{d}x = 0$$
$$\left(\frac{1}{\sqrt{2}}, \sin(j\cdot)\right) = \frac{1}{\sqrt{2}\pi} \int_0^{2\pi} \sin(jx)\,\mathrm{d}x = 0$$

for $j = 1, \ldots, n$, so that the functions in (4.15) form an *orthogonal system*.

The *orthonormality* of the functions in (4.15) finally follows from

$$\left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right) = \frac{1}{2\pi} \int_0^{2\pi} 1\,\mathrm{d}x = 1$$

and

$$(\cos(j\,\cdot),\cos(j\,\cdot)) = \frac{1}{2\pi}\int_0^{2\pi} [1+\cos(2jx)]\,\mathrm{d}x = 1$$
$$(\sin(j\,\cdot),\sin(j\,\cdot)) = \frac{1}{2\pi}\int_0^{2\pi} [1-\cos(2jx)]\,\mathrm{d}x = 1$$

where we use the representations in (4.16) and (4.17) yet once more. ■

We now connect to the results of Theorems 4.5 and 4.11, where we can, for any function $f \in \mathscr{C}_{2\pi}$, represent its unique best approximation $s^* \in \mathcal{T}_n$ by

$$s^*(x) = \left(f, \frac{1}{\sqrt{2}}\right)\frac{1}{\sqrt{2}} + \sum_{j=1}^{n} \left[(f,\cos(j\cdot))\cos(jx) + (f,\sin(j\cdot))\sin(jx)\right]. \quad (4.19)$$

We reformulate the representation for s^* in (4.19) and introduce on this occasion the important notion of *Fourier partial sums*.

Corollary 4.12. *For $f \in \mathscr{C}_{2\pi}$, the unique best approximation $s^* \in \mathcal{T}_n$ to f is given by the n-th* `Fourier partial sum` *of f,*

$$(F_n f)(x) = \frac{a_0}{2} + \sum_{j=1}^{n} \left[a_j\cos(jx) + b_j\sin(jx)\right]. \quad (4.20)$$

The coefficients $a_0 = (f,1)$ and

$$a_j \equiv a_j(f) = (f,\cos(j\cdot)) = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(jx)\,\mathrm{d}x \quad (4.21)$$

$$b_j \equiv b_j(f) = (f,\sin(j\cdot)) = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(jx)\,\mathrm{d}x \quad (4.22)$$

for $1 \le j \le n$ are called `Fourier coefficients` *of f.* ■

The Fourier partial sum (4.20) is split into an *even* part, given by the partial sum of the *even* trigonometric polynomials $\{\cos(j\cdot), 0 \le j \le n\}$ with "even" Fourier coefficients a_j, and into an *odd* part, given by the partial sum of the *odd* trigonometric polynomials $\{\sin(j\cdot), 1 \le j \le n\}$ with "odd" Fourier coefficients b_j. We can show that for any even function $f \in \mathscr{C}_{2\pi}$, all odd Fourier coefficients b_j vanish. Likewise, for an odd function $f \in \mathscr{C}_{2\pi}$, all even Fourier coefficients a_j vanish. On this occasion, we recall the result of Proposition 3.42, from which these statements immediately follow. But we wish to compute the Fourier coefficients explicitly.

Corollary 4.13. *For $f \in \mathscr{C}_{2\pi}$, the following statements are true.*

(a) If f is even, then the Fourier partial sum $F_n f$ in (4.20) is even and the Fourier coefficients a_j in (4.21) have the representation

$$a_j = \frac{2}{\pi} \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x \qquad \text{for } 0 \le j \le n.$$

(b) If f is odd, then the Fourier partial sum $F_n f$ in (4.20) is odd and the Fourier coefficients b_j in (4.22) have the representation

$$b_j = \frac{2}{\pi} \int_0^\pi f(x) \sin(jx)\, \mathrm{d}x \qquad \text{for } 1 \le j \le n.$$

Proof. For an even function $f \in \mathscr{C}_{2\pi}$ we have $b_j = 0$, for all $1 \le j \le n$, since

$$\begin{aligned} b_j &= \frac{1}{\pi} \int_0^{2\pi} f(x) \sin(jx)\, \mathrm{d}x = -\frac{1}{\pi} \int_0^{-2\pi} f(-x) \sin(-jx)\, \mathrm{d}x \\ &= -\frac{1}{\pi} \int_{-2\pi}^0 f(x) \sin(jx)\, \mathrm{d}x = -\frac{1}{\pi} \int_0^{2\pi} f(x) \sin(jx)\, \mathrm{d}x = -b_j, \end{aligned}$$

and so the Fourier partial sum $F_n f$ in (4.20) is even. Moreover, we have

$$a_0 = \frac{1}{\pi} \int_0^{2\pi} f(x)\, \mathrm{d}x = \frac{2}{\pi} \int_0^\pi f(x)\, \mathrm{d}x$$

and, for $1 \le j \le n$,

$$\begin{aligned} \pi a_j &= \int_{-\pi}^\pi f(x) \cos(jx)\, \mathrm{d}x = \int_{-\pi}^0 f(x) \cos(jx)\, \mathrm{d}x + \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x \\ &= \int_0^\pi f(-x) \cos(-jx)\, \mathrm{d}x + \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x = 2 \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x. \end{aligned}$$

This completes our proof for (a). We can prove (b) analogously. ∎

Example 4.14. We consider approximating the periodic function $f \in \mathscr{C}_{2\pi}$, defined as $f(x) = \pi - |x|$, for $x \in [-\pi, \pi]$. To this end, we determine for $n \in \mathbb{N}$ the Fourier coefficients a_j, b_j of the Fourier partial sum $F_n f$. Since f is an even function, we can apply Corollary 4.13, statement (a). From this, we see that $b_j = 0$, for all $1 \le j \le n$, and, moreover,

$$a_j = \frac{2}{\pi} \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x \qquad \text{for } 0 \le j \le n.$$

Integration by parts gives

$$\begin{aligned} \int_0^\pi f(x) \cos(jx)\, \mathrm{d}x &= \frac{1}{j} f(x) \sin(jx) \Big|_0^\pi - \frac{1}{j} \int_0^\pi f'(x) \sin(jx)\, \mathrm{d}x \\ &= \frac{1}{j} \int_0^\pi \sin(jx)\, \mathrm{d}x = -\frac{1}{j^2} \cos(jx) \Big|_0^\pi \qquad \text{for } 1 \le j \le n, \end{aligned}$$

and so we have $a_j = 0$ for all *even* indices $j \in \{1, \dots, n\}$, while

$$a_j = \frac{4}{\pi j^2} \qquad \text{for all } \textit{odd} \text{ indices } j \in \{1, \dots, n\}.$$

We finally compute the Fourier coefficient a_0 by

$$a_0 = (f, 1) = \frac{1}{\pi} \int_0^{2\pi} f(x)\, \mathrm{d}x = \frac{2}{\pi} \int_0^{\pi} (\pi - x)\, \mathrm{d}x = \frac{2}{\pi} \left[-\frac{1}{2}(\pi - x)^2 \right]_0^{\pi} = \pi.$$

Altogether, we obtain the representation

$$\begin{aligned}(F_n f)(x) = \frac{\pi}{2} + \sum_{j=1}^{n} a_j \cos(jx) &= \frac{\pi}{2} + \frac{4}{\pi} \sum_{\substack{j=1 \\ j \text{ odd}}}^{n} \frac{1}{j^2} \cos(jx) \\ &= \frac{\pi}{2} + \frac{4}{\pi} \sum_{k=0}^{\lfloor \frac{n-1}{2} \rfloor} \frac{\cos((2k+1)x)}{(2k+1)^2}\end{aligned}$$

for the n-th Fourier partial sum of f.

For illustration the function graphs of the Fourier partial sums $F_n f$ and of the error functions $F_n f - f$, for $n = 2, 4, 16$, are shown in Figures 4.1-4.3. ♢

As we have seen in Section in 2.6, the *real-valued* Fourier partial sum $F_n f$ in (4.20) can be represented as *complex* Fourier partial sum of the form

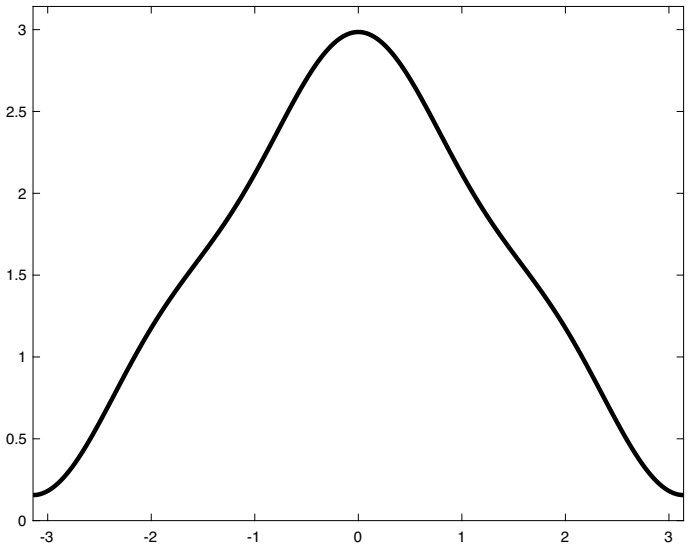
$$(F_n f)(x) = \sum_{j=-n}^{n} c_j \mathrm{e}^{\mathrm{i}jx}. \tag{4.23}$$

For the conversion of the Fourier coefficients, we apply the linear mapping in (2.69), whereby, with using the Eulerean formula (2.67), we obtain for the complex Fourier coefficients in (4.23) the representation

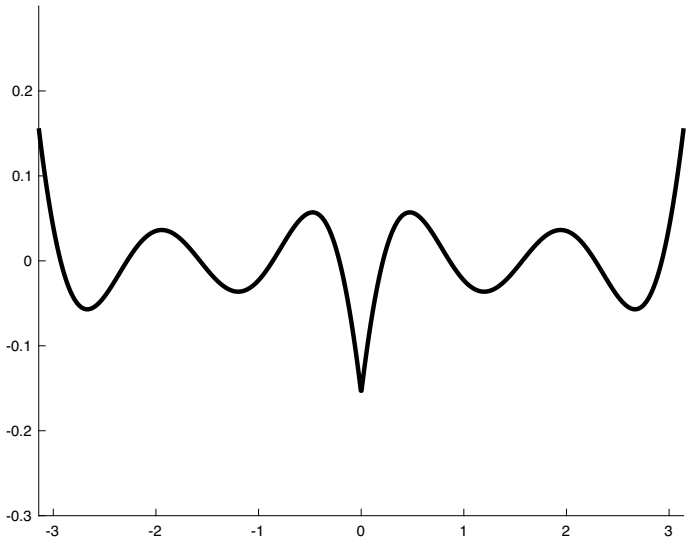
$$c_j = \frac{1}{2\pi} \int_0^{2\pi} f(x) \mathrm{e}^{-\mathrm{i}jx}\, \mathrm{d}x \qquad \text{for } j = -n, \dots, n. \tag{4.24}$$

We remark that the complex Fourier coefficients c_j in (4.24) can also, like the real Fourier coefficients a_j in (4.21) and b_j in (4.22), be expressed via inner products. In fact, by using the *complex* inner product $(\cdot, \cdot)_{\mathbb{C}}$ in (4.14), we can rewrite the representation in (4.24) as

$$c_j = (f, \exp(\mathrm{i}j\cdot))_{\mathbb{C}} \qquad \text{for } j = -n, \dots, n.$$

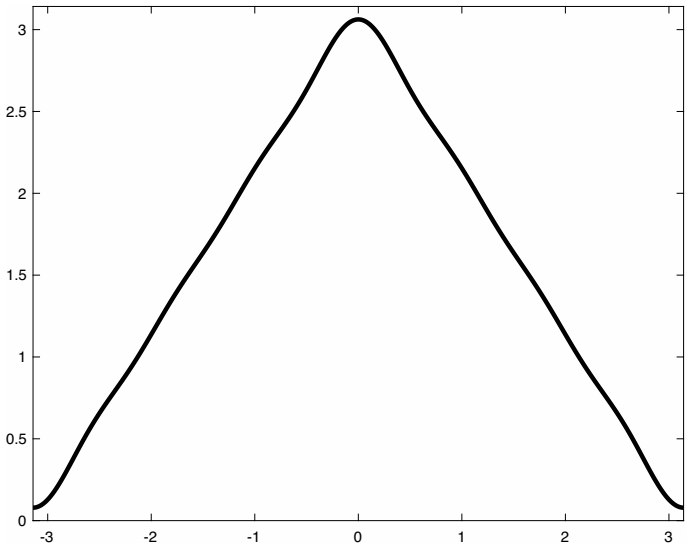


Fourier partial sum $F_2 f$

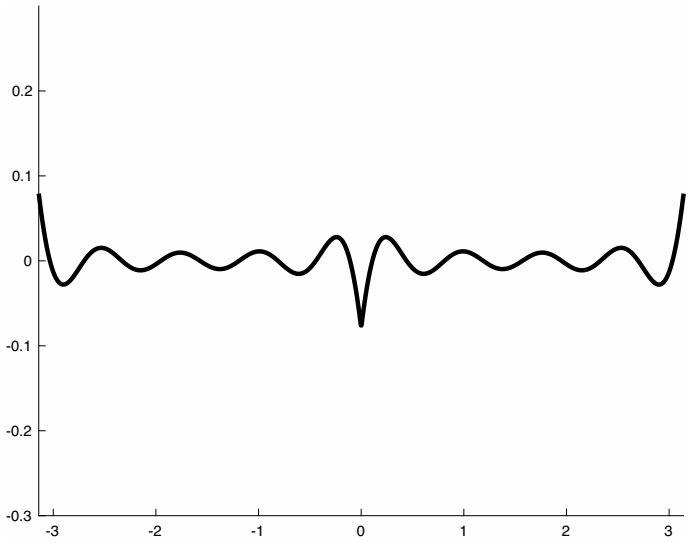


error function $F_2 f - f$

Fig. 4.1. Approximation of the function $f(x) = \pi - |x|$ on the interval $[-\pi, \pi]$ by the Fourier partial sum $(F_2 f)(x)$ (see Example 4.14).

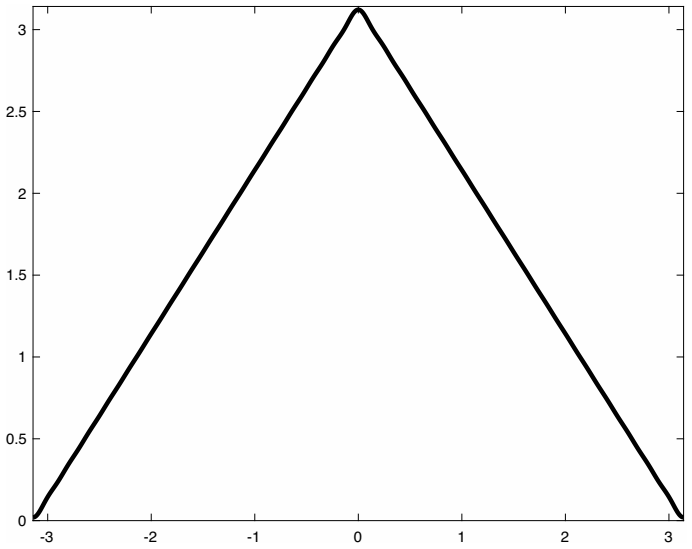


Fourier partial sum $F_4 f$

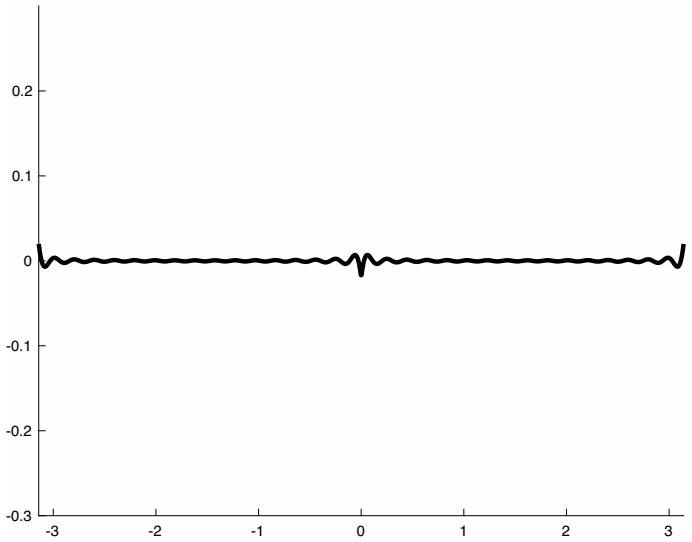


error function $F_4 f - f$

Fig. 4.2. Approximation of the function $f(x) = \pi - |x|$ on the interval $[-\pi, \pi]$ by the Fourier partial sum $(F_4 f)(x)$ (see Example 4.14).



Fourier partial sum $F_{16}f$



error function $F_{16}f - f$

Fig. 4.3. Approximation of the function $f(x) = \pi - |x|$ on the interval $[-\pi, \pi]$ by the Fourier partial sum $(F_{16}f)(x)$ (see Example 4.14).

Now we wish to approximate the complex Fourier coefficients c_j.

To this end, we apply the *composite trapezoidal rule* with $N = 2n + 1$ equidistant knots

$$x_k = \frac{2\pi}{N} k \in [0, 2\pi) \qquad \text{for } k = 0, \ldots, N-1,$$

so that

$$c_j \approx \frac{1}{N} \sum_{k=0}^{N-1} f(x_k) \mathrm{e}^{-\mathrm{i} j x_k} = \frac{1}{N} \sum_{k=0}^{N-1} f(x_k) \omega_N^{-jk} \tag{4.25}$$

where $\omega_N = \mathrm{e}^{2\pi \mathrm{i}/N}$ denotes the N-th root of unity in (2.74). In this way, the vector $c = (c_{-n}, \ldots, c_n)^T \in \mathbb{C}^N$ of the complex Fourier coefficients (4.24) is approximated by the discrete Fourier transform (2.79) from the data vector

$$f = (f_0, \ldots, f_{N-1})^T \in \mathbb{R}^N,$$

where $f_k = f(x_k)$ for $k = 0, \ldots, N-1$. In order to compute the Fourier coefficients $c \in \mathbb{C}^N$ efficiently, we apply the fast Fourier transform (FFT) from Section 2.7. According to Theorem 2.46, the FFT can be performed in $\mathcal{O}(N \log(N))$ steps.

We close this section with the following remark.

The `Fourier operator` $F_n : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ gives the orthogonal projection of $\mathscr{C}_{2\pi}$ onto $\mathcal{T}_n$. In Chapter 6, we will analyze the asymptotic behaviour of the operator F_n, for $n \to \infty$, in more detail, where we will address the following fundamental questions.

- Is the `Fourier series`

$$(F_\infty f)(x) = \frac{a_0}{2} + \sum_{j=1}^{\infty} [a_j \cos(jx) + b_j \sin(jx)] \qquad \text{for } f \in \mathscr{C}_{2\pi}$$

  of f convergent?
- If so, does the Fourier series $F_\infty f$ converge to f?
- If so, how fast does the Fourier series $F_\infty f$ converge to f?

In particular, we will investigate, if at all, *in which sense* (e.g. pointwise, or uniformly, or with respect to the Euclidean norm $\|\cdot\|$) the convergence of the Fourier series $F_\infty f$ holds. In Chapter 6, we will give answers, especially for the asymptotic behaviour of the approximation error

$$\eta(f, \mathcal{T}_n) = \|F_n f - f\| \quad \text{and} \quad \eta_\infty(f, \mathcal{T}_n) = \|F_n f - f\|_\infty \qquad \text{for } n \to \infty.$$

This will lead us to specific conditions on the smoothness of f.

4.4 Orthogonal Polynomials

Now we study another important special case of Euclidean approximation. In this particular case, we wish to approximate continuous functions from $\mathscr{C}[a,b]$, where $[a,b] \subset \mathbb{R}$ denotes a compact interval.

For the approximation space, we take, for fixed $n \in \mathbb{N}_0$, the linear space $\mathcal{P}_n$ of algebraic polynomials of degree at most $n \in \mathbb{N}_0$, where $\dim(\mathcal{P}_n) = n+1$, i.e., throughout this section, we regard the special case where $\mathcal{S} = \mathcal{P}_n$ and $\mathcal{F} = \mathscr{C}[a,b]$.

We introduce an inner product for the function space $\mathscr{C}[a,b]$ as follows. For a positive and integrable *weight function* $w \in \mathscr{C}(a,b)$, satisfying

$$\int_a^b w(x)\,\mathrm{d}x < \infty,$$

the function space $\mathscr{C}[a,b]$ is by

$$(f,g)_w = \int_a^b f(x)g(x)w(x)\,\mathrm{d}x \qquad \text{for } f,g \in \mathscr{C}[a,b]$$

equipped with an inner product, yielding the Euclidean norm $\|\cdot\|_w = (\cdot,\cdot)_w^{1/2}$, so that

$$\|f\|_w^2 = \int_a^b |f(x)|^2 w(x)\,\mathrm{d}x \qquad \text{for } f \in \mathscr{C}[a,b].$$

Later in this section, we make concrete examples for the weight function w.

To approximate functions from $\mathscr{C}[a,b]$, we apply Theorem 4.5, so that we can, for $f \in \mathscr{C}[a,b]$, represent the unique best approximation $s^* \in \mathcal{P}_n$ to f explicitly. In order to do so, we need an orthogonal system for $\mathcal{P}_n$. To this end, we propose an algorithm, which constructs for any weighted inner product $(\cdot,\cdot)_w$ an orthogonal basis

$$\{p_0, p_1, \ldots, p_n\} \subset \mathcal{P}_n$$

for the polynomial space $\mathcal{P}_n$.

The following orthogonalization algorithm by *Gram-Schmidt*[7] belongs to the standard repertoire of linear algebra. In this iterative method, a given basis $\mathcal{B} \subset \mathcal{S}$ of a finite-dimensional Euclidean space $\mathcal{S}$ is, by successive orthogonal projections of the basis elements from $\mathcal{B}$, transformed into an orthogonal basis of $\mathcal{S}$. The formulation of this constructive method is detailed in the `Gram-Schmidt algorithm`, Algorithm 4, which we here apply to the monomial basis $\mathcal{B} = \{1, x, x^2, \ldots, x^n\}$ of $\mathcal{S} = \mathcal{P}_n$.

[7] Erhard Schmidt (1876-1959), German mathematician

Algorithm 4 Gram-Schmidt algorithm

1: **function** GRAM-SCHMIDT
2: **let** $p_0 := 1$;
3: **for** $k = 0, \ldots, n-1$ **do**
4: **let**

$$p_{k+1} := x^{k+1} - \sum_{j=0}^{k} \frac{(x^{k+1}, p_j)_w}{\|p_j\|_w^2} p_j;$$

5: **end for**
6: **end function**

Proposition 4.15. *The polynomials $p_0, \ldots, p_n \in \mathcal{P}_n$, output by the Gram-Schmidt algorithm, Algorithm 4, form an orthogonal basis for $\mathcal{P}_n$.*

Proof. Obviously, $p_k \in \mathcal{P}_k \subset \mathcal{P}_n$, for all $0 \le k \le n$. Moreover, the orthogonality relation

$$p_{k+1} = x^{k+1} - \Pi(x^{k+1}) \perp \mathcal{P}_k \qquad \text{for all } k = 0, \ldots, n-1,$$

holds, where

$$\Pi(x^{k+1}) = \sum_{j=0}^{k} \frac{(x^{k+1}, p_j)_w}{\|p_j\|_w^2} p_j$$

is the orthogonal projection of the monomial x^{k+1} onto $\mathcal{P}_k$ w.r.t. $(\cdot,\cdot)_w$. Therefore, the polynomials $p_0, \ldots, p_n$ form an orthogonal basis for $\mathcal{P}_n$. ∎

Note that the orthogonalization method of Gram-Schmidt guarantees, for any weighted inner product $(\cdot,\cdot)_w$, the existence of an orthogonal basis for $\mathcal{P}_n$ with respect to $(\cdot,\cdot)_w$. Moreover, the Gram-Schmidt construction of orthogonal polynomials in Algorithm 4 is unique up to $n+1$ (non-vanishing) scaling factors, one for the initialization (in line 2) and one for each of the n for loop cycles (line 4). The scaling factors could be used to normalize the orthogonal system of polynomials, where the following options are commonly used.

- *Normalization of the leading coefficient*
  $p_0 \equiv 1$ and $p_k(x) = x^k + q_{k-1}(x)$ for some $q_{k-1} \in \mathcal{P}_{k-1}$ for $k = 1, \ldots, n$
- *Normalization at one*
  $p_k(1) = 1$ for all $k = 0, \ldots, n$
- *Normalization of norm (orthonormalization)*
  Let $p_0 := p_0/\|p_0\|_w$ (line 2) and $p_k := p_k/\|p_k\|_w$ (line 4), $k = 1, \ldots, n$.

However, the Gram-Schmidt algorithm is problematic for numerical reasons. In fact, on the one hand, it is *unstable*, especially for input bases $\mathcal{B}$ with almost linearly dependent basis elements. On the other hand, the Gram-Schmidt algorithm is very *inefficient*. In contrast, the following *three-term recursion* is much more suitable for efficient and stable constructions of orthogonal polynomials.

Theorem 4.16. *For any weighted inner product $(\cdot,\cdot)_w$, there are unique orthogonal polynomials $p_k \in \mathcal{P}_k$, for $k \geq 0$, with leading coefficient one. The orthogonal polynomials $(p_k)_{k\in\mathbb{N}_0}$ satisfy the* `three-term recursion`

$$p_k(x) = (x + a_k)p_{k-1}(x) + b_k p_{k-2}(x) \qquad \text{for } k \geq 1 \tag{4.26}$$

for initial values $p_{-1} \equiv 0$, $p_0 \equiv 1$ and coefficients

$$a_k = -\frac{(xp_{k-1}, p_{k-1})_w}{\|p_{k-1}\|_w^2} \text{ for } k \geq 1 \quad \text{and} \quad b_1 = 1,\ b_k = -\frac{\|p_{k-1}\|_w^2}{\|p_{k-2}\|_w^2} \text{ for } k \geq 2.$$

Proof. We prove the statement by induction on k.

Initial step: For $k = 0$, the constant $p_0 \equiv 1$ is the unique polynomial in $\mathcal{P}_0$ with leading coefficient one.

Induction hypothesis: Assume that $p_0, \ldots, p_{k-1}$, $k \geq 1$, are unique orthogonal polynomials with leading coefficient one, where $p_j \in \mathcal{P}_j$ for $j = 0, \ldots, k-1$.

Induction step $(k-1 \longrightarrow k)$*:* Let $p_k \in \mathcal{P}_k \setminus \mathcal{P}_{k-1}$ be a polynomial with leading coefficient one. Then, the difference $p_k - xp_{k-1}$ lies in $\mathcal{P}_{k-1}$, so that we have (with using the orthogonal basis $p_0, \ldots, p_{k-1}$ of $\mathcal{P}_{k-1}$) the representation

$$p_k(x) - xp_{k-1}(x) = \sum_{j=0}^{k-1} c_j p_j(x) \qquad \text{with } c_j = \frac{(p_k - xp_{k-1}, p_j)_w}{\|p_j\|_w^2}.$$

We now formulate necessary conditions for the coefficients c_j, under which the orthogonality $p_k \perp \mathcal{P}_{k-1}$ holds. From $p_k \perp \mathcal{P}_{k-1}$ we get

$$c_j = -\frac{(xp_{k-1}, p_j)_w}{\|p_j\|_w^2} = -\frac{(p_{k-1}, xp_j)_w}{\|p_j\|_w^2} \qquad \text{for } j = 0, \ldots, k-1.$$

This in turn implies $c_0 = \ldots = c_{k-3} = 0$ and, moreover,

$$c_{k-1} = -\frac{(xp_{k-1}, p_{k-1})_w}{\|p_{k-1}\|_w^2} \text{ and } c_{k-2} = -\frac{(p_{k-1}, xp_{k-2})_w}{\|p_{k-2}\|_w^2} = -\frac{(p_{k-1}, p_{k-1})_w}{\|p_{k-2}\|_w^2}.$$

Therefore, all coefficients $c_0, \ldots, c_{k-1}$ are uniquely determined, whereby p_k is uniquely determined. Moreover, the stated three-term recursion in (4.26),

$$p_k(x) = (x + c_{k-1})p_{k-1}(x) + c_{k-2}p_{k-2}(x) = (x + a_k)p_{k-1}(x) + b_k p_{k-2}(x)$$

holds with $a_k = c_{k-1}$, for $k \geq 1$, $b_k = c_{k-2}$, for $k \geq 2$, and where $b_1 := 1$. ■

Remark 4.17. Due to the uniqueness of the coefficients a_k, for $k \geq 1$, and b_k, for $k \geq 2$, the conditions of the three-term recursion (4.26) are also sufficient, i.e., the three-term recursion in (4.26) generates the unique orthogonal polynomials $p_k \in \mathcal{P}_k$, for $k \geq 0$, w.r.t. the weighted inner product $(\cdot,\cdot)_w$. □

Next, we discuss important properties of orthogonal polynomials, where their zeros are of particular interest. We can show that orthogonal polynomials have only simple zeros. To this end, we first prove a more general result for continuous functions.

Theorem 4.18. *Let $g \in \mathscr{C}[a,b]$ satisfy $(g,p)_w = 0$ for all $p \in \mathcal{P}_n$, i.e., $g \perp \mathcal{P}_n$, for $n \in \mathbb{N}_0$. Then, either g vanishes identically on $[a,b]$ or g has at least $n+1$ zeros with changing sign in (a,b).*

Proof. Suppose $g \in \mathscr{C}[a,b] \setminus \{0\}$ has only $k < n+1$ zeros

$$a < x_1 < \ldots < x_k < b$$

with changing sign. Then, the product $g \cdot p$ between g and the polynomial

$$p(x) = \prod_{j=1}^{k} (x - x_j) \in \mathcal{P}_k \subset \mathcal{P}_n$$

has no sign change on (a,b). Therefore, the inner product

$$(g,p)_w = \int_a^b g(x)p(x)w(x)\,\mathrm{d}x$$

cannot vanish. This is in contradiction to the assumed orthogonality $g \perp \mathcal{P}_n$. Therefore, g has at least $n+1$ zeros with changing sign in (a,b). ∎

Corollary 4.19. *Suppose $p_n \in \mathcal{P}_n$ is a polynomial satisfying $p_n \perp \mathcal{P}_{n-1}$, for $n \in \mathbb{N}$. Then, either $p_n \equiv 0$ or p_n has exactly n simple zeros in (a,b).*

Proof. On the one hand, by Theorem 4.18, p_n has *at least* n pairwise distinct zeros in (a,b). Now suppose $p_n \not\equiv 0$. Since p_n is an algebraic polynomial in $\mathcal{P}_n \setminus \{0\}$, p_n has, on the other hand, *at most* n zeros. Altogether, p_n has *exactly* n zeros in (a,b), where each zero must be simple. ∎

Corollary 4.20. *Let $p_n^* \in \mathcal{P}_n$ be a best approximation to $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_n$. Then, the error function $p_n^* - f$ has at least $n+1$ zeros with changing sign in (a,b).*

Proof. According to Remark 4.2, we have the orthogonality $p_n^* - f \perp \mathcal{P}_n$. Since $f \not\equiv p_n^*$, the error function $p_n^* - f$ has, due to Theorem 4.18, at least $n+1$ zeros with changing sign in (a,b). ∎

We remark that Corollary 4.20 yields a necessary condition to characterize the best approximation $p^* \in \mathcal{P}_n$ to $f \in \mathscr{C}[a,b]$. This condition could *a posteriori* be used for consistency check. If we knew the $n+1$ simple zeros $X = \{x_1, \ldots, x_{n+1}\} \subset (a,b)$ of the error function $p^* - f$ *a priori*, then we

would be able to compute the best approximation $p^* \in \mathcal{P}_n$ via the interpolation conditions $p_X^* = f_X$. However, in the general case, the zeros of $p^* - f$ are usually unknown.

We finally introduce three relevant families of orthogonal polynomials. For a more comprehensive discussion on orthogonal polynomials, we refer to the classical textbook [67] by Gábor Szegő[8] or, for numerical aspects, to the textbook [57].

4.4.1 Chebyshev Polynomials

We have studied the *Chebyshev polynomials*

$$T_n(x) = \cos(n \arccos(x)) \qquad \text{for } n \in \mathbb{N}_0 \tag{4.27}$$

already in Section 2.5. Let us first recall some of the basic properties of the Chebyshev polynomials $T_n \in \mathcal{P}_n$, in particular the three-term recursion from Theorem 2.27,

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \qquad \text{for } n \in \mathbb{N} \tag{4.28}$$

with initial values $T_0 \equiv 1$ and $T_1(x) = x$.

Now we show that the Chebyshev polynomials $\{T_0, \ldots, T_n\} \subset \mathcal{P}_n$ are orthogonal polynomials w.r.t. the weight function $w : (-1, 1) \longrightarrow (0, \infty)$, defined as

$$w(x) = \frac{1}{\sqrt{1 - x^2}} \qquad \text{for } x \in (-1, 1). \tag{4.29}$$

Theorem 4.21. *For $n \in \mathbb{N}_0$ the set of Chebyshev polynomials $\{T_0, \ldots, T_n\}$ form an orthogonal basis for $\mathcal{P}_n$ w.r.t. the weight function w in (4.29), where*

$$(T_j, T_k)_w = \begin{cases} 0 & \text{for } j \neq k \\ \pi & \text{for } j = k = 0 \\ \pi/2 & \text{for } j = k > 0 \end{cases} \qquad \text{for } 0 \leq j, k \leq n. \tag{4.30}$$

Proof. By the substitution $\phi = \arccos(x)$ we show the orthogonality

$$\begin{aligned} (T_j, T_k)_w &= \int_{-1}^{1} \frac{T_j(x) T_k(x)}{\sqrt{1 - x^2}} \, \mathrm{d}x = \int_{-1}^{1} \frac{\cos(j \arccos(x)) \cos(k \arccos(x))}{\sqrt{1 - x^2}} \, \mathrm{d}x \\ &= \int_{\pi}^{0} \frac{\cos(j\phi) \cos(k\phi)}{\sqrt{1 - \cos^2(\phi)}} (-\sin(\phi)) \, \mathrm{d}\phi = \int_{0}^{\pi} \cos(j\phi) \cos(k\phi) \, \mathrm{d}\phi \\ &= \|T_k\|_w^2 \, \delta_{jk} \end{aligned}$$

by using Theorem 4.11. Theorem 4.11 also yields the stated values for the squared norms $\|T_k\|_w^2 = (T_k, T_k)_w$. ∎

[8] Gábor Szegő (1895-1985), Hungarian mathematician

We remark that the Chebyshev polynomials are normalized by

$$T_n(1) = 1 \qquad \text{for all } n \geq 0.$$

Indeed, this follows directly by induction from the three-term recursion (4.28). Due to Corollary 2.28, the n-th Chebyshev polynomial T_n has, for $n \geq 1$, the leading coefficient 2^{n-1}, and so the scaled polynomial

$$p_n(x) = 2^{1-n} T_n(x) \qquad \text{for } n \geq 1 \tag{4.31}$$

has leading coefficient one. Thus, the orthogonal polynomials $p_0, \ldots, p_n \in \mathcal{P}_n$ satisfy the three-term recursion (4.26) in Theorem 4.16. We now show that the three-term recursion in (4.26) is consistent with the three-term recursion (4.28) for the Chebyshev polynomials.

To this end, we compute the coefficients a_k and b_k from Theorem 4.16. We first remark that the coefficients a_k are invariant under scalings of the basis elements p_k. Thus we can show that for the case of the Chebyshev polynomials all coefficients a_k in the three-term recursion (4.26) must vanish, since by the substitution $\phi = \arccos(x)$ we have

$$(xT_k(x), T_k(x))_w = \int_0^\pi \cos(\phi) \cos^2(k\phi) \, \mathrm{d}\phi = 0 \qquad \text{for all } k \geq 0$$

for the nominator of a_{k+1} in (4.26). For the coefficients b_k, we get $b_1 = 1$, $b_2 = -1/2$, and

$$b_{k+1} = -\frac{\|p_k\|_w^2}{\|p_{k-1}\|_w^2} = -\frac{\|2^{1-k} T_k\|_w^2}{\|2^{2-k} T_{k-1}\|_w^2} = -\frac{1}{4} \frac{\|T_k\|_w^2}{\|T_{k-1}\|_w^2} = -\frac{1}{4} \qquad \text{for } k \geq 2.$$

From Theorem 4.16, we obtain $p_0 \equiv 1$, $p_1(x) = x$, $p_2(x) = x^2 - 1/2$ and

$$p_{k+1}(x) = x p_k(x) - \frac{1}{4} p_{k-1}(x) \qquad \text{for } k \geq 2.$$

Rescaling with (4.31) finally yields the three-term recursion (4.28).

In our above computations, we rely on structural advantages of Chebyshev polynomials: On the one hand, the *degree-independent* representation of the squared norms $\|T_k\|_w^2$ in (4.30) simplifies the calculations of the coefficients b_k significantly. On the other hand, for the calculations of the coefficients a_k we can take advantage of the orthonormality of the even trigonometric polynomials $\cos(k\cdot)$. We wish to further discuss this important relation between the orthonormal system $(\cos(k\cdot))_{k \in \mathbb{N}_0}$ and the orthogonal system $(T_k)_{k \in \mathbb{N}_0}$.

By our previous (more general) investigations in Section 4.2 the unique best approximation $p_n^* \in \mathcal{P}_n$ to a function $f \in \mathscr{C}[-1, 1]$ is given by the orthogonal projection $\mathit{\Pi}_n f \equiv \mathit{\Pi}_{\mathcal{P}_n} f$ of f onto $\mathcal{P}_n$,

$$\mathit{\Pi}_n f = \sum_{k=0}^{n} \frac{(f, T_k)_w}{\|T_k\|_w^2} T_k = \frac{1}{\pi}(f, 1)_w + \frac{2}{\pi} \sum_{k=1}^{n} (f, T_k)_w T_k, \tag{4.32}$$

where the form of `Chebyshev partial sum` in (4.32) reminds us on the form of the Fourier partial sum $F_n f$ from Corollary 4.12. Indeed, the coefficients in the series expansion for the best approximation $\Pi_n f$ in (4.32) can be identified as Fourier coefficients.

Theorem 4.22. *For $f \in \mathscr{C}[-1,1]$ the coefficients of the Chebyshev partial sum (4.32) coincide with the Fourier coefficients $a_k \equiv a_k(g)$ of the even function $g(x) = f(\cos(x))$, so that*

$$\Pi_n f = \frac{a_0}{2} + \sum_{k=1}^{n} a_k T_k. \tag{4.33}$$

Proof. For $f \in \mathscr{C}[-1,1]$, the coefficients $(f, T_k)_w$ in (4.32) can be computed by using the substitution $\phi = \arccos(x)$:

$$\begin{aligned}(f, T_k)_w &= \int_{-1}^{1} \frac{f(x)T_k(x)}{\sqrt{1-x^2}}\,\mathrm{d}x = \int_0^{\pi} f(\cos(\phi))\cos(k\phi)\,\mathrm{d}\phi \\ &= \frac{\pi}{2}\frac{1}{\pi}\int_0^{2\pi} f(\cos(x))\cos(kx)\,\mathrm{d}x = \frac{\pi}{2}a_k(g),\end{aligned}$$

where $a_k(g)$, $k \geq 1$, denotes the k-th Fourier coefficient of $g(x) = f(\cos(x))$. For $k = 0$ we finally get the Fourier coefficient a_0 of g by

$$\frac{(f, T_0)_w}{\|T_0\|_w^2} = \frac{\pi}{2}a_0(g)\frac{1}{\pi} = \frac{a_0(g)}{2}.$$

■

As we had remarked at the end of Section 4.3, the Fourier coefficients $a_k \equiv a_k(g)$ can efficiently be approximated by the fast Fourier transform (FFT). Now we introduce the *Clenshaw algorithm* [15], Algorithm 5, which, on input coefficients $a = (a_0, \ldots, a_n)^T \in \mathbb{R}^{n+1}$, yields an efficient and stable evaluation of the Chebyshev partial sum (4.33) at $x \in [-1,1]$.

Algorithm 5 Clenshaw algorithm

function CLENSHAW(a, x)
  Input: coefficients $a = (a_0, \ldots, a_n)^T \in \mathbb{R}^{n+1}$ and $x \in [-1,1]$.

  let $z_{n+1} := 0$; $z_n := a_n$;
  for $k = n-1, \ldots, 0$ **do**
    let $z_k := a_k + 2xz_{k+1} - z_{k+2}$;
  end for
  return $(\Pi_n f)(x) = (z_0 - z_2)/2$.
end function

To verify the Clenshaw algorithm, we use the recursion for the Chebyshev polynomials in (4.28). By the assignment in line 6 of the Clenshaw algorithm, we get the representation

$$a_k = z_k - 2xz_{k+1} + z_{k+2} \qquad \text{for } k = n-1, \ldots, 0 \tag{4.34}$$

for the coefficients of the Chebyshev partial sum, where for $k = n$ with $z_{n+1} = 0$ and $z_n = a_n$ we get $z_{n+2} = 0$. The sum over the last n terms of the Chebyshev partial sum (4.33) can be rewritten by using the representation in (4.34) in combination with the recursion (4.28):

$$\begin{aligned}
\sum_{k=1}^{n} a_k T_k(x) &= \sum_{k=1}^{n} (z_k - 2xz_{k+1} + z_{k+2}) T_k(x) \\
&= \sum_{k=1}^{n} z_k T_k(x) - \sum_{k=2}^{n+1} 2xz_k T_{k-1}(x) + \sum_{k=3}^{n+2} z_k T_{k-2}(x) \\
&= z_1 T_1(x) + z_2 T_2(x) - 2xz_2 T_1(x) \\
&\qquad + \sum_{k=3}^{n} z_k \left[T_k(x) - 2xT_{k-1}(x) + T_{k-2}(x) \right] \\
&= z_1 x + z_2 (2x^2 - 1) - 2xz_2 x \\
&= z_1 x - z_2 .
\end{aligned}$$

With $a_0 = z_0 - 2xz_1 + z_2$ we get the representation

$$(\Pi_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} a_k T_k(x) = \frac{1}{2} \left(z_0 - 2xz_1 + z_2 + 2z_1 x - 2z_2 \right) = \frac{1}{2} (z_0 - z_2).$$

Finally, we provide a *memory efficient* implementation of the Clenshaw algorithm, Algorithm 6.

Algorithm 6 Clenshaw algorithm (memory efficient)

```
1: function CLENSHAW(a, x)
2:     Input: coefficients a = (a_0, ..., a_n)^T ∈ R^{n+1} and x ∈ [−1, 1].
3:
4:     let z ≡ (z_0, z_1, z_2) := (a_n, 0, 0);
5:     for k = n − 1, ..., 0 do
6:         let z_2 = z_1; z_1 = z_0;
7:         let z_0 = a_k + 2x · z_1 − z_2;
8:     end for
9:     return (Π_n f)(x) = (z_0 − z_2)/2.
10: end function
```

4.4.2 Legendre Polynomials

Now we discuss another example for orthogonal polynomials on $[-1, 1]$.

Definition 4.23. *For $n \in \mathbb{N}_0$, the* `Rodrigues`[9] `formula`

$$L_n(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left((x^2-1)^n\right)\frac{n!}{(2n)!} \qquad \text{for } n \geq 0 \tag{4.35}$$

defines the n-th `Legendre`[10] `polynomial`. ◯

We show that the Legendre polynomials are the (unique) orthogonal polynomials with leading coefficient one, belonging to the weight function $w \equiv 1$. Therefore, we regard the usual (unweighted) L^2 inner product on $\mathscr{C}[-1, 1]$, defined as

$$(f, g)_w := (f, g) = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x \qquad \text{for } f, g \in \mathscr{C}[-1, 1].$$

Theorem 4.24. *For $n \in \mathbb{N}_0$, the Legendre polynomials $L_0, \ldots, L_n$ form an orthogonal basis for $\mathcal{P}_n$ with respect to the weight function $w \equiv 1$ on $[-1, 1]$.*

Proof. Obviously, $L_k \in \mathcal{P}_k \subset \mathcal{P}_n$ for all $0 \leq k \leq n$.

Now for $0 \leq k \leq n$ we consider the integral

$$I_{nk} = \int_{-1}^{1} \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left((x^2-1)^n\right)\frac{\mathrm{d}^k}{\mathrm{d}x^k}\left((x^2-1)^k\right)\,\mathrm{d}x.$$

For $0 \leq i \leq n$, we have the representation

$$I_{nk} = (-1)^i \int_{-1}^{1} \frac{\mathrm{d}^{n-i}}{\mathrm{d}x^{n-i}}\left((x^2-1)^n\right)\frac{\mathrm{d}^{k+i}}{\mathrm{d}x^{k+i}}\left((x^2-1)^k\right)\,\mathrm{d}x, \tag{4.36}$$

as can be shown by induction (using integration by parts).

For $i = n$ in (4.36), we have

$$I_{nk} = (-1)^n \int_{-1}^{1} (x^2-1)^n \frac{\mathrm{d}^{k+n}}{\mathrm{d}x^{k+n}}\left((x^2-1)^k\right)\,\mathrm{d}x = 0 \qquad \text{for } n > k \tag{4.37}$$

which implies

$$(L_n, L_k) = \frac{n!k!}{(2n)!(2k)!} I_{nk} = 0 \qquad \text{for } n > k. \tag{4.38}$$

■

[9] Benjamin Olinde Rodrigues (1795-1851), French mathematician and banker
[10] Adrien-Marie Legendre (1752-1833), French mathematician

We note two more important properties of the Legendre polynomials.

Theorem 4.25. *The Legendre polynomials L_n in (4.35) satisfy the following properties.*

(a) L_n has leading coefficient one.
(b) We have $L_n(-x) = (-1)^n L_n(x)$ for all $x \in [-1, 1]$.

Proof. For $n \geq 0$, by using (4.35), we get the representation

$$\frac{(2n)!}{n!} L_n(x) = \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left((x^2-1)^n\right) = \frac{\mathrm{d}^n}{\mathrm{d}x^n}\left(\sum_{j=0}^{n} \binom{n}{j} x^{2j}(-1)^{n-j}\right)$$
$$= n! \sum_{n/2 \leq j \leq n} \binom{n}{j}\binom{2j}{n} x^{2j-n}(-1)^{n-j}. \tag{4.39}$$

(a) By (4.39) the Legendre polynomial L_n has leading coefficient one.

(b) For even n, we have that $2j - n$ is even, and so in this case all terms in (4.39) are even, which implies that L_n is even. Likewise, we can show that L_n is odd for odd n (by analogy). Altogether, we see that statement (b) holds. ∎

In conclusion, the Legendre polynomials are, for the L^2 inner product $(\cdot, \cdot)$ on $[-1, 1]$, the unique orthogonal polynomials with leading coefficient one. Finally, we derive a three-term recursion for the Legendre polynomials from (4.26).

Theorem 4.26. *The Legendre polynomials satisfy the three-term recursion*

$$L_{n+1}(x) = xL_n(x) - \frac{n^2}{4n^2-1} L_{n-1}(x) \qquad \text{for } n \geq 1 \tag{4.40}$$

with initial values $L_0 \equiv 1$ and $L_1(x) = x$.

Proof. Obviously, $L_0 \equiv 1$ and $L_1(x) = x$.

By Theorem 4.16, the sought three-term recursion has the form (4.26), where

$$a_n = -\frac{(xL_{n-1}, L_{n-1})}{\|L_{n-1}\|^2} \text{ for } n \geq 1 \quad \text{and} \quad b_1 = 1,\ b_n = -\frac{\|L_{n-1}\|^2}{\|L_{n-2}\|^2} \text{ for } n \geq 2.$$

By statement (b) in Theorem 4.25, the Legendre polynomial L_n^2 is, for any $n \in \mathbb{N}_0$, even, and therefore $xL_n^2(x)$ is odd, so that $a_n = 0$ for all $n \geq 0$.

Table 4.1. The Legendre polynomials L_n in monomial form, for $n = 1, \ldots, 10$.

$$
\begin{aligned}
L_1(x) &= x \\
L_2(x) &= x^2 - \frac{1}{3} \\
L_3(x) &= x^3 - \frac{3}{5}x \\
L_4(x) &= x^4 - \frac{6}{7}x^2 + \frac{3}{35} \\
L_5(x) &= x^5 - \frac{10}{9}x^3 + \frac{5}{21}x \\
L_6(x) &= x^6 - \frac{15}{11}x^4 + \frac{5}{11}x^2 - \frac{5}{231} \\
L_7(x) &= x^7 - \frac{21}{13}x^5 + \frac{105}{143}x^3 - \frac{35}{429}x \\
L_8(x) &= x^8 - \frac{28}{15}x^6 + \frac{14}{13}x^4 - \frac{28}{143}x^2 + \frac{7}{1287} \\
L_9(x) &= x^9 - \frac{36}{17}x^7 + \frac{126}{85}x^5 - \frac{84}{221}x^3 + \frac{63}{2431}x \\
L_{10}(x) &= x^{10} - \frac{45}{19}x^8 + \frac{630}{323}x^6 - \frac{210}{323}x^4 + \frac{315}{4199}x^2 - \frac{63}{46189}
\end{aligned}
$$

Now we compute the coefficients b_n for $n \geq 2$.
By the representation (4.37) for the integral I_{nk} we have, for $k = n$,

$$
\begin{aligned}
I_{nn} &= (-1)^n (2n)! \int_{-1}^{1} (x^2 - 1)^n \, \mathrm{d}x = (2n)! \int_{-1}^{1} (1 - x^2)^n \, \mathrm{d}x \\
&= (2n)! \int_{-1}^{1} (1 - x)^n (1 + x)^n \, \mathrm{d}x \\
&= (2n)! \cdot \frac{n!}{(n+1) \cdot \ldots \cdot (2n)} \int_{-1}^{1} (1 + x)^{2n} \, \mathrm{d}x \\
&= (n!)^2 \cdot \left[\frac{1}{2n+1} (1 + x)^{2n+1} \right]_{x=-1}^{x=1} \\
&= (n!)^2 \cdot \frac{2^{2n+1}}{2n+1}
\end{aligned}
$$

after n-fold integration by parts. This gives the representation

$$\|L_n\|^2 = \frac{(n!)^2}{((2n)!)^2} \cdot I_{nn} = \frac{(n!)^4}{((2n)!)^2} \cdot \frac{2^{2n+1}}{2n+1} \qquad \text{for } n \geq 0$$

and therefore

$$\begin{aligned} b_{n+1} &= -\frac{\|L_n\|^2}{\|L_{n-1}\|^2} = -\frac{n^4}{(2n)^2(2n-1)^2} \frac{2^2(2n-1)}{2n+1} \\ &= -\frac{n^2}{(2n-1)(2n+1)} = -\frac{n^2}{4n^2-1} \qquad \text{for } n \geq 1, \end{aligned}$$

which proves the stated three-term recursion. ■

The Legendre polynomials L_n, for $n = 1, \ldots, 10$, are, in their monomial form, shown in Table 4.1. To compute the entries for Table 4.1, we have used the three-term recursion (4.40) in Theorem 4.26, with initial values $L_0 \equiv 1$ and $L_1(x) = x$. In summary, we see from Theorem 4.25 that

- L_n has leading coefficient one;
- L_{2k} is even for $k \in \mathbb{N}_0$;
- L_{2k+1} is odd for $k \in \mathbb{N}$.

Note that the above properties of the Legendre polynomials are consistent with the representations of L_n for $n = 2, \ldots, 10$, in Table 4.1.

4.4.3 Hermite Polynomials

We finally discuss one example of orthogonal polynomials on $\mathbb{R}$.

Definition 4.27. *For $n \in \mathbb{N}_0$, we let $H_n : \mathbb{R} \longrightarrow \mathbb{R}$, defined as*

$$H_n(x) = (-1)^n \mathrm{e}^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2} \qquad \textit{for } n \geq 0, \tag{4.41}$$

denote the n-th `Hermite`[11] `polynomial`. ○

Next, we show that the Hermite polynomials H_n are orthogonal polynomials on $\mathbb{R}$ with leading coefficient 2^n with respect to the weight function

$$w(x) = \mathrm{e}^{-x^2}.$$

Therefore, in this case, we work with the weighted L^2 inner product

$$(f, g)_w = (f, g) = \int_{\mathbb{R}} f(x) g(x)\, \mathrm{e}^{-x^2}\, \mathrm{d}x \qquad \text{for } f, g \in \mathscr{C}(\mathbb{R}). \tag{4.42}$$

Theorem 4.28. *For $n \in \mathbb{N}_0$, the Hermite polynomials $H_0, \ldots, H_n$ form an orthogonal basis for $\mathcal{P}_n$ with respect to the weighted inner product $(\cdot, \cdot)_w$.*

[11] Charles Hermite (1822-1901), French mathematician

Proof. We first show for $n \in \mathbb{N}_0$ the representation

$$w^{(n)}(x) = P_n(x) \cdot \mathrm{e}^{-x^2} \qquad \text{for some } P_n \in \mathcal{P}_n \setminus \mathcal{P}_{n-1}. \tag{4.43}$$

We prove the representation in (4.43) by induction on $n \geq 0$.
Initial step: For $n = 0$, we have (4.43) with $P_0 \equiv 1 \in \mathcal{P}_0$.
Induction hypothesis: Suppose the representation in (4.43) holds for $n \in \mathbb{N}_0$.
Induction step ($n \longrightarrow n+1$)*:* We get the stated representation by

$$\begin{aligned} w^{(n+1)}(x) &= \frac{\mathrm{d}}{\mathrm{d}x} w^{(n)}(x) = \frac{\mathrm{d}}{\mathrm{d}x} \left(P_n(x) \cdot \mathrm{e}^{-x^2} \right) \\ &= (P_n'(x) - 2xP_n(x)) \cdot \mathrm{e}^{-x^2} = P_{n+1}(x) \cdot \mathrm{e}^{-x^2} \end{aligned}$$

with $P_{n+1}(x) = P_n'(x) - 2xP_n(x)$, where $P_{n+1} \in \mathcal{P}_{n+1} \setminus \mathcal{P}_n$ for $P_n \in \mathcal{P}_n \setminus \mathcal{P}_{n-1}$.

Due to (4.43), the Hermite polynomial H_n, $n \geq 0$, has the representation

$$H_n(x) = (-1)^n \mathrm{e}^{x^2} \cdot P_n(x) \cdot \mathrm{e}^{-x^2} = (-1)^n P_n(x) \qquad \text{for } x \in \mathbb{R},$$

so that $H_n \in \mathcal{P}_n \setminus \mathcal{P}_{n-1}$. Moreover, by (4.43) we have

$$w^{(n)}(x) = (-1)^n \mathrm{e}^{-x^2} \cdot H_n(x) \qquad \text{for } x \in \mathbb{R}.$$

Now we consider for fixed $x \in \mathbb{R}$ the function $g_x : \mathbb{R} \longrightarrow \mathbb{R}$, defined as

$$g_x(t) := w(x+t) = \mathrm{e}^{-(x+t)^2} \qquad \text{for } t \in \mathbb{R}.$$

By Taylor series expansion on the analytic function g_x around zero, we get

$$w(x+t) = g_x(t) = \sum_{k=0}^{\infty} \frac{g_x^{(k)}(0)}{k!} t^k = \sum_{k=0}^{\infty} \frac{t^k}{k!} w^{(k)}(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} (-1)^k \mathrm{e}^{-x^2} H_k(x).$$

This yields, for the function $h(x,t) = \mathrm{e}^{2xt - t^2}$, the series expansion

$$h(x,t) = w(x-t) \cdot \mathrm{e}^{x^2} = \sum_{k=0}^{\infty} \frac{t^k}{k!} H_k(x) \qquad \text{for all } x, t \in \mathbb{R}. \tag{4.44}$$

Now on the one hand we have for $s, t \in \mathbb{R}$ the representation

$$\begin{aligned} \int_{\mathbb{R}} \mathrm{e}^{-x^2} h(x,t) h(x,s) \, \mathrm{d}x &= \int_{\mathbb{R}} \mathrm{e}^{-x^2} \mathrm{e}^{2x(t+s)} \mathrm{e}^{-(t^2+s^2)} \, \mathrm{d}x \\ &= \int_{\mathbb{R}} \mathrm{e}^{-(x-(t+s))^2} \mathrm{e}^{2ts} \, \mathrm{d}x \\ &= \mathrm{e}^{2ts} \int_{\mathbb{R}} \mathrm{e}^{-x^2} \, \mathrm{d}x = \sqrt{\pi} \cdot \mathrm{e}^{2ts} \\ &= \sqrt{\pi} \cdot \sum_{k=0}^{\infty} \frac{(2ts)^k}{k!}. \end{aligned} \tag{4.45}$$

On the other hand, with using the uniform convergence of the series for $h(x,t)$ in (4.44), we have the representation

$$\int_{\mathbb{R}} \mathrm{e}^{-x^2} h(x,t)h(x,s)\,\mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{-x^2} \left(\sum_{k=0}^{\infty} \frac{t^k}{k!} H_k(x)\right) \left(\sum_{j=0}^{\infty} \frac{s^j}{j!} H_j(x)\right) \mathrm{d}x$$

$$= \sum_{k,j=0}^{\infty} \frac{t^k s^j}{k!j!} \int_{\mathbb{R}} \mathrm{e}^{-x^2} H_k(x)H_j(x)\,\mathrm{d}x. \tag{4.46}$$

By comparing the coefficients in (4.45) and (4.46), we get

$$\int_{\mathbb{R}} \mathrm{e}^{-x^2} H_k(x)H_j(x)\,\mathrm{d}x = 2^k\sqrt{\pi}k! \cdot \delta_{jk} \qquad \text{for all } j,k \in N_0, \tag{4.47}$$

and so in particular

$$\|H_k\|_w^2 = 2^k\sqrt{\pi}k! \qquad \text{for all } k \in \mathbb{N}_0.$$

This completes our proof. ■

Now we proof a three-term recursion for the Hermite polynomials.

Theorem 4.29. *The Hermite polynomials satisfy the three-term recursion*

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x) \qquad \textit{for } n \geq 0 \tag{4.48}$$

with the initial values $H_{-1} \equiv 0$ *and* $H_0(x) \equiv 1$.

Proof. Obviously, we have $H_0 \equiv 1$. By applying partial differentiation to the series expansion for $h(x,t)$ in (4.44) with respect to variable t we get

$$\frac{\partial}{\partial t} h(x,t) = 2(x-t)h(x,t) = \sum_{k=1}^{\infty} \frac{t^{k-1}}{(k-1)!} H_k(x)$$

and this implies

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} 2xH_k(x) - \sum_{k=0}^{\infty} 2\frac{t^{k+1}}{k!} H_k(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} H_{k+1}(x). \tag{4.49}$$

Moreover, we have

$$\sum_{k=0}^{\infty} \frac{t^{k+1}}{k!} H_k(x) = \sum_{k=0}^{\infty} \frac{t^{k+1}}{(k+1)!}(k+1)H_k(x) = \sum_{k=0}^{\infty} \frac{t^k}{k!} kH_{k-1}(x) \tag{4.50}$$

with $H_{-1} \equiv 0$. Inserting (4.50) into (4.49) gives the identity

$$\sum_{k=0}^{\infty} \frac{t^k}{k!} \left(2xH_k(x) - 2kH_{k-1}(x)\right) = \sum_{k=0}^{\infty} \frac{t^k}{k!} H_{k+1}(x). \tag{4.51}$$

By comparing the coefficients in (4.51), we finally get the stated three-term recursion in (4.48) with the initial values $H_{-1} \equiv 0$ and $H_0 \equiv 1$. ■

Table 4.2. The Hermite polynomials H_n in monomial form, for $n = 1, \ldots, 8$.

$$
\begin{aligned}
H_1(x) &= 2x \\
H_2(x) &= 4x^2 - 2 \\
H_3(x) &= 8x^3 - 12x \\
H_4(x) &= 16x^4 - 48x^2 + 12 \\
H_5(x) &= 32x^5 - 160x^3 + 120x \\
H_6(x) &= 64x^6 - 480x^4 + 720x^2 - 120 \\
H_7(x) &= 128x^7 - 1344x^5 + 3360x^3 - 1680x \\
H_8(x) &= 256x^8 - 3584x^6 + 13440x^4 - 13440x + 1680
\end{aligned}
$$

By Theorem 4.29, we get another recursion for the Hermite polynomials.

Corollary 4.30. *The Hermite polynomials H_n satisfy the recursion*

$$H_n'(x) = 2nH_{n-1}(x) \qquad \text{for } n \in \mathbb{N}. \tag{4.52}$$

Proof. Differentiation of H_n in (4.41) yields

$$H_n'(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left((-1)^n \mathrm{e}^{x^2} \frac{\mathrm{d}^n}{\mathrm{d}x^n} \mathrm{e}^{-x^2}\right) = 2xH_n(x) - H_{n+1}(x),$$

whereby (4.52) follows from the three-term recursion for H_{n+1} in (4.48). ■

Further properties of the Hermite polynomials H_n follow immediately from the recursions in Theorem 4.29 and in Corollary 4.30.

Corollary 4.31. *The Hermite polynomials H_n in (4.41) satisfy the following properties.*

(a) H_n has leading coefficient 2^n, for $n \geq 0$.
(b) H_{2n} is even and H_{2n+1} is odd, for $n \geq 0$.

Proof. Statement (a) follows by induction from the three-term recursion (4.48), whereas statement (b) follows from (4.52) with $H_0 \equiv 1$ and $H_1(x) = 2x$. ■

We can conclude that for the weighted L^2 inner product $(\cdot, \cdot)_w$ in (4.42) the Hermite polynomials H_n are the unique orthogonal polynomials with leading coefficient 2^n. The Hermite polynomials H_n are, for $n = 1, \ldots, 8$, shown in their monomial form in Table 4.2.

4.5 Exercises

Exercise 4.32. Let $\mathcal{F} = \mathscr{C}[-1,1]$ be equipped with the Euclidean norm $\|\cdot\|_2$, defined by the inner product

$$(f,g) = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x \qquad \text{for } f,g \in \mathscr{C}[-1,1],$$

so that $\|\cdot\|_2 = (\cdot,\cdot)^{1/2}$. Compute on given coefficients $a,b,c,d \in \mathbb{R}$, $a \neq 0$, of a cubic polynomial

$$f(x) = a\,x^3 + b\,x^2 + c\,x + d \in \mathcal{P}_3 \setminus \mathcal{P}_2 \qquad \text{for } x \in [-1,1]$$

the unique best approximation p_2^* to f from $\mathcal{P}_2$ with respect to $\|\cdot\|_2$.

Exercise 4.33. Compute for $n \in \mathbb{N}_0$ the Fourier coefficients $a_0,\ldots,a_n \in \mathbb{R}$ and $b_1,\ldots,b_n \in \mathbb{R}$ of the Fourier partial sum

$$F_n(x) = \frac{a_0}{2} + \sum_{j=1}^{n}[a_j\,\cos(jx) + b_j\,\sin(jx)] \qquad \text{for } x \in [0,2\pi)$$

(a) for the *rectangular wave*

$$R(x) = \begin{cases} 0 & \text{for } x \in \{0,\pi,2\pi\} \\ 1 & \text{for } x \in (0,\pi) \\ -1 & \text{for } x \in (\pi,2\pi); \end{cases}$$

(b) for the *saw-tooth function*

$$S(x) = \begin{cases} 0 & \text{for } x \in \{0,2\pi\} \\ \frac{1}{2}(\pi - x) & \text{for } x \in (0,2\pi). \end{cases}$$

Plot the graphs of R and the best approximation $F_{10}R$ to R in one figure. Plot the graphs of S and the best approximation $F_{10}S$ to S in one figure.

Exercise 4.34. Approximate the function $f(x) = 2x-1$ on the unit interval $[0,1]$ by a trigonometric polynomial of the form

$$T_n(x) = \frac{c_0}{2} + \sum_{k=1}^{n} c_k \cos(k\pi x) \qquad \text{for } x \in [0,1]. \tag{4.53}$$

Compute (for arbitrary $n \in \mathbb{N}_0$) the unique best approximation T_n^* of the form (4.53) to f with respect to the Euclidean norm $\|\cdot\|_2$ on $[0,1]$. Then determine the *smallest* $m \in \mathbb{N}$, satisfying

$$\int_0^1 |f(x) - T_m^*(x)|^2\,\mathrm{d}x \leq 10^{-4},$$

and give the best approximation T_m^* to f in explicit form.

Exercise 4.35. For a continuous, positive and integrable weight function $w : (a,b) \longrightarrow (0,\infty)$, let $\mathscr{C}[a,b]$ be equipped with the weighted Euclidean norm $\|\cdot\|_w = (\cdot,\cdot)_w^{1/2}$, defined by

$$(f,g)_w = \int_a^b f(x)\, g(x)\, w(x)\, \mathrm{d}x \quad \text{for } f,g \in \mathscr{C}[a,b].$$

Moreover, let $(p_k)_{k\in\mathbb{N}_0}$, with $p_k \in \mathcal{P}_k$, be the unique sequence of orthogonal polynomials with respect to $(\cdot,\cdot)_w$ with leading coefficient one. According to Theorem 4.16, the orthogonal polynomials p_k satisfy the three-term recursion

$$p_k(x) = (x + a_k)\, p_{k-1}(x) + b_k\, p_{k-2}(x) \quad \text{for } k \geq 1$$

with initial values $p_{-1} \equiv 0$, $p_0 \equiv 1$, and with the coefficients

$$a_k = -\frac{(xp_{k-1}, p_{k-1})_w}{\|p_{k-1}\|_w^2} \text{ for } k \geq 1 \quad \text{and} \quad b_1 = 1,\ b_k = -\frac{\|p_{k-1}\|_w^2}{\|p_{k-2}\|_w^2} \text{ for } k \geq 2.$$

Prove the following statements for $k \in \mathbb{N}_0$.

(a) Among all polynomials $p \in \mathcal{P}_k$ with leading coefficient one, the orthogonal polynomial p_k is *norm-minimal* with respect to $\|\cdot\|_w$, i.e.,

$$\|p_k\|_w = \min\left\{\|p\|_w \,|\, p \in \mathcal{P}_k \text{ with } p(x) = x^k + q(x) \text{ for } q \in \mathcal{P}_{k-1}\right\}.$$

(b) For all $x, y \in [a,b]$, where $x \neq y$, we have

$$\sum_{j=0}^{k} \frac{p_j(x)\, p_j(y)}{\|p_j\|_w^2} = \frac{1}{\|p_k\|_w^2}\, \frac{p_{k+1}(x)\, p_k(y) - p_k(x)\, p_{k+1}(y)}{x - y}$$

and, moreover,

$$\sum_{j=0}^{k} \frac{(p_j(x))^2}{\|p_j\|_w^2} = \frac{p'_{k+1}(x)\, p_k(x) - p'_k(x)\, p_{k+1}(x)}{\|p_k\|_w^2} \qquad \text{for all } x \in [a,b].$$

(c) Conclude from (b) that all zeros of p_k are simple. Moreover, conclude that p_{k+1} and p_k have no common zeros.

Exercise 4.36. Show the following identities of the Chebyshev polynomials.

(a) $T_k \cdot T_\ell = \frac{1}{2}\left(T_{k+\ell} + T_{|k-\ell|}\right)$ for all $k, \ell \in \mathbb{N}_0$.
(b) $T_k(-x) = (-1)^k\, T_k(x)$ for all $k \in \mathbb{N}_0$.
(c) $T_k \circ T_\ell = T_{k\ell}$ for all $k, \ell \in \mathbb{N}_0$.

Exercise 4.37. In this problem, make use of the results in Exercise 4.36.

(a) Prove for $g \in \mathscr{C}[-1,1]$ and $h(x) = x \cdot g(x)$, for $x \in [-1,1]$, the relation

$$c_0(h) = c_1(g) \quad \text{and} \quad c_k(h) = \frac{1}{2}\left(c_{k-1}(g) + c_{k+1}(g)\right) \quad \text{for all } k \geq 1$$

between the Chebyshev coefficients $c_k(g)$ of g and $c_k(h)$ of h.

(b) Conclude from the relation in Exercise 4.36 (c) the representation

$$T_{2k}(x) = T_k(2x^2 - 1) \qquad \text{for all } x \in [-1,1] \text{ and } k \in \mathbb{N}_0. \tag{4.54}$$

(c) Can the representation in (4.54) be used to simplify the evaluation of a Chebyshev partial sum for an *even* function in the Clenshaw algorithm, Algorithm 5? If so, how could this simplification be used for the implementation of the Clenshaw algorithm?

Exercise 4.38. On given coefficient functions $a_k \in \mathscr{C}[a,b]$, for $k \geq 1$, and $b_k \in \mathscr{C}[a,b]$, for $k \geq 2$, let $p_k \in \mathscr{C}[a,b]$, for $k \geq 0$, be a function sequence satisfying the three-term recursion

$$p_{k+1}(x) = a_{k+1}(x)\, p_k(x) + b_{k+1}(x)\, p_{k-1}(x) \qquad \text{for } k \geq 1$$

with initial functions $p_0 \in \mathscr{C}[a,b]$ and $p_1 = a_1 p_0 \in \mathscr{C}[a,b]$. Show that the sum

$$f_n(x) = \sum_{j=0}^{n} c_j\, p_j(x) \qquad \text{for } x \in [a,b]$$

can, on given coefficients $c = (c_0, \ldots, c_n)^T \in \mathbb{R}^{n+1}$, be evaluated by the following generalization of the *Clenshaw algorithm*, Algorithm 7.

Algorithm 7 Clenshaw algorithm

```
1: function CLENSHAW(c, x)
2:     Input: coefficients c = (c_0, ..., c_n)^T ∈ R^{n+1} and x ∈ [a, b].
3:
4:     let z_{n+1} = 0; z_n = c_n;
5:     for k = n − 1, ..., 0 do
6:         let z_k = c_k + a_{k+1}(x) z_{k+1} + b_{k+2}(x) z_{k+2};
7:     end for
8:     return f_n(x) = p_0(x) z_0.
9: end function
```

Which algorithm results especially for the evaluation of a *Legendre partial sum*

$$f_n(x) = \sum_{j=0}^{n} c_j L_j(x) \qquad \text{for } x \in [-1,1]$$

with the Legendre polynomials $L_0, \ldots, L_n$ (from Definition 4.23)?

Exercise 4.39. In this problem, we consider approximating the exponential function $f(x) = \mathrm{e}^{-x}$ on the interval $[-1, 1]$ by polynomials from $\mathcal{P}_n$, for $n \in \mathbb{N}_0$, with respect to the weighted norm $\|\cdot\|_w = (\cdot,\cdot)_w^{1/2}$, where

$$w(x) = \frac{1}{\sqrt{1-x^2}} \qquad \text{for } x \in (-1, 1).$$

To this end, we use the Chebyshev polynomials $T_k(x) = \cos(k \arccos(x))$. Compute the coefficients $c^* = (c_0^*, \ldots, c_n^*)^T \in \mathbb{R}^{n+1}$ of the best approximation

$$p_n^*(x) = \sum_{k=0}^{n} c_k^* \, T_k(x) \in \mathcal{P}_n \qquad \text{for } x \in [-1, 1] \text{ and } n \in \mathbb{N}_0.$$

Exercise 4.40. In this problem, we use the Legendre polynomials

$$L_k(x) = \frac{\mathrm{d}^k}{\mathrm{d}x^k}\left((x^2-1)^k\right) \frac{k!}{(2k)!} \qquad \text{for } 0 \le k \le n$$

to determine the best approximation $p_n^* \in \mathcal{P}_n$, $n \in \mathbb{N}_0$, to the exponential function $f(x) = \mathrm{e}^{-x}$ on $[-1, 1]$ w.r.t. the (unweighted) Euclidean norm $\|\cdot\|_2$.

Compute the first eight coefficients $c^* = (c_0^*, \ldots, c_7^*)^T \in \mathbb{R}^8$ of the sought best approximation

$$p_n^*(x) = \sum_{k=0}^{n} c_k^* \, L_k(x) \qquad \text{for } x \in [-1, 1].$$

Exercise 4.41. In this programming problem, we compare the two approximations to $f(x) = \mathrm{e}^{-x}$ from the previous Exercises 4.39 and 4.40.

(a) Evaluate the two best approximations $p_n^* \in \mathcal{P}_n$ (from Exercises 4.39 and 4.40, respectively) for $n = 3, 4, 5, 6, 7$ at $N + 1$ equidistant points

$$x_j = -1 + \frac{2j}{N} \qquad \text{for } j = 0, \ldots, N$$

for a *suitable* $N \ge 1$ by the modified Clenshaw algorithm, Algorithm 6. Plot the graphs of the functions p_n^* and f in one figure, for $n = 3, 4, 5, 6, 7$.

(b) Record for your computations in (a) the approximation errors

$$\varepsilon_2 = \sqrt{\sum_{j=0}^{N} |p_n^*(x_j) - f(x_j)|^2} \quad \text{and} \quad \varepsilon_\infty = \max_{0 \le j \le N} |p_n^*(x_j) - f(x_j)|,$$

for $n = 3, 4, 5, 6, 7$. Display your results in one table.

(c) Compare the approximation by Chebyshev polynomials (Exercise 4.39) with the approximation by Legendre polynomials (Exercise 4.40). Take notes of your numerical observations. Did the observed numerical results match your perception?

Exercise 4.42. Consider for $n \in \mathbb{N}_0$ the *Hermite function*

$$h_n(x) := H_n(x) \cdot \mathrm{e}^{-x^2/2} \qquad \text{for } x \in \mathbb{R}, \tag{4.55}$$

where H_n denotes the n-th Hermite polynomial in (4.41).

Show that the Hermite functions h_n satisfy the differential equation

$$h_n''(x) - \left(x^2 - 2n - 1\right) h_n(x) = 0 \qquad \text{for } n \geq 0.$$

Moreover, prove the recursion

$$h_{n+1}(x) = x h_n(x) - h_n'(x) \qquad \text{for } n \geq 0.$$

Hint: Use the recursions from Theorem 4.29 and Corollary 4.30.

5 Chebyshev Approximation

In this chapter, we study, for a compact domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, the approximation of continuous functions from the linear space

$$\mathscr{C}(\Omega) = \{u : \Omega \longrightarrow \mathbb{R} \,|\, u \text{ continuous}\}$$

with respect to the maximum norm

$$\|u\|_\infty = \max_{x \in \Omega} |u(x)| \qquad \text{for } u \in \mathscr{C}(\Omega).$$

The maximum norm $\|\cdot\|_\infty$ is also referred to as *Chebyshev*[1] *norm*, and so in this chapter, we are concerned with *Chebyshev approximation*, i.e., approximation with respect to $\|\cdot\|_\infty$.

To approximate functions from $\mathscr{C}(\Omega)$, we work with *finite-dimensional* linear subspaces $\mathcal{S} \subset \mathscr{C}(\Omega)$. Under this assumption, there is for any $f \in \mathscr{C}(\Omega)$ a best approximation $s^* \in \mathcal{S}$ to f, see Corollary 3.8. Further in Chapter 3, we analyzed the problem of Chebyshev approximation from a more general viewpoint. Recall that we have made a negative observation: The Chebyshev norm $\|\cdot\|_\infty$ is *not* strictly convex, as shown in Example 3.34. According to Theorem 3.37, however, strictly convex norms guarantee (for convex $\mathcal{S} \subset \mathcal{F}$) the uniqueness of best approximations. Therefore, the problem of approximation with respect to the Chebyshev norm $\|\cdot\|_\infty$ appears to be, at first sight, rather critical.

But we should not be too pessimistic. In this chapter, we will derive suitable conditions on the approximation space $\mathcal{S} \subset \mathscr{C}(\Omega)$, under which we can even guarantee *strong* uniqueness of best approximations. However, according to the *Mairhuber-Curtis theorem*, Theorem 5.25, strong uniqueness can only be achieved for the univariate case. Therefore, the case $d = 1$, where $\Omega = [a, b] \subset \mathbb{R}$ is a compact interval, is of primary importance. In fact, we will study Chebyshev approximation to continuous functions in $\mathscr{C}[a, b]$ by algebraic polynomials from $\mathcal{P}_n$, for $n \in \mathbb{N}_0$, in more detail.

Furthermore, we derive suitable characterizations for best approximations for the particular case $\|\cdot\| = \|\cdot\|_\infty$, where we can rely on our previous results in Chapter 3. This finally leads us to the *Remez algorithm*, an iterative numerical method to compute best approximations with respect to the Chebyshev norm $\|\cdot\|_\infty$. We show linear convergence for the Remez iteration.

[1] Pafnuty Lvovich Chebyshev (1821-1894), Russian mathematician

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_5

5.1 Approaches to Construct Best Approximations

For a compact domain $\Omega \subset \mathbb{R}^d$, $d \geq 1$, we denote by $\mathscr{C}(\Omega)$ the linear space of all continuous functions on Ω. Moreover, we assume throughout this chapter that $\mathcal{S} \subset \mathscr{C}(\Omega)$ is a finite-dimensional linear subspace of $\mathscr{C}(\Omega)$. Under these assumptions, there exists, according to Corollary 3.8, for any $f \in \mathscr{C}(\Omega)$ a best approximation $s^* \in \mathcal{S}$ to f with respect to the Chebyshev norm $\|\cdot\|_\infty$. However, s^* is not necessarily unique, since $\|\cdot\|_\infty$ is not strictly convex.

Now we apply the characterizations for best approximations from Chapter 3 to the special case of the Chebyshev norm $\|\cdot\|_\infty$. We begin with the direct characterizations from Section 3.4, where we had proven the Kolmogorov criterion, Corollary 3.55. We can adapt the Kolmogorov criterion to the Chebyshev norm $\|\cdot\|_\infty$ as follows.

Theorem 5.1. *Let $\mathcal{S} \subset \mathscr{C}(\Omega)$ be a linear subspace of $\mathscr{C}(\Omega)$ and suppose $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$. Then $s^* \in \mathcal{S}$ is a best approximation to f with respect to $\|\cdot\|_\infty$, if and only if*

$$\max_{x \in E_{s^*-f}} s(x) \operatorname{sgn}((s^* - f)(x)) \geq 0 \qquad \text{for all } s \in \mathcal{S}, \tag{5.1}$$

where

$$E_{s^*-f} = \{x \in \Omega : |(s^* - f)(x)| = \|s^* - f\|_\infty\} \subset \Omega$$

denotes the set of `extremal points` *of $s^* - f$ in Ω.*

Proof. By the equivalence of the Kolmogorov criterion, Corollary 3.55, s^* is a best approximation to f with respect to $\|\cdot\| = \|\cdot\|_\infty$, if and only if

$$\|'_+(s^* - f, s - s^*) = \max_{x \in E_{s^*-f}} (s - s^*)(x) \operatorname{sgn}((s^* - f)(x)) \geq 0 \quad \text{for all } s \in \mathcal{S},$$

where we have used the Gâteaux derivative of the norm $\|\cdot\|_\infty$ from Theorem 3.64. By the linearity of $\mathcal{S}$, this condition is equivalent to (5.1). ∎

Given the result of Theorem 5.1, we can immediately solve one simple problem of Chebyshev approximation. To this end, we regard the univariate case, $d = 1$, where $\Omega = [a, b] \subset \mathbb{R}$ for a compact interval. In this case, we wish to approximate continuous functions from $\mathscr{C}[a, b]$ by constants.

Corollary 5.2. *Let $[a, b] \subset \mathbb{R}$ be compact and $f \in \mathscr{C}[a, b]$. Then*

$$c^* = \frac{f_{\min} + f_{\max}}{2} \in \mathcal{P}_0$$

is the unique best approximation to f from $\mathcal{P}_0$ with respect to $\|\cdot\|_\infty$, where

$$f_{\min} = \min_{x \in [a,b]} f(x) \qquad \text{and} \qquad f_{\max} = \max_{x \in [a,b]} f(x).$$

Proof. For $f \in \mathcal{P}_0$, the statement is trivial. Now suppose $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_0$. The continuous function $f \in \mathscr{C}[a,b]$ attains its minimum and maximum on the compact interval $[a,b]$. Therefore, there are $x_{\min}, x_{\max} \in [a,b]$ satisfying

$$f_{\min} = f(x_{\min}) \qquad \text{and} \qquad f_{\max} = f(x_{\max}).$$

Obviously, $x_{\min}, x_{\max}$ lie in the set of extremal points E_{c^*-f}, where

$$\begin{aligned} c^* - f(x_{\min}) &= \eta \\ c^* - f(x_{\max}) &= -\eta \end{aligned}$$

with $\eta = \|c^* - f\|_\infty = (f_{\max} - f_{\min})/2 > 0$. Moreover, in this case we have

$$\max_{x \in E_{c^*-f}} c\,\mathrm{sgn}(c^* - f(x)) = c\,\mathrm{sgn}(c^* - f(x_{\min})) = c \geq 0$$

for $c \geq 0$ on the one hand, and

$$\max_{x \in E_{c^*-f}} c\,\mathrm{sgn}(c^* - f(x)) = c\,\mathrm{sgn}(c^* - f(x_{\max})) = -c > 0$$

for $c < 0$ on the other hand. Altogether, the Kolmogorov criterion from Theorem 5.1,

$$\max_{x \in E_{c^*-f}} c\,\mathrm{sgn}(c^* - f(x)) \geq 0 \qquad \text{for all } c \in \mathcal{P}_0,$$

is satisfied. Therefore, c^* is a best approximation to f from $\mathcal{P}_0$ with respect to $\|\cdot\|_\infty$.

Finally, c^* is the *unique* best approximation to f, since for $c \neq c^*$ we have

$$\begin{aligned} \|c - f\|_\infty \geq |c - f_{\min}| > |c^* - f_{\min}| = \|c^* - f\|_\infty \quad &\text{for } c > c^*; \\ \|c - f\|_\infty \geq |c - f_{\max}| > |c^* - f_{\max}| = \|c^* - f\|_\infty \quad &\text{for } c < c^*. \end{aligned}$$

■

Observe from the above construction of the unique best approximation $c^* \in \mathcal{P}_0$ to $f \in \mathscr{C}[a,b]$ that there are *at least* two different extremal points $x_1, x_2 \in E_{c^*-f}$ which satisfy the *alternation condition*

$$(c^* - f)(x_k) = (-1)^k \sigma \|c^* - f\|_\infty \qquad \text{for } k = 1, 2 \tag{5.2}$$

for some $\sigma \in \{\pm 1\}$. The alternation condition (5.2) is necessary and sufficient for a best approximation from $\mathcal{P}_0$. Moreover, there is no upper bound for the number of alternation points. We can further explain this by the following simple example.

Example 5.3. We approximate $f_m(x) = \cos(mx)$, for $m \in \mathbb{N}$, on the interval $[-\pi, \pi]$ by constants. According to Corollary 5.2, $c^* \equiv 0$ is the unique best approximation to f_m from $\mathcal{P}_0$, for all $m \in \mathbb{N}$. We get $\|c^* - f_m\|_\infty = 1$ for the minimal distance between f_m and $\mathcal{P}_0$, and the error function $c^* - f_m$ has $2m + 1$ alternation points $x_k = \pi k/m$, for $k = -m, \ldots, m$. ♢

For the approximation of $f \in \mathscr{C}[a,b]$ by polynomials from $\mathcal{P}_{n-1}$, there are *at least* $n+1$ alternation points. Moreover, the best approximation $p^* \in \mathcal{P}_{n-1}$ to f is unique. We can prove these two statements by another corollary from the Kolmogorov criterion, Theorem 5.1.

Corollary 5.4. *Let $[a,b] \subset \mathbb{R}$ be compact and $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_{n-1}$, where $n \in \mathbb{N}$. Then there is a unique best approximation $p^* \in \mathcal{P}_{n-1}$ to f from $\mathcal{P}_{n-1}$ with respect to $\|\cdot\|_\infty$. Moreover, there are at least $n+1$ extremal points $\{x_1, \ldots, x_{n+1}\} \subset E_{p^*-f}$ with $a \le x_1 < \ldots < x_{n+1} \le b$, that are satisfying the alternation condition*

$$(p^* - f)(x_k) = (-1)^k \sigma \|p^* - f\|_\infty \qquad \text{for } k = 1, \ldots, n+1 \tag{5.3}$$

for some $\sigma \in \{\pm 1\}$.

Proof. The existence of a best approximation is covered by Corollary 3.8.

Now let $p^* \in \mathcal{P}_{n-1}$ be a best approximation to f. We decompose the set of extremal points E_{p^*-f} into m pairwise disjoint, non-empty, and monotonically increasing subsets, $E_1, \ldots, E_m \subset E_{p^*-f}$, i.e.,

$$a \le x_1 < x_2 < \ldots < x_m \le b \qquad \text{for all } x_k \in E_k \text{ and } k = 1, \ldots, m, \tag{5.4}$$

so that the sign of the error $p^* - f$ is alternating on the sets $E_k \subset E_{p^*-f}$, $1 \le k \le m$, i.e., we have, for some $\sigma \in \{\pm 1\}$,

$$\operatorname{sgn}((p^* - f)(x_k)) = (-1)^k \sigma \quad \text{for all } x_k \in E_k \qquad \text{for } k = 1, \ldots, m. \tag{5.5}$$

We denote the order relation in (5.4) in short by $E_1 < \ldots < E_m$.

Note that there are at least two extremal points in E_{p^*-f}, at which the error function $p^* - f$ has different signs. Indeed, this is because the continuous function $p^* - f$ attains its minimum and maximum on $[a,b]$, so that

$$(p^* - f)(x_{\min}) = -\|p^* - f\|_\infty \qquad \text{and} \qquad (p^* - f)(x_{\max}) = \|p^* - f\|_\infty$$

for $x_{\min}, x_{\max} \in [a,b]$. Otherwise, p^* cannot be a best approximation to f.

Therefore, there are at least two subsets E_k in the above decomposition of E_{p^*-f}, i.e., $m \ge 2$. We now show that we even have $m \ge n+1$ for the number of subsets E_k.

Suppose $m < n+1$, or, $m \le n$. Then there is a set $X^* = \{x_1^*, \ldots, x_{m-1}^*\}$ of size $m-1$, whose points x_k^* are located *between* the points from neighbouring subsets $E_k < E_{k+1}$, respectively, so that

$$x_k < x_k^* < x_{k+1} \qquad \text{for all } x_k \in E_k, x_{k+1} \in E_{k+1} \text{ and } k = 1, \ldots, m-1.$$

In this case, the corresponding knot polynomial

$$\omega_{X^*}(x) = \prod_{k=1}^{m-1} (x - x_k^*) \in \mathcal{P}_{m-1} \subset \mathcal{P}_{n-1}$$

has on the subsets E_k alternating signs, where

$$\operatorname{sgn}(\omega_{X^*}(x_k)) = (-1)^{m-k} \qquad \text{for all } k = 1, \ldots, m.$$

Now for the polynomial $p = p^* + \hat{\sigma}\,\omega_{X^*} \in \mathcal{P}_{n-1}$, with $\hat{\sigma} \in \{\pm 1\}$, we have

$$\operatorname{sgn}((p - p^*)(x_k)(p^* - f)(x_k)) = \hat{\sigma}(-1)^{m-k}(-1)^k\sigma = \hat{\sigma}(-1)^m\sigma$$

for all $x_k \in E_k$ and for $k = 1, \ldots, m$. Letting $\hat{\sigma} = -(-1)^m\sigma$, we have

$$\max_{x_k \in E_{p^*-f}} (p - p^*)(x_k)\operatorname{sgn}((p^* - f)(x_k) < 0.$$

This, however, is in contradiction to the Kolmogorov criterion, Theorem 5.1. Therefore, there are at least $m \geq n + 1$ monotonically increasing non-empty subsets $E_1 < \ldots < E_m$ of E_{p^*-f}, for which the sign of the error function $p^* - f$ is alternating, i.e., we have (5.5) with $m \geq n + 1$, which implies the alternation condition (5.3).

Now we prove the uniqueness of p^* by contradiction.

Suppose there is another best approximation $q^* \in \mathcal{P}_{n-1}$ to f, $p^* \neq q^*$. Then, the convex combination $p = (p^* + q^*)/2 \in \mathcal{P}_{n-1}$ is according to Theorem 3.16 yet another best approximation to f. In this case, there are for p at least $n + 1$ alternation points $x_1 < \ldots < x_{n+1}$, so that

$$(p - f)(x_k) = (-1)^k\sigma\|p - f\|_\infty \qquad \text{for } k = 1, \ldots, n + 1$$

for some $\sigma \in \{\pm 1\}$, where $\{x_1, \ldots, x_{n+1}\} \subset E_{p-f}$.

But the $n + 1$ alternation points $x_1, \ldots, x_{n+1}$ of p are also contained in each of the extremal point sets E_{p^*-f} and E_{q^*-f}. Indeed, this is because in

$$\begin{aligned}\|p - f\|_\infty = |(p - f)(x_k)| &\leq \frac{1}{2}|(p^* - f)(x_k)| + \frac{1}{2}|(q^* - f)(x_k)| \\ &\leq \frac{1}{2}\|p^* - f\|_\infty + \frac{1}{2}\|q^* - f\|_\infty = \|p^* - f\|_\infty = \|q^* - f\|_\infty,\end{aligned}$$

equality holds for $k = 1, \ldots, n + 1$. In particular, we have

$$|(p^* - f)(x_k) + (q^* - f)(x_k)| = |(p^* - f)(x_k)| + |(q^* - f)(x_k)|$$

for all $1 \leq k \leq n + 1$.

Due to the strict convexity of the norm $|\cdot|$ (see Remark 3.27) and by the equivalence statement (d) in Theorem 3.26, the signs of the error functions $p^* - f$ and $q^* - f$ must agree on $\{x_1, \ldots, x_{n+1}\}$, i.e.,

$$\operatorname{sgn}((p^* - f)(x_k)) = \operatorname{sgn}((q^* - f)(x_k)) \qquad \text{for all } k = 1, \ldots, n + 1.$$

Altogether, the values of the polynomials $p^*, q^* \in \mathcal{P}_{n-1}$ coincide on the $n+1$ points $x_1, \ldots, x_{n+1}$, which implies $p^* \equiv q^*$. ■

Now we note another important corollary, which directly follows from our observation in Proposition 3.42 and from Exercise 3.73.

Corollary 5.5. *For $L > 0$ let $f \in \mathscr{C}[-L, L]$. Moreover, let $p^* \in \mathcal{P}_n$, for $n \in \mathbb{N}_0$, be the unique best approximation to f from $\mathcal{P}_n$ with respect to $\|\cdot\|_\infty$. Then the following statements are true.*

(a) If f is even, then its best approximation $p^* \in \mathcal{P}_n$ is even.
(b) If f is odd, then its best approximation $p^* \in \mathcal{P}_n$ is odd.

Proof. The linear space $\mathcal{P}_n$ of algebraic polynomials is *reflection-invariant*, i.e., for $p(x) \in \mathcal{P}_n$, we have $p(-x) \in \mathcal{P}_n$. Moreover, by Corollary 5.4 there exists for any $f \in \mathscr{C}[-L, L]$ a unique best approximation $p^* \in \mathcal{P}_n$ to f from $\mathcal{P}_n$ with respect to $\|\cdot\|_\infty$. Without loss of generality, we assume $L = 1$. By Proposition 3.42 and Exercise 3.73, both statements (a) and (b) hold. ■

For illustration, we apply Corollary 5.5 in the following two examples.

Example 5.6. We approximate $f_m(x) = \sin(mx)$, for $m \in \mathbb{N}$, on $[-\pi, \pi]$ by linear polynomials. The function f_m is odd, for all $m \in \mathbb{N}$, and so is the best approximation $p_m^* \in \mathcal{P}_1$ to f_m odd. Therefore, p_m^* has the form $p_m^*(x) = \alpha_m x$ for a slope $\alpha_m \geq 0$, which is yet to be determined.

Case 1: For $m = 1$, the constant $c \equiv 0$ cannot be a best approximation to $f_1(x) = \sin(x)$, since $c - f_1$ has only *two* alternation points $\pm\pi/2$. By symmetry, we can restrict our following investigations to the interval $[0, \pi]$. The function $p_1^*(x) - f_1(x) = \alpha_1 x - \sin(x)$, with $\alpha_1 > 0$, has two alternation points $\{x^*, \pi\}$ on $[0, \pi]$,

$$(p_1^* - f_1)(x^*) = \alpha_1 x^* - \sin(x^*) = -\eta \qquad \text{and} \qquad (p_1^* - f_1)(\pi) = \alpha_1 \pi = \eta,$$

where $\eta = \|p_1^* - f_1\|_\infty$ is the minimal distance between f_1 and $\mathcal{P}_1$. Moreover, the alternation point x^* satisfies the condition

$$0 = (p_1^* - f_1)'(x^*) = \alpha_1 - \cos(x^*) \qquad \text{which implies } \alpha_1 = \cos(x^*).$$

Therefore, x^* is a solution of the nonlinear equation

$$\cos(x^*)(x^* + \pi) = \sin(x^*),$$

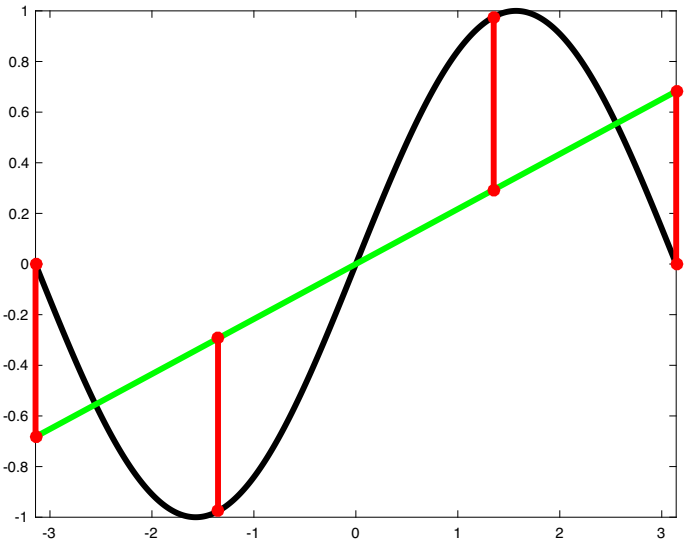
which we can solve numerically, whereby we obtain the alternation point $x^* \approx 1.3518$, the slope $\alpha_1 = \cos(x^*) \approx 0.2172$ and the minimal distance $\eta \approx 0.6825$. Altogether, the best approximation $p_1^*(x) = \alpha_1 x$ with $\{-\pi, -x^*, x^*, \pi\}$ gives *four* alternation points for $p_1^* - f_1$ on $[-\pi, \pi]$, see Figure 5.1 (a).

Case 2: For $m > 1$, $p_m^* \equiv 0$ is the unique best approximation to f_m.

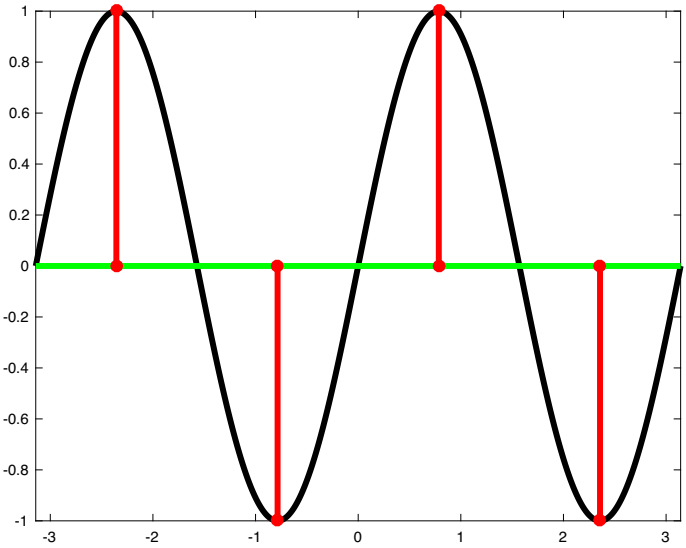
For the minimal distance, we get $\|p_m^* - f_m\|_\infty = 1$ and the error function $p_m^* - f_m$ has $2m$ alternation points

$$x_k = \pm\frac{2k-1}{2m}\pi \qquad \text{for } k = 1, 2, \ldots, m,$$

see Figure 5.1 (b) for the case $m = 2$. ♢



(a) approximation of the function $f_1(x) = \sin(x)$



(b) approximation of the function $f_2(x) = \sin(2x)$

Fig. 5.1. Approximation of the function $f_m(x) = \sin(mx)$ on $[-\pi, \pi]$ by linear polynomials for (a) $m = 1$ and (b) $m = 2$. The best approximation $p_m^* \in \mathcal{P}_1$ to f_m, $m = 1, 2$, is odd. In Example 5.6, we determine the best approximation $p_m^* \in \mathcal{P}_1$ to f_m and the corresponding alternation points for all $m \in \mathbb{N}$.

Regrettably, the characterization of best approximations in Corollary 5.4 is not constructive, since neither do we know the set of extremal points E_{p^*-f} nor do we know the minimal distance $\|p^*-f\|_\infty$ a priori. Otherwise, we could immediately compute the best approximating polynomial $p^* \in \mathcal{P}_{n-1}$ from the interpolation conditions

$$p^*(x_k) = f(x_k) + (-1)^k \eta \qquad \text{where } \eta = \sigma \|p^* - f\|_\infty,$$

for $k = 1, \ldots, n+1$. For further illustration, we discuss the following example, where we can predetermine some of the extremal points.

Example 5.7. We approximate the absolute-value function $f(x) = |x|$ on $[-1, 1]$ by quadratic polynomials. To construct the best approximation $p_2^* \in \mathcal{P}_2$ to $f \in \mathscr{C}[-1, 1]$ we first note the following observations.

- The function f is even, and so p_2^* must be even, by Corollary 5.5.
- By Corollary 5.4 there are *at least* four extremal points, $|E_{p_2^*-f}| \geq 4$.
- The error function $e = p_2^* - f$ on $[0, 1]$ is a quadratic polynomial. Therefore, e has on $(0, 1)$ at most one local extremum $x^* \in (0, 1)$. This local extremum must lie in the set of extremal points $E_{p_2^*-f}$. By symmetry, $-x^* \in (-1, 0)$ is also contained in the set of extremal points $E_{p_2^*-f}$.
- Further extrema of the error function e can only be at the origin or at the boundary points ± 1. Since $|E_{p_2^*-f}| \geq 4$ and due to symmetry, *both* boundary points ± 1 must lie in $E_{p_2^*-f}$.
- To satisfy the alternation condition, the origin must necessarily lie in $E_{p_2^*-f}$. Indeed, for the subset $E = \{-1, -x^*, x^*, 1\} \subset E_{p_2^*-f}$ the four signs of e on E are symmetric, in particular *not* alternating. Therefore, we have $E_{p_2^*-f} = \{-1, -x^*, 0, x^*, 1\}$ for some $x^* \in (0, 1)$.
- By symmetry we can restrict ourselves in the following investigations to the unit interval $[0, 1]$: Since the error function $e = p_2^* - f$ has three extrema $\{0, x^*, 1\}$ in $[0, 1]$, e has two zeros in $(0, 1)$, i.e., the function graphs of f and p_2^* intersect in $(0, 1)$ at two points. Hence, p_2^* is convex, where $p_2^*(0) > 0$.

We can now sketch the function graphs of f and p_2^* (see Figure 5.2).

By our above observations the best approximation p_2^* has the form

$$p_2^*(x) = \eta + \alpha x^2$$

with the minimal distance $\eta = \|p_2^* - f\|_\infty$, for some positive slope $\alpha > 0$. Moreover, $e = p_2^* - f$ has on the set of extremal points $E_{p_2^*-f} = \{-1, -x^*, 0, x^*, 1\}$ alternating signs $\varepsilon = (1, -1, 1, -1, 1)$. We compute α by the alternation condition at $x = 1$,

$$(p_2^* - f)(1) = \eta + \alpha - 1 = \eta,$$

and so we obtain $\alpha = 1$, so that $p_2^*(x) = \eta + x^2$. The local minimum x^* of the error function $e = p_2^* - f$ satisfies the necessary condition

$$e'(x^*) = 2x^* - 1 = 0,$$

whereby $x^* = 1/2$, so that $E_{p_2^*-f} = \{-1, -1/2, 0, 1/2, 1\}$. Finally, at $x^* = 1/2$ we have the alternation condition

$$(p_2^* - f)(1/2) = \eta + 1/4 - 1/2 = -\eta,$$

holds, whereby $\eta = 1/8$. Hence, the quadratic polynomial $p_2^*(x) = 1/8 + x^2$ is the unique best approximation to f from $\mathcal{P}_2$ with respect to $\|\cdot\|_\infty$. ♢

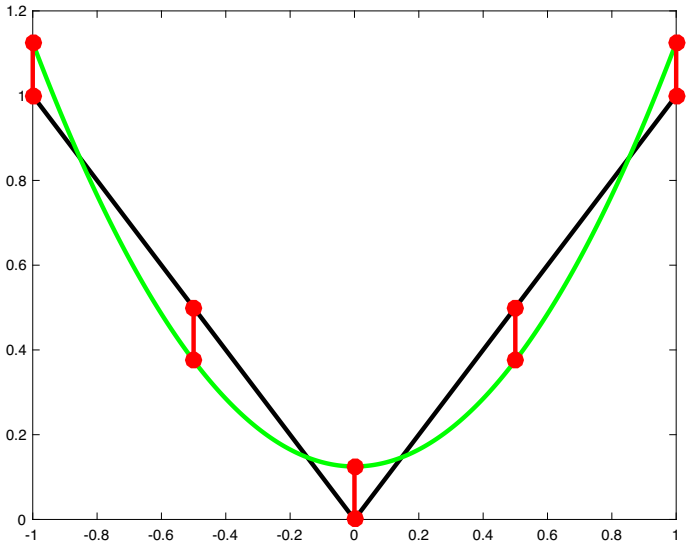


Fig. 5.2. Approximation of the function $f(x) = |x|$ on $[-1, 1]$ by quadratic polynomials. The best approximation $p_2^* \in \mathcal{P}_2$ to f is even and convex. The set of extremal points $E_{p_2^*-f} = \{-1, -x^*, 0, x^*, 1\}$ has five alternation points.

A more constructive account for the computation of best approximations relies on the dual characterizations from Section 3.3. To this end, we recall the necessary and sufficient condition from Theorem 3.48. According to Theorem 3.48, $s^* \in \mathcal{S} \subset \mathscr{C}(\Omega)$ is a best approximation to $f \in \mathscr{C}(\Omega)$ with respect to $\|\cdot\|_\infty$, if and only if there is a dual functional $\varphi \in (\mathscr{C}(\Omega))'$ satisfying

(a) $\|\varphi\|_\infty = 1$.
(b) $\varphi(s^* - f) = \|s^* - f\|_\infty$.
(c) $\varphi(s - s^*) \geq 0$ for all $s \in \mathcal{S}$.

To construct such a characterizing dual functional, we use the assumption

$$\varphi(u) = \sum_{k=1}^{m} \lambda_k \varepsilon_k u(x_k) \qquad \text{for } u \in \mathscr{C}(\Omega) \tag{5.6}$$

with coefficient vector $\lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m$, lying at the boundary

$$\Lambda_m = \left\{ (\lambda_1, \ldots, \lambda_m)^T \in \mathbb{R}^m \;\middle|\; \lambda_k \in [0,1], 1 \le k \le m, \sum_{k=1}^{m} \lambda_k = 1 \right\} \tag{5.7}$$

of the standard simplex $\Delta_m \subset \mathbb{R}^m$ from (2.38). Moreover,

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^T \in \{\pm 1\}^m$$

denotes a sign vector and $X = \{x_1, \ldots, x_m\} \subset \Omega$ is a point set.

Assuming (5.6) condition (a) is already satisfied, since

$$|\varphi(u)| = \left| \sum_{k=1}^{m} \lambda_k \varepsilon_k u(x_k) \right| \le \|u\|_\infty \qquad \text{for all } u \in \mathscr{C}(\Omega) \tag{5.8}$$

and so $\|\varphi\|_\infty \le 1$. Moreover, for any $u \in \mathscr{C}(\Omega)$ satisfying $\|u\|_\infty = 1$ and $u(x_k) = \varepsilon_k$, for all $1 \le k \le m$, we have equality in (5.8), so that φ has norm length one by $\|\varphi\|_\infty = \|\lambda\|_1 = 1$.

To satisfy condition (b), we choose $X = E_{s^*-f}$, i.e., $E_{s^*-f} = \{x_1, \ldots, x_m\}$. In this case, we get, in combination with $\varepsilon_k = \operatorname{sgn}((s^* - f)(x_k))$, the identity

$$\varphi(s^* - f) = \sum_{k=1}^{m} \lambda_k \varepsilon_k (s^* - f)(x_k) = \sum_{k=1}^{m} \lambda_k |(s^* - f)(x_k)| = \|s^* - f\|_\infty.$$

But the set of extremal points E_{s^*-f} is unknown a priori. Moreover, it remains to satisfy condition (c).

From now on we study the construction of coefficients $\lambda \in \Lambda_m$, signs $\varepsilon \in \{\pm 1\}^m$ and points $X = \{x_1, \ldots, x_m\}$ in more detail. In order to do so, we need some technical preparations. We begin with the representation of convex hulls.

Definition 5.8. *Let $\mathcal{F}$ be a linear space and $\mathcal{M} \subset \mathcal{F}$. Then the* `convex hull` $\operatorname{conv}(\mathcal{M})$ *of $\mathcal{M}$ is the smallest convex set in $\mathcal{F}$ containing $\mathcal{M}$, i.e.,*

$$\operatorname{conv}(\mathcal{M}) = \bigcap_{\substack{\mathcal{M} \subset \mathcal{K} \subset \mathcal{F} \\ \mathcal{K} \text{ convex}}} \mathcal{K}.$$

○

The following representation for $\operatorname{conv}(\mathcal{M})$ is much more useful in practice.

Theorem 5.9. *Let $\mathcal{F}$ be a linear space and $\mathcal{M} \subset \mathcal{F}$. Then we have*

$$\operatorname{conv}(\mathcal{M}) = \left\{ \sum_{j=1}^{m} \lambda_j x_j \,\middle|\, x_j \in \mathcal{M} \text{ and } \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m \text{ for } m \in \mathbb{N} \right\}.$$

Proof. Let us consider the set

$$\mathcal{K} = \left\{ \sum_{j=1}^{m} \lambda_j x_j \,\middle|\, x_j \in \mathcal{M} \text{ and } \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m \text{ for } m \in \mathbb{N} \right\}. \quad (5.9)$$

We now show the following properties for $\mathcal{K}$.

(a) $\mathcal{K}$ is convex.
(b) $\mathcal{M} \subset \mathcal{K}$.
(c) $\operatorname{conv}(\mathcal{M}) \subset \mathcal{K}$.

(a): For $x, y \in \mathcal{K}$ we have the representations

$$x = \sum_{j=1}^{m} \lambda_j x_j \quad \text{with } \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m, \{x_1, \ldots, x_m\} \subset \mathcal{M}, m \in \mathbb{N}$$

$$y = \sum_{k=1}^{n} \mu_k y_k \quad \text{with } \mu = (\mu_1, \ldots, \mu_n)^T \in \Lambda_n, \{y_1, \ldots, y_n\} \subset \mathcal{M}, n \in \mathbb{N}.$$

Note that any convex combination $\alpha x + (1-\alpha)y$, $\alpha \in [0,1]$, can be written as a convex combination of the points $x_1, \ldots, x_m, y_1, \ldots, y_n$,

$$\alpha x + (1-\alpha)y = \alpha \sum_{j=1}^{m} \lambda_j x_j + (1-\alpha) \sum_{k=1}^{n} \mu_k y_k = \sum_{j=1}^{m} \alpha \lambda_j x_j + \sum_{k=1}^{n} (1-\alpha)\mu_k y_k,$$

so that $\alpha x + (1-\alpha)y \in \mathcal{K}$ for all $\alpha \in [0,1]$.

(b): Any point $x \in \mathcal{M}$ lies in $\mathcal{K}$, by $m = 1$, $\lambda_1 = 1$ and $x_1 = x$ in (5.9). Therefore, the inclusion $\mathcal{M} \subset \mathcal{K}$ holds.

(c): By (a) and (b) $\mathcal{K}$ is a convex set containing $\mathcal{M}$. From the minimality of $\operatorname{conv}(\mathcal{M})$ we can conclude $\operatorname{conv}(\mathcal{M}) \subset \mathcal{K}$.

We now show the inclusion $\mathcal{K} \subset \operatorname{conv}(\mathcal{M})$. To this end, we first note that any convex $\mathcal{L}$ containing $\mathcal{M}$, i.e., $\mathcal{M} \subset \mathcal{L}$, is necessarily a superset of $\mathcal{K}$, i.e., $\mathcal{K} \subset \mathcal{L}$. Indeed, this is because $\mathcal{L}$ contains all finite convex combinations of points from $\mathcal{M}$. This immediately implies

$$\mathcal{K} \subset \bigcap_{\substack{\mathcal{M} \subset \mathcal{L} \\ \mathcal{L} \text{ convex}}} \mathcal{L} = \operatorname{conv}(\mathcal{M}).$$

Altogether, we have $\mathcal{K} = \operatorname{conv}(\mathcal{M})$. ∎

By the characterization in Theorem 5.9, we can identify the convex hull $\mathrm{conv}(\mathcal{M})$, for any set $\mathcal{M} \subset \mathcal{F}$, as the set of all *finite* convex combinations of points from $\mathcal{M}$. For *finite-dimensional* linear spaces $\mathcal{F}$, the length of the convex combinations can uniformly be bounded above according to the *Carathéodory*[2] *theorem.*

Theorem 5.10. (Carathéodory).
Let $\mathcal{F}$ be a linear space of finite dimension $\dim(\mathcal{F}) = n < \infty$. Moreover, suppose $\mathcal{M} \subset \mathcal{F}$. Then we have the representation

$$\mathrm{conv}(\mathcal{M}) = \left\{ \sum_{j=1}^{m} \lambda_j x_j \,\middle|\, x_j \in \mathcal{M}, \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m \text{ for } m \le n+1 \right\}.$$

Proof. For $x \in \mathrm{conv}(\mathcal{M})$ we consider a representation of the form

$$x = \sum_{j=1}^{m} \lambda_j x_j \qquad \text{with } \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m \text{ and } x_1, \ldots, x_m \in \mathcal{M}$$

but with *minimal* $m \in \mathbb{N}$. Then $\lambda_j > 0$, i.e., $\lambda_j \in (0, 1]$ for all $1 \le j \le m$. From the assumed representation we get

$$\sum_{j=1}^{m} \lambda_j (x - x_j) = 0\,,$$

i.e., the elements $x - x_j \in \mathcal{F}$, $1 \le j \le m$, are linearly dependent in $\mathcal{F}$.

Now suppose $m > n + 1$, or, $m - 1 > n$. Then there are $\alpha_2, \ldots, \alpha_m \in \mathbb{R}$, that are not all vanishing, with

$$\sum_{j=1}^{m} \alpha_j (x - x_j) = 0,$$

where we let $\alpha_1 = 0$. This gives the representation

$$0 = \sum_{j=1}^{m} (\lambda_j + t\alpha_j)(x - x_j) = \sum_{j=1}^{m} \mu_j(t)(x - x_j) \qquad \text{for all } t \in \mathbb{R},$$

with $\mu_j(t) = (\lambda_j + t\alpha_j)$ and therefore $\mu_j(0) = \lambda_j > 0$.

Now we choose one $t^* \in \mathbb{R}$ satisfying

$$\mu_j(t^*) = \lambda_j + t^*\alpha_j \ge 0 \qquad \text{for all } j = 1, \ldots, m,$$

and $\mu_k(t^*) = 0$ for some $k \in \{1, \ldots, m\}$. By

[2] Constantin Carathéodory (1873-1950), Greek mathematician

$$\rho_j = \frac{\mu_j(t^*)}{\sum_{k=1}^{m} \mu_k(t^*)} \geq 0 \qquad \text{for } j = 1, \ldots, m$$

we have

$$\sum_{j=1}^{m} \rho_j = 1$$

and

$$\sum_{j=1}^{m} \rho_j (x - x_j) = 0 \qquad \Longleftrightarrow \qquad x = \sum_{j=1}^{m} \rho_j x_j.$$

Note that $\rho_k = 0$. But this is in contradiction to the minimality of m. ■

The Carathéodory theorem implies the following important result.

Corollary 5.11. *Let $\mathcal{F}$ be a normed linear space of finite dimension $n < \infty$. Suppose $\mathcal{M} \subset \mathcal{F}$ is a compact subset of $\mathcal{F}$. Then* conv$(\mathcal{M})$ *is compact.*

Proof. We regard on the compact set $\mathcal{L}_{n+1} = \Lambda_{n+1} \times \mathcal{M}^{n+1}$ the continuous mapping $\varphi : \mathcal{L}_{n+1} \longrightarrow \mathcal{F}$, defined as

$$\varphi(\lambda, X) = \sum_{j=1}^{n+1} \lambda_j x_j$$

for $\lambda = (\lambda_1, \ldots, \lambda_{n+1})^T \in \Lambda_{n+1}$ and $X = (x_1, \ldots, x_{n+1}) \in \mathcal{M}^{n+1}$. According to the Carathéodory theorem, Theorem 5.10, we have $\varphi(\mathcal{L}_{n+1}) = \text{conv}(\mathcal{M})$. Therefore, conv$(\mathcal{M})$ is also compact, since conv$(\mathcal{M})$ is the image of the compact set $\mathcal{L}_{n+1}$ under the continuous mapping $\varphi : \mathcal{L}_{n+1} \longrightarrow \mathcal{F}$. ■

From Corollary 5.11, we gain the following separation theorem.

Theorem 5.12. *Let $\mathcal{M} \subset \mathbb{R}^d$ be compact. Then the following two statements are equivalent.*

(a) There is no $\beta \in \mathbb{R}^d \setminus \{0\}$ satisfying $\beta^T x > 0$ for all $x \in \mathcal{M}$.
(b) $0 \in \text{conv}(\mathcal{M})$.

Proof. $(b) \Rightarrow (a)$**:** Let $0 \in \text{conv}(\mathcal{M})$. Then we have the representation

$$0 = \sum_{j=1}^{m} \lambda_j x_j \qquad \text{with } \lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m \text{ and } x_1, \ldots, x_m \in \mathcal{M}.$$

Suppose there is one $\beta \in \mathbb{R}^d \setminus \{0\}$ satisfying $\beta^T x > 0$ for all $x \in \mathcal{M}$. Then we immediately get a contradiction by

$$\beta^T 0 = 0 = \sum_{j=1}^{m} \lambda_j \beta^T x_j > 0.$$

$(a) \Rightarrow (b)$**:** Suppose statement (a) holds. Further suppose that $0 \notin \operatorname{conv}(\mathcal{M})$.

Since $\operatorname{conv}(\mathcal{M})$ is compact, by Corollary 5.11, there is one $\beta_* \in \operatorname{conv}(\mathcal{M})$, $\beta_* \neq 0$, of minimal Euclidean norm in $\operatorname{conv}(\mathcal{M})$. This minimum β_*, viewed as a best approximation from $\operatorname{conv}(\mathcal{M})$ to the origin with respect to $\|\cdot\|_2$, is characterized by

$$(\beta_* - 0, x - \beta_*) \geq 0 \qquad \text{for all } x \in \operatorname{conv}(\mathcal{M})$$

according to the Kolmogorov theorem, Corollary 3.55, in combination with the Gâteaux derivative for Euclidean norms in Theorem 3.62. But this condition is equivalent to

$$\beta_*^T x = (\beta_*, x) \geq (\beta_*, \beta_*) = \|\beta_*\|_2^2 > 0 \qquad \text{for all } x \in \operatorname{conv}(\mathcal{M}).$$

But this is in contradiction to our assumption in (a). ■

Remark 5.13. The equivalence statement (a) in Theorem 5.12 says that the Euclidean space $\mathbb{R}^d$ cannot be split by a separating hyperplane through the origin into two half-spaces, such that $\mathcal{M}$ is entirely contained in one of the two half-spaces. □

5.2 Strongly Unique Best Approximations

Now we wish to further develop the characterizations for best approximations from Sections 3.3 and 3.4 for the special case of the Chebyshev norm $\|\cdot\|_\infty$. In the following discussion, $\{s_1, \ldots, s_n\} \subset \mathcal{S}$, for $n \in \mathbb{N}$, denotes a basis for the finite-dimensional approximation space $\mathcal{S} \subset \mathscr{C}(\Omega)$. To characterize a best approximation $s^* \in \mathcal{S}$ to some $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$, we work with the compact point set

$$\mathcal{M}_{s^*-f} = \left\{ (s^* - f)(x)(s_1(x), \ldots, s_n(x))^T \,\middle|\, x \in E_{s^*-f} \right\} \subset \mathbb{R}^n,$$

where we can immediately draw the following conclusion from Theorem 5.12.

Corollary 5.14. *For $s^* \in \mathcal{S}$ the following statements are equivalent.*

(a) s^* is a best approximation to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$.
(b) $0 \in \operatorname{conv}(\mathcal{M}_{s^*-f})$.

Proof. In this proof, we use the notation

$$s_\beta = \sum_{j=1}^{n} \beta_j s_j \in \mathcal{S} \qquad \text{for } \beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n. \tag{5.10}$$

$(b) \Rightarrow (a)$**:** Let $0 \in \operatorname{conv}(\mathcal{M}_{s^*-f})$. Suppose $s^* \in \mathcal{S}$ is *not* a best approximation to f. Then there is one $\beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n \setminus \{0\}$ satisfying

$$\|s^* - f - s_\beta\|_\infty < \|s^* - f\|_\infty$$

In this case, we have

$$|(s^* - f)(x) - s_\beta(x)|^2 < |(s^* - f)(x)|^2 \qquad \text{for all } x \in E_{s^*-f}.$$

But this is equivalent to

$$|(s^* - f)(x)|^2 - 2(s^* - f)(x)s_\beta(x) + s_\beta^2(x) < |(s^* - f)(x)|^2$$

for all $x \in E_{s^*-f}$, so that

$$(s^* - f)(x)s_\beta(x) > \frac{1}{2}s_\beta^2(x) \geq 0 \qquad \text{for all } x \in E_{s^*-f},$$

i.e.,

$$\beta^T(s^* - f)(x)(s_1(x), \ldots, s_n(x))^T > 0 \qquad \text{for all } x \in E_{s^*-f}.$$

By the equivalence statements in Theorem 5.12, we see that the origin is in this case *not* contained in the convex hull $\operatorname{conv}(\mathcal{M}_{s^*-f})$. But this is in contradiction to statement (b).

$(a) \Rightarrow (b)$**:** Let s^* be a best approximation to f. Suppose $0 \notin \operatorname{conv}(\mathcal{M}_{s^*-f})$. Due to Theorem 5.12, there is one $\beta = (\beta_1, \ldots, \beta_n)^T \in \mathbb{R}^n \setminus \{0\}$ satisfying $\beta^T u > 0$, or, $-\beta^T u < 0$, for all $u \in \mathcal{M}_{s^*-f}$. But this is equivalent to

$$(s^* - f)(x)\, s_{-\beta}(x) < 0 \qquad \text{for all } x \in E_{s^*-f},$$

i.e., $s^* - f$ and $s_{-\beta}$ have opposite signs on E_{s^*-f}, whereby

$$\operatorname{sgn}((s^* - f)(x))\, s_{-\beta}(x) < 0 \qquad \text{for all } x \in E_{s^*-f}.$$

In particular (by using the compactness of E_{s^*-f}), we have

$$\max_{x \in E_{s^*-f}} \operatorname{sgn}((s^* - f)(x))\, s_{-\beta}(x) < 0.$$

But this is, due to the Kolmogorov criterion, Theorem 5.1, in contradiction to the optimality of s^* in (a), ■

Corollary 5.14 yields an important result concerning the characterization of best approximations.

Corollary 5.15. *For $s^* \in \mathcal{S}$ the following statements are equivalent.*

(a) s^* is a best approximation to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$.
(b) There are $m \leq n + 1$

- *pairwise distinct extremal points $x_1, \ldots, x_m \in E_{s^*-f}$*
- *signs $\varepsilon_j = \operatorname{sgn}((s^* - f)(x_j))$, for $j = 1, \ldots, m$,*
- *coefficients $\lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m$ with $\lambda_j > 0$ for all $1 \leq j \leq m$, satisfying*

$$\varphi(s) := \sum_{j=1}^{m} \lambda_j \varepsilon_j s(x_j) = 0 \qquad \text{for all } s \in \mathcal{S}. \tag{5.11}$$

Proof. $(a) \Rightarrow (b)$**:** Let s^* be a best approximation to f.

Then we have $0 \in \operatorname{conv}(\mathcal{M}_{s^*-f})$ by Corollary 5.14. According to the Carathéodory theorem, Theorem 5.10, there are $m \le n+1$ extremal points $x_1, \ldots, x_m \in E_{s^*-f}$ and coefficients $\lambda = (\lambda_1, \ldots, \lambda_m)^T \in \Lambda_m$ satisfying

$$0 = \sum_{j=1}^{m} \lambda_j ((s^* - f)(x_j)) s_k(x_j) = \sum_{j=1}^{m} \lambda_j \varepsilon_j \|s^* - f\|_\infty s_k(x_j) = \sum_{j=1}^{m} \lambda_j \varepsilon_j s_k(x_j)$$

for all basis elements $s_k \in \mathcal{S}$, $k = 1, \ldots, n$.

$(b) \Rightarrow (a)$**:** Under the assumption in (b), we have $0 \in \operatorname{conv}(\mathcal{M}_{s^*-f})$, whereby s^* is a best approximation to f, due to Corollary 5.14. ■

Remark 5.16. In statement (b) of Corollary 5.15 the alternation condition

$$\varepsilon_j \cdot \varepsilon_{j+1} = -1 \qquad \text{for } j = 1, \ldots, m-1$$

is *not* necessarily satisfied. In Corollary 5.4, we considered the special case of polynomial approximation by $\mathcal{S} = \mathcal{P}_{n-1} \subset \mathscr{C}[a,b]$. In that case, the alternation condition (5.3) is satisfied with *at least* $n+1$ extremal points. But in Corollary 5.15 only *at most* $n+1$ extremal points are allowed. □

In the following discussion, we will see how the characterizations of Corollary 5.4 and Corollary 5.15 can be combined. To this end, the following result is of central importance, whereby we can even prove *strong* uniqueness for best approximations.

Theorem 5.17. *For $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$, let $s^* \in \mathcal{S}$ be a best approximation to f. Moreover, suppose that $\varphi : \mathscr{C}(\Omega) \longrightarrow \mathbb{R}$ is a linear functional of the form*

$$\varphi(u) = \sum_{k=1}^{m} \lambda_k \varepsilon_k u(x_k) \qquad \text{for } u \in \mathscr{C}(\Omega) \tag{5.12}$$

satisfying the dual characterization (5.11) of Corollary 5.15 for a point set $X = \{x_1, \ldots, x_m\} \subset E_{s^*-f}$, where $2 \le m \le n+1$. Then, we have for any $s \in \mathcal{S}$ the estimates

$$\|s - f\|_\infty \ge \|s - f\|_{\infty,X} \ge \|s^* - f\|_\infty + \frac{\lambda_{\min}}{1 - \lambda_{\min}} \|s^* - s\|_{\infty,X}, \tag{5.13}$$

where $\lambda_{\min} := \min_{1 \le j \le m} \lambda_j > 0$.

Proof. Suppose $s \in \mathcal{S}$. Then, the first estimate in (5.13) is trivial. To show the second estimate in (5.13), we use the ingredients $\varepsilon \in \{\pm 1\}^m$, $\lambda \in \Lambda_m$, and $X \subset E_{s^*-f}$ for the functional φ in (5.12) from the dual characterization of Corollary 5.15. Note that we have

$$\|s - f\|_{\infty,X} \ge \varepsilon_j (s - f)(x_j) = \varepsilon_j (s - s^*)(x_j) + \varepsilon_j (s^* - f)(x_j)$$

and, moreover, $\varepsilon_j(s^* - f)(x_j) = \|s^* - f\|_\infty$, for all $j = 1, \ldots, m$, so that

$$\|s - f\|_{\infty,X} \geq \|s^* - f\|_\infty + \varepsilon_j(s - s^*)(x_j) \qquad \text{for all } 1 \leq j \leq m. \quad (5.14)$$

Since $m \geq 2$, we have $\lambda_{\min} \in (0, 1/2]$ and so $\lambda_{\min}/(1 - \lambda_{\min}) \in (0, 1]$.

Now let $x_{j^*} \in X$ be a point satisfying $|(s - s^*)(x_{j^*})| = \|s - s^*\|_{\infty,X}$. If $\varepsilon_j(s - s^*)(x_{j^*}) = \|s - s^*\|_{\infty,X}$, then the second estimate in (5.13) is satisfied, with $\lambda_{\min}/(1-\lambda_{\min}) \leq 1$. Otherwise, we have $\varepsilon_j(s-s^*)(x_{j^*}) = -\|s-s^*\|_{\infty,X}$, whereby with $\varphi(s - s^*) = 0$ the estimate

$$\lambda_{j^*}\|s-s^*\|_{\infty,X} = \sum_{\substack{k=1 \\ k\neq j^*}}^{m} \lambda_k \varepsilon_k(s-s^*)(x_k) \leq (1-\lambda_{j^*}) \max_{k\neq j^*} \varepsilon_k(s-s^*)(x_k) \quad (5.15)$$

follows. Now then, for $k^* \in \{1, \ldots, m\} \setminus \{j^*\}$ satisfying

$$\varepsilon_{k^*}(s - s^*)(x_{k^*}) = \max_{k\neq j^*} \varepsilon_k(s - s^*)(x_k)$$

we find, due to (5.15), the estimate

$$\frac{\lambda_{\min}}{1 - \lambda_{\min}}\|s^* - s\|_{\infty,X} \leq \frac{\lambda_{j^*}}{1 - \lambda_{j^*}}\|s^* - s\|_{\infty,X} \leq \varepsilon_{k^*}(s - s^*)(x_{k^*}),$$

which implies, in combination with (5.14), the second estimate in (5.13). ■

Note that for any best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$, the estimates in (5.13) yield the inequality

$$\|s - f\|_\infty - \|s^* - f\|_\infty \geq \frac{\lambda_{\min}}{1 - \lambda_{\min}}\|s - s^*\|_{\infty,X} \qquad \text{for all } s \in \mathcal{S}. \quad (5.16)$$

Given this result, we can further analyze the question for the (strong) uniqueness of best approximations to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$. To this end, we first take note of the following simple observation.

Remark 5.18. Let $s^* \in \mathcal{S}$ be a best approximation to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$. Then, for any other best approximation $s^{**} \in \mathcal{S}$ to $f \in \mathscr{C}(\Omega)$ we have

$$0 = \|s^{**} - f\|_\infty - \|s^* - f\|_\infty \geq \frac{\lambda_{\min}}{1 - \lambda_{\min}}\|s^{**} - s^*\|_{\infty,X}$$

by (5.16), and this implies, for $\lambda_{\min} \in (0, 1)$, the identity

$$\|s^{**} - s^*\|_{\infty,X} = 0.$$

In conclusion, all best approximations to f must coincide on X. Now if $\|\cdot\|_{\infty,X}$ is a *norm* on $\mathcal{S}$, then s^* will be the *unique* best approximation to f. □

In the following Section 5.3, we develop suitable conditions on $\mathcal{S} \subset \mathscr{C}(\Omega)$, under which we can guarantee the uniqueness of best approximations. In our developments the definiteness of $\|\cdot\|_{\infty,X}$ plays an important role. As we can show already now, the definiteness of $\|\cdot\|_{\infty,X}$ guarantees the *strong* uniqueness of a best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$.

Theorem 5.19. *Under the assumptions of Theorem 5.17, let $\|\cdot\|_{\infty,X}$ be a norm on $\mathcal{S}$. Then there exists for any $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$ a strongly unique best approximation $s^* \in \mathcal{S}$ to f.*

Proof. The approximation space $\mathcal{S} \subset \mathscr{C}(\Omega)$ is finite-dimensional. Therefore, there exists for any $f \in \mathscr{C}(\Omega)$ a best approximation $s^* \in \mathcal{S}$ to f, according to Corollary 3.8. Moreover, all norms on $\mathcal{S}$ are equivalent. In particular, the two norms $\|\cdot\|_\infty$ and $\|\cdot\|_{\infty,X}$ are on $\mathcal{S}$ equivalent, so that there is a constant $\beta > 0$ satisfying

$$\|s\|_{\infty,X} \geq \beta \|s\|_\infty \qquad \text{for all } s \in \mathcal{S}. \tag{5.17}$$

By (5.13), the best approximation s^* to f is strongly unique, since

$$\|s-f\|_\infty - \|s^*-f\|_\infty \geq \frac{\lambda_{\min}}{1-\lambda_{\min}} \|s-s^*\|_{\infty,X} \geq \alpha \|s-s^*\|_\infty \qquad \text{for all } s \in \mathcal{S},$$

where $\alpha = \beta\lambda_{\min}/(1-\lambda_{\min}) > 0$. ∎

Before we continue our analysis, we first discuss two examples.

Example 5.20. Let $\mathcal{F} = \mathscr{C}[-1,1]$, $\mathcal{S} = \mathcal{P}_1 \subset \mathcal{F}$ and $f(x) = x^2$. Then, $c^* \equiv 1/2$ is according to Corollary 5.2 the unique best approximation to f from $\mathcal{P}_0$. Since f is even, the unique best approximation $p_1^* \in \mathcal{P}_1$ to f from $\mathcal{P}_1$ is also even, due to Corollary 5.5. In this case, p_1^* is necessarily constant, and so c^* is also the unique best approximation to f from $\mathcal{P}_1$. Moreover, the error function $c^* - f$ has on the interval $[-1,1]$ exactly three extremal points $X = \{x_1, x_2, x_3\} = \{-1, 0, 1\}$, where the alternation conditions are satisfied,

$$c^* - f(x_j) = (-1)^j \|c^* - f\|_\infty = (-1)^j \cdot \frac{1}{2} \qquad \text{for } j = 1, 2, 3.$$

For $\lambda_1 = 1/4$, $\lambda_2 = 1/2$, $\lambda_3 = 1/4$ and $\varepsilon_j = (-1)^j$, for $j = 1, 2, 3$, we have

$$\sum_{j=1}^{3} \lambda_j \varepsilon_j p(x_j) = 0 \quad \text{for all } p \in \mathcal{P}_1.$$

Moreover, $\|\cdot\|_{\infty,X}$ is a norm on $\mathcal{P}_1$. According to Theorem 5.19, the constant c^* is the strongly unique best approximation to f from $\mathcal{P}_1$. By $\lambda_{\min} = \min_{1 \leq j \leq 3} \lambda_j = 1/4$ we get, like in the proof of Theorem 5.19 (with $\beta = 1$), the estimate

$$\|p-f\|_\infty - \|c^*-f\|_\infty \geq \frac{\lambda_{\min}}{1-\lambda_{\min}} \|p-c^*\|_{\infty,X} = \frac{1}{3}\|p-c^*\|_\infty \quad \text{for all } p \in \mathcal{P}_1$$

for the strong uniqueness of c^* with the constant $\alpha = 1/3$. ♢

For further illustration, we make the following link to Example 5.7.

Example 5.21. Let $\mathcal{F} = \mathscr{C}[-1,1]$, $\mathcal{S} = \mathcal{P}_2 \subset \mathcal{F}$ and $f(x) = |x|$. From Example 5.7 the function $p_2^*(x) = 1/8 + x^2$ is the unique best approximation to f from $\mathcal{P}_2$ with extremal point set $E_{p_2^*-f} = \{0, \pm 1/2, \pm 1\}$.

For the dual characterization of the best approximation $p_2^* \in \mathcal{P}_2$ to f, we seek, according to Corollary 5.15, a set $X \subset E_{p_2^*-f}$ of extremal points, where $2 \le m = |X| \le \dim(\mathcal{P}_2) + 1 = 4$, signs $\varepsilon_j = \operatorname{sgn}((p_2^* - f)(x_j))$, $1 \le j \le m$, and coefficients $\lambda = (\lambda_1, \ldots, \lambda_m) \in \Lambda_m$ satisfying

$$\sum_{j=1}^{m} \lambda_j \varepsilon_j p(x_j) = 0 \qquad \text{for all } p \in \mathcal{P}_2. \tag{5.18}$$

This results in $\dim(\mathcal{P}_2) = 3$ linear equations. Together with

$$\lambda_1 + \ldots + \lambda_m = 1 \tag{5.19}$$

we get a total number of *four* linear equation conditions for $\lambda \in \Lambda_m$. Therefore, we let $m = 4$ and, moreover, we take $X = \{-1/2, 0, 1/2, 1\} \subset E_{p_2^*-f}$ with signs $\varepsilon = (-1, 1, -1, 1)$. In this way, we reformulate (5.18) as follows.

$$-\lambda_1 p(-1/2) + \lambda_2 p(0) - \lambda_3 p(1/2) + \lambda_4 p(1) = 0 \qquad \text{for all } p \in \mathcal{P}_2. \tag{5.20}$$

We pose the conditions from (5.20) to the three elements of the monomial basis $\{1, x, x^2\}$ of $\mathcal{P}_2$. For $p \equiv 1$ we get $-\lambda_1 + \lambda_2 - \lambda_3 + \lambda_4 = 0$, whereby from (5.19) we get

$$\lambda_2 + \lambda_4 = 1/2 \qquad \text{and} \qquad \lambda_1 + \lambda_3 = 1/2. \tag{5.21}$$

For $p(x) = x$ and $p(x) = x^2$, we get by (5.20) the conditions

$$\lambda_4 = \frac{1}{2}(\lambda_3 - \lambda_1) \qquad \text{and} \qquad \lambda_4 = \frac{1}{4}(\lambda_1 + \lambda_3). \tag{5.22}$$

Then, (5.21) implies $\lambda_4 = 1/8$ and moreover $\lambda_2 = 3/8$. From (5.21) and (5.22) we finally compute $\lambda_3 = 3/8$ and $\lambda_1 = 1/8$. Therefore,

$$\lambda_{\min} = \frac{1}{8} \qquad \text{and} \qquad \frac{\lambda_{\min}}{1 - \lambda_{\min}} = \frac{1}{7}.$$

The characterization (5.13) in Theorem 5.17 implies the estimate

$$\|p - f\|_\infty - \|p_2^* - f\|_\infty \ge \frac{1}{7}\|p - p_2^*\|_{\infty,X} \qquad \text{for all } p \in \mathcal{P}_2. \tag{5.23}$$

Next, we show the strong uniqueness of p_2^*, where we use Theorem 5.19. To this end, note that $\|\cdot\|_{\infty,X}$ is a norm on $\mathcal{P}_2$. Therefore, it remains to determine an equivalence constant $\beta > 0$, like in (5.17), satisfying

$$\|p\|_{\infty,X} \geq \beta \|p\|_\infty \qquad \text{for all } p \in \mathcal{P}_2. \tag{5.24}$$

We choose for $p \in \mathcal{P}_2$ the monomial representation $p(x) = a_0 + a_1 x + a_2 x^2$. By evaluation of p on the point set $X = \{-1/2, 0, 1/2, 1\}$, we get

$$a_0 = p(0), \quad a_1 = p(1/2) - p(-1/2), \quad a_2 = p(1) - p(0) - p(1/2) + p(-1/2),$$

and therefore the (rough) estimate

$$\|p\|_\infty \leq |a_0| + |a_1| + |a_2| \leq \|p\|_{\infty,X} + 2\|p\|_{\infty,X} + 4\|p\|_{\infty,X} = 7\|p\|_{\infty,X}$$

for all $p \in \mathcal{P}_2$, whereby (5.24) holds for $\beta = 1/7$. Together with (5.23), this finally yields the sought estimate

$$\|p - f\|_\infty - \|p_2^* - f\|_\infty \geq \frac{1}{7}\|p - p_2^*\|_{\infty,X} \geq \frac{1}{49}\|p - p_2^*\|_\infty \qquad \text{for all } p \in \mathcal{P}_2.$$

Therefore, p_2^* is the strongly unique best approximation to f. ♢

5.3 Haar Spaces

In this section, we develop sufficient conditions for the approximation space $\mathcal{S} \subset \mathscr{C}(\Omega)$ under which a best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}(\Omega) \setminus \mathcal{S}$ is strongly unique. To this end, we can rely on the result of Theorem 5.19, according to which we need to ensure the definiteness of $\|\cdot\|_{\infty,X}$ on $\mathcal{S}$ for any $X \subset E_{s^*-f}$.

We continue to use the assumptions and notations from the previous section, where $(s_1, \ldots, s_n) \in \mathcal{S}^n$, for $n \in \mathbb{N}$, denotes an *ordered* basis of a finite-dimensional linear approximation space $\mathcal{S} \subset \mathscr{C}(\Omega)$. By the introduction of *Haar*[3] *spaces* we specialize our assumptions on $\mathcal{S}$ and $(s_1, \ldots, s_n)$ as follows.

Definition 5.22. *A linear space $\mathcal{S} \subset \mathscr{C}(\Omega)$ with $\dim(\mathcal{S}) = n < \infty$ is called a* `Haar space` *of dimension $n \in \mathbb{N}$ on Ω, if any $s \in \mathcal{S} \setminus \{0\}$ has at most $n-1$ zeros on Ω. A basis $\mathcal{H} = (s_1, \ldots, s_n) \in \mathcal{S}^n$ for a Haar space $\mathcal{S}$ on Ω is called a* `Haar system` *on Ω.* ◯

In Haar spaces $\mathcal{S}$ of dimension $n \in \mathbb{N}$, we can solve interpolation problems for a discrete set $X \subset \Omega$ containing $|X| = n$ pairwise distinct points. In this case, $\|\cdot\|_{\infty,X}$ is a norm on $\mathcal{S}$, and so a solution of the interpolation problem is unique. We can further characterize Haar spaces as follows.

[3] Alfréd Haar (1885-1933), Hungarian mathematician

Theorem 5.23. *Let $\mathcal{S} \subset \mathscr{C}(\Omega)$ be a linear space of dimension $n \in \mathbb{N}$ and $X = \{x_1, \ldots, x_n\} \subset \Omega$ a set of n pairwise distinct points. Then the following statements are equivalent.*

(a) Any $s \in \mathcal{S} \setminus \{0\}$ has at most $n-1$ zeros on X.
(b) For $s \in S$, we have the implication

$$s_X = 0 \qquad \Longrightarrow \qquad s \equiv 0 \text{ on } \Omega,$$

i.e., $\|\cdot\|_{\infty,X}$ is a norm on $\mathcal{S}$.
(c) For any $f_X \in \mathbb{R}^n$, there is one unique $s \in \mathcal{S}$ satisfying $s_X = f_X$.
(d) For any basis $\mathcal{H} = (s_1, \ldots, s_n) \in \mathcal{S}^n$ of $\mathcal{S}$, the `Vandermonde matrix`

$$V_{\mathcal{H},X} = \begin{bmatrix} s_1(x_1) & \cdots & s_1(x_n) \\ \vdots & & \vdots \\ s_n(x_1) & \cdots & s_n(x_n) \end{bmatrix} \in \mathbb{R}^{n \times n}$$

is regular, where in particular $\det(V_{\mathcal{H},X}) \neq 0$.

If one of the statements (a)-(d) holds for all sets $X = \{x_1, \ldots, x_n\} \subset \Omega$ of n pairwise distinct points, then all of the remaining three statements (a)-(d) are satisfied for all X. In this case, $\mathcal{S}$ is a Haar space of dimension n on Ω.

Proof. Let $X = \{x_1, \ldots, x_n\} \subset \Omega$ be a set of n pairwise distinct points in Ω. Obviously, the statements (a) and (b) are equivalent. By statement (b), the linear mapping $L_X : s \longmapsto s_X$ is injective. Since $n = \dim(\mathcal{S}) = \dim(\mathbb{R}^n)$, this is equivalent to statement (c), i.e., L_X is surjective, and, moreover, also equivalent to statement (d), i.e., L_X is bijective. This completes our proof for the equivalence of statements (a)-(d).

If one of the statements in (a)-(d) holds for *all* sets $X = \{x_1, \ldots, x_n\} \subset \Omega$ of n pairwise distinct points, then all of the remaining three statements in (a)-(d) are satisfied, due to the equivalence of statements (a)-(d). In this case, statement (a) holds in particular, for *all* sets $X = \{x_1, \ldots, x_n\} \subset \Omega$, i.e., any $s \in \mathcal{S} \setminus \{0\}$ has at most $n-1$ zeros on Ω, whereby $\mathcal{S}$ is, according Definition 5.22, a Haar space of dimension n on Ω. ■

According to the *Mairhuber*[4]*-Curtis*[5] *theorem* [17, 48] there are *no* non-trivial Haar systems on *multivariate* connected domains $\Omega \subset \mathbb{R}^d$, $d > 1$. Before we prove the Mairhuber-Curtis theorem, we introduce a few notions.

Definition 5.24. *A domain $\Omega \subset \mathbb{R}^d$ is said to be* `connected`*, if for any pair of two points $x, y \in \Omega$ there is a continuous mapping $\gamma : [0,1] \longrightarrow \Omega$ satisfying $\gamma(0) = x$ and $\gamma(1) = y$, i.e., the points x and y can be connected by a continuous path in Ω.* ◯

[4] John C. Mairhuber (1922-2007), US American mathematician
[5] Philip C. Curtis, Jr. (1928-2016), US-American mathematician

Moreover, we call a domain $\Omega \subset \mathbb{R}^d$ `homeomorphic` to a subset of the sphere $\mathbb{S}^1 := \{x \in \mathbb{R}^2 \,|\, \|x\|_2 = 1\} \subset \mathbb{R}^2$, if for a non-empty and connected subset $U \subset \mathbb{S}^1$ there is a bijective continuous mapping $\varphi : \Omega \longrightarrow U$ with continuous inverse $\varphi^{-1} : U \longrightarrow \Omega$.

Theorem 5.25. (Mairhuber-Curtis, 1956/1959).
Let $\mathcal{H} = (s_1, \ldots, s_n) \in (\mathscr{C}(\Omega))^n$ be a Haar system of dimension $n \geq 2$ on a connected set $\Omega \subset \mathbb{R}^d$, $d > 1$. Then, Ω contains no bifurcation, i.e., Ω is homeomorphic to a subset of the sphere $\mathbb{S}^1 \subset \mathbb{R}^2$.

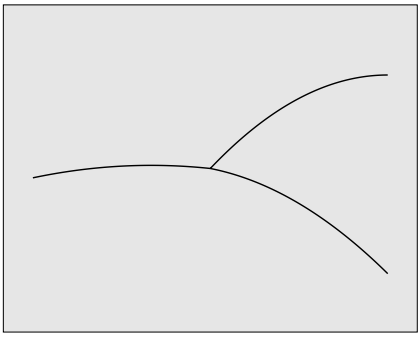


Fig. 5.3. According to the Mairhuber-Curtis theorem, Theorem 5.25, there are *no* non-trivial Haar systems $\mathcal{H}$ on domains Ω containing bifurcations.

Proof. Suppose Ω contains a bifurcation (see Figure 5.3 for illustration). Moreover, $X = \{x_1, \ldots, x_n\} \subset \Omega$ be a subset of $n \geq 2$ pairwise distinct points in Ω. Now regard the determinant

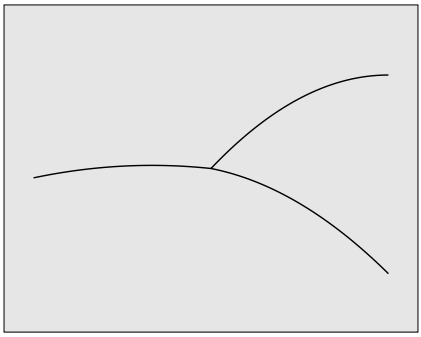
$$d_{\{x_1,x_2,x_3\ldots,x_n\}} = \det(V_{\mathcal{H},X}) = \det \begin{bmatrix} s_1(x_1) & s_1(x_2) & s_1(x_3) & \cdots & s_1(x_n) \\ s_2(x_1) & s_2(x_2) & s_2(x_3) & \cdots & s_2(x_n) \\ \vdots & \vdots & \vdots & & \vdots \\ s_n(x_1) & s_n(x_2) & s_n(x_3) & \cdots & s_n(x_n) \end{bmatrix}.$$

If $d_{\{x_1,x_2,x_3\ldots,x_n\}} = 0$, then $\mathcal{H}$, by Theorem 5.23, is not a Haar system.

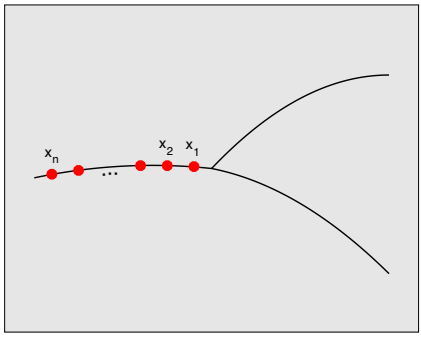
Otherwise, we can shift the two points x_1 und x_2 by a *continuous* mapping along the two branches of the bifurcation, without any coincidence between points in X (see Figure 5.4).

Therefore, the determinant $d_{\{x_2,x_1,x_3,\ldots,x_n\}}$ has, by swapping the first two columns in matrix $V_{\mathcal{H},X}$, opposite sign to $d_{\{x_1,x_2,x_3,\ldots,x_n\}}$, i.e.,

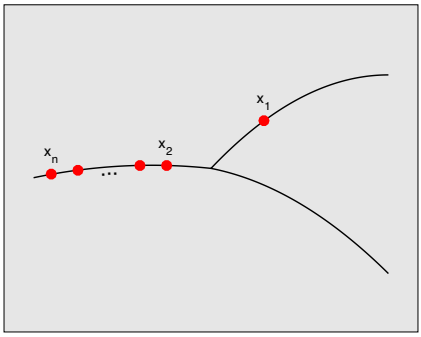
$$\operatorname{sgn}\left(d_{\{x_1,x_2,x_3,\ldots,x_n\}}\right) = -\operatorname{sgn}\left(d_{\{x_2,x_1,x_3,\ldots,x_n\}}\right).$$



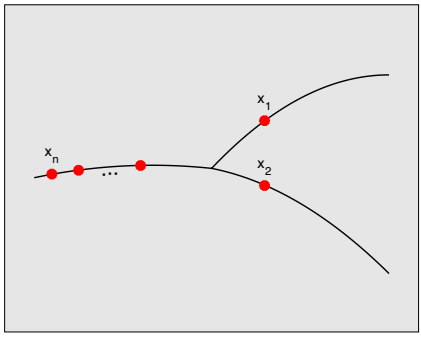
domain Ω



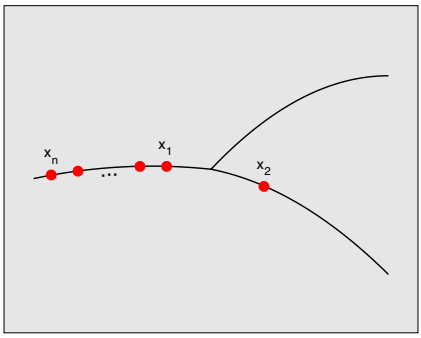
$X = (x_1, \ldots, x_n) \in \Omega^n$



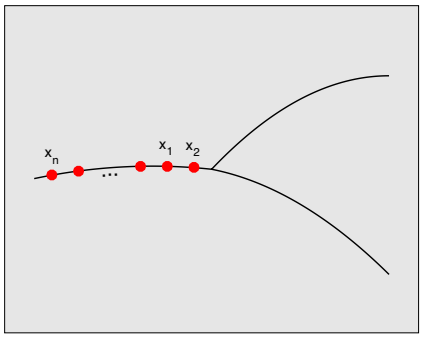
step 1: shift of x_1



step 2: shift of x_2



step 3: back-shift of x_1



step 4: back-shift of x_2

Fig. 5.4. Illustration of the Mairhuber-Curtis theorem, Theorem 5.25. The two points x_1 and x_2 can be swapped by a *continuous* mapping, i.e., by shifts along the branches of the bifurcation without coinciding with any other point from X.

Due to the continuity of the determinant, there must be a sign change of the determinant during the (continuous) swapping between x_1 and x_2. In this case, $\mathcal{H} = \{s_1, \ldots, s_n\}$ cannot be a Haar system, by Theorem 5.23. But this is in contradiction to our assumption to $\mathcal{H}$. ■

Due to the result of the Mairhuber-Curtis theorem, Theorem 5.25, we restrict ourselves from now to the *univariate* case, $d = 1$. Moreover, we assume from now that the domain Ω is a compact interval, i.e.,

$$\Omega = [a, b] \subset \mathbb{R} \qquad \text{for } -\infty < a < b < \infty.$$

Before we continue our analysis on strongly unique best approximations, we first give a few elementary examples for Haar spaces.

Example 5.26. For $n \in \mathbb{N}_0$ and $[a, b] \subset \mathbb{R}$ the linear space of polynomials $\mathcal{P}_n$ is a Haar space of dimension $n+1$ on $[a, b]$, since according to the fundamental theorem of algebra any non-trivial polynomial from $\mathcal{P}_n$ has at most n zeros. ♢

Example 5.27. For $N \in \mathbb{N}_0$ the linear space $\mathcal{T}_N^{\mathbb{C}}$ of all complex trigonometric polynomials of degree at most N is a Haar space of dimension $N + 1$ on $[0, 2\pi)$, since $\mathcal{T}_N^{\mathbb{C}}$ is, by Theorem 2.36, a linear space of dimension $N+1$, and, moreover, the linear mapping $p \longmapsto p_X$, for $p \in \mathcal{T}_N^{\mathbb{C}}$ is, due to Theorem 2.39, for all sets $X \subset [0, 2\pi)$ of $|X| = N + 1$ pairwise distinct points bijective.

Likewise, we can show, by using Corollaries 2.38 and 2.40, that the linear space $\mathcal{T}_n^{\mathbb{R}}$ of all real trigonometric polynomials of degree at most $n \in \mathbb{N}_0$ is a Haar space of dimension $2n + 1$ on $[0, 2\pi)$. ♢

Example 5.28. For $[a, b] \subset \mathbb{R}$ and $\lambda_0 < \ldots < \lambda_n$ the functions

$$\left\{ \mathrm{e}^{\lambda_0 x}, \ldots, \mathrm{e}^{\lambda_n x} \right\}$$

are a Haar system on $[a, b]$. We can show this by induction on n.

Initial step: For $n = 0$ the statement is trivial.

Induction hypothesis: Suppose the statement is true for $n - 1 \in \mathbb{N}$.

Induction step $(n - 1 \longrightarrow n)$*:* If a function of the form

$$u(x) \in \operatorname{span} \left\{ \mathrm{e}^{\lambda_0 x}, \ldots, \mathrm{e}^{\lambda_n x} \right\}$$

has $n + 1$ zeros in $[a, b]$, then the function

$$v(x) = \frac{\mathrm{d}}{\mathrm{d}x} \left(\mathrm{e}^{-\lambda_0 x} \cdot u(x) \right) \qquad \text{for } x \in [a, b]$$

has, according to the *Rolle*[6] *theorem*, at least n zeros in $[a, b]$. However,

$$v(x) \in \operatorname{span} \left\{ \mathrm{e}^{(\lambda_1 - \lambda_0)x}, \ldots, \mathrm{e}^{(\lambda_n - \lambda_0)x} \right\},$$

which implies $v \equiv 0$ by the induction hypothesis, and so $u \equiv 0$. ♢

[6] Michel Rolle (1652-1719), French mathematician

Example 5.29. The functions $f_1(x) = x$ and $f_2(x) = \mathrm{e}^x$ are *not* a Haar system on $[0, 2]$. This is because $\dim(\mathcal{S}) = 2$ for $\mathcal{S} = \operatorname{span}\{f_1, f_2\}$, but the continuous function

$$f(x) = \mathrm{e}^x - 3x \not\equiv 0$$

has by $f(0) = 1$, $f(1) = \mathrm{e} - 3 < 0$ and $f(2) > 0$ at least two zeros in $[0, 2]$. Therefore, $\mathcal{S}$ cannot be a Haar space on $[0, 2]$. ♢

Example 5.30. For $[a, b] \subset \mathbb{R}$ let $g \in \mathscr{C}^{n+1}[a, b]$ satisfy $g^{(n+1)}(x) > 0$ for all $x \in [a, b]$. Then, the functions $\{1, x, \ldots, x^n, g\}$ are a Haar system on $[a, b]$: First note that the functions $1, x, \ldots, x^n, g(x)$ are linearly independent, since from

$$\alpha_0 1 + \alpha_1 x + \ldots + \alpha_n x^n + \alpha_{n+1} g(x) \equiv 0 \qquad \text{for } x \in [a, b]$$

we can conclude $\alpha_{n+1} g^{(n+1)}(x) \equiv 0$ after $(n+1)$-fold differentiation, whereby $\alpha_{n+1} = 0$. The remaining coefficients $\alpha_0, \ldots, \alpha_n$ do also vanish, since the monomials $1, x, \ldots, x^n$ are linearly independent. Moreover, we can show that any function $u \in \operatorname{span}\{1, x, \ldots, x^n, g\} \setminus \{0\}$ has at most $n + 1$ zeros in $[a, b]$: Suppose

$$u(x) = \sum_{j=0}^{n} \alpha_j x^j + \alpha_{n+1} g(x) \not\equiv 0$$

has $n + 2$ zeros in $[a, b]$. Then, the $(n + 1)$-th derivative

$$u^{(n+1)}(x) = \alpha_{n+1} g^{(n+1)}(x)$$

has, due to the Rolle theorem, at least one zero in $[a, b]$. But this implies $\alpha_{n+1} = 0$, since $g^{(n+1)}$ is positive on $[a, b]$. In this case, $u \in \mathcal{P}_n$ is a polynomial of degree at most n, which, according to the fundamental theorem of algebra, vanishes identically on $[a, b]$. But this is in contraction to our assumption. ♢

Now we return to the dual characterization of (strongly) unique best approximations. According to Corollary 5.15, there is for any best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}[a, b]$ a characterizing dual functional $\varphi : \mathscr{C}(\Omega) \longrightarrow \mathcal{S}$ of the form

$$\varphi(u) = \sum_{j=1}^{m} \lambda_j \varepsilon_j u(x_j) \qquad \text{for } u \in \mathscr{C}[a, b] \tag{5.25}$$

satisfying $\varphi(\mathcal{S}) = 0$, where $m \leq n+1$. For the case of Haar spaces $\mathcal{S} \subset \mathscr{C}[a, b]$ the length of the dual functional in (5.25) is necessarily $m = n + 1$. Let us take note of this important observation.

Proposition 5.31. *Let $\varphi : \mathscr{C}[a, b] \longrightarrow \mathbb{R}$ be a functional of the form (5.25), where $m \leq n + 1$. Moreover, let $\mathcal{S} \subset \mathscr{C}[a, b]$ be a Haar space of dimension $\dim(\mathcal{S}) = n \in \mathbb{N}$ on $[a, b]$. If $\varphi(\mathcal{S}) = \{0\}$, then we have $m = n + 1$.*

Proof. Suppose $m \leq n$. Then, due to Theorem 5.23 (c), the Haar space $\mathcal{S}$ contains one element $s \in \mathcal{S}$ satisfying $s(x_j) = \varepsilon_j$, for all $1 \leq j \leq m$. But for this s, we find $\varphi(s) = \|\lambda\|_1 = 1$, in contradiction to $\varphi(\mathcal{S}) = \{0\}$. ■

In the following discussion, we consider, for a fixed basis $\mathcal{H} = (s_1, \ldots, s_n)$ of the Haar space $\mathcal{S}$, points $X = (x_1, \ldots, x_{n+1}) \in I^{n+1}$, and sign vectors $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n+1}) \in \{\pm 1\}^{n+1}$, the *non-singular* Vandermonde matrices

$$V_{\mathcal{H},X\setminus\{x_k\}} = \begin{bmatrix} s_1(x_1) & \cdots & s_1(x_{k-1}) & s_1(x_{k+1}) & \cdots & s_1(x_{n+1}) \\ \vdots & & \vdots & \vdots & & \vdots \\ s_n(x_1) & \cdots & s_n(x_{k-1}) & s_n(x_{k+1}) & \cdots & s_n(x_{n+1}) \end{bmatrix} \in \mathbb{R}^{n\times n} \quad (5.26)$$

for $1 \le k \le n+1$, and the *alternation matrix*

$$A_{\varepsilon,\mathcal{H},X} = \left[\frac{\varepsilon}{V_{\mathcal{H},X}}\right] = \left[\begin{array}{ccc} \varepsilon_1 & \cdots & \varepsilon_{n+1} \\ \hline s_1(x_1) & \cdots & s_1(x_{n+1}) \\ \vdots & & \vdots \\ s_n(x_1) & \cdots & s_n(x_{n+1}) \end{array}\right] \in \mathbb{R}^{(n+1)\times(n+1)}. \quad (5.27)$$

We first take note of a few properties for $V_{\mathcal{H},X\setminus\{x_k\}}$ and $A_{\varepsilon,\mathcal{H},X}$.

Proposition 5.32. *Let $\mathcal{H} = (s_1, \ldots, s_n)$ be a Haar system on an interval $I \subset \mathbb{R}$ and $X = (x_1, \ldots, x_{n+1}) \in I^{n+1}$ be a vector of $n+1$ pairwise distinct points. Then, the following statements are true.*

(a) For the Vandermonde matrices $V_{\mathcal{H},X\setminus\{x_k\}}$ in (5.26), the signs of the $n+1$ determinants

$$d_k = \det(V_{\mathcal{H},X\setminus\{x_k\}}) \neq 0 \qquad \text{for } 1 \le k \le n+1$$

are constant, i.e., $\operatorname{sgn}(d_k) = \sigma$, for all $1 \le k \le n+1$, for some $\sigma \in \{\pm 1\}$.

(b) If the signs in $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n+1}) \in \{\pm 1\}^{n+1}$ are alternating, i.e., if

$$\varepsilon_k = (-1)^{k-1}\sigma \qquad \text{for } 1 \le k \le n+1$$

for some $\sigma \in \{\pm 1\}$, then the matrix $A_{\varepsilon,\mathcal{H},X}$ in (5.27) is non-singular.

Proof. (a): Suppose we have $\operatorname{sgn}(d_k) \neq \operatorname{sgn}(d_{k+1})$ for some $1 \le k \le n$.

We consider a *continuous* mapping $\gamma : [0,1] \longrightarrow I$ satisfying $\gamma(0) = x_k$ and $\gamma(1) = x_{k+1}$. In this case, the *continuous* determinant mapping

$$d(\alpha) = \det(V_{\mathcal{H},(x_1,\ldots,x_{k-1},\gamma(\alpha),x_{k+2},\ldots,x_{n+1})}) \qquad \text{for } \alpha \in [0,1]$$

satisfying $d(0) = d_{k+1}$ and $d(1) = d_k$ must have a sign change on $(0,1)$. Due to the continuity of d there is one $\alpha^* \in (0,1)$ satisfying $d(\alpha^*) = 0$. However, in this case, the Vandermonde matrix $V_{\mathcal{H},(x_1,\ldots,x_{k-1},\gamma(\alpha^*),x_{k+2},\ldots,x_{n+1})} \in \mathbb{R}^{n\times n}$ is singular. Due to Theorem 5.23 (d), the elements in $(s_1, \ldots, s_n)$ are *not* a Haar system on $I \subset \mathbb{R}$. But this is in contradiction to our assumption.

(b): According to the Laplace[7] expansion (here with respect to the first row), the determinant of $A_{\varepsilon,\mathcal{H},X}$ has the representation

$$\det(A_{\varepsilon,\mathcal{H},X}) = \sum_{k=1}^{n+1}(-1)^{k+1}(-1)^{k-1}\sigma \cdot d_k = \sigma \sum_{k=1}^{n+1} d_k.$$

Due to statement (a), the signs of the determinants d_k, $1 \leq k \leq n+1$, are constant, which implies $\det(A_{\varepsilon,\mathcal{H},X}) \neq 0$. ■

By using the results of Propositions 5.31 and 5.32, we can prove the *alternation theorem*, being the central result of this chapter. According to the alternation theorem, the signs $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n+1})$ of the dual characterization in (5.25) are for the case of Haar spaces $\mathcal{S}$ alternating. Before we prove the alternation theorem, we first give a formal definition for *alternation sets*.

Definition 5.33. *Let $\mathcal{S} \subset \mathscr{C}(I)$ be a Haar space of dimension $n \in \mathbb{N}$ on an interval $I \subset \mathbb{R}$. Moreover, suppose $s^* \in \mathcal{S}$ and $f \in \mathscr{C}(I) \setminus \mathcal{S}$. Then, an ordered set $X = (x_1, \ldots, x_{n+1}) \in E^{n+1}_{s^*-f} \subset I^{n+1}$ of $n+1$ monotonically increasing extremal points $x_1 < \ldots < x_{n+1}$ is called an* `alternation set` *for s^* and f, if*

$$\varepsilon_j = \operatorname{sgn}((s^* - f)(x_j)) = (-1)^j\sigma \qquad \text{for all } j = 1, \ldots, n+1$$

for some $\sigma \in \{\pm 1\}$, i.e., if the signs of $s^* - f$ are alternating on X. ◯

Theorem 5.34. (Alternation theorem).
Let $\mathcal{S} \subset \mathscr{C}(I)$ be a Haar space of dimension $n \in \mathbb{N}$ on an interval $I \subset \mathbb{R}$. Moreover, let $I_K \subset I$ be a compact subset containing at least $n+1$ elements. Then, there is for any $f \in \mathscr{C}(I_K) \setminus \mathcal{S}$ a strongly unique best approximation $s^* \in \mathcal{S}$ to f with respect to $\|\cdot\|_{\infty,I_K}$. The best approximation s^* is characterized by the existence of an alternation set $X \in E^{n+1}_{s^*-f} \subset I_K^{n+1}$ for s^* and f.

Proof. Due to Corollary 3.8, any $f \in \mathscr{C}(I_K)$ has a best approximation $s^* \in \mathcal{S}$. Moreover, the strong uniqueness of s^* follows from Theorem 5.19, where the assumptions required therein for Theorem 5.17 are covered by Corollary 5.15.

Now we prove the stated characterization for s^*.

To this end, let $X = (x_1, \ldots, x_{n+1}) \in E^{n+1}_{s^*-f} \subset I_K^{n+1}$ be an alternation set for s^* and f with (alternating) signs $\varepsilon_j = \operatorname{sgn}((s^* - f)(x_j)) = (-1)^j\sigma$, for $1 \leq j \leq n+1$, and some $\sigma \in \{\pm 1\}$. Then, we consider the linear system

$$\left[\begin{array}{ccc} \varepsilon_1 & \cdots & \varepsilon_{n+1} \\ \hline s_1(x_1) & \cdots & s_1(x_{n+1}) \\ \vdots & & \vdots \\ s_n(x_1) & \cdots & s_n(x_{n+1}) \end{array}\right] \cdot \begin{bmatrix} \varepsilon_1\lambda_1 \\ \varepsilon_2\lambda_2 \\ \vdots \\ \varepsilon_{n+1}\lambda_{n+1} \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{5.28}$$

[7] Pierre-Simon Laplace (1749-1827), French mathematician and physicist

with the alternation matrix $A_{\varepsilon,\mathcal{H},X}$ on the left hand side in (5.28). According to Proposition 5.32 (a), the matrix $A_{\varepsilon,\mathcal{H},X}$ is non-singular. Therefore, the products $\varepsilon_k\lambda_k$, for $1 \le k \le n+1$, uniquely solve the linear system (5.28).

Due to the *Cramer*[8] *rule* we have the representation

$$\varepsilon_k\lambda_k = \frac{(-1)^{k-1}d_k}{\det(A_{\varepsilon,\mathcal{H},X})} \qquad \text{for all } 1 \le k \le n+1,$$

where according to Proposition 5.32 (a) the signs of the $n+1$ determinants $d_k = \det(V_{\mathcal{H},X\setminus\{x_k\}})$, for $1 \le k \le n+1$, are constant. This implies $\varepsilon_k\lambda_k \neq 0$, and, moreover, there is one unique vector $\lambda = (\lambda_1, \ldots, \lambda_{n+1})^T \in \Lambda_{n+1}$ with *positive* coefficients

$$\lambda_k = \frac{d_k}{\sum_{j=1}^{n+1} d_j} > 0 \qquad \text{for all } 1 \le k \le n+1$$

which solves the linear system (5.28). This solution $\lambda \in \Lambda_{n+1}$ of (5.28) finally yields the characterizing functional (according to Corollary 5.15),

$$\varphi(u) = \sum_{j=1}^{n+1} \lambda_j\varepsilon_j u(x_j) \qquad \text{for } u \in \mathscr{C}(I_K), \tag{5.29}$$

satisfying $\varphi(\mathcal{S}) = \{0\}$. Due to Corollary 5.15, s^* is the (strongly unique) best approximation to f.

Now suppose that $s^* \in \mathcal{S}$ is the strongly unique best approximation to $f \in \mathscr{C}(I_K)\setminus\mathcal{S}$. Recall that the dual characterization in Corollary 5.15 proves the existence of a functional $\varphi : \mathscr{C}(I_K) \longrightarrow \mathbb{R}$ of the form (5.25) satisfying $\varphi(\mathcal{S}) = \{0\}$, where φ has, according to Proposition 5.31, length $m = n+1$. We show that the point set $X = (x_1, \ldots, x_{n+1}) \in E^{n+1}_{s^*-f}$ (from the dual characterization in Corollary 5.15) is an alternation set for s^* and f, where our proof is by contradiction. To this end, let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_{n+1}) \in \{\pm 1\}^{n+1}$ denote the sign vector of $s^* - f$ with $\varepsilon_j = \operatorname{sgn}((s^*-f)(x_j))$, for $1 \le j \le n+1$. Now suppose there is one index $k \in \{1, \ldots, n\}$ satisfying $\varepsilon_k = \varepsilon_{k+1}$. Then there is one $s \in \mathcal{S} \setminus \{0\}$ satisfying $s(x_j) = 0$ for all $j \notin \{k, k+1\}$ and $s(x_k) = \varepsilon_k$. Since s cannot have more than $n-1$ zeros in I, we necessarily have

$$\varepsilon_k = \operatorname{sgn}(s(x_k)) = \operatorname{sgn}(s(x_{k+1})) = \varepsilon_{k+1}.$$

This particularly implies

$$\varphi(s) = \lambda_k + \lambda_{k+1}|s(x_{k+1})| > 0,$$

which, however, is in contradiction to $\varphi(s) = 0$. ■

We finally remark that the characterizing alternation set $X \in E^{n+1}_{s^*-f}$ for s^* and f in the alternation theorem, Theorem 5.34, is *not* necessarily unique. This is because the set E_{s^*-f} of extremal points can be arbitrarily large (see Example 5.3).

[8] Gabriel Cramer (1704-1752), Swiss mathematician

5.4 The Remez Algorithm

In this section, we discuss the *Remez*[9] *algorithm* [59, 60], an iterative method to numerically compute the (strongly unique) best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}[a,b] \setminus \mathcal{S}$, where $[a,b] \subset \mathbb{R}$ is a compact interval. Moreover, $\mathcal{S} \subset \mathscr{C}[a,b]$ denotes a Haar space of dimension $n \in \mathbb{N}$ on $[a,b]$.

In any of its iteration steps, the Remez algorithm computes, for an ordered (i.e., monotonically increasing) *reference set* $X = (x_1, \ldots, x_{n+1}) \in [a,b]^{n+1}$ of length $|X| = n+1$, the corresponding (strongly unique) best approximation s_X^* to f with respect to $\|\cdot\|_{\infty,X}$, so that

$$\|s_X^* - f\|_{\infty,X} < \|s - f\|_{\infty,X} \qquad \text{for all } s \in \mathcal{S} \setminus \{s_X^*\}.$$

To compute s_X^*, we first fix an ordered basis $\mathcal{H} = (s_1, \ldots, s_n)$ of the Haar space $\mathcal{S}$, so that s_X^* can be represented as linear combination

$$s_X^* = \sum_{j=1}^{n} \alpha_j^* s_j \in \mathcal{S} \tag{5.30}$$

of the Haar system $\mathcal{H}$ with coefficients $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T \in \mathbb{R}^n$. According to the alternation theorem, Theorem 5.34, the sought best approximation s_X^* necessarily satisfies the *alternation condition*

$$(s_X^* - f)(x_k) = (-1)^{k-1}\sigma\|s_X^* - f\|_{\infty,X} \qquad \text{for } 1 \le k \le n+1 \tag{5.31}$$

for some $\sigma \in \{\pm 1\}$. Letting $\eta_X = \sigma\|s_X^* - f\|_{\infty,X}$ we rewrite (5.31) as

$$s_X^*(x_k) + (-1)^k \eta_X = f(x_k) \qquad \text{for } 1 \le k \le n+1. \tag{5.32}$$

Therefore, η_X and the unknown coefficients $\alpha^* \in \mathbb{R}^n$ of s_X^* are the solution of the linear equation system

$$A_{\varepsilon,\mathcal{H},X}^T \cdot \begin{bmatrix} \eta_X \\ \alpha^* \end{bmatrix} = f_X \tag{5.33}$$

with the right hand side $f_X = (f(x_1), \ldots, f(x_{n+1}))^T \in \mathbb{R}^{n+1}$ and the alternation matrix $A_{\varepsilon,\mathcal{H},X} \in \mathbb{R}^{(n+1)\times(n+1)}$ in (5.27), containing the sign vector $\varepsilon = (-1, 1, \ldots, (-1)^{n+1}) \in \{\pm 1\}^{n+1}$, or,

$$\left[\begin{array}{c|ccc} -1 & s_1(x_1) & \cdots & s_n(x_1) \\ 1 & s_1(x_2) & \cdots & s_n(x_2) \\ \vdots & \vdots & & \vdots \\ (-1)^{n+1} & s_1(x_{n+1}) & \cdots & s_n(x_{n+1}) \end{array}\right] \left[\begin{array}{c} \eta_X \\ \hline \alpha_1^* \\ \vdots \\ \alpha_n^* \end{array}\right] = \begin{bmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{n+1}) \end{bmatrix}.$$

By Proposition 5.32 (b), the matrix $A_{\varepsilon,\mathcal{H},X}$ is non-singular, and so the solution of the linear system (5.33) is unique. By the solution of (5.33), we

[9] Evgeny Yakovlevich Remez (1896-1975), mathematician

do not only obtain the coefficients $\alpha^* = (\alpha_1^*, \ldots, \alpha_n^*)^T \in \mathbb{R}^n$ of the best approximation s_X^* in (5.30), but also by $|\eta_X| = \|s_X^* - f\|_{\infty,X}$ we get the minimal distance and the sign $\sigma = \operatorname{sgn}(\eta_X)$ in (5.31).

The following observation concerning Chebyshev approximation to a function $f \in \mathscr{C}[a,b]$ by algebraic polynomials from $\mathcal{P}_{n-1}$ shows that, in this special case, the linear system (5.33) can be avoided. To this end, we use the Newton representation (2.34) for the interpolation polynomial in Theorem 2.13. Recall that the *Newton polynomials* are given as

$$\omega_k(x) = \prod_{j=1}^{k}(x - x_j) \in \mathcal{P}_k \qquad \text{for } 0 \le k \le n-1.$$

In particular, we apply the linear operator $[x_1, \ldots, x_{n+1}] : \mathscr{C}[a,b] \longrightarrow \mathbb{R}$ of the *divided differences* to f (see Definition 2.10). To evaluate $[x_1, \ldots, x_{n+1}](f)$, we apply the recursion in Theorem 2.14. The recursion in Theorem 2.14 operates only on the vector of function values $f_X = (f(x_1), \ldots, f(x_{n+1}))^T \in \mathbb{R}^{n+1}$. Therefore, the application of $[x_1, \ldots, x_{n+1}]$ is also well-defined for any sign vector $\varepsilon \in \{\pm 1\}^{n+1}$ of length $n+1$. In particular, the divided difference $[x_1, \ldots, x_{n+1}](\varepsilon)$ can also be evaluated by the recursion in Theorem 2.14. In the formulation of the following result, we apply divided differences to vectors ε with alternating signs.

Proposition 5.35. *For $n \in \mathbb{N}$, let $X = (x_1, \ldots, x_{n+1})$ be an ordered set of $n+1$ points in $[a,b] \subset \mathbb{R}$, $\varepsilon = (-1, 1, \ldots, (-1)^{n+1}) \in \{\pm 1\}^{n+1}$ a sign vector, and $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_{n-1}$. Then,*

$$s_X^* = \sum_{k=0}^{n-1} [x_1, \ldots, x_{k+1}](f - \eta_X \varepsilon)\omega_k \in \mathcal{P}_{n-1} \tag{5.34}$$

is the strongly unique best approximation $s_X^* \in \mathcal{P}_{n-1}$ to f w.r.t. $\|\cdot\|_{\infty,X}$, where

$$\eta_X = \frac{[x_1, \ldots, x_{n+1}](f)}{[x_1, \ldots, x_{n+1}](\varepsilon)}. \tag{5.35}$$

The minimal distance is given as

$$\|s_X^* - f\|_{\infty,X} = |\eta_X|.$$

Proof. Application of the linear operator $[x_1, \ldots, x_{n+1}] : \mathscr{C}[a,b] \longrightarrow \mathbb{R}$ to the alternation condition (5.32) immediately gives the representation

$$\eta_X = \frac{[x_1, \ldots, x_{n+1}](f)}{[x_1, \ldots, x_{n+1}](\varepsilon)}.$$

Indeed, due to Corollary 2.18 (b), all polynomials from $\mathcal{P}_{n-1}$ are contained in the kernel of $[x_1, \ldots, x_{n+1}]$. In particular, we have $[x_1, \ldots, x_{n+1}](s_X^*) = 0$.

Under the alternation condition (5.32), $s_X^* \in \mathcal{P}_{n-1}$ is the unique solution of the interpolation problem

$$s_X^*(x_k) = f(x_k) - (-1)^k \eta_X \qquad \text{for } 1 \le k \le n,$$

already for the first n alternation points $(x_1, \ldots, x_n) \in E^n_{s_X^* - f}$. This gives the stated Newton representation of s_X^* in (5.34). ■

Remark 5.36. Note that all divided differences

$$[x_1, \ldots, x_{k+1}](f - \eta_X \varepsilon) = [x_1, \ldots, x_{k+1}](f) - \eta_X [x_1, \ldots, x_{k+1}](\varepsilon)$$

in (5.34) are readily available from the computation of η_X in (5.35). Therefore, we can compute the best approximation $s_X^* \in \mathcal{P}_{n-1}$ to f with respect to $\|\cdot\|_{\infty,X}$ by divided differences in only $\mathcal{O}(n^2)$ steps, where the computation is efficient and stable. □

To show how the result of Proposition 5.35, in combination with Remark 5.36, can be applied, we make the following concrete example.

Example 5.37. Let $\mathcal{F} = \mathscr{C}[0,2]$ and $\mathcal{S} = \mathcal{P}_1 \subset \mathcal{F}$. We approximate the exponential function $f(x) = \mathrm{e}^x$ on the reference set $X = \{0,1,2\}$. To compute the minimal distance η_X and the best approximation $s_X^* \in \mathcal{P}_1$ we use Proposition 5.35 with $n = 2$. We apply divided differences to the sign vector $\varepsilon = (-1,1,-1)$ and to the data vector $f_X = (1, e, e^2)$, where e denotes the *Euler number*. By the recursion in Theorem 2.14, we obtain the following triangular scheme for divided differences (see Table 2.1).

X	f_X		
0	1		
1	e	$e-1$	
2	e^2	$e(e-1)$	$(e-1)^2/2$

X	ε_X		
0	-1		
1	1	2	
2	-1	-2	-2

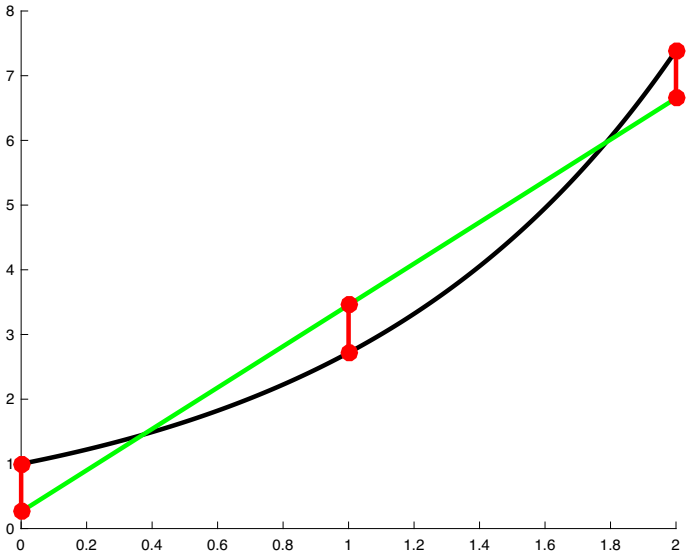
Hereby we obtain

$$\eta_X = \frac{[0,1,2](f)}{[0,1,2](\varepsilon)} = -\left(\frac{e-1}{2}\right)^2 \quad \text{and so} \quad \|s_X^* - f\|_{\infty,X} = \left(\frac{e-1}{2}\right)^2.$$

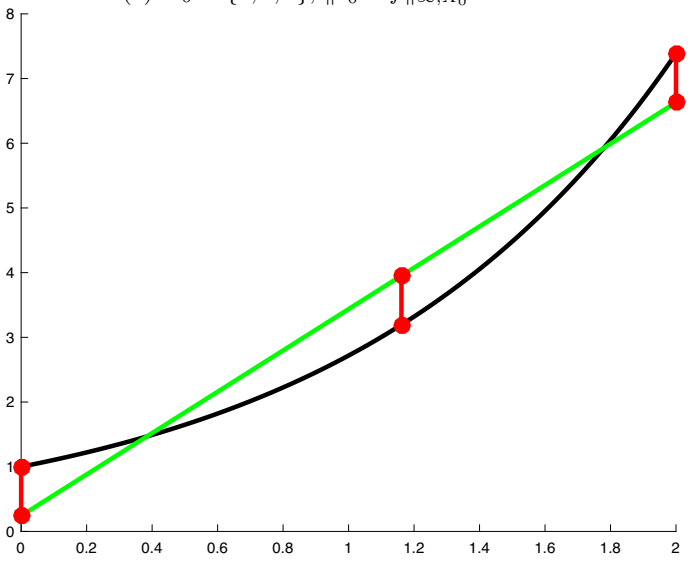
Moreover,

$$s_X^* = [0](f - \eta_X\varepsilon) + [0,1](f - \eta_X\varepsilon)x = 1 - \left(\frac{e-1}{2}\right)^2 + \frac{e^2-1}{2}x$$

is the unique best approximation to f from $\mathcal{P}_1$ w.r.t. $\|\cdot\|_{\infty,X}$ (see Fig. 5.5 (a)). ♢



(a) $X_0 = \{0, 1, 2\}$, $\|s_0^* - f\|_{\infty, X_0} \approx 0.7381$



(b) $X_1 = \{0, x^*, 2\}$, $\|s_1^* - f\|_{\infty, X_1} \approx 0.7579$

Fig. 5.5. Approximation to $f(x) = \mathrm{e}^x$ on $[0, 2]$ by linear polynomials from $\mathcal{P}_1$. (a) Initial reference set $X_0 = \{0, 1, 2\}$ with minimal distance $|\eta_0| = (e-1)^2/4 \approx 0.7381$. (b) Reference set $X_1 = \{0, x^*, 2\}$, where $x^* = \log((e^2 - 1)/2) \approx 1.1614$, with minimal distance $|\eta_1| = \frac{1}{4}\left[(e^2 - 1)(x^* - 1) + 2\right] \approx 0.7579$ (see Example 5.45).

We present another example, where we link with Example 5.7.

Example 5.38. We approximate the absolute-value function $f(x) = |x|$ on $[-1, 1]$ by quadratic polynomials, i.e., $\mathcal{S} = \mathcal{P}_2$. By our previous investigations in Example 5.7, $E_{p_2^*-f} = \{-1, -1/2, 0, 1/2, 1\}$ is the set of extremal points for the best approximation $p_2^* \in \mathcal{P}_2$ to f. To compute p_2^*, we apply Proposition 5.35 with $n = 3$. We let $\varepsilon = (-1, 1, -1, 1)$ and we choose $X = \{-1, -1/2, 0, 1/2\} \subset E_{p_2^*-f}$ as the reference set. By the recursion in Theorem 2.14, we obtain the following triangular scheme (see Table 2.1).

X	f_X			
-1	1			
$-\frac{1}{2}$	$\frac{1}{2}$	-1		
0	0	-1	0	
$\frac{1}{2}$	$\frac{1}{2}$	1	2	$\frac{4}{3}$

X	ε_X			
-1	-1			
$-\frac{1}{2}$	1	4		
0	-1	-4	-8	
$\frac{1}{2}$	1	4	8	$\frac{32}{3}$

Hereby we obtain $\eta_X = 1/8$, and so $\|s_X^* - f\|_{\infty,X} = 1/8$, along with

$$[x_1](f - \eta_X \varepsilon) = \frac{9}{8}, \quad [x_1, x_2](f - \eta_X \varepsilon) = -\frac{3}{2}, \quad [x_1, x_2, x_3](f - \eta_X \varepsilon) = 1.$$

Therefore,

$$s_X^*(x) = \frac{9}{8} - \frac{3}{2}(x + 1) + (x + 1)\left(x + \frac{1}{2}\right) = \frac{1}{8} + x^2$$

is the unique best approximation to f from $\mathcal{P}_2$ with respect to $\|\cdot\|_\infty$. Note that this is consistent with our observations in Example 5.7, since $p_2^* \equiv s_X^*$. ♢

Now we describe the iteration steps of the Remez algorithm. At any *Remez step* the current reference set (in increasing order)

$$X = (x_1, \ldots, x_{n+1}) \in [a, b]^{n+1}$$

is modified. This is done by a *Remez exchange* of one point $\hat{x} \in X$ for one point $x^* \in [a, b] \setminus X$, where

$$|(s_X^* - f)(x^*)| = \|s_X^* - f\|_\infty,$$

so that the next reference set is

$$X_+ = (X \setminus \{\hat{x}\}) \cup \{x^*\} = (x_1^+, \ldots, x_{n+1}^+) \in [a, b]^{n+1}.$$

With the Remez exchange, the point x^* is swapped for the point $\hat{x} \in X$, such that the points of the new reference set X_+ are in increasing order, i.e.,

$$a \le x_1^+ < x_2^+ < \ldots < x_n^+ < x_{n+1}^+ \le b,$$

and with maintaining the alternation condition, i.e.,

$$\operatorname{sgn}((s_X^* - f)(x_j^+)) = (-1)^j \sigma \qquad \text{for } 1 \le j \le n+1$$

for some $\sigma \in \{\pm 1\}$. The exchange for the point pair $(\hat{x}, x^*) \in X \times [a,b] \setminus X$ is described by the *Remez exchange*, Algorithm 8.

Algorithm 8 Remez exchange

function REMEZ EXCHANGE(X,s_X^*)
  Input: reference set $X = (x_1, \ldots, x_{n+1}) \in [a,b]^{n+1}$;
    best approximation s_X^* to f with respect to $\|\cdot\|_{\infty,X}$;

  find $x^* \in [a,b]$ satisfying $|(s_X^* - f)(x^*)| = \|s_X^* - f\|_\infty$;
  let $\sigma^* := \operatorname{sgn}((s_X^* - f)(x^*))$;

  if $x^* \in X$ **then return** X; ▷ best approximation found
  else if $x^* < x_1$ **then**
    if $\operatorname{sgn}((s_X^* - f)(x_1)) = \sigma^*$ **then** $X_+ = (x^*, x_2, \ldots, x_{n+1})$;
    else $X_+ = (x^*, x_1, \ldots, x_n)$;
    end if
  else if $x^* > x_{n+1}$ **then**
    if $\operatorname{sgn}((s_X^* - f)(x_{n+1})) = \sigma^*$ **then** $X_+ = (x_1, \ldots, x_n, x^*)$;
    else $X_+ = (x_2, \ldots, x_{n+1}, x^*)$;
    end if
  else
    find $j \in \{1, \ldots, n\}$ satisfying $x_j < x^* < x_{j+1}$;
    if $\operatorname{sgn}((s_X^* - f)(x_j)) = \sigma^*$ **then** $X_+ = (x_1, \ldots, x_{j-1}, x^*, x_{j+1}, \ldots, x_{n+1})$;
    else $X_+ = (x_1, \ldots, x_j, x^*, x_{j+2}, \ldots, x_{n+1})$;
    end if
  end if
  return X_+;
end function

Remark 5.39. The reference set $X_+ = (x_1^+, \ldots, x_{n+1}^+) \in [a,b]^{n+1}$, after the application of one Remez exchange, Algorithm 8, to the previous reference set $X = (x_1, \ldots, x_{n+1}) \in [a,b]^{n+1}$, satisfies the following three conditions.

- $|(s_X^* - f)(x^*)| = \|s_X^* - f\|_\infty$ for one $x^* \in X_+$;
- $|(s_X^* - f)(x)| \ge \|s_X^* - f\|_{\infty,X}$ for all $x \in X_+$;
- $\operatorname{sgn}((s_X^* - f)(x_j^+)) = (-1)^j \sigma$ for $1 \le j \le n+1$ and some $\sigma \in \{\pm 1\}$;

These conditions are required for the performance of the Remez algorithm. □

Now we formulate the Remez algorithm, Algorithm 9, as an iterative method to numerically compute the (strongly unique) best approximation $s^* \in \mathcal{S}$ to $f \in \mathscr{C}[a,b] \setminus \mathcal{S}$ satisfying

$$\eta = \|s^* - f\|_\infty < \|s - f\|_\infty \qquad \text{for all } s \in \mathcal{S} \setminus \{s^*\}.$$

The Remez algorithm generates a sequence $(X_k)_{k\in\mathbb{N}_0} \subset [a,b]^{n+1}$ of reference sets, so that for the transition from $X = X_k$ to $X_+ = X_{k+1}$, for any $k \in \mathbb{N}_0$, all three conditions in Remark 5.39 are satisfied. The corresponding sequence of best approximations $s_k^* \in \mathcal{S}$ to f with respect to $\|\cdot\|_{\infty,X_k}$ satisfying

$$\eta_k = \|s_k^* - f\|_{\infty,X_k} < \|s - f\|_{\infty,X_k} \qquad \text{for all } s \in \mathcal{S} \setminus \{s_k^*\}$$

converges to s^*, i.e., $s_k^* \longrightarrow s^*$ and $\eta_k \longrightarrow \eta$, for $k \to \infty$, as we will prove in Theorem 5.43.

Algorithm 9 Remez algorithm

1: **function** REMEZ ALGORITHM
2: **Input:** Haar space $\mathcal{S}$ of dimension $n \in \mathbb{N}$; $f \in \mathscr{C}[a,b] \setminus \mathcal{S}$;
3:
4: **find** initial reference set $X_0 = (x_1^{(0)}, \ldots, x_{n+1}^{(0)}) \in [a,b]^{n+1}$;
5: **for** $k = 0, 1, 2, \ldots$ **do**
6: **compute** best approximation $s_k^* \in \mathcal{S}$ to f with respect to $\|\cdot\|_{\infty,X_k}$;
7: **let** $\eta_k := \|s_k^* - f\|_{\infty,X_k}$;
8: **compute** $\rho_k = \|s_k^* - f\|_\infty$;
9: **if** $\rho_k \le \eta_k$ **then return** s_k^* ▷ best approximation found
10: **else**
11: **find** reference set $X_{k+1} = (x_1^{(k+1)}, \ldots, x_{n+1}^{(k+1)}) \in [a,b]^{n+1}$ satisfying
12: • $|(s_k^* - f)(x^*)| = \rho_k$ for some $x^* \in X_{k+1}$;
13: • $|(s_k^* - f)(x)| \ge \eta_k$ for all $x \in X_{k+1}$;
14: • $\operatorname{sgn}((s_k^* - f)(x_j^{(k+1)})) = (-1)^j \sigma_k$
15: for $1 \le j \le n+1$ and some $\sigma_k \in \{\pm 1\}$. ▷ alternation condition
16: **end if**
17: **end for**
18: **end function**

Remark 5.40. We remark that the construction of the reference set X_{k+1} in line 11 of Algorithm 9 can be accomplished by a Remez exchange step, Algorithm 8. In this case, all three conditions in lines 12-15 of Algorithm 9 are satisfied, according to Remark 5.39. □

Next, we analyze the convergence of the Remez algorithm. To this end, we first remark that at any step k in the Remez algorithm, we have

- a current reference set $X_k = (x_1^{(k)}, \dots, x_{n+1}^{(k)}) \subset [a,b]^{n+1}$,
- *alternating* signs $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \dots, \varepsilon_{n+1}^{(k)}) \in \{\pm 1\}^{n+1}$,
- and *positive* coefficients $\lambda^{(k)} = (\lambda_1^{(k)}, \dots, \lambda_{n+1}^{(k)})^T \in \Lambda_{n+1}$,

so that the dual functional $\varphi : [a,b] \longrightarrow \mathbb{R}$, defined as

$$\varphi(u) = \sum_{j=1}^{n+1} \lambda_j^{(k)} \varepsilon_j^{(k)} u(x_j^{(k)}) \qquad \text{for } u \in \mathscr{C}[a,b],$$

satisfies the characterization (5.29) in the alternation theorem, Theorem 5.34. In particular, by the alternation theorem, the following properties hold.

- $X_k \subset E_{s_k^* - f}$;
- $\varepsilon_j^{(k)} = \operatorname{sgn}((s_k^* - f)(x_j^{(k)})) = (-1)^j \sigma_k$ for all $1 \le j \le n+1$ with $\sigma_k \in \{\pm 1\}$;
- $\varepsilon_j^{(k)} (s_k^* - f)(x_j^{(k)}) = \|s_k^* - f\|_{\infty, X_k} = \eta_k$ for all $1 \le j \le n+1$;
- $\varphi(s) = 0$ for all $s \in \mathcal{S}$.

Now we prove the monotonicity of the minimal distances η_k.

Proposition 5.41. *Let the assumptions from the Remez algorithm be satisfied. Then, for any step $k \in \mathbb{N}_0$, where the Remez iteration does not terminate, we have the monotonicity of the minimal distances,*

$$\eta_{k+1} > \eta_k.$$

Proof. The representation

$$\begin{aligned}
\eta_{k+1} &= \sum_{j=1}^{n+1} \lambda_j^{(k+1)} \varepsilon_j^{(k+1)} (s_{k+1}^* - f)(x_j^{(k+1)}) \\
&= \sum_{j=1}^{n+1} \lambda_j^{(k+1)} \varepsilon_j^{(k+1)} (s_k^* - f)(x_j^{(k+1)}) \\
&= \sum_{j=1}^{n+1} \lambda_j^{(k+1)} |(s_k^* - f)(x_j^{(k+1)})|
\end{aligned}$$

holds. Moreover, we have

$$\varepsilon_j^{(k+1)} = \operatorname{sgn}(s_k^* - f)(x_j^{(k+1)}) \qquad \text{for all } 1 \le j \le n+1$$

after the Remez exchange (see Algorithm 9, line 14).

Moreover, we have $|(s_k^* - f)(x_j^{(k+1)})| \ge \eta_k$, for all $1 \le j \le n+1$ (cf. line 13), and there is one index $j^* \in \{1, \dots, n+1\}$ (cf. line 12) satisfying

$$|(s_k^* - f)(x_{j^*}^{(k+1)})| = \rho_k = \|s_k^* - f\|_\infty.$$

But this implies

$$\eta_{k+1} \geq \lambda_{j^*}^{(k+1)}\rho_k + (1 - \lambda_{j^*}^{(k+1)})\eta_k > \lambda_{j^*}^{(k+1)}\eta_k + (1 - \lambda_{j^*}^{(k+1)})\eta_k = \eta_k, \quad (5.36)$$

which already completes our proof. ■

Next, we show that the coefficients $\lambda_j^{(k)}$ are uniformly bounded away from zero.

Lemma 5.42. *Let $f \in \mathscr{C}[a,b]\backslash\mathcal{S}$. Then, under the assumptions of the Remez algorithm, the uniform bound*

$$\lambda_j^{(k)} \geq \alpha > 0 \qquad \text{for all } 1 \leq j \leq n+1 \text{ and all } k \in \mathbb{N}_0,$$

holds for some $\alpha > 0$ which is independent of $1 \leq j \leq n$ and $k \in \mathbb{N}_0$.

Proof. We have

$$\eta_k = -\sum_{j=1}^{n+1} \lambda_j^{(k)}\varepsilon_j^{(k)} f(x_j^{(k)}) = \sum_{j=1}^{n+1} \lambda_j^{(k)}\varepsilon_j^{(k)}(s_k^* - f)(x_j^{(k)}) \geq \eta_0 = \|s^* - f\|_{\infty,X_0}.$$

Suppose the statement is false. Then, there are sequences of reference sets $(X_k)_k$, signs $(\varepsilon^{(k)})_k$, and coefficients $(\lambda^{(k)})_k$ satisfying

$$\eta_k = -\sum_{j=1}^{n+1} \lambda_j^{(k)}\varepsilon_j^{(k)} f(x_j^{(k)}) \geq \eta_0 > 0 \qquad \text{for all } k \in \mathbb{N}_0, \quad (5.37)$$

where one index $j^* \in \{1,\ldots,n+1\}$ satisfies $\lambda_{j^*}^{(k)} \longrightarrow 0$, for $k \to \infty$.

But the elements of the sequences $(X_k)_k$, $(\varepsilon^{(k)})_k$, and $(\lambda^{(k)})_k$ lie in compact sets, respectively. Therefore, there are convergent subsequences with

$$\begin{aligned} x_j^{(k_\ell)} &\longrightarrow x_j \in [a,b] \quad \text{for } \ell \to \infty, \\ \varepsilon_j^{(k_\ell)} &\longrightarrow \varepsilon_j \in \{\pm 1\} \quad \text{for } \ell \to \infty, \\ \lambda_j^{(k_\ell)} &\longrightarrow \lambda_j \in [0,1] \quad \text{for } \ell \to \infty, \end{aligned}$$

for all $1 \leq j \leq n+1$, where $\lambda_{j^*} = 0$ for one index $j^* \in \{1,\ldots,n+1\}$.

Now we regard an interpolant $s \in \mathcal{S}$ satisfying $s(x_j) = f(x_j)$ for all $1 \leq j \leq n+1$, $j \neq j^*$. Then, we have

$$\begin{aligned} \eta_{k_\ell} &= \sum_{j=1}^{n+1} \lambda_j^{(k_\ell)}\varepsilon_j^{(k_\ell)}(s_{k_\ell}^* - f)(x_j^{(k_\ell)}) = \sum_{j=1}^{n+1} \lambda_j^{(k_\ell)}\varepsilon_j^{(k_\ell)}(s - f)(x_j^{(k_\ell)}) \\ &= \sum_{\substack{j=1 \\ j\neq j^*}}^{n+1} \lambda_j^{(k_\ell)}\varepsilon_j^{(k_\ell)}(s - f)(x_j^{(k_\ell)}) + \lambda_{j^*}^{(k_\ell)}\varepsilon_{j^*}^{(k_\ell)}(s - f)(x_{j^*}^{(k_\ell)}) \\ &\longrightarrow \sum_{\substack{j=1 \\ j\neq j^*}}^{n+1} \lambda_j\varepsilon_j(s - f)(x_j) + \lambda_{j^*}\varepsilon_{j^*}(s - f)(x_{j^*}) = 0 \quad \text{for } \ell \to \infty. \end{aligned}$$

But this is in contradiction to (5.37). ■

Now we can finally prove convergence for the Remez agorithm.

Theorem 5.43. *Either the Remez algorithm, Algorithm 9, terminates after $k \in \mathbb{N}$ steps with returning the best approximation $s_k^* = s^*$ to $f \in \mathscr{C}[a,b]$ or the Remez algorithm generates convergent sequences of minimal distances $(\eta_k)_k$ and best approximations $(s_k^*)_k$ with limit elements*

$$\lim_{k\to\infty} \eta_k = \eta = \|s^* - f\|_\infty \qquad \text{and} \qquad \lim_{k\to\infty} s_k^* = s^* \in \mathcal{S},$$

where $s^* \in \mathcal{S}$ is the strongly unique best approximation to $f \in \mathscr{C}[a,b]$ with minimal distance η. The sequence $(\eta_k)_k$ of minimal distances converges linearly to η by the contraction

$$\eta - \eta_{k+1} < \theta(\eta - \eta_k) \qquad \text{for some } \theta \in (0,1). \tag{5.38}$$

Proof. Let $f \in \mathscr{C}[a,b] \setminus \mathcal{S}$ (for $f \in \mathcal{S}$ the statement is trivial).

If the Remez algorithm terminates after $k \in \mathbb{N}$ steps, in line 9 of Algorithm 9, with $s_k^* \in \mathcal{S}$, then $s_k^* = s^*$ is, according to the alternation theorem, Theorem 5.34, the strongly unique best approximation to f.

Now suppose the Remez algorithm does not terminate after finitely many steps. For this case, we first show the contraction property (5.38).

By the estimate in (5.36),

$$\eta_{k+1} \geq \lambda_{j^*}^{(k+1)} \rho_k + (1 - \lambda_{j^*}^{(k+1)})\eta_k \tag{5.39}$$

and from $\rho_k = \|s_k^* - f\|_\infty > \|s^* - f\|_\infty = \eta > 0$ it follows that

$$\eta_{k+1} > \lambda_{j^*}^{(k+1)} \eta + (1 - \lambda_{j^*}^{(k+1)})\eta_k$$

and so

$$\eta - \eta_{k+1} < (1 - \lambda_{j^*}^{(k+1)})(\eta - \eta_k).$$

By Lemma 5.42, there is one $\alpha > 0$ satisfying $\lambda_j^{(k+1)} \geq \alpha$, for all $1 \leq j \leq n+1$ and all $k \in \mathbb{N}_0$. Therefore, the stated contraction (5.38) holds for $\theta = 1 - \alpha \in (0,1)$. From this, we get the estimate

$$\eta - \eta_k < \theta^k(\eta - \eta_0) \qquad \text{for all } k \in \mathbb{N}_0$$

by induction on k. Therefore, the sequence $(\eta_k)_k$ of minimal distances is convergent with limit element η, i.e., $\eta_k \longrightarrow \eta$, for $k \to \infty$.

From estimate (5.39), we can conclude

$$\rho_k \leq \frac{\eta_{k+1} - \eta_k}{\lambda_{j^*}^{(k+1)}} + \eta_k < \frac{\eta_{k+1} - \eta_k}{1 - \theta} + \eta_k,$$

and this gives the estimates

$$\eta_k < \rho_k < \frac{\eta_{k+1} - \eta_k}{1 - \theta} + \eta_k.$$

This implies the convergence of the distances ρ_k to η, i.e.,

$$\lim_{k\to\infty} \rho_k = \lim_{k\to\infty} \|s_k^* - f\|_\infty = \|s^* - f\|_\infty = \eta.$$

We can conclude that the sequence $(s_k^*)_k \subset \mathcal{S}$ of the strongly unique best approximations to f on X_k converges to the strongly unique best approximation s^* to f. ∎

Finally, we discuss one important observation. We note that for the approximation of *strictly convex* functions $f \in \mathscr{C}[a,b]$ by linear polynomials, the Remez algorithm may return the best approximation $s^* \in \mathcal{P}_1$ to f after only one step.

Proposition 5.44. *Let $f \in \mathscr{C}[a,b]$ be a strictly convex function on a compact interval $[a,b]$ and $\mathcal{S} = \mathcal{P}_1$. Moreover, let $X_0 = (a, x_0, b)$, for $x_0 \in (a,b)$, be an initial reference set for the Remez algorithm. Then, the Remez algorithm terminates after at most one Remez exchange.*

Proof. Regard $s \in \mathcal{P}_1$ in its monomial representation $s(x) = m \cdot x + c$ for $m, c \in \mathbb{R}$. Then, we have for $x, y \in [a,b]$, $x \neq y$, and $\lambda \in (0,1)$ the strict inequality

$$\begin{aligned}
&(f-s)(\lambda x + (1-\lambda)y) \\
&= f(\lambda x + (1-\lambda)y) - m \cdot (\lambda x + (1-\lambda)y) - c \\
&< \lambda f(x) + (1-\lambda) f(y) - m \cdot (\lambda x + (1-\lambda)y) - c \\
&= \lambda f(x) - \lambda m x - \lambda c + (1-\lambda) f(y) - (1-\lambda) m y - (1-\lambda) c \\
&= \lambda (f-s)(x) + (1-\lambda)(f-s)(y)
\end{aligned}$$

by the strict convexity of f, i.e., $f - s$ is also strictly convex.

Now let $s^* \in \mathcal{P}_1$ be the strongly unique best approximation to f. Due to the alternation theorem, Theorem 5.34, the error function $f - s^*$ has at least *three* extremal points with alternating signs in $[a,b]$. Since $f - s^*$ is strictly convex and continuous, $f - s^*$ has exactly one global minimum x^* on (a,b). Moreover, two global maxima of $f - s^*$ are at the boundary of $[a,b]$, i.e., we have $\{a,b\} \subset E_{s^*-f}$ with

$$(f - s^*)(a) = \|f - s^*\|_\infty = (f - s^*)(b).$$

From the representation $s^*(x) = m^* \cdot x + c^*$, we obtain the slope

$$m^* = \frac{f(b) - f(a)}{b - a} = [a,b](f).$$

Let $s_0^* \in \mathcal{P}_1$ be the best approximation to f with respect to $X_0 = (a, x_0, b)$. Then, according to the alternation theorem, we have

$$(f - s_0^*)(a) = \sigma \|f - s_0^*\|_{\infty, X_0} = (f - s_0^*)(b) \qquad \text{for some } \sigma \in \{\pm 1\},$$

whereby the representation $s_0^*(x) = m_0 \cdot x + c_0$ implies $m_0 = [a, b](f) = m^*$, i.e., s^* and s_0^* differ by at most one constant.

If $x_0 \in E_{s^* - f}$, then $x_0 = x^*$, and the best approximation s^* to f is already found by s_0^*, due to the alternation theorem. In this case, the Remez algorithm terminates immediately with returning $s^* = s_0^*$.

If $x_0 \notin E_{s^* - f}$, then the Remez algorithm selects the unique global minimum $x^* \neq x_0$ of $f - s^*$ for the exchange with x_0: Since s^* and s_0^* differ by at most a constant, x^* is also the unique global minimum of $f - s_0^*$ on (a, b), i.e., we have

$$(f - s_0^*)(x^*) < (f - s_0^*)(x) \qquad \text{for all } x \in [a, b], \text{ where } x \neq x^*. \tag{5.40}$$

By the strict convexity of $f - s_0^*$, we can further conclude the strict inequality

$$(f - s_0^*)(x^*) < (f - s_0^*)(x_0) < 0,$$

or,

$$\rho_0 = \|f - s_0^*\|_\infty = |(f - s_0^*)(x^*)| > |(f - s_0^*)(x_0)| = \|f - s_0^*\|_{\infty, X_0} = \eta_0. \tag{5.41}$$

By (5.40) and (5.41), the point x^* is the unique global maximum of $|f - s_0^*|$ on $[a, b]$. Therefore, x^* is the only candidate for the required Remez exchange (in line 5 of Algorithm 8) for x_0. After the execution of the Remez exchange, we have $X_1 = (a, x^*, b)$, so that the Remez algorithm immediately terminates with returning $s_1^* = s^*$. ■

For further illustration, we make an example linked with Example 5.37.

Example 5.45. We approximate the strictly convex exponential function $f(x) = \exp(x)$ on the interval $[0, 2]$ by linear polynomials, i.e., $\mathcal{F} = \mathscr{C}[0, 2]$ and $\mathcal{S} = \mathcal{P}_1$. We take $X_0 = (0, 1, 2)$ as the initial reference set in the Remez algorithm, Algorithm 9. According to Example 5.37,

$$s_0^*(x) = 1 - \left(\frac{e - 1}{2}\right)^2 + \frac{e^2 - 1}{2} x$$

is the unique best approximation to f from $\mathcal{P}_1$ with respect to $\|\cdot\|_{\infty, X_0}$, with minimal distance $|\eta_0| = (e - 1)^2/4 \approx 0.7381$, where e is the Euler number.

The error function $|s_0^*(x) - \exp(x)|$ attains on $[0, 2]$ its unique maximum $\rho_0 = \|s_0^* - \exp\|_\infty \approx 0.7776$ at $x^* = \log((e^2 - 1)/2) > 1$. We have $\rho_0 > \eta_0$, and so one Remez exchange leads to the new reference set $X_1 = (0, x^*, 2)$.

According to Proposition 5.44, the Remez algorithm returns already after the next iteration the best approximation s_1^* to f.

Finally, we compute s_1^*, the best approximation to f for the reference set $X_1 = (0, x^*, 2)$. To this end, we proceed as in Example 5.37, where we first determine the required divided differences for f and $\varepsilon = (-1, 1, -1)$ by using the recursion in Theorem 2.14:

X	f_X		
0	1		
x^*	$\frac{e^2-1}{2}$	$\frac{e^2-3}{2x^*}$	
2	e^2	$\frac{e^2+1}{2(2-x^*)}$	$\frac{(e^2-1)(x^*-1)+2}{2(2-x^*)x^*}$

X	ε_X		
0	-1		
x^*	1	$\frac{2}{x^*}$	
2	-1	$-\frac{2}{2-x^*}$	$-\frac{2}{(2-x^*)x^*}$

From this we compute the minimal distance $\|s_1^* - f\|_{\infty,X_1} = -\eta_1 \approx 0.7579$ by

$$\eta_1 = -\frac{1}{4}\left[(e^2-1)(x^*-1)+2\right]$$

and the best approximation to f from $\mathcal{P}_1$ with respect to $\|\cdot\|_{\infty,X_1}$ by

$$s_1^* = [0](f - \eta_X\varepsilon) + [0, x^*](f - \eta_X\varepsilon)x = 1 + \eta_1 + \frac{e^2 - 4\eta_1 - 3}{2x^*}x.$$

By Proposition 5.44, the Remez algorithm terminates with the reference set $X_1 = E_{s_1^*-f}$, so that by $s_1^* \in \mathcal{P}_1$ the unique best approximation to f with respect to $\|\cdot\|_\infty$ is found. Figure 5.5 shows the best approximations $s_j^* \in \mathcal{P}_1$ to f for the reference sets X_j, for $j = 0, 1$. ♢

5.5 Exercises

Exercise 5.46. Let $\mathcal{F} = \mathscr{C}[-1, 1]$ be equipped with the maximum norm $\|\cdot\|_\infty$. Moreover, let $f \in \mathcal{P}_3 \setminus \mathcal{P}_2$ be a cubic polynomial, i.e., has the form

$$f(x) = a\,x^3 + b\,x^2 + c\,x + d \quad \text{for } x \in [-1, 1]$$

with coefficients $a, b, c, d \in \mathbb{R}$, where $a \neq 0$.

(a) Compute a best approximation $p_2^* \in \mathcal{P}_2$ to f from $\mathcal{P}_2$ w.r.t. $\|\cdot\|_\infty$.
(b) Is the best approximation p_2^* from (a) unique?

Exercise 5.47. Let $P_\infty : \mathscr{C}[a, b] \longrightarrow \mathcal{P}_n$ denote the operator, which assigns every $f \in \mathscr{C}[a, b]$ to its best approximation $p_\infty^*(f) \in \mathcal{P}_n$ from $\mathcal{P}_n$ w.r.t. $\|\cdot\|_\infty$, i.e.,

$$P_\infty(f) = p_\infty^*(f) \qquad \text{for } f \in \mathscr{C}[a, b].$$

(a) Show that P_∞ is well-defined.
(b) Is P_∞ linear or non-linear?

Exercise 5.48. For a compact interval $[a,b] \subset \mathbb{R}$, let $\mathcal{F} = \mathscr{C}[a,b]$ be equipped with the maximum norm $\|\cdot\|_\infty$. Moreover, let $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_{n-1}$, for $n \in \mathbb{N}$. Then, there is a strongly unique best approximation $p^* \in \mathcal{P}_{n-1}$ to f from $\mathcal{P}_{n-1}$ w.r.t. $\|\cdot\|_\infty$ and an alternation set $X = (x_1, \ldots, x_{n+1}) \in E^{n+1}_{s^*-f}$ for s^* and f (see Corollary 5.4). For the dual characterization of the best approximation $p^* \in \mathcal{P}_{n-1}$ we use, as in (5.6), a linear functional $\varphi \in \mathcal{F}'$ of the form

$$\varphi(u) = \sum_{k=1}^{n+1} \lambda_k \, \varepsilon_k \, u(x_k) \qquad \text{for } u \in \mathscr{C}[a,b]$$

with coefficients $\lambda = (\lambda_1, \ldots, \lambda_{n+1})^T \in \Lambda_{n+1}$ and alternating signs

$$\varepsilon_k = \operatorname{sgn}(p^* - f)(x_k) = \sigma\,(-1)^k \qquad \text{for } k = 1, \ldots, n+1.$$

By these assumptions on φ, two conditions of the dual characterization (according to Theorem 3.48) are already satisfied, that is (a) $\|\varphi\|_\infty = 1$ and (b) $\varphi(p^* - f) = \|p^* - f\|_\infty$.

Now this problem is concerning condition (c) of the dual characterization in Theorem 3.48. To this end, consider using divided differences (cf. Definition 2.10) to construct, from given alternation points

$$a \le x_1 < \ldots < x_n \le b$$

and $\sigma \in \{\pm 1\}$, a coefficient vector $\lambda = (\lambda_1, \ldots, \lambda_{n+1})^T \in \Lambda_{n+1}$ satisfying

$$\varphi(p) = 0 \qquad \text{for all } p \in \mathcal{P}_{n-1}.$$

Exercise 5.49. Let $\mathcal{F} = \mathscr{C}[0, 2\pi]$ and $\mathcal{S} = \mathcal{P}_1$. Moreover, for $n \in \mathbb{N}$, let

$$f_n(x) = \sin(nx) \qquad \text{for } x \in [0, 2\pi].$$

(a) Compute the unique best approximation $s_n^* \in \mathcal{P}_1$ to f_n w.r.t. $\|\cdot\|_\infty$.
(b) How many alternation points occur for the error function $s_n^* - f_n$ in (a)? But should not there only be *three* alternation points?

Exercise 5.50. Let $\mathcal{F} = \mathscr{C}[-2,1]$ be equipped with the maximum norm $\|\cdot\|_\infty$. Compute the unique best approximation $p^* \in \mathcal{P}_2$ from $\mathcal{P}_2$ to the function $f \in \mathscr{C}[-2,1]$, defined as

$$f(x) = |x+1| \quad \text{for } x \in [-2,1]$$

with respect to the maximum norm $\|\cdot\|_\infty$.

Moreover, determine the set of extremal points $X = E_{p^*-f}$, along with a constant $K > 0$ satisfying

$$\|p - f\|_{\infty,X} \ge \|p^* - f\|_\infty + K \cdot \|p - p^*\|_{\infty,X} \qquad \text{for all } p \in \mathcal{P}_2.$$

Plot the graphs of f and the best approximation p^* to f in one figure.

Exercise 5.51. Let $\mathcal{F} = \mathscr{C}[0,2]$ be equipped with the maximum norm $\|\cdot\|_\infty$. Determine the strongly unique best approximation $p^* \in \mathcal{P}_1$ from $\mathcal{P}_1$ to the function $f \in \mathscr{C}[0,2]$, defined as

$$f(x) = \exp\bigl(-(x-1)^2\bigr) \qquad \text{for } x \in [0,2]$$

with respect to the maximum norm $\|\cdot\|_\infty$.

Moreover, determine a constant $K > 0$ satisfying

$$\|p - f\|_\infty - \|p^* - f\|_\infty \geq K \cdot \|p - p^*\|_\infty \qquad \text{for all } p \in \mathcal{P}_1.$$

Use this inequality to conclude the uniqueness of the best approximation $p^* \in \mathcal{P}_1$ yet once more.

Exercise 5.52. Let $\mathcal{S} \subset \mathscr{C}[a,b]$ be a Haar space with $\dim(\mathcal{S}) = n+1 \in \mathbb{N}$.

Prove the *Haar condition*: If $s \in \mathcal{S} \setminus \{0\}$ has on the interval $[a,b]$ exactly m zeros from which k zeros are without sign change, then we have $m+k \leq n$.

Exercise 5.53. In this problem, let $I \subset \mathbb{R}$ be a compact set containing sufficiently many points, respectively. Analyze whether or not the following function systems $\mathcal{H} = (s_1, \ldots, s_n) \in (\mathscr{C}(I))^n$ are a Haar system on I.

(a) $\mathcal{H} = (x, 1/x)$ for $I \subset (0,\infty)$.
(b) $\mathcal{H} = (1/(x-c_0), 1/(x-c_1))$ for $I \subset \mathbb{R} \setminus \{c_0, c_1\}$, where $c_0 \neq c_1$.
(c) $\mathcal{H} = (1, x^2, x^4, \ldots, x^{2n})$ for $I = [-1,1]$.
(d) $\mathcal{H} = (1, x, \ldots, x^n, g(x))$ for a compact interval $I = [a,b]$, where $g \in \mathscr{C}^{n+1}[a,b]$ with $g^{(n+1)} \geq 0$ and $g^{(n+1)} \not\equiv 0$ on $[a,b]$.

Exercise 5.54. For $n \in \mathbb{N}_0$, let $\mathcal{T}_n^e$ be the linear space of all *even* real-valued trigonometric polynomials of degree at most n, and let $\mathcal{T}_n^o$ be the linear space of all *odd* real-valued trigonometric polynomials of degree at most n.

(a) Show that $\mathcal{T}_n^e$ is a Haar space on the interval $[0,\pi)$.
(b) Determine the dimension of $\mathcal{T}_n^e$.
(c) Is $\mathcal{T}_n^o$ a Haar space on the interval $[0,\pi)$?
(d) Is $\mathcal{T}_n^o$ a Haar space on the open interval $(0,\pi)$?
(e) Determine the dimension of $\mathcal{T}_n^o$.

Exercise 5.55. Prove the following results.

(a) The functions

$$s_0(x) = 1, \quad s_1(x) = x\cos(x), \quad s_2(x) = x\sin(x)$$

are a Haar system on $[0,\pi]$.
(b) There is *no* two-dimensional subspace of

$$\mathcal{S} = \operatorname{span}\{s_0, s_1, s_2\} \subset \mathscr{C}[0,\pi],$$

which is a Haar space on $[0,\pi]$.

Exercise 5.56. For $n \in \mathbb{N}_0$, let $\mathcal{S} \subset \mathscr{C}[a,b]$ be a $(n+1)$-dimensional linear subspace of $\mathscr{C}[a,b]$. Moreover, let $\mathcal{S}$ satisfy the *weak Haar condition* on $[a,b]$, according to which any $s \in \mathcal{S}$ has at most n sign changes in $[a,b]$.

Prove the following statements for $f \in \mathscr{C}[a,b]$.

(a) If there is an alternation set for $s \in \mathcal{S}$ and f of length $n+2$, so that there are $n+2$ pairwise distinct alternation points $a \le x_0 < \ldots < x_{n+1} \le b$ and one sign $\sigma \in \{\pm 1\}$ satisfying

$$(s-f)(x_k) = \sigma\,(-1)^k\,\|s-f\|_\infty \qquad \text{for all } k = 0,\ldots,n+1,$$

then s is a best approximation to f from $\mathcal{S}$ with respect to $\|\cdot\|_\infty$.

(b) The converse of statement (a) is false (for the general case).

Exercise 5.57. Let $\mathcal{F} = \mathscr{C}[a,b]$ and $\mathcal{S} \subset \mathcal{F}$ be a Haar space on $[a,b]$ of dimension $n+1$ containing the constant functions. Moreover, let $f \in \mathcal{F} \setminus \mathcal{S}$, such that

$$\operatorname{span}\big\{\mathcal{S} \cup \{f\}\big\}$$

is a Haar space on $[a,b]$. Finally, $s^* \in \mathcal{S}$ be the unique best approximation to f from $\mathcal{S}$ with respect to $\|\cdot\|_\infty$.

Show that the error function $f - s^*$ has exactly $n+2$ extremal points

$$a = x_0 < \ldots < x_{n+1} = b,$$

where $f - s^*$ is strictly monotone between neighbouring extremal points.

Exercise 5.58. In this programming exercise, we wish to compute for any $n \in \mathbb{N}$ the strongly unique best approximation $p^* \in \mathcal{P}_{n-1}$ to $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_{n-1}$ from $\mathcal{P}_{n-1}$ w.r.t. $\|\cdot\|_{\infty,X}$ on a point set $X = (x_1,\ldots,x_{n+1}) \in [a,b]^{n+1}$, so that

$$\|p^* - f\|_{\infty,X} < \|p - f\|_{\infty,X} \qquad \text{for all } p \in \mathcal{P}_{n-1} \setminus \{p^*\}.$$

To this end, implement a function called `mybestpoly` with header

```
[alpha,eta] = mybestpoly(f,X),
```

which returns on input point set X (of length $|X| = n+1$) the Newton coefficients $\alpha = (\alpha_0,\ldots,\alpha_{n-1}) \in \mathbb{R}^n$ of the best approximation

$$p^*(x) = \sum_{k=0}^{n-1} \alpha_k\,\omega_k(x) \quad \text{where } \omega_k(x) = \prod_{j=1}^{k}(x - x_j) \in \mathcal{P}_k \text{ for } 0 \le k \le n-1$$

to f w.r.t. $\|\cdot\|_{\infty,X}$, along with the minimal distance $\eta_X = \|f - s^*\|_{\infty,X}$.

Exercise 5.59. To *efficiently* evaluate the best approximation $p^* \in \mathcal{P}_{n-1}$ from Exercise 5.58, we use the *Horner*[10] *scheme* (a standard numerical method, see e.g. [28, Section 5.3.3].

[10] William George Horner (1786-1837), English mathematician

To this end, implement a function called `mynewtonhorner` with header

```
[p] = mynewtonhorner(X,alpha,x),
```

which returns on an input point set $X = \{x_1, \ldots, x_{n+1}\} \subset [a,b]$, Newton coefficients $\alpha = (\alpha_0, \ldots, \alpha_{n-1}) \in \mathbb{R}^n$ and $x \in \mathbb{R}$ the value

$$p(x) = \sum_{k=0}^{n-1} \alpha_k \, \omega_k(x) \in \mathcal{P}_{n-1},$$

where the evaluation of p at x should rely on the Horner scheme.

Exercise 5.60. Implement the Remez exchange, Algorithm 8. To this end, write a function called `myremezexchange` with header

```
[X] = myremezexchange(X,epsilon,x),
```

which returns, on input reference set $X = (x_1, \ldots, x_{n+1}) \in [a,b]^{n+1}$, an extremal point $x = x^* \in [a,b] \setminus X$ satisfying

$$|(p^* - f)(x^*)| = \|p^* - f\|_\infty$$

and a sign vector $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \{\pm 1\}^2$ satisfying

$$\varepsilon_1 = \operatorname{sgn}(p^* - f)(x_1) \quad \text{and} \quad \varepsilon_2 = \operatorname{sgn}(p^* - f)(x^*)$$

the updated reference set X_+ (as output by Algorithm 8), i.e.,

$$X_+ = (X \setminus \{x_j\}) \cup \{x^*\} \quad \text{for one } 1 \le j \le n+1.$$

Exercise 5.61. Implement the Remez algorithm, Algorithm 9. To this end, write a function `myremez` with header

```
[alpha,eta,X,its] = myremez(f,X),
```

which returns, on input function $f \in \mathscr{C}[a,b] \setminus \mathcal{P}_{n-1}$ and an initial reference set $X = (x_1, \ldots, x_{n+1}) \in [a,b]^{n+1}$, the Newton coefficients $\alpha = (\alpha_0, \ldots, \alpha_{n-1})$ of the (strongly unique) best approximation $p^* \in \mathcal{P}_{n-1}$ to f from $\mathcal{P}_{n-1}$ w.r.t. $\|\cdot\|_\infty$, the minimal distance $\eta = \|p^* - f\|_\infty$, a set of alternation points $X \subset E_{p^*-f}$, and the number `its` of the performed Remez iterations. For your implementation, use the functions `mybestpoly` (from Exercise 5.58), `mynewtonhorner` (Exercise 5.59) and `myremezexchange` (Exercise 5.60).

Verify your function `myremez` by using the following examples.

(a) $f(x) = \sqrt[3]{x}$, $[a,b] = [0,1]$, $X = \left(0, \frac{1}{2}, \frac{3}{4}, 1\right)$;
(b) $f(x) = \sin(5x) + \cos(6x)$, $[a,b] = [0,\pi]$, $X = \left(0, \frac{1}{2}, \frac{3}{2}, \frac{5}{2}, \pi\right)$.

Exercise 5.62. Analyze for the case $\mathcal{S} = \mathcal{P}_{n-1}$ the asymptotic computational complexity for only *one* iteration of the Remez algorithm, Algorithm 9.

(a) Determine the costs for the minimal distance $\eta_k = \|s_k^* - f\|_{\infty,X_k}$.
Hint: Use divided differences (according to Proposition 5.35).
(b) Determine the costs for computing the Newton coefficients of s_k^*.
Hint: Reuse the divided differences from (a).
(c) Sum up the required asymptotic costs in (a) and (b).

How do you *efficiently* compute the update η_{k+1} from information that is required to compute η_k?

Exercise 5.63. Assuming the notations of the *Remez algorithm*, Algorithm 9, we consider the (global) distance

$$\rho_k = \|s_k^* - f\|_\infty \qquad \text{for } k \in \mathbb{N}_0$$

between $f \in \mathscr{C}[a,b]$ and the current best approximation $s_k^* \in \mathcal{S}$ to f, for the current reference set $X_k = (x_1^{(k)}, \ldots, x_{n+1}^{(k)}) \in [a,b]^{n+1}$ and w.r.t. $\|\cdot\|_{\infty,X_k}$.

Show that the sequence $(\rho_k)_{k\in\mathbb{N}_0}$ is *not necessarily* strictly increasing. To this end, construct a simple (but non-trivial) counterexample.

6 Asymptotic Results

In this chapter, we prove *asymptotic* statements to quantify the convergence behaviour of both *algebraic* and *trigonometric* approximation by partial sums.

For the trigonometric case, the analysis of Fourier partial sums plays a central role. Recall that we have studied Fourier partial sums,

$$(F_n f)(x) = \frac{(f,1)}{2} + \sum_{j=1}^{n} \left[(f, \cos(j\cdot)) \cos(jx) + (f, \sin(j\cdot)) \sin(jx) \right],$$

for $f \in \mathscr{C}_{2\pi}$, already in Chapter 4: According to Corollary 4.12, $F_n f$ is the unique best approximation to f from the linear space $\mathcal{T}_n$ of trigonometric polynomials of degree at most $n \in \mathbb{N}_0$ with respect to the *Euclidean* norm $\|\cdot\|$.

As we proceed in this chapter, we will analyze the asymptotic behaviour of the minimal distances with respect to both the Euclidean norm $\|\cdot\|$,

$$\eta(f, \mathcal{T}_n) := \inf_{T \in \mathcal{T}_n} \|T - f\| = \|F_n f - f\| \qquad \text{for } n \to \infty,$$

and with respect to the maximum norm $\|\cdot\|_\infty$. To this end, we first show for *continuous* functions $f \in \mathscr{C}_{2\pi}$ convergence of $F_n f$ to f with respect to $\|\cdot\|$, and then we prove *convergence rates*, for $f \in \mathscr{C}^k_{2\pi}$, $k \in \mathbb{N}_0$, of the form

$$\eta(f, \mathcal{T}_n) = o(n^{-k}) \quad \text{for } n \to \infty.$$

Finally, we analyze the *uniform* convergence of Fourier partial sums, i.e., we study the asymptotic behaviour of the distances

$$\|F_n f - f\|_\infty \qquad \text{for } n \to \infty.$$

In this chapter, we prove the following classical results of approximation:

- **The Weierstrass theorem**, according to which any function $f \in \mathscr{C}_{2\pi}$ can, w.r.t. $\|\cdot\|_\infty$, be approximated arbitrarily well by trigonometric polynomials.
- **The Jackson inequalities**, which allow us to quantify the asymptotic behaviour of the minimal distances

$$\eta_\infty(f, \mathcal{T}_n) := \inf_{T \in \mathcal{T}_n} \|T - f\|_\infty \qquad \text{for } n \to \infty.$$

Likewise, we will also discuss the algebraic case for the approximation to $f \in \mathscr{C}[a,b]$ by partial sums $P_n f$ from $\mathcal{P}_n$.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_6

6.1 The Weierstrass Theorem

We analyze the following two fundamental questions of approximation:

Question 1: Can we approximate any function $f \in \mathscr{C}[a,b]$ on a compact interval $[a,b] \subset \mathbb{R}$ with respect to $\|\cdot\|_\infty$ arbitrarily well by algebraic polynomials?

Question 2: Can we approximate any continuous 2π-periodic function $f \in \mathscr{C}_{2\pi}$ with respect to $\|\cdot\|_\infty$ arbitrarily well by trigonometric polynomials?

Not too surprisingly, the two questions are related. In fact, a positive answer to both questions was given already in 1885 by Weierstrass[1], who also discovered the intrinsic relation between the two problems of these questions. As we show in this section, the answer to the trigonometric case (Question 2) can be concluded from the solution for the algebraic case (Question 1). The solutions given by Weierstrass were celebrated as *the birth of approximation*.

In the following discussion, we will be more precise about the above two questions. To this end, we need only a few preparations.

Definition 6.1. *Let $\mathcal{F}$ be a normed linear space with norm $\|\cdot\|$. Then a subset $\mathcal{S} \subset \mathcal{F}$ is said to lie* `dense` *in $\mathcal{F}$ with respect to $\|\cdot\|$, if there exists, for any $f \in \mathcal{F}$ and any $\varepsilon > 0$, an element $s \equiv s(f,\varepsilon) \in \mathcal{S}$ satisfying*

$$\|s - f\| < \varepsilon.$$

○

Now we can give a more concise formulation for the above two questions.

- Are the algebraic polynomials $\mathcal{P}$ dense in $\mathscr{C}[a,b]$ with respect to $\|\cdot\|_\infty$?
- Are the trigonometric polynomials $\mathcal{T}$ dense in $\mathscr{C}_{2\pi}$ with respect to $\|\cdot\|_\infty$?

Remark 6.2. If $\mathcal{S} \subset \mathcal{F}$ is dense in $\mathcal{F}$ with respect to $\|\cdot\|$, then the topological closure $\overline{\mathcal{S}}$ of $\mathcal{S}$ (with respect to $\|\cdot\|$) coincides with $\mathcal{F}$, i.e.,

$$\overline{\mathcal{S}} = \mathcal{F},$$

or, in other words: For any $f \in \mathcal{F}$, there is a convergent sequence $(s_n)_{n\in\mathbb{N}}$ in $\mathcal{S}$ with limit f, so that $\|s_n - f\| \longrightarrow 0$ for $n \to \infty$. □

Remark 6.3. For a linear subspace $\mathcal{S} \subset \mathcal{F}$, $\mathcal{S} \neq \mathcal{F}$, Definition 6.1 does only make sense, if $\mathcal{S}$ is *infinite-dimensional*. Otherwise, if $\mathcal{S} \neq \mathcal{F}$ is only *finite-dimensional*, then there is, according to Corollary 3.8, for any $f \in \mathcal{F} \setminus \mathcal{S}$ a best approximation $s^* \in \mathcal{S}$ to f at a *positive* minimal distance $\eta(f,\mathcal{S}) > 0$, i.e., f *cannot* be approximated arbitrarily well by elements from $\mathcal{S}$, since the closest distance between f and $\mathcal{S}$ is $\eta(f,\mathcal{S})$. In this case, $\mathcal{S}$ is *not* dense in $\mathcal{F}$. □

[1] Karl Weierstrass (1815-1897), German mathematician

Example 6.4. The set $\mathbb{Q}$ of rational numbers is dense in the set $\mathbb{R}$ of real numbers with respect to the absolute-value function $|\cdot|$. ♢

Now let us turn to the Weierstrass theorems, for which there exist many different proofs (see, e.g. [33]). Our constructive proof for the algebraic case of the Weierstrass theorem relies on a classical account via *Korovkin sequences*.

Definition 6.5. *A sequence* $(K_n)_{n\in\mathbb{N}}$ *of linear and monotone operators* $K_n : \mathscr{C}[a,b] \longrightarrow \mathscr{C}[a,b]$ *is called a* `Korovkin`[2] `sequence` *on* $\mathscr{C}[a,b]$*, if*

$$\lim_{n\to\infty} \|K_n p - p\|_\infty = 0 \qquad \textit{for all } p \in \mathcal{P}_2.$$

◯

To further explain the utilized terminology, we recall a standard characterization for monotone linear operators.

Remark 6.6. A linear operator $K : \mathscr{C}[a,b] \longrightarrow \mathscr{C}[a,b]$ is *monotone* on $\mathscr{C}[a,b]$, if and only if K is *positive* on $\mathscr{C}[a,b]$, i.e., the following two statements are equivalent.

(a) For any $f, g \in \mathscr{C}[a,b]$ satisfying $f \le g$, we have $Kf \le Kg$;
(b) For any $f \in \mathscr{C}[a,b]$ satisfying $f \ge 0$, we have $Kf \ge 0$;

where all inequalities in (a) and (b) are taken *pointwise* on $[a,b]$. □

Next, we study an important special case for a Korovkin sequence. To this end, we restrict ourselves to the continuous functions $\mathscr{C}[0,1]$ on the unit interval $[0,1]$. This is without loss of generality, since otherwise, i.e., for any other compact interval $[a,b] \subset \mathbb{R}$, we may apply the affine-linear mapping $x \longmapsto (x-a)/(b-a)$, for $x \in [a,b]$.

Now we consider the `Bernstein`[3] `polynomials`

$$\beta_j^{(n)}(x) = \binom{n}{j} x^j (1-x)^{n-j} \in \mathcal{P}_n \qquad \text{for } 0 \le j \le n. \tag{6.1}$$

Let us note a few elementary properties of Bernstein polynomials.

Remark 6.7. The Bernstein polynomials $\beta_0^{(n)}, \ldots, \beta_n^{(n)} \in \mathcal{P}_n$, for $n \in \mathbb{N}_0$,

(a) form a basis for the polynomial space $\mathcal{P}_n$,
(b) are positive on $[0,1]$, i.e., $\beta_j^{(n)}(x) \ge 0$ for all $x \in [0,1]$,
(c) are on $[0,1]$ a *partition of unity*, i.e.,

$$\sum_{j=0}^{n} \beta_j^{(n)}(x) = 1 \qquad \text{for all } x \in [0,1].$$

[2] Pavel Petrovich Korovkin (1913-1985), Russian mathematician
[3] Sergei Natanovich Bernstein (1880-1968), Russian mathematician

Note that property (c) holds by the binomial theorem, whereas properties (a) and (b) can be verified by elementary calculations (cf. Exercise 6.83). □

By using the Bernstein polynomials in (6.1) we can make an important example for monotone linear operators on $\mathscr{C}[0,1]$.

Definition 6.8. *For $n \in \mathbb{N}$, the* `Bernstein operator` *$B_n : \mathscr{C}[0,1] \longrightarrow \mathcal{P}_n$ is defined as*

$$(B_n f)(x) = \sum_{j=0}^{n} f(j/n)\beta_j^{(n)}(x) \qquad \text{for } f \in \mathscr{C}[0,1], \tag{6.2}$$

where $\beta_0^{(n)}, \ldots, \beta_n^{(n)} \in \mathcal{P}_n$ are the `Bernstein polynomials` *in (6.1).* ◯

The Bernstein operators B_n are obviously linear on $\mathscr{C}[0,1]$. By the positivity of the Bernstein polynomials $\beta_j^{(n)}$, Remark 6.7 (b), the Bernstein operators B_n are, moreover, positive (and therefore monotone) on $\mathscr{C}[0,1]$. We note yet another elementary property of the operators B_n.

Remark 6.9. The Bernstein operators $B_n : \mathscr{C}[0,1] \longrightarrow \mathcal{P}_n$ in (6.2) are bounded on $\mathscr{C}[0,1]$ with respect to $\|\cdot\|_\infty$, since for any $f \in \mathscr{C}[0,1]$, we have

$$\|B_n f\|_\infty = \left\| \sum_{j=0}^{n} f(j/n)\beta_j^{(n)}(x) \right\|_\infty \leq \|f\|_\infty \left\| \sum_{j=0}^{n} \beta_j^{(n)}(x) \right\|_\infty = \|f\|_\infty$$

and so

$$\|B_n f\|_\infty \leq \|f\|_\infty \qquad \text{for all } f \in \mathscr{C}[0,1].$$

In particular, by transferring the result of Theorem 3.45 from linear functionals to linear operators, we can conclude that the Bernstein operators $B_n : \mathscr{C}[0,1] \longrightarrow \mathcal{P}_n$ are continuous on $\mathscr{C}[0,1]$. □

Now we prove the Korovkin property for the Bernstein operators.

Theorem 6.10. *The sequence of Bernstein operators $B_n : \mathscr{C}[0,1] \longrightarrow \mathcal{P}_n$, for $n \in \mathbb{N}$, is a Korovkin sequence on $\mathscr{C}[0,1]$.*

Proof. The Bernstein operators B_n, $n \in \mathbb{N}$, reproduce linear polynomials. Indeed, on the one hand, we have $B_n 1 \equiv 1$, for all $n \in \mathbb{N}$, by the partition of unity, according to Remark 6.7 (c). On the other hand, we find for $p_1(x) = x$ the identity $B_n p_1 = p_1$, for any $n \in \mathbb{N}$, since we get

$$\begin{aligned}(B_n p_1)(x) &= \sum_{j=0}^{n} \frac{j}{n} \binom{n}{j} x^j (1-x)^{n-j} = \sum_{j=1}^{n} \binom{n-1}{j-1} x^j (1-x)^{n-j} \\ &= x \sum_{j=0}^{n-1} \binom{n-1}{j} x^j (1-x)^{n-j-1} = x.\end{aligned}$$

According to Definition 6.5, it remains to show the uniform convergence

$$\lim_{n\to\infty} \|B_n p_2 - p_2\|_\infty = 0$$

for the quadratic monomial $p_2(x) = x^2$. To this end, we apply the Bernstein operators B_n to the sequence of functions

$$f_n(x) = x^2 \frac{n}{n-1} - \frac{x}{n-1} \in \mathcal{P}_2 \qquad \text{for } n \geq 2,$$

where for $n \geq 2$ we have

$$\begin{aligned}
(B_n f_n)(x) &= \sum_{j=0}^{n} \binom{n}{j} \left(\frac{j^2}{n^2} \frac{n}{n-1} - \frac{j}{n(n-1)} \right) x^j (1-x)^{n-j} \\
&= \sum_{j=0}^{n} \frac{n!}{(n-j)!j!} \frac{j(j-1)}{n(n-1)} x^j (1-x)^{n-j} \\
&= \sum_{j=2}^{n} \frac{(n-2)!}{(n-j)!(j-2)!} x^j (1-x)^{n-j} \\
&= x^2 \sum_{j=0}^{n-2} \binom{n-2}{j} x^j (1-x)^{n-j-2} = p_2(x).
\end{aligned}$$

Together with the boundedness of the Bernstein operators B_n (according to Remark 6.9), this finally implies

$$\|B_n p_2 - p_2\|_\infty = \|B_n(p_2 - f_n)\|_\infty \leq \|p_2 - f_n\|_\infty,$$

whereby through $\|p_2 - f_n\|_\infty \longrightarrow 0$, for $n \to \infty$, the statement is proven. ■

The following result of Korovkin is of fundamental importance.

Theorem 6.11. (Korovkin, 1953). *For a compact interval $[a,b] \subset \mathbb{R}$, let $(K_n)_{n\in\mathbb{N}}$ be a Korovkin sequence on $\mathscr{C}[a,b]$. Then, we have*

$$\lim_{n\to\infty} \|K_n f - f\|_\infty = 0 \qquad \textit{for all } f \in \mathscr{C}[a,b]. \tag{6.3}$$

Proof. Suppose $f \in \mathscr{C}[a,b]$. Then, f is bounded on $[a,b]$, i.e., there is some $M > 0$ with $\|f\|_\infty \leq M$. Moreover, f is *uniformly continuous* on the compact interval $[a,b]$, i.e., for any $\varepsilon > 0$ there is some $\delta > 0$ satisfying

$$|x - y| < \delta \quad \Longrightarrow \quad |f(x) - f(y)| < \varepsilon/2 \qquad \text{for all } x, y \in [a,b].$$

Now let $t \in [a,b]$ be fixed. Then, we have for $x \in [a,b]$ the two estimates

$$\begin{aligned}
f(x) - f(t) &\leq \frac{\varepsilon}{2} + 2M \left(\frac{x-t}{\delta} \right)^2 = \frac{\varepsilon}{2} + \frac{2M}{\delta^2} \left[x^2 - 2xt + t^2 \right] \\
f(x) - f(t) &\geq -\frac{\varepsilon}{2} - 2M \left(\frac{x-t}{\delta} \right)^2 = -\frac{\varepsilon}{2} - \frac{2M}{\delta^2} \left[x^2 - 2xt + t^2 \right],
\end{aligned}$$

where ε, δ and M are independent of x. If we apply the linear and monotone operator K_n, for $n \in \mathbb{N}$, to both sides of these inequalities (with respect to variable x), then this implies

$$\begin{aligned}(K_n f)(x) - f(t)(K_n 1)(x) \leq \\ \frac{\varepsilon}{2}(K_n 1)(x) + \frac{2M}{\delta^2}\left[(K_n x^2)(x) - 2t(K_n x)(x) + t^2(K_n 1)(x)\right]\end{aligned}$$

$$\begin{aligned}(K_n f)(x) - f(t)(K_n 1)(x) \geq \\ -\frac{\varepsilon}{2}(K_n 1)(x) - \frac{2M}{\delta^2}\left[(K_n x^2)(x) - 2t(K_n x)(x) + t^2(K_n 1)(x)\right]\end{aligned}$$

for all $x \in [a, b]$. Therefore, we have the estimate

$$\begin{aligned}|(K_n f)(x) - f(t)(K_n 1)(x)| \leq \\ \frac{\varepsilon}{2}|(K_n 1)(x)| + \frac{2M}{\delta^2}|(K_n x^2)(x) - 2t(K_n x)(x) + t^2(K_n 1)(x)|. \qquad (6.4)\end{aligned}$$

By assumption, there is for any $\tilde{\varepsilon} > 0$ some $N \equiv N(\tilde{\varepsilon}) \in \mathbb{N}$ satisfying

$$\|(K_n x^k) - x^k\|_\infty < \tilde{\varepsilon} \qquad \text{for } k = 0, 1, 2,$$

for all $n \geq N$. This in particular implies

$$|(K_n 1)(x)| \leq \|K_n 1\|_\infty = \|(K_n 1 - 1) + 1\|_\infty \leq \tilde{\varepsilon} + 1 \qquad (6.5)$$

as well as

$$\begin{aligned}&|(K_n x^2)(x) - 2t(K_n x)(x) + t^2(K_n 1)(x)| = \\ &\quad |((K_n x^2)(x) - x^2) - 2t((K_n x)(x) - x) + t^2((K_n 1)(x) - 1) + x^2 - 2tx + t^2| \\ &\quad \leq \tilde{\varepsilon}(1 + 2|t| + t^2) + (x - t)^2 \qquad (6.6)\end{aligned}$$

for all $n \geq N$. From (6.4), (6.5) and (6.6), we obtain the estimate

$$\begin{aligned}|(K_n f)(x) - f(t)| &\leq |(K_n f)(x) - f(t)(K_n 1)(x)| + |f(t)(K_n 1)(x) - f(t)| \\ &\leq \frac{\varepsilon}{2}(\tilde{\varepsilon} + 1) + \frac{2M}{\delta^2}\left[\tilde{\varepsilon}(1 + 2|t| + t^2) + (x - t)^2\right] + M\tilde{\varepsilon},\end{aligned}$$

where for $x = t$, the inequality

$$|(K_n f)(t) - f(t)| \leq \frac{\varepsilon}{2}(\tilde{\varepsilon} + 1) + \frac{2M}{\delta^2}\left[\tilde{\varepsilon}(1 + 2|t| + t^2)\right] + M\tilde{\varepsilon} \qquad (6.7)$$

follows for all $n \geq N$.

Now the right hand side in (6.7) can uniformly be bounded from above by an arbitrarily small $\hat{\varepsilon} > 0$, so that we have, for some $N \equiv N(\hat{\varepsilon}) \in \mathbb{N}$,

$$\|K_n f - f\|_\infty < \hat{\varepsilon} \qquad \text{for all } n \geq N.$$

This proves the uniform convergence in (6.3), as stated. ■

Now we can prove the *density theorem of Weierstrass*.

Corollary 6.12. (Weierstrass theorem for algebraic polynomials). *The algebraic polynomials $\mathcal{P}$ are, w.r.t. the maximum norm $\|\cdot\|_\infty$ on a compact interval $[a,b] \subset \mathbb{R}$, dense in $\mathscr{C}[a,b]$. In particular, any $f \in \mathscr{C}[a,b]$ can, w.r.t. $\|\cdot\|_\infty$, be approximated arbitrarily well by algebraic polynomials, i.e., for any $f \in \mathscr{C}[a,b]$ and $\varepsilon > 0$, there is a polynomial $p \in \mathcal{P}$ satisfying*

$$\|p - f\|_\infty < \varepsilon.$$

Proof. We use the Bernstein operators $(B_n)_{n\in\mathbb{N}}$, which are a Korovkin sequence on $\mathscr{C}[0,1]$. Suppose $f \in \mathscr{C}[0,1]$ and $\varepsilon > 0$. Then, according to the Korovkin theorem, there is one $n \equiv n(\varepsilon) \in \mathbb{N}$, satisfying $\|B_n f - f\|_\infty < \varepsilon$. By $p = B_n f \in \mathcal{P}_n \subset \mathcal{P}$, the statement follows immediately from Theorem 6.11. ■

Note that the Weierstrass theorem gives a positive answer to Question 1, as posed at the outset of this section. Next, we specialize the density theorem of Weierstrass, Corollary 6.12, to even (or odd) functions.

Corollary 6.13. *Any even continuous function $f \in \mathscr{C}[-1,1]$ can, w.r.t. the norm $\|\cdot\|_\infty$, be approximated arbitrarily well by an even algebraic polynomial.*

Likewise, any odd continuous function $f \in \mathscr{C}[-1,1]$ can, with respect to $\|\cdot\|_\infty$, be approximated arbitrarily well by an odd algebraic polynomial.

Proof. Let $f \in \mathscr{C}[-1,1]$ be even and $\varepsilon > 0$. Then, due to the Weierstrass theorem, Corollary 6.12, and Proposition 3.42, there is an even algebraic polynomial $p \in \mathcal{P}$ satisfying $\|p - f\|_\infty < \varepsilon$. Likewise, for odd $f \in \mathscr{C}[-1,1]$, the second statement follows by similar arguments (cf. Exercise 3.73). ■

Now from our observations in Corollary 6.13, we wish to conclude a corresponding density result for the case of trigonometric polynomials $\mathcal{T} \subset \mathscr{C}_{2\pi}$. In preparation, we first prove two lemmas.

Lemma 6.14. *The linear space of real-valued trigonometric polynomials*

$$\mathcal{T} = \operatorname{span}_{\mathbb{R}} \left\{ \frac{1}{\sqrt{2}}, \cos(jx), \sin(jx) \,\middle|\, j \in \mathbb{N} \right\}$$

is a unital commutative algebra over $\mathbb{R}$. In particular, $\mathcal{T}$ is closed under the multiplication, i.e., the product of two real-valued trigonometric polynomials is a real-valued trigonometric polynomial.

Proof. The statement follows directly from the trigonometric addition formulas, in particular from the representations (4.16)-(4.18), for $j,k \in \mathbb{Z}$, i.e.,

$$\begin{aligned} 2\cos(jx)\cos(kx) &= \cos((j-k)x) + \cos((j+k)x) \\ 2\sin(jx)\sin(kx) &= \cos((j-k)x) - \cos((j+k)x) \\ 2\sin(jx)\cos(kx) &= \sin((j-k)x) + \sin((j+k)x). \end{aligned}$$

The remaining properties for a unital commutative algebra $\mathcal{T}$ are trivial. ■

Remark 6.15. Let $p \in \mathcal{P}$ be an algebraic polynomial. Then,

$$p(\sin(jx)\cos(kx)) \in \mathcal{T} \qquad \text{for } j,k \in \mathbb{N}_0$$

is a trigonometric polynomial. Moreover, every trigonometric polynomial

$$p(\cos(kx)) \in \mathcal{T} \qquad \text{for } k \in \mathbb{N}_0$$

is an even function. □

We now show that the *even* trigonometric polynomials are, with respect to the maximum norm $\|\cdot\|_\infty$, *dense* in $\mathscr{C}[0,\pi]$.

Lemma 6.16. *For any $f \in \mathscr{C}[0,\pi]$ and $\varepsilon > 0$, there is one even trigonometric polynomial $T_g \in \mathcal{T}$ satisfying*

$$\|T_g - f\|_\infty < \varepsilon.$$

Proof. Suppose $f \in \mathscr{C}[0,\pi]$. Then, $g(t) = f(\arccos(t)) \in \mathscr{C}[-1,1]$. Therefore, according to the Weierstrass theorem, Corollary 6.12, there is one algebraic polynomial $p \in \mathcal{P}$ satisfying $\|p-g\|_{\infty,[-1,1]} < \varepsilon$. This implies

$$\|p(\cos(\cdot)) - f\|_{\infty,[0,\pi]} = \|p-g\|_{\infty,[-1,1]} < \varepsilon$$

with the (bijective) variable transformation $x = \arccos(t)$, or, $t = \cos(x)$. Letting $T_g(x) = p(\cos(x)) \in \mathcal{T}$, our proof is complete. ■

Now we transfer the Weierstrass theorem for algebraic polynomials, Corollary 6.12, to the case of trigonometric polynomials. To this end, we consider the linear space $\mathscr{C}_{2\pi} \subset \mathscr{C}(\mathbb{R})$ of all continuous 2π-periodic target functions. Due to the periodicity of the elements in $\mathscr{C}_{2\pi}$, we can restrict ourselves to the compact interval $[0,2\pi]$.

This results in the formulation of the *Weierstrass density theorem.*

Corollary 6.17. (Weierstrass theorem for the trigonometric case). *The trigonometric polynomials $\mathcal{T}$ are, w.r.t. the maximum norm $\|\cdot\|_\infty$, dense in $\mathscr{C}_{2\pi}$. In particular, any function $f \in \mathscr{C}_{2\pi}$ can, w.r.t. $\|\cdot\|_\infty$, be approximated arbitrarily well by trigonometric polynomials, i.e., for any $f \in \mathscr{C}_{2\pi}$ and $\varepsilon > 0$ there is a trigonometric polynomial $T_f \in \mathcal{T}$ satisfying $\|T_f - f\|_\infty < \varepsilon$.*

Proof. Any $f \in \mathscr{C}_{2\pi}$ can be decomposed as a sum

$$f(x) = \frac{1}{2}(f(x) + f(-x)) + \frac{1}{2}(f(x) - f(-x)) = f_e(x) + f_o(x)$$

of an even function $f_e \in \mathscr{C}_{2\pi}$ and an odd function $f_o \in \mathscr{C}_{2\pi}$. Now the two *even* functions

$$f_e(x) \qquad \text{and} \qquad g_e(x) = \sin(x) f_o(x)$$

can be approximated arbitrarily well on $[0, \pi]$ by *even* trigonometric polynomials $T_{f_e}, T_{g_e} \in \mathcal{T}$, so that we have

$$\|T_{f_e} - f_e\|_\infty = \|T_{f_e} - f_e\|_{\infty,[-\pi,\pi]} = \|T_{f_e} - f_e\|_{\infty,[0,\pi]} < \varepsilon/4$$
$$\|T_{g_e} - g_e\|_\infty = \|T_{g_e} - g_e\|_{\infty,[-\pi,\pi]} = \|T_{g_e} - g_e\|_{\infty,[0,\pi]} < \varepsilon/4.$$

Therefore, we have, everywhere on $\mathbb{R}$, the representations

$$f_e = T_{f_e} + \eta_{f_e} \qquad \text{and} \qquad g_e = T_{g_e} + \eta_{g_e}$$

with (even) error functions $\eta_{f_e}, \eta_{g_e} \in \mathscr{C}_{2\pi}$, where $\|\eta_{f_e}\|_\infty, \|\eta_{g_e}\|_\infty < \varepsilon/4$. From these two representations, we obtain the identity

$$\begin{aligned}\sin^2(x)f(x) &= \sin^2(x)(f_e(x) + f_o(x)) \\ &= \sin^2(x)T_{f_e}(x) + \sin(x)T_{g_e}(x) + \sin^2(x)\eta_{f_e}(x) + \sin(x)\eta_{g_e}(x) \\ &= T_f^s(x) + \eta_f^s(x),\end{aligned}$$

where

$$\begin{aligned}T_f^s(x) &= \sin^2(x)T_{f_e}(x) + \sin(x)T_{g_e}(x) \in \mathcal{T} \\ \eta_f^s(x) &= \sin^2(x)\eta_{f_e}(x) + \sin(x)\eta_{g_e}(x) \qquad \text{with } \|\eta_f^s\|_\infty < \varepsilon/2.\end{aligned}$$

Using similar arguments we can derive, for the phase-shifted function

$$\tilde{f}(x) = f\,(x + \pi/2) \in \mathscr{C}_{2\pi},$$

a representation of the form

$$\sin^2(x)\tilde{f}(x) = T_{\tilde{f}}^s(x) + \eta_{\tilde{f}}^s(x) \qquad \text{with } \|\eta_{\tilde{f}}^s\|_\infty < \varepsilon/2$$

with $T_{\tilde{f}}^s \in \mathcal{T}$, so that after reversion of the translation $x \longmapsto x - \pi/2$, we have

$$\cos^2(x)f(x) = T_{\tilde{f}}^s(x-\pi/2) + \eta_{\tilde{f}}^s(x-\pi/2) = T_f^c(x) + \eta_f^c(x) \text{ with } \|\eta_f^c\|_\infty < \varepsilon/2,$$

where $T_f^c(x) \in \mathcal{T}$. By summation of the two representations for f, we obtain by

$$f(x) = T_f^s(x) + T_f^c(x) + \eta_f^s(x) + \eta_f^c(x) = T_f(x) + \eta_f(x) \qquad \text{with } \|\eta_f\|_\infty < \varepsilon$$

the stated estimate

$$\|T_f - f\|_\infty < \varepsilon$$

for the so constructed trigonometric polynomial $T_f = T_f^s + T_f^c \in \mathcal{T}$. ■

This gives a positive answer to Question 2 from the outset of this section.

Finally, we remark that the maximum norm $\|\cdot\|_\infty$ is in the following sense *stronger* than any p-norm $\|\cdot\|_p$, $1 \le p < \infty$.

Corollary 6.18. *The algebraic polynomials $\mathcal{P}$ are, w.r.t. any p-norm $\|\cdot\|_p$, $1 \le p < \infty$, and for compact $[a,b] \subset \mathbb{R}$, dense in $\mathscr{C}[a,b]$. Likewise, the trigonometric polynomials $\mathcal{T}$ are, w.r.t. $\|\cdot\|_p$, dense in $\mathscr{C}_{2\pi}$ for all $1 \le p < \infty$.*

Proof. For $f \in \mathscr{C}[a,b]$ and $\varepsilon > 0$ there is one $p \in \mathcal{P}$ satisfying $\|p - f\|_\infty < \varepsilon$. This immediately implies the estimate

$$\|p-f\|_p^p = \int_a^b |p(x)-f(x)|^p \,\mathrm{d}x \le (b-a)\|p-f\|_\infty^p < (b-a)\varepsilon^p \quad \text{for } 1 \le p < \infty,$$

i.e., any $f \in \mathscr{C}[a,b]$ can, w.r.t. $\|\cdot\|_p$, be approximated arbitrarily well by algebraic polynomials. The case of trigonometric polynomials $\mathcal{T}$, i.e., the second statement, can be covered by using similar arguments. ■

Remark 6.19. Corollary 6.18 states that convergence in the maximum norm $\|\cdot\|_\infty$ implies convergence in any p-norm $\|\cdot\|_p$, $1 \le p < \infty$. The converse, however, does not hold in general. In this sense, the maximum norm $\|\cdot\|_\infty$ is the *strongest* among all p-norms, for $1 \le p \le \infty$. □

A corresponding statement holds for weighted Euclidean norms.

Corollary 6.20. *Let $w : (a,b) \longrightarrow (0,\infty)$ be a continuous and integrable weight function, so that w defines on $\mathscr{C}[a,b]$, for compact $[a,b] \subset \mathbb{R}$, the inner product*

$$(f,g)_w = \int_a^b f(x)g(x)w(x)\,\mathrm{d}x \qquad \text{for } f,g \in \mathscr{C}[a,b] \tag{6.8}$$

and the Euclidean norm $\|\cdot\|_w = (\cdot,\cdot)_w^{1/2}$. Then, any function $f \in \mathscr{C}[a,b]$ can, w.r.t. $\|\cdot\|_w$, be approximated arbitrarily well by algebraic polynomials, i.e., the polynomial space $\mathcal{P}$ is, with respect to $\|\cdot\|_w$, dense in $\mathscr{C}[a,b]$.

Proof. For $f \in \mathscr{C}[a,b]$, we have

$$\|f\|_w^2 = \int_a^b |f(x)|^2 w(x)\,\mathrm{d}x \le \|f\|_\infty^2 \int_a^b w(x)\,\mathrm{d}x = C_w\|f\|_\infty^2,$$

where $C_w = \|1\|_w < \infty$. Now let $\varepsilon > 0$ and $p \in \mathcal{P}$ with $\|p - f\|_\infty < \varepsilon/\sqrt{C_w}$. Then,

$$\|p - f\|_w \le \sqrt{C_w}\|p - f\|_\infty < \varepsilon,$$

i.e., f can, with respect to $\|\cdot\|_w$, be approximated arbitrarily well by $p \in \mathcal{P}$. ■

6.2 Complete Orthogonal Systems and Riesz Bases

We recall the notion and properties of *orthogonal* (and *orthonormal*) systems from Section 4.2. In the following discussion, we consider a Euclidean space $\mathcal{F}$ with inner product $(\cdot,\cdot)$ and norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$. Moreover, let $\mathcal{S}_n \subset \mathcal{F}$ be a *finite-dimensional* linear subspace of dimension $\dim(\mathcal{S}_n) = n \in \mathbb{N}$ with an (ordered) orthogonal basis $(s_j)_{j=1}^n$ in $\mathcal{S}_n$, so that the *orthogonality relation*

$$(s_j, s_k) = \delta_{jk} \cdot \|s_j\|^2 \qquad \text{for } 1 \le j,k \le n$$

holds. According to Theorem 4.5, the unique best approximation to $f \in \mathcal{F}$ is given by the *orthogonal projection*

$$\Pi_n f = \sum_{j=1}^{n} \frac{(f,s_j)}{\|s_j\|^2} s_j \in \mathcal{S}_n \tag{6.9}$$

of f onto $\mathcal{S}_n$, obtained by the orthogonal projection operator $\Pi_n : \mathcal{F} \longrightarrow \mathcal{S}_n$.

In the following discussion, we investigate approximation properties of the partial sums $\Pi_n f$ in (6.9), where our particular interest is placed on their asymptotic behaviour. To this end, we analyze convergence for the sequence $(\Pi_n f)_{n\in\mathbb{N}}$, for $n \to \infty$, where we link to our discussion in Section 4.2. On this occasion, we recall the *Pythagoras theorem* (4.6), the *Bessel inequality* (4.12), and the *Parseval identity* (4.10), or, (4.11), according to which we have

$$\|\Pi_n f\|^2 = \sum_{j=1}^{n} \frac{|(f,s_j)|^2}{\|s_j\|^2} \qquad \text{for all } f \in \mathcal{F}. \tag{6.10}$$

6.2.1 Complete Orthogonal Systems

We wish to transfer our results from Section 4.2 to *infinite* (countable and ordered) orthogonal systems (and orthonormal systems) $(s_j)_{j\in\mathbb{N}}$ in $\mathcal{F}$. Our first result on this is based on the following characterization.

Theorem 6.21. *Let $(s_j)_{j\in\mathbb{N}}$ be an orthogonal system in a Euclidean space $\mathcal{F}$ with inner product $(\cdot,\cdot)$ and norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$. Then, the following statements are equivalent.*

(a) The span of $(s_j)_{j\in\mathbb{N}}$ is dense in $\mathcal{F}$, i.e., $\mathcal{F} = \overline{\operatorname{span}\{s_j \mid j \in \mathbb{N}\}}$.

(b) For any $f \in \mathcal{F}$ the sequence $(\Pi_n f)_{n\in\mathbb{N}}$ of partial sums $\Pi_n f$ in (6.9) converges to f with respect to the norm $\|\cdot\|$, i.e.,

$$\Pi_n f \longrightarrow f \quad \text{for } n \to \infty. \tag{6.11}$$

(c) For any $f \in \mathcal{F}$ we have the `Parseval identity`

$$\|f\|^2 = \sum_{j=1}^{\infty} \frac{|(f,s_j)|^2}{\|s_j\|^2}. \tag{6.12}$$

Proof. For any $f \in \mathcal{F}$, the n-th partial sum $\Pi_n f$ is the unique best approximation to f from $\mathcal{S}_n = \operatorname{span}\{s_1, \ldots, s_n\}$ with respect to $\|\cdot\|$.

$(a) \Rightarrow (b)$: Suppose for $f \in \mathcal{F}$ and $\varepsilon > 0$, there is one $N \in \mathbb{N}$ and $s_N \in \mathcal{S}_N$ satisfying $\|s_N - f\| < \varepsilon$. Then, we have for $n \geq N$

$$\|\Pi_n f - f\| = \inf_{s \in \mathcal{S}_n} \|s - f\| \leq \inf_{s \in \mathcal{S}_N} \|s - f\| \leq \|s_N - f\| < \varepsilon,$$

and so the sequence $(\Pi_n f)_{n \in \mathbb{N}}$ converges, with respect to $\|\cdot\|$, to f, i.e.,

$$\|\Pi_n f - f\| \longrightarrow 0 \quad \text{for } n \to \infty,$$

or, in short, $\Pi_n(f) \longrightarrow f$ for $n \to \infty$.

$(b) \Rightarrow (c)$: Suppose the sequence $(\Pi_n f)_{n \in \mathbb{N}}$ of the partial sums $\Pi_n f$ converges to $f \in \mathcal{F}$, so that $\|\Pi_n f - f\| \longrightarrow 0$ for $n \to \infty$. Then, by the Pythagoras theorem,

$$\|f\|^2 = \|\Pi_n f - f\|^2 + \|\Pi_n f\|^2, \tag{6.13}$$

in combination with the Parseval identity (6.10) we obtain, for $n \to \infty$, the representation

$$\|f\|^2 = \lim_{n \to \infty} \|\Pi_n f\|^2 = \sum_{j=1}^{\infty} \frac{|(f, s_j)|^2}{\|s_j\|^2}.$$

$(c) \Rightarrow (a)$: From the Pythagoras theorem (6.13) and by (6.10), we obtain

$$\|\Pi_n f - f\|^2 = \|f\|^2 - \sum_{j=1}^{n} \frac{|(f, s_j)|^2}{\|s_j\|^2} \longrightarrow 0 \quad \text{for } n \to \infty$$

and so there is, for any $\varepsilon > 0$, one $N \equiv N(\varepsilon)$ satisfying $\|\Pi_N f - f\| < \varepsilon$. ■

Definition 6.22. *An orthogonal system $(s_j)_{j \in \mathbb{N}}$ satisfying one of the properties (a), (b), or (c) in Theorem 6.21 (and so all three properties), is called a* `complete orthogonal system` *in $\mathcal{F}$. The notion of a* `complete orthonormal system` *is defined accordingly.* ◯

Remark 6.23. For a complete orthogonal system $(s_j)_{j \in \mathbb{N}}$ in $\mathcal{F}$ we have, according to property (b) in Theorem 6.21, the series representation

$$f = \sum_{j=1}^{\infty} \frac{(f, s_j)}{\|s_j\|^2} s_j \qquad \text{for } f \in \mathcal{F} \tag{6.14}$$

by convergence of $(\Pi_n f)_{n \in \mathbb{N}}$ to f with respect to $\|\cdot\|$. The series in (6.14) is often referred to as *(generalized) Fourier series* of f with *(generalized) Fourier coefficients* $(f, s_j)/\|s_j\|^2$. □

From the equivalence in Theorem 6.21, we can conclude a useful result.

Corollary 6.24. *Under the assumptions in Theorem 6.21, we have*

$$\|\Pi_n f - f\|^2 = \sum_{j=n+1}^{\infty} \frac{|(f, s_j)|^2}{\|s_j\|^2} \qquad \text{for all } f \in \mathcal{F} \tag{6.15}$$

for the representation of the squared error norm of $\Pi_n(f) - f$.

Proof. The representation (6.15) follows from property (c) in Theorem 6.21 by the Pythagoras theorem (6.13) and the Parseval identity (6.10). ■

By using the Weierstrass density theorems for algebraic and trigonometric polynomials, in Corollaries 6.12 and 6.17, we can give examples for complete orthogonal systems.

Our first example draws a link to Corollary 6.20.

Example 6.25. Let $w : (a, b) \longrightarrow (0, \infty)$ be a continuous weight function, so that w defines on $\mathscr{C}[a, b]$, for compact $[a, b] \subset \mathbb{R}$, an inner product $(\cdot, \cdot)_w$, see (6.8). Moreover, suppose $(p_j)_{j\in\mathbb{N}_0}$ is a sequence of orthogonal polynomials with respect to $(\cdot, \cdot)_w$ (cf. our construction in Theorem 4.16). Then, $(p_j)_{j\in\mathbb{N}_0}$ is a complete orthogonal system in $\mathscr{C}[a, b]$ with respect to the Euclidean norm $\|\cdot\|_w = (\cdot, \cdot)_w^{1/2}$. Indeed, this is because the algebraic polynomials $\mathcal{P}$ are, according to the Weierstrass theorem, Corollary 6.12, dense in $\mathscr{C}[a, b]$ with respect to the maximum norm $\|\cdot\|_\infty$, and so, by Corollary 6.20, $\mathcal{P}$ is also dense in $\mathscr{C}[a, b]$ with respect to $\|\cdot\|_w$. ♢

Next, we prove a useful criterion for the completeness of systems $(s_j)_{j\in\mathbb{N}}$ in Hilbert spaces $\mathcal{F}$, in particular for the completeness of orthogonal systems.

Theorem 6.26. (Completeness criterion). *For a system* $(s_j)_{j\in\mathbb{N}}$ *of elements in a Hilbert space* $\mathcal{F}$*, the following statements are equivalent.*

(a) The system $(s_j)_{j\in\mathbb{N}}$ *is complete in* $\mathcal{F}$*, i.e.,* $\mathcal{F} = \overline{\operatorname{span}\{s_j \,|\, j \in \mathbb{N}\}}$.
(b) If $f \in \mathcal{F}$ *is orthogonal to all elements* s_j*, then* $f = 0$*, i.e., we have the implication*

$$(f, s_j) = 0 \text{ for all } j \in \mathbb{N} \quad \Longrightarrow \quad f = 0.$$

Proof. Without loss of generality, we suppose that $(s_j)_{j\in\mathbb{N}}$ is an orthonormal system in $\mathcal{F}$. Otherwise, we can choose a subsequence $(s_{j_k})_{k\in\mathbb{N}}$ of linearly independent elements, which we then orthonormalize (as in the Gram-Schmidt algorithm, Algorithm 4). In the following, we use the notation

$$\mathcal{S} := \overline{\operatorname{span}\{s_j \,|\, j \in \mathbb{N}\}} \subset \mathcal{F}$$

for the closure of $\operatorname{span}\{s_j \,|\, j \in \mathbb{N}\}$ in $\mathcal{F}$ and, moreover,

$$\mathcal{S}^\perp := \{u \in \mathcal{F} \,|\, (u, s) = 0 \text{ for all } s \in \mathcal{S}\} \subset \mathcal{F}$$

for the orthogonal complement of $\mathcal{S}$ in $\mathcal{F}$, so that $\mathcal{F} = \mathcal{S} \oplus \mathcal{S}^\perp$.

$(a) \Rightarrow (b)$**:** Let $(s_j)_{j\in\mathbb{N}}$ be complete in $\mathcal{F}$. Then, the Parseval identity (6.12) holds according to Theorem 6.21. From this, we see that $(f, s_j) = 0$, for all $j \in \mathbb{N}$, implies $\|f\| = 0$, and so $f = 0$.

$(b) \Rightarrow (a)$**:** Let $f \in \mathcal{F}$ satisfy $(f, s_j) = 0$ for all $j \in \mathbb{N}$. In this case, we have $f \in \mathcal{S}^\perp$ by the linearity and the continuity of the inner product. Conversely, for $f \in \mathcal{S}^\perp$ the orthogonality relation $(f, s_j) = 0$ holds for all $j \in \mathbb{N}$. Therefore, f is contained in $\mathcal{S}^\perp$, if and only if $(f, s_j) = 0$ for all $j \in \mathbb{N}$. With the assumed implication in (b), we have $\mathcal{S}^\perp = \{0\}$ and so $\mathcal{S} = \mathcal{F}$. ■

6.2.2 Riesz Bases and Frames

Next, we extend the concept of complete orthonormal systems. To this end, we fix a Hilbert space $\mathcal{F}$ with inner product $(\cdot, \cdot)$ and norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$. In the following discussion, we regard systems $(s_n)_{n\in\mathbb{Z}}$ with the *bi-infinite* index set $\mathbb{Z}$. Recall that for a complete orthonormal system $(s_n)_{n\in\mathbb{Z}}$ in $\mathcal{F}$, we have, for any $f \in \mathcal{F}$, the series representation

$$f = \sum_{n\in\mathbb{Z}} (f, s_n) s_n$$

according to Remark 6.23. Moreover, the *Parseval identity* in (6.12) holds, which we represent as

$$\|f\| = \|((f, s_n))_{n\in\mathbb{Z}}\|_{\ell^2}\,, \tag{6.16}$$

where ℓ^2 denotes the linear space of all square summable sequences with indices in $\mathbb{Z}$ (cf. Remark 3.15).

Definition 6.27. *A system $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of elements in a Hilbert space $\mathcal{F}$ is called a* `Riesz`[4] `basis` *of $\mathcal{F}$, if the following properties are satisfied.*

(a) The span of $\mathcal{B}$ is dense in $\mathcal{F}$, i.e.,

$$\mathcal{F} = \overline{\operatorname{span}\{u_n \,|\, n \in \mathbb{Z}\}}. \tag{6.17}$$

(b) There are constants $0 < A \le B < \infty$ satisfying

$$A\|c\|_{\ell^2}^2 \le \left\| \sum_{n\in\mathbb{Z}} c_n u_n \right\|^2 \le B\|c\|_{\ell^2}^2 \qquad \text{for all } c = (c_n)_{n\in\mathbb{Z}} \in \ell^2. \tag{6.18}$$

For a Riesz basis $\mathcal{B}$, the "best possible" constants, i.e., the largest A and the smallest B satisfying (6.18), are called `Riesz constants` *of $\mathcal{B}$.* ◯

[4] Frigyes Riesz (1880-1956), Hungarian mathematician

Remark 6.28. Every complete orthonormal system in $\mathcal{F}$ is a Riesz basis of $\mathcal{F}$. Indeed, in this case, we have the Parseval identity in (6.16), whereby equality holds in (6.18) for $A = B = 1$. Moreover, the completeness in (6.17) holds by Theorem 6.21 (a).

We remark that the Riesz estimates in (6.18), often written in short as

$$\left\| \sum_{n\in\mathbb{Z}} c_n u_n \right\| \sim \|c\|_{\ell^2} \qquad \text{for all } c = (c_n)_{n\in\mathbb{Z}} \in \ell^2,$$

describe the stability of the Riesz basis representation with respect to perturbations of the coefficients in $c \in \ell^2$. Therefore, Riesz bases are also often referred to as ℓ^2-*stable bases* of $\mathcal{F}$. □

In the following analysis concerning Riesz bases $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$, the linear `synthesis operator` $G : \ell^2 \longrightarrow \mathcal{F}$, defined as

$$G(c) = \sum_{n\in\mathbb{Z}} c_n u_n \in \mathcal{F} \qquad \text{for } c = (c_n)_{n\in\mathbb{Z}} \in \ell^2, \tag{6.19}$$

plays an important role. We note the following properties of G.

Proposition 6.29. *Let $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ be a Riesz basis of $\mathcal{F}$ with Riesz constants $0 < A \le B < \infty$. Then, the synthesis operator $G : \ell^2 \longrightarrow \mathcal{F}$ in (6.19) has the following properties.*

(a) The operator G is continuous, where G has operator norm $\|G\| = \sqrt{B}$.
(b) The operator G is bijective.
(c) The inverse G^{-1} of G is continuous with operator norm $\|G^{-1}\| = 1/\sqrt{A}$.

Proof. Statement (a) follows directly from the upper Riesz estimate in (6.18).

As for the proof of (b), note that G is surjective, since $\operatorname{span}\{u_n \,|\, n \in \mathbb{Z}\}$ is by (6.17) dense in $\mathcal{F}$. Moreover, G is injective, since by (6.18) the kernel of G can only contain the zero element. Altogether, the operator G is bijective.

Finally, for the inverse $G^{-1} : \mathcal{F} \longrightarrow \ell^2$ of G we find by (6.18) the estimate

$$\|G^{-1}(f)\|^2 \le \frac{1}{A}\|f\|^2 \qquad \text{for all } f \in \mathcal{F}$$

and this implies the continuity of G^{-1} at operator norm $\|G^{-1}\| = 1/\sqrt{A}$. This proves property (c). ■

Now we consider the *dual* `analysis operator` $G^* : \mathcal{F} \longrightarrow \ell^2$ of G in (6.19), where G^* is characterized by the duality relation

$$(G^*(f), c)_{\ell^2} = (f, G(c)) \qquad \text{for all } c \in \ell^2 \text{ and all } f \in \mathcal{F}. \tag{6.20}$$

We note the following properties of G^*.

Proposition 6.30. *The pair of dual operators G in (6.19) and G^* in (6.20) satisfies the following properties.*

(a) The operator G^* has the representation

$$G^*(f) = ((f, u_n))_{n\in\mathbb{Z}} \in \ell^2 \qquad \text{for all } f \in \mathcal{F}.$$

(b) The operator G^* is bijective and has the inverse $(G^*)^{-1} = (G^{-1})^*$.
(c) The operators G^* and $(G^*)^{-1}$ are continuous via the isometries

$$\|G\| = \|G^*\| \quad \text{and} \quad \|G^{-1}\| = \|(G^*)^{-1}\|.$$

Proof. By (6.20), we find for the dual operator $G^* : \mathcal{F} \longrightarrow \ell^2$ the identty

$$(G^*(f), c)_{\ell^2} = (f, G(c)) = \sum_{n\in\mathbb{Z}} c_n (f, u_n) = (((f, u_n))_{n\in\mathbb{Z}}, c)_{\ell^2}$$

for all $c \in \ell^2$, and this already implies the stated representation in (a).

By the representation in (a) in combination with the Riesz basis property of $\mathcal{B}$, we see that G^* is bijective. Moreover, for $f, g \in \mathcal{F}$ the representation

$$((G^{-1})^* G^*(f), g) = (G^*(f), G^{-1}(g))_{\ell^2} = (f, GG^{-1}(g)) = (f, g)$$

holds. Therefore, $(G^{-1})^* G^*$ is the identity on $\mathcal{F}$. Likewise, we see that $G^*(G^{-1})^*$ is the identity on ℓ^2. This proves statement (b).

As regards statement (c), we find on the one hand

$$\|G^*(f)\|_{\ell^2}^2 = (G^*(f), G^*(f))_{\ell^2} = (f, GG^*(f)) \le \|f\| \cdot \|G\| \cdot \|G^*(f)\|_{\ell^2}$$

by letting $c = G^*(f)$ in (6.20), and this implies $\|G^*\| \le \|G\|$. On the other hand, we have

$$\|G(c)\|^2 = (G(c), G(c)) = (G^*G(c), c)_{\ell^2} \le \|G^*\| \cdot \|G(c)\| \cdot \|c\|_{\ell^2}$$

by letting $f = G(c)$ in (6.20), which implies $\|G\| \le \|G^*\|$. Altogether, we have $\|G\| = \|G^*\|$. The other statement in (c) follows from similar arguments. ∎

Now we explain a fundamental duality property for Riesz bases.

Theorem 6.31. *For any Riesz basis $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$ with Riesz constants $0 < A \le B < \infty$, there is a unique Riesz basis $\tilde{\mathcal{B}} = (\tilde{u}_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$, such that*

(a) the elements in $\mathcal{B}$ and $\tilde{\mathcal{B}}$ are mutually orthonormal, i.e.,

$$(u_n, \tilde{u}_m) = \delta_{nm} \qquad \text{for all } n, m \in \mathbb{Z}. \tag{6.21}$$

(b) the Riesz basis $\tilde{\mathcal{B}}$ has Riesz constants $0 < 1/B \le 1/A < \infty$.
(c) any $f \in \mathcal{F}$ can uniquely be represented w.r.t. $\mathcal{B}$ or $\tilde{\mathcal{B}}$, respectively, as

$$f = \sum_{n\in\mathbb{Z}} (f, \tilde{u}_n) u_n = \sum_{n\in\mathbb{Z}} (f, u_n) \tilde{u}_n. \tag{6.22}$$

The Riesz basis $\tilde{\mathcal{B}}$ is called the `dual Riesz basis` *of $\mathcal{B}$ in $\mathcal{F}$.*

Proof. We consider the linear operator $G : \ell^2 \longrightarrow \mathcal{F}$ in (6.19) associated with the Riesz basis $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ and its dual operator $G^* : \mathcal{F} \longrightarrow \ell^2$ in (6.20). According to Propositions 6.29 and 6.30 each of the linear operators G and G^* is continuous and has a continuous inverse. Therefore, their composition $GG^* : \mathcal{F} \longrightarrow \mathcal{F}$ is continuous and has a continuous inverse.

Now we consider $\tilde{\mathcal{B}} = (\tilde{u}_n)_{n\in\mathbb{Z}}$, where

$$\tilde{u}_n := (GG^*)^{-1}u_n \qquad \text{for } n \in \mathbb{Z}.$$

The elements in $\tilde{\mathcal{B}}$ satisfy the orthonormality relation (6.21) in (a), since

$$(u_n, \tilde{u}_m) = (u_n, (GG^*)^{-1}u_m) = (G^{-1}u_n, G^{-1}u_m)_{\ell^2} = \delta_{mn} \tag{6.23}$$

holds for any $m, n \in \mathbb{Z}$. Moreover, for $c = (c_n)_{n\in\mathbb{Z}} \in \ell^2$, we have the identity

$$\left\| \sum_{n\in\mathbb{Z}} c_n \tilde{u}_n \right\| = \left\| (GG^*)^{-1} \left(\sum_{n\in\mathbb{Z}} c_n u_n \right) \right\| = \left\| (G^*)^{-1} c \right\|.$$

By $\|G^*\|^2 = B$ and $\|(G^*)^{-1}\|^2 = 1/A$, we get the Riesz stability for $\tilde{\mathcal{B}}$, i.e.,

$$\frac{1}{B}\|c\|_{\ell^2}^2 \le \left\| \sum_{n\in\mathbb{Z}} c_n \tilde{u}_n \right\|^2 \le \frac{1}{A}\|c\|_{\ell^2}^2 \qquad \text{for all } c = (c_n)_{n\in\mathbb{Z}} \in \ell^2. \tag{6.24}$$

Now the continuity of $(GG^*)^{-1}$ and the completeness of $\mathcal{B}$ in (6.17) implies

$$\mathcal{F} = \overline{\operatorname{span}\{\tilde{u}_n \,|\, n \in \mathbb{Z}\}},$$

i.e., $\tilde{\mathcal{B}}$ is a Riesz basis of $\mathcal{F}$ with Riesz constants $0 < 1/B \le 1/A < \infty$. The stated uniqueness of $\tilde{\mathcal{B}}$ follows from the orthonormality relation (6.23).

Let us finally show property (c). Since G is surjective, any $f \in \mathcal{F}$ can be represented as

$$f = \sum_{n\in\mathbb{Z}} c_n u_n \qquad \text{for some } c = (c_n)_{n\in\mathbb{Z}} \in \ell^2.$$

But this implies

$$(f, \tilde{u}_m) = \left(\sum_{n\in\mathbb{Z}} c_n u_n, \tilde{u}_m \right) = c_m,$$

whereby the stated (unique) representation in (6.22) holds, i.e.,

$$f = \sum_{n\in\mathbb{Z}} (f, \tilde{u}_n) u_n \qquad \text{for all } f \in \mathcal{F}.$$

Likewise, the stated representation in (6.22) with respect to the Riesz basis $\tilde{\mathcal{B}}$ can be shown by similar arguments. ∎

From the estimates in (6.24) and the representation in (6.22), we get the stability of the coefficients $(f, u_n))_{n\in\mathbb{Z}} \in \ell^2$ under perturbations of $f \in \mathcal{F}$.

Corollary 6.32. *Let $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ be a Riesz basis of $\mathcal{F}$ with Riesz constants $0 < A \le B < \infty$. Then, the stability estimates*

$$A\|f\|^2 \le \|((f, u_n))_{n\in\mathbb{Z}}\|_{\ell^2}^2 \le B\|f\|^2 \qquad \text{for all } f \in \mathcal{F} \tag{6.25}$$

hold. ■

Remark 6.33. Every Riesz basis $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$ yields a system of ℓ^2-*linearly independent* elements in $\mathcal{F}$, i.e., for $c = (c_n)_{n\in\mathbb{Z}} \in \ell^2$ the implication

$$\sum_{n\in\mathbb{Z}} c_n u_n = 0 \quad \Longrightarrow \quad c = 0$$

holds. In other words, $G(c) = 0$ implies $c = 0$, as this is covered by Proposition 6.29 (b). Moreover, Corollary 6.32 gives the stability estimates in (6.25). □

The required conditions for a Riesz basis $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ (according to Definition 6.27) often appear as too restrictive. In fact, relevant applications work with weaker conditions on $\mathcal{B}$, where they merely require the stability in (6.25), but not the ℓ^2-*linearly independence* of $\mathcal{B}$.

Definition 6.34. *A system $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of elements in a Hilbert space $\mathcal{F}$ is called a* `frame` *of $\mathcal{F}$, if for $0 < A \le B < \infty$ the estimates*

$$A\|f\|^2 \le \|((f, u_n))_{n\in\mathbb{Z}}\|_{\ell^2}^2 \le B\|f\|^2 \qquad \text{for all } f \in \mathcal{F} \tag{6.26}$$

hold, where the "best possible" constants, i.e., the largest A and the smallest B satisfying (6.26), are called `frame constants` *of $\mathcal{B}$.* ◯

Remark 6.35. Any frame $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$ is *complete* in $\mathcal{F}$, i.e., the span of $\mathcal{B}$ is dense in $\mathcal{F}$,

$$\mathcal{F} = \overline{\operatorname{span}\{u_n \,|\, n \in \mathbb{Z}\}}.$$

This immediately follows from the completeness criterion, Theorem 6.26, by using the lower estimate in (6.26). □

Remark 6.36. Every Riesz basis $\mathcal{B}$ is a frame, but the converse is general not true. Indeed, a frame $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ allows ambiguities in the representation

$$f = \sum_{n\in\mathbb{Z}} c_n u_n \qquad \text{for } f \in \mathcal{F},$$

due to a possible ℓ^2-*linear dependence* of the elements in $\mathcal{B}$. □

Remark 6.37. For any frame $\mathcal{B} = (u_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$, there exists a dual frame $\tilde{\mathcal{B}} = (\tilde{u}_n)_{n\in\mathbb{Z}}$ of $\mathcal{F}$ satisfying

$$f = \sum_{n\in\mathbb{Z}} (f, u_n)\tilde{u}_n = \sum_{n\in\mathbb{Z}} (f, \tilde{u}_n)u_n \qquad \text{for all } f \in \mathcal{F}.$$

However, the duality relation $(u_n, \tilde{u}_m) = \delta_{nm}$ in (6.21) does not in general hold, since otherwise the elements of $\mathcal{B}$ and the elements of $\tilde{\mathcal{B}}$ would be ℓ^2-*linearly independent*, respectively. □

For further illustration, we discuss the following examples.

Example 6.38. The three vectors

$$u_1 = (0, 1)^T, \quad u_2 = (-\sqrt{3}/2, -1/2)^T, \quad u_3 = (\sqrt{3}/2, -1/2)^T$$

form a frame in $\mathcal{F} = \mathbb{R}^2$, since for $f = (f_1, f_2)^T \in \mathbb{R}^2$, we have

$$\begin{aligned}\sum_{j=1}^{3} (f, u_j)^2 &= f_2^2 + \left(-\frac{\sqrt{3}}{2} f_1 - \frac{1}{2} f_2\right)^2 + \left(\frac{\sqrt{3}}{2} f_1 - \frac{1}{2} f_2\right)^2 \\ &= \frac{3}{2}(f_1^2 + f_2^2) = \frac{3}{2}\|f\|_2^2,\end{aligned}$$

and so the stability in (6.25) holds with $A = B = 3/2$. However, note that the vectors u_1, u_2, u_3 are ℓ^2-linearly dependent, since $u_1 + u_2 + u_3 = 0$. ♢

In our next example we discuss Riesz bases in *finite-dimensional* Euclidean spaces for the prototypical case of the Euclidean space $\mathcal{F} = \mathbb{R}^d$, where $d \in \mathbb{N}$.

Example 6.39. For the Euclidean space $\mathbb{R}^d$, where $d \in \mathbb{N}$, equipped with the Euclidean norm $\|\cdot\|_2$, any basis $\mathcal{B} = \{u_1, \ldots, u_d\}$ of $\mathbb{R}^d$ is a Riesz basis of $\mathbb{R}^d$. Indeed, in this case, we have for the regular matrix $U = (u_1, \ldots, u_d) \in \mathbb{R}^{d\times d}$ and for any vector $c = (c_1, \ldots, c_d)^T \in \mathbb{R}^d$ the stability estimates

$$\|U^{-1}\|_2^{-1}\|c\|_2 \le \left\|\sum_{n=1}^{N} c_n u_n\right\|_2 = \|Uc\|_2 \le \|U\|_2\|c\|_2.$$

Therefore, the Riesz constants $0 < A \le B < \infty$ of $\mathcal{B}$ are given by the spectral norms of the matrices U and U^{-1}, so that $A = \|U^{-1}\|_2^{-2}$ and $B = \|U\|_2^2$. The unique dual Riesz basis $\tilde{\mathcal{B}}$ of $\mathcal{B}$ is given by the rows of the inverse U^{-1}. This immediately follows by $UU^{-1} = I$ from Theorem 6.31 (a). ♢

We close this section by studying an example for a frame of $\mathbb{R}^d$.

Example 6.40. We continue to work with the Euclidean space $\mathbb{R}^d$, where $d \in \mathbb{N}$, whose inner product is denoted by $(\cdot,\cdot)$. For a frame $\mathcal{B} = (u_n)_{n=1}^N$ of $\mathbb{R}^d$, where $N > d$, we consider the dual operator $G^* : \mathbb{R}^d \longrightarrow \mathbb{R}^N$ in (6.20). According to Proposition 6.30 (a), the representation

$$G^*(f) = ((f,u_n))_{n=1}^N = (u_n^T f)_{n=1}^N \in \mathbb{R}^N \qquad \text{for } f \in \mathbb{R}^d$$

holds, or, in matrix notation,

$$G^* f = U^T f = c_f \qquad \text{for } f \in \mathbb{R}^d,$$

where $U = (u_1, \ldots, u_N) \in \mathbb{R}^{d\times N}$ and $c_f = ((f,u_n))_{n=1}^N \in \mathbb{R}^N$. Due to the completeness of $\mathcal{B}$, according to Definition 6.34 (a), we see that the columns $(u_1, \ldots, u_N)$ of U must contain a basis of $\mathbb{R}^d$. Hence, U has full rank, $d = \operatorname{rank}(U)$, and $U^T \in \mathbb{R}^{N\times d}$ is injective. But this is consistent with the injectivity of the dual operator G^*, as established in the lower estimate in (6.26), for $A > 0$.

Now we consider the *dual* frame $\tilde{\mathcal{B}} = (\tilde{u}_n)_{n=1}^N$ of $\mathcal{B}$, as characterized by

$$f = \sum_{n=1}^N (f,u_n)\tilde{u}_n \qquad \text{for all } f \in \mathbb{R}^d.$$

By $U^T f = c_f$, we have $UU^T f = U c_f$ and so

$$f = (UU^T)^{-1} U c_f \qquad \text{for all } f \in \mathbb{R}^d,$$

i.e., the dual frame $\tilde{\mathcal{B}} = (\tilde{u}_n)_{n=1}^N$ is determined by the columns of $(UU^T)^{-1}U$. However, the elements in $\mathcal{B}$ and $\tilde{\mathcal{B}}$ do not satisfy the orthonormality relation in Theorem 6.31 (a). ♢

6.3 Convergence of Fourier Partial Sums

In this section, we analyze the approximation behaviour of Fourier partial sums in more detail. To this end, we recall our discussion from Section 4.3, where, in particular, we had proven the orthonormality of the real-valued trigonometric polynomials in $\mathscr{C}_{2\pi} \equiv \mathscr{C}_{2\pi}^{\mathbb{R}}$, see Theorem 4.11. By using the Weierstrass theorem for trigonometric polynomials, Corollary 6.17, we can prove the following result.

Corollary 6.41. *The real-valued trigonometric polynomials*

$$\left\{ \frac{1}{\sqrt{2}}, \cos(j\cdot), \sin(j\cdot) \,\middle|\, j \in \mathbb{N} \right\} \subset \mathscr{C}_{2\pi}^{\mathbb{R}} \tag{6.27}$$

form a complete orthonormal system in $\mathscr{C}_{2\pi}^{\mathbb{R}}$ with respect to the Euclidean norm $\|\cdot\|_{\mathbb{R}} = (\cdot,\cdot)_{\mathbb{R}}^{1/2}$, as defined by the (real) inner product

$$(f,g)_{\mathbb{R}} = \frac{1}{\pi}\int_0^{2\pi} f(x)g(x)\,\mathrm{d}x \qquad \text{for } f,g \in \mathscr{C}_{2\pi}^{\mathbb{R}}.$$

Proof. The orthonormality of the trigonometric polynomials in (6.27) holds by Theorem 4.11. Moreover, due to the trigonometric version of the Weierstrass theorem, Corollary 6.17, the real-valued trigonometric polynomials $\mathcal{T} \equiv \mathcal{T}^{\mathbb{R}}$ are dense in $\mathscr{C}_{2\pi} = \mathscr{C}_{2\pi}^{\mathbb{R}}$ with respect to the maximum norm $\|\cdot\|_\infty$, and so $\mathcal{T}$ is a dense subset of $\mathscr{C}_{2\pi}$ also with respect to the weaker Euclidean norm $\|\cdot\|_{\mathbb{R}}$, cf. Corollary 6.18. ■

Remark 6.42. The result of Corollary 6.41 can directly be transferred to the complex case, whereby the complex-valued trigonometric polynomials

$$\{\mathrm{e}^{\mathrm{i}j\cdot} \mid j \in \mathbb{Z}\} \subset \mathscr{C}_{2\pi}^{\mathbb{C}}$$

form a complete orthonormal system in $\mathscr{C}_{2\pi}^{\mathbb{C}}$ with respect to the Euclidean norm $\|\cdot\|_{\mathbb{C}} = (\cdot,\cdot)_{\mathbb{C}}^{1/2}$, defined by the (complex) inner product

$$(f,g)_{\mathbb{C}} = \frac{1}{2\pi}\int_0^{2\pi} f(x)\overline{g(x)}\,\mathrm{d}x \qquad \text{for } f,g \in \mathscr{C}_{2\pi}^{\mathbb{C}}, \tag{6.28}$$

cf. Remark 4.10. □

Now we consider, for $n \in \mathbb{N}_0$, real-valued Fourier partial sums of the form

$$(F_n f)(x) = \frac{a_0}{2} + \sum_{j=1}^{n}(a_j\cos(jx) + b_j\sin(jx)) \qquad \text{for } f \in \mathscr{C}_{2\pi}^{\mathbb{R}} \tag{6.29}$$

with Fourier coefficients $a_0 = (f,1)_{\mathbb{R}}$, $a_j = (f,\cos(j\cdot))_{\mathbb{R}}$, and $b_j = (f,\sin(j\cdot))_{\mathbb{R}}$, for $j \in \mathbb{N}$, see Corollary 4.12. As we noticed in Section 4.3, the Fourier operator $F_n : \mathscr{C}_{2\pi}^{\mathbb{R}} \longrightarrow \mathcal{T}_n^{\mathbb{R}}$ gives the orthogonal projection of $\mathscr{C}_{2\pi}^{\mathbb{R}}$ onto $\mathcal{T}_n^{\mathbb{R}}$. In particular, $F_n f \in \mathcal{T}_n^{\mathbb{R}}$ is the unique best approximation to $f \in \mathscr{C}_{2\pi}^{\mathbb{R}}$ from $\mathcal{T}_n^{\mathbb{R}}$ with respect to the Euclidean norm $\|\cdot\|_{\mathbb{R}}$.

As regards our notations concerning real-valued against complex-valued functions, we recall Remark 4.10: For *real-valued* functions $f \in \mathscr{C}_{2\pi} \equiv \mathscr{C}_{2\pi}^{\mathbb{R}}$, we apply the inner product $(\cdot,\cdot) = (\cdot,\cdot)_{\mathbb{R}}$ and the norm $\|\cdot\| = \|\cdot\|_{\mathbb{R}}$. In contrast, for *complex-valued* functions $f \in \mathscr{C}_{2\pi}^{\mathbb{C}}$, we use $(\cdot,\cdot)_{\mathbb{C}}$ and $\|\cdot\|_{\mathbb{C}}$.

6.3.1 Convergence in Quadratic Mean

From our above discussion, we can conclude the following convergence result.

Corollary 6.43. *For the approximation to $f \in \mathscr{C}_{2\pi}$ by Fourier partial sums $F_n f$ we have* `convergence in quadratic mean`*, i.e.,*

$$\lim_{n\to\infty} \|F_n f - f\| = 0 \qquad \textit{for all } f \in \mathscr{C}_{2\pi}.$$

Proof. The statement follows immediately from property (b) in Theorem 6.21 in combination with Corollary 6.41. ■

Next, we quantify the *speed of convergence* for the Fourier partial sums $F_n f$. To this end, the complex representation in (4.23),

$$(F_n f)(x) = \sum_{j=-n}^{n} c_j \mathrm{e}^{\mathrm{i}jx}, \tag{6.30}$$

with the complex Fourier coefficients $c_j \equiv c_j(f) = (f, \exp(\mathrm{i}j\,\cdot))_{\mathbb{C}}$, i.e.,

$$c_j = \frac{1}{2\pi} \int_0^{2\pi} f(x) \mathrm{e}^{-\mathrm{i}jx}\,\mathrm{d}x \qquad \text{for } -n \le j \le n,$$

and the orthonormal system $\{\exp(\mathrm{i}j\,\cdot) \mid -n \le j \le n\} \subset \mathcal{T}_n^{\mathbb{C}}$ with respect to the *complex* inner product $(\cdot,\cdot)_{\mathbb{C}}$ in (6.28) turns out to be particularly useful. Indeed, from the representation in (6.30) we can prove the following result.

Theorem 6.44. *For $f \in \mathscr{C}^k_{2\pi}$ the Fourier partial sums $F_n f$ converge to f at* `convergence rate` *$k \in \mathbb{N}_0$ according to*

$$\|F_n f - f\| \le \frac{1}{(n+1)^k} \left\| F_n f^{(k)} - f^{(k)} \right\| = o(n^{-k}) \quad \text{for } n \to \infty. \tag{6.31}$$

Proof. For $k = 0$, we obtain the stated convergence result from Corollary 6.43.

For $k = 1$, we apply integration by parts to obtain, for $j \neq 0$ and $f \in \mathscr{C}^1_{2\pi}$, by the identity

$$\begin{aligned} c_j(f) &= \frac{1}{2\pi} \int_0^{2\pi} f(x) \mathrm{e}^{-\mathrm{i}jx}\,\mathrm{d}x = \frac{\mathrm{i}}{j} \frac{1}{2\pi} \left[f(x) \mathrm{e}^{-\mathrm{i}jx} \right]_0^{2\pi} - \frac{\mathrm{i}}{j} \frac{1}{2\pi} \int_0^{2\pi} f'(x) \mathrm{e}^{-\mathrm{i}jx}\,\mathrm{d}x \\ &= -\frac{\mathrm{i}}{j} \frac{1}{2\pi} \int_0^{2\pi} f'(x) \mathrm{e}^{-\mathrm{i}jx}\,\mathrm{d}x = -\frac{\mathrm{i}}{j} (f', \mathrm{e}^{-\mathrm{i}j\cdot}) = -\frac{\mathrm{i}}{j} c_j(f') \end{aligned}$$

an alternative representation for the complex Fourier coefficients c_j in (6.30).

By induction on k, we obtain for $f \in \mathscr{C}^k_{2\pi}$ the representation

$$c_j(f) = (-\mathrm{i})^k \frac{1}{j^k} c_j(f^{(k)}) \qquad \text{for all } j \in \mathbb{Z} \setminus \{0\},$$

and so in this case, we find the estimate

$$|c_j(f)| \le \frac{1}{|j|^k} |c_j(f^{(k)})| \qquad \text{for all } j \in \mathbb{Z} \setminus \{0\} \text{ and } k \in \mathbb{N}_0. \tag{6.32}$$

By the representation of the error in Corollary 6.24, this in turn implies

$$\begin{aligned} \|F_n f - f\|_{\mathbb{C}}^2 &= \sum_{|j| \ge n+1} |c_j(f)|^2 \le \sum_{|j| \ge n+1} \frac{1}{j^{2k}} |c_j(f^{(k)})|^2 \\ &\le \frac{1}{(n+1)^{2k}} \sum_{|j| \ge n+1} |c_j(f^{(k)})|^2 \\ &= \frac{1}{(n+1)^{2k}} \left\| F_n f^{(k)} - f^{(k)} \right\|_{\mathbb{C}}^2 \end{aligned}$$

for $f \in \mathscr{C}_{2\pi}^k$ and therefore

$$\|F_n f - f\| \leq \frac{1}{(n+1)^k} \left\| F_n f^{(k)} - f^{(k)} \right\| = o(n^{-k}) \quad \text{for } n \to \infty,$$

where we use the convergence

$$\left\| F_n f^{(k)} - f^{(k)} \right\| \longrightarrow 0 \quad \text{for } n \to \infty$$

for $f^{(k)} \in \mathscr{C}_{2\pi}$ according to Corollary 6.43. ■

Remark 6.45. The convergence rate $k \in \mathbb{N}_0$, as achieved in Theorem 6.44, follows from the asymptotic decay of the Fourier coefficients $c_j(f)$ of f in (6.32), whereby

$$|c_j(f)| = \mathcal{O}\left(|j|^{-k}\right) \quad \text{for } |j| \to \infty.$$

Further note that the decay of $c_j(f)$ follows from the assumption $f \in \mathscr{C}_{2\pi}^k$.

As for the converse, we can determine the *smoothness* of f from the asymptotic decay of the Fourier coefficients $c_j(f)$. More precisely: If the Fourier coefficients $c_j(f)$ of f have the asymptotic decay

$$|c_j(f)| = \mathcal{O}\left(|j|^{-(k+1+\varepsilon)}\right) \quad \text{for } |j| \to \infty$$

for some $\varepsilon > 0$, then this implies $f \in \mathscr{C}_{2\pi}^k$ (see Exercise 6.91).

Conclusion: The smoother $f \in \mathscr{C}_{2\pi}$, the faster the convergence of the Fourier partial sums $F_n f$ to f, and vice versa. □

6.3.2 Uniform Convergence

Next, we analyze the *uniform* convergence of the Fourier partial sums $F_n f$. Although we have proven convergence in quadratic mean, i.e., convergence with respect to the Euclidean norm $\|\cdot\|$, we cannot expect convergence in the stronger maximum norm $\|\cdot\|_\infty$, due to Remark 6.19. In fact, to prove uniform convergence we need to assume further conditions on $f \in \mathscr{C}_{2\pi}$, especially concerning its smoothness. As we show now, it is sufficient to require that f has a continuous derivative, i.e., $f \in \mathscr{C}_{2\pi}^1$.

Corollary 6.46. *For $f \in \mathscr{C}_{2\pi}^1$, the Fourier partial sums $F_n f$ converge uniformly to f, i.e., we have*

$$\lim_{n\to\infty} \|F_n f - f\|_\infty = 0.$$

Proof. For any $n \in \mathbb{N}$, the orthogonality $F_n f - f \perp 1$ holds, i.e., we have

$$\int_0^{2\pi} (F_n f - f)(x)\,\mathrm{d}x = 0 \qquad \text{for all } n \in \mathbb{N}.$$

Therefore, the error function $F_n f - f$ has at least one zero x_n in the open interval $(0, 2\pi)$, whereby for $x \in [0, 2\pi]$ we obtain the representation

$$(F_n f - f)(x) = \int_{x_n}^{x} (F_n f - f)'(\xi)\,\mathrm{d}\xi = \int_{x_n}^{x} (F_n f' - f')(\xi)\,\mathrm{d}\xi,$$

where we used the identity $(F_n f)' = F_n f'$ (see Exercise 6.92). By the Cauchy-Schwarz inequality, we further obtain

$$\begin{aligned} |(F_n f - f)(x)|^2 &\le \left| \int_{x_n}^{x} 1\,\mathrm{d}\xi \right| \cdot \left| \int_{x_n}^{x} |(F_n f' - f')(\xi)|^2\,\mathrm{d}\xi \right| \\ &\le (2\pi)^2 \|F_n f' - f'\|^2 \longrightarrow 0 \quad \text{for } n \to \infty, \end{aligned} \tag{6.33}$$

which already proves the stated uniform convergence. ■

Now we conclude from Theorem 6.44 a corresponding result concerning the convergence rate of $(F_n f)_{n\in\mathbb{N}_0}$ with respect to the maximum norm $\|\cdot\|_\infty$.

Corollary 6.47. *For $f \in \mathscr{C}_{2\pi}^k$, where $k \ge 1$, the Fourier partial sums $F_n f$ converge uniformly to f at convergence rate $k-1$, according to*

$$\|F_n f - f\|_\infty = o(n^{-(k-1)}) \quad \text{for } n \to \infty.$$

Proof. For $f' \in \mathscr{C}_{2\pi}^{k-1}$, we have by (6.33) and (6.31) the estimate

$$\|F_n f - f\|_\infty \le 2\pi \|F_n f' - f'\| \le \frac{2\pi}{(n+1)^{k-1}} \left\| F_n f^{(k)} - f^{(k)} \right\|,$$

whereby we obtain for $f^{(k)} \in \mathscr{C}_{2\pi}$ the asymptotic convergence behaviour

$$\|F_n f - f\|_\infty = o(n^{-(k-1)}) \quad \text{for } n \to \infty$$

according to Corollary 6.43. ■

6.3.3 Pointwise Convergence

Next, we analyze *pointwise* convergence for the Fourier partial sums $F_n f$. To this end, we first derive for $x \in \mathbb{R}$ a suitable representation for the pointwise error $(F_n f)(x) - f(x)$ at x. We utilize, for $f \in \mathscr{C}_{2\pi}$, the *real* representation of $F_n f$, whereby we obtain

$$\begin{aligned} (F_n f)(x) &= \frac{a_0}{2} + \sum_{j=1}^{n} [a_j \cos(jx) + b_j \sin(jx)] \\ &= \frac{1}{\pi} \int_0^{2\pi} f(\tau) \left[\frac{1}{2} + \sum_{j=1}^{n} (\cos(j\tau)\cos(jx) + \sin(j\tau)\sin(jx)) \right] \mathrm{d}\tau \\ &= \frac{1}{\pi} \int_0^{2\pi} f(\tau) \left[\frac{1}{2} + \sum_{j=1}^{n} \cos(j(\tau - x)) \right] \mathrm{d}\tau. \end{aligned} \tag{6.34}$$

Note that in the last line we applied the trigonometric addition formula

$$\cos(u+v) = \cos(u)\cos(v) - \sin(u)\sin(v)$$

for $u = j\tau$ and $v = -jx$. Now we simplify the integrand in (6.34) by applying the substitution $z = \tau - x$ along with the representation

$$\begin{aligned}
&\left[\frac{1}{2} + \sum_{j=1}^{n} \cos(jz)\right] 2\sin(z/2) \\
&= \sin(z/2) + \sum_{j=1}^{n} 2\cos(jz)\sin(z/2) \\
&= \sin(z/2) + \sum_{j=1}^{n} \left[\sin\left(\left(j+\frac{1}{2}\right)z\right) - \sin\left(\left(j-\frac{1}{2}\right)z\right)\right] \\
&= \sin\left(\left(n+\frac{1}{2}\right)z\right), \qquad (6.35)
\end{aligned}$$

where we used the trigonometric identity

$$\sin(u) - \sin(v) = 2\cos\left(\frac{u+v}{2}\right)\sin\left(\frac{u-v}{2}\right)$$

for $u = (j+1/2)z$ and $v = (j-1/2)z$. This implies the representation

$$(F_n f)(x) = \frac{1}{\pi}\int_0^{2\pi} f(\tau) D_n(\tau - x)\,\mathrm{d}\tau, \qquad (6.36)$$

where the function

$$D_n(z) = \frac{1}{2}\frac{\sin((n+1/2)z)}{\sin(z/2)} \qquad \text{for } n \in \mathbb{N}_0 \qquad (6.37)$$

is called *Dirichlet*[5] *kernel*. Note that the Dirichlet kernel is 2π-periodic and even, so that we can further simplify the representation in (6.36) to obtain

$$\begin{aligned}
(F_n f)(x) &= \frac{1}{\pi}\int_0^{2\pi} f(\tau) D_n(\tau - x)\,\mathrm{d}\tau \\
&= \frac{1}{\pi}\int_{-x}^{2\pi - x} f(x+\sigma) D_n(\sigma)\,\mathrm{d}\sigma \\
&= \frac{1}{\pi}\int_{-\pi}^{\pi} f(x+\sigma) D_n(\sigma)\,\mathrm{d}\sigma. \qquad (6.38)
\end{aligned}$$

Since $F_n 1 \equiv 1$, for $n \in \mathbb{N}_0$, we further obtain by (6.38) the representation

[5] Peter Gustav Lejeune Dirichlet (1805-1859), German mathematician

$$(F_n f)(x) - f(x) = \frac{1}{\pi}\int_{-\pi}^{\pi} [f(x+\sigma) - f(x)]\, D_n(\sigma)\,\mathrm{d}\sigma$$
$$= \frac{1}{\pi}\int_{-\pi}^{\pi} g_x(\sigma)\cdot \sin((n+1/2)\sigma)\,\mathrm{d}\sigma$$

for the pointwise error at $x \in \mathbb{R}$, where

$$g_x(\sigma) := \frac{f(x+\sigma) - f(x)}{2\sin(\sigma/2)}. \tag{6.39}$$

By using the trigonometric addition formula

$$\sin(n\sigma + \sigma/2) = \sin(n\sigma)\cos(\sigma/2) + \cos(n\sigma)\sin(\sigma/2)$$

we can rewrite the representation for the pointwise error as a sum of the form

$$\begin{aligned}&(F_n f)(x) - f(x)\\ &= \frac{1}{\pi}\int_{-\pi}^{\pi} g_x(\sigma)\cos(\sigma/2)\cdot\sin(n\sigma)\,\mathrm{d}\sigma + \frac{1}{\pi}\int_{-\pi}^{\pi} g_x(\sigma)\sin(\sigma/2)\cdot\cos(n\sigma)\,\mathrm{d}\sigma\\ &= b_n(g_x(\cdot)\cos(\cdot/2)) + a_n(g_x(\cdot)\sin(\cdot/2))\end{aligned}$$

with the Fourier coefficients $b_n(v_x)$ and $a_n(w_x)$ of the 2π-periodic functions

$$v_x(\sigma) = g_x(\sigma)\cos(\sigma/2)$$
$$w_x(\sigma) = g_x(\sigma)\sin(\sigma/2).$$

Suppose $g_x(\sigma)$ is a continuous function. Then, $v_x, w_x \in \mathscr{C}_{2\pi}$. Moreover, by the Parseval identity, we have in this case

$$\|v_x\|_{\mathbb{C}}^2 = \sum_{n\in\mathbb{Z}} |(v_x, \exp(\mathrm{i}n\cdot))|^2 < \infty \quad \text{and} \quad \|w_x\|_{\mathbb{C}}^2 = \sum_{n\in\mathbb{Z}} |(w_x, \exp(\mathrm{i}n\cdot))|^2 < \infty,$$

so that the Fourier coefficients $(b_n(v_x))_{n\in\mathbb{Z}}$ and $(a_n(w_x))_{n\in\mathbb{Z}}$ are a zero sequence, respectively, whereby the pointwise convergence of $(F_n f)(x)$ to $f(x)$ at x would follow.

Now we are in a position where we can, from our above investigations, formulate a sufficient condition for $f \in \mathscr{C}_{2\pi}$ which guarantees pointwise convergence of $(F_n f)(x)$ to $f(x)$ at $x \in \mathbb{R}$.

Theorem 6.48. *Let $f \in \mathscr{C}_{2\pi}$ be differentiable at $x \in \mathbb{R}$. Then, we have pointwise convergence of $(F_n f)(x)$ to $f(x)$ at x, i.e.,*

$$(F_n f)(x) \longrightarrow f(x) \quad \text{for } n \to \infty.$$

Proof. First note that the function g_x in (6.39) can only have singularities at $\sigma_k = 2\pi k$, for $k \in \mathbb{Z}$. Now we analyze the behaviour of g_x around zero, where we find

$$\lim_{\sigma\to 0} g_x(\sigma) = \lim_{\sigma\to 0} \frac{f(x+\sigma)-f(x)}{2\sin(\sigma/2)} = \lim_{\sigma\to 0} \frac{f(x+\sigma)-f(x)}{\sigma} \cdot \lim_{\sigma\to 0} \frac{\sigma}{2\sin(\sigma/2)}$$
$$= f'(x),$$

by using *L'Hôpital's* [6] *rule*. Therefore, the function g_x is continuous at $\sigma = 0$. By the periodicity of g_x and f, we see that the function g_x is also continuous at $\sigma = 2\pi k$, for all $k \in \mathbb{Z}$, whereby g_x is continuous on $\mathbb{R}$. ■

6.3.4 Asymptotic Behaviour of the Fourier Operator Norms

Now let us return to the uniform convergence of Fourier partial sums, where the following question is of particular importance.

Question: Can we, under *mild as possible* conditions on $f \in \mathscr{C}_{2\pi} \setminus \mathscr{C}^1_{2\pi}$, prove statements concerning *uniform* convergence of the Fourier partial sums $F_n f$?

To answer this question, we need to analyze the norm $\|F_n\|_\infty$ of the Fourier operator F_n with respect to the maximum norm $\|\cdot\|_\infty$. To this end, we first derive a suitable representation for the operator norm

$$\|F_n\|_\infty := \sup_{f\in\mathscr{C}_{2\pi}\setminus\{0\}} \frac{\|F_n f\|_\infty}{\|f\|_\infty} \qquad \text{for } n \in \mathbb{N}_0, \tag{6.40}$$

before we study the asymptotic behaviour of $\|F_n\|_\infty$ for $n \to \infty$.

From (6.38), we obtain the uniform estimate

$$|(F_n f)(x)| \le \frac{1}{\pi}\|f\|_\infty \int_{-\pi}^{\pi} |D_n(\sigma)|\,\mathrm{d}\sigma = \|f\|_\infty \cdot \frac{2}{\pi}\int_0^{\pi} |D_n(\sigma)|\,\mathrm{d}\sigma. \tag{6.41}$$

This leads us to a suitable representation for the norm $\|F_n\|_\infty$ of F_n in (6.40).

Theorem 6.49. *The norm of the Fourier operator F_n has the representation*

$$\|F_n\|_\infty = \lambda_n \qquad \textit{for all } n \in \mathbb{N}_0,$$

where

$$\lambda_n := \frac{2}{\pi}\int_0^{\pi} |D_n(\sigma)|\,\mathrm{d}\sigma = \frac{1}{\pi}\int_0^{\pi} \left|\frac{\sin((n+1/2)\sigma)}{\sin(\sigma/2)}\right| \mathrm{d}\sigma \tag{6.42}$$

is called `Lebesgue`[7] `constant`.

Proof. From (6.41), we immediately obtain

$$\|F_n f\|_\infty \le \|f\|_\infty \cdot \lambda_n$$

[6] Marquis de L'Hôpital (1661-1704), French mathematician
[7] Henri Léon Lebesgue (1875-1941), French mathematician

and so we have, on the one hand, the upper bound $\|F_n\|_\infty \le \lambda_n$.

On the other hand, we can choose, for any $\varepsilon > 0$ an even 2π-periodic continuous function f satisfying $\|f\|_\infty = 1$, such that f approximates the *even* step function $\operatorname{sgn}(D_n(x))$ arbitrarily well, i.e.,

$$\begin{aligned}\|F_n\|_\infty \ge \|F_n f\|_\infty \ge |(F_n f)(0)| &= \frac{1}{\pi}\left|\int_{-\pi}^{\pi} f(\sigma) D_n(\sigma)\,\mathrm{d}\sigma\right| \\ &\ge \frac{1}{\pi}\left|\int_{-\pi}^{\pi} \operatorname{sgn}(D_n(\sigma)) D_n(\sigma)\,\mathrm{d}\sigma\right| - \varepsilon \\ &= \frac{2}{\pi}\int_0^{\pi} |D_n(\sigma)|\,\mathrm{d}\sigma - \varepsilon \\ &= \lambda_n - \varepsilon,\end{aligned}$$

whereby for $\varepsilon \to 0$ we have the lower bound $\|F_n\|_\infty \ge \lambda_n$.

Altogether, we find $\|F_n\|_\infty = \lambda_n$, as stated. ■

Remark 6.50. To obtain uniform convergence,

$$\|F_n f - f\|_\infty \longrightarrow 0 \quad \text{for } n \to \infty,$$

for *all* $f \in \mathscr{C}_{2\pi}$, we require the Fourier operator norms $\|F_n\|_\infty = \lambda_n$ to be *uniformly* bounded from above. We can see this from the triangle inequality

$$\|F_n f\|_\infty \le \|F_n f - f\|_\infty + \|f\|_\infty.$$

Indeed, if the norms $\|F_n\|_\infty$ are *not* uniformly bounded from above, then there must be at least one $f \in \mathscr{C}_{2\pi}$ yielding divergence $\|F_n f\|_\infty \longrightarrow \infty$ for $n \to \infty$, in which case the sequence of error norms $\|F_n f - f\|_\infty$ must be divergent, i.e., $\|F_n f - f\|_\infty \longrightarrow \infty$ for $n \to \infty$. □

Unfortunately, the operator norms $\|F_n\|_\infty$ are *not* uniformly bounded from above. This is because we have the following estimates for $\lambda_n = \|F_n\|_\infty$.

Theorem 6.51. *For the Lebesgue constants* λ_n *in (6.42), we have*

$$\frac{4}{\pi^2}\log(n+1) \le \lambda_n \le 1 + \log(2n+1) \qquad \textit{for all } n \in \mathbb{N}_0. \tag{6.43}$$

Proof. For $n = 0$, the estimates in (6.43) are satisfied by $\lambda_0 = 1$.

Now suppose $n \ge 1$. For the zeros

$$\sigma_k = \frac{k\pi}{n + 1/2} \qquad \text{for } k \in \mathbb{Z}$$

of $D_n(\sigma)$ in (6.37) we obtain, on the one hand, the lower estimates

$$\begin{aligned}\lambda_n &\geq \frac{1}{\pi}\sum_{k=0}^{n-1}\int_{\sigma_k}^{\sigma_{k+1}}\left|\frac{\sin((n+1/2)\sigma)}{\sin(\sigma/2)}\right|\mathrm{d}\sigma \\ &\geq \frac{2}{\pi}\sum_{k=0}^{n-1}\frac{1}{\sigma_{k+1}}\int_{\sigma_k}^{\sigma_{k+1}}|\sin((n+1/2)\sigma)|\,\mathrm{d}\sigma \qquad (6.44)\\ &= \frac{4}{\pi^2}\sum_{k=0}^{n-1}\frac{1}{k+1} \\ &\geq \frac{4}{\pi^2}\log(n+1), \qquad (6.45)\end{aligned}$$

where we have used the estimate

$$|\sin(\sigma/2)| \leq |\sigma/2| \qquad \text{for all } \sigma \in \mathbb{R}$$

in (6.44) and, moreover, we have used the estimate

$$\sum_{k=0}^{n-1}\frac{1}{k+1} \geq \log(n+1) \qquad \text{for all } n \in \mathbb{N}$$

in (6.45).

On the other hand, we have for the integrand in (6.42) the estimates

$$\left|\frac{\sin((n+1/2)\sigma)}{\sin(\sigma/2)}\right| = \left|2\left[\frac{1}{2}+\sum_{j=1}^{n}\cos(j\sigma)\right]\right| = \left|1+2\sum_{j=1}^{n}\cos(j\sigma)\right| \leq 1+2n,$$

see (6.35), and, moreover,

$$\left|\frac{\sin((n+1/2)\sigma)}{\sin(\sigma/2)}\right| \leq \frac{1}{\sigma/\pi} = \frac{\pi}{\sigma} \qquad \text{for } \pi \geq \sigma \geq \frac{\pi}{2n+1} =: \mu_n,$$

where we have used the estimate

$$\sin(\sigma/2) \geq \sigma/\pi \qquad \text{for all } \sigma \in [0,\pi].$$

But this already implies the upper bound

$$\begin{aligned}\lambda_n &\leq \frac{1}{\pi}\left(\int_0^{\mu_n}(2n+1)\,\mathrm{d}\sigma + \int_{\mu_n}^{\pi}\frac{\pi}{\sigma}\,\mathrm{d}\sigma\right) \\ &= (2n+1)\frac{\mu_n}{\pi} + \log(\pi/\mu_n) = 1+\log(2n+1).\end{aligned}$$

■

Since $\|F_n\|_\infty$ is unbounded, we can conclude from Remark 6.50, that there exists at least one function $f \in \mathscr{C}_{2\pi}$ for which the sequence of Fourier partial sums $F_n f$ does *not* converge uniformly to f. This fundamental insight is based on the important *uniform boundedness principle* of Banach-Steinhaus.

6.3.5 Uniform Boundedness Principle

Let us first quote the Banach[8]-Steinhaus[9] theorem, a well-known result from functional analysis, before we draw relevant conclusions. We will not prove the Banach-Steinhaus theorem, but rather refer the reader to the textbook [33].

Theorem 6.52. (Banach-Steinhaus, 1927).
Let $(L_n)_{n\in\mathbb{N}}$ be a sequence of bounded linear operators

$$L_n : \mathcal{B}_1 \longrightarrow \mathcal{B}_2 \qquad \text{for } n \in \mathbb{N}$$

between two Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$. Moreover, suppose the operators L_n are pointwise bounded, i.e., for any $f \in \mathcal{B}_1$ we have

$$\sup_{n\in\mathbb{N}} \|L_n f\| < \infty.$$

Then, the `uniform boundedness principle` *holds for the operators L_n, i.e.,*

$$\sup_{n\in\mathbb{N}} \|L_n\| < \infty.$$

□

In conclusion, by the Banach-Steinhaus theorem, the pointwise boundedness of the operators $(L_n)_{n\in\mathbb{N}}$ implies their uniform boundedness. But this has negative consequences for the approximation with Fourier partial sums. We can further explain this by providing the following corollary.

Corollary 6.53. *There is a function $f \in \mathscr{C}_{2\pi}$ for which the sequence $(F_n f)_{n\in\mathbb{N}}$ of Fourier partial sums does not converge uniformly to f, i.e.,*

$$\|F_n f - f\|_\infty \longrightarrow \infty \quad \text{for } n \to \infty.$$

Moreover, for this f, we have the divergence

$$\|F_n f\|_\infty \longrightarrow \infty \quad \text{for } n \to \infty.$$

Proof. The function space $\mathscr{C}_{2\pi}$, equipped with the maximum norm $\|\cdot\|_\infty$, is a Banach space. By the divergence $\|F_n\|_\infty = \lambda_n \longrightarrow \infty$ for $n \to \infty$, there is one $f \in \mathscr{C}_{2\pi}$ with $\|F_n f\|_\infty \longrightarrow \infty$ for $n \to \infty$. Indeed, otherwise this would contradict the Banach-Steinhaus theorem. Now the estimate

$$\|F_n f - f\|_\infty \geq \|F_n f\|_\infty - \|f\|_\infty$$

immediately implies, for this f, the stated divergence $\|F_n f - f\|_\infty \longrightarrow \infty$, for $n \to \infty$, of the Fourier partial sums' maximum norms. ■

[8] Stefan Banach (1892-1945), Polish mathematician
[9] Hugo Steinhaus (1887-1972), Polish mathematician

Next, we show the norm minimality of the Fourier operator F_n among all surjective projection operators onto the linear space of trigonometric polynomials $\mathcal{T}_n$. The following result dates back to Charshiladse-Losinski.

Theorem 6.54. (Charshiladse-Losinski).
For $n \in \mathbb{N}_0$, let $L : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ be a continuous linear projection operator, i.e.,

$$L(Lf) = L(f) \qquad \text{for all } f \in \mathscr{C}_{2\pi}.$$

Moreover, suppose L is surjective, i.e., $L(\mathscr{C}_{2\pi}) = \mathcal{T}_n$. Then, we have

$$\|L\|_\infty \geq \|F_n\|_\infty.$$

Proof. We define for $s \in \mathbb{R}$ the translation operator T_s by

$$(T_s f)(x) := f(x+s) \qquad \text{for } f \in \mathscr{C}_{2\pi} \text{ and } x \in \mathbb{R}.$$

Note that $\|T_s\|_\infty = 1$. Moreover, we define a linear operator G by

$$(Gf)(x) := \frac{1}{2\pi} \int_{-\pi}^{\pi} (T_{-s} L T_s f)(x)\, \mathrm{d}s \qquad \text{for } f \in \mathscr{C}_{2\pi} \text{ and } x \in \mathbb{R}. \tag{6.46}$$

Then, $G : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ is bounded (i.e., continuous) on $\mathscr{C}_{2\pi}$, since we have

$$|(Gf)(x)| \leq \|T_{-s} L T_s f\|_\infty \leq \|T_{-s}\|_\infty \|L\|_\infty \|T_s\|_\infty \|f\|_\infty = \|L\|_\infty \|f\|_\infty$$

and so $\|Gf\|_\infty \leq \|L\|_\infty \|f\|_\infty$ for all $f \in \mathcal{T}_n$, or,

$$\|G\|_\infty \leq \|L\|_\infty.$$

Now the operator G coincides on $\mathscr{C}_{2\pi}$ with the Fourier operator F_n, as we will show by the following lemma. This then completes our proof. ∎

Lemma 6.55. *Suppose the operator $L : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ satisfies the assumptions in Theorem 6.54. Then, the operator G in (6.46) coincides on $\mathscr{C}_{2\pi}$ with the Fourier operator $F_n : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$, i.e., we have*

$$Gf = F_n f \qquad \text{for all } f \in \mathscr{C}_{2\pi}.$$

Proof. We obtain the extension $L : \mathscr{C}_{2\pi}^{\mathbb{C}} \longrightarrow \mathcal{T}_n^{\mathbb{C}}$ of the operator L by letting

$$Lf := Lu + \mathrm{i}Lv \qquad \text{for } f = u + \mathrm{i}v \in \mathscr{C}_{2\pi}^{\mathbb{C}} \text{ where } u, v \in \mathscr{C}_{2\pi}^{\mathbb{R}} = \mathscr{C}_{2\pi}.$$

In this way, the extension of G in (6.46) from $\mathscr{C}_{2\pi}$ to $\mathscr{C}_{2\pi}^{\mathbb{C}}$ is well-defined. Moreover, we work with the extension of F_n from $\mathscr{C}_{2\pi}$ to $\mathscr{C}_{2\pi}^{\mathbb{C}}$.

Since the orthonormal system $\{\mathrm{e}^{\mathrm{i}j\cdot} \mid j \in \mathbb{Z}\}$ is complete in $\mathscr{C}_{2\pi}^{\mathbb{C}}$ (cf. Remark 6.42 and Exercise 6.89) and by the continuity of the linear operators $F_n : \mathscr{C}_{2\pi}^{\mathbb{C}} \longrightarrow \mathcal{T}_n^{\mathbb{C}}$ and $G : \mathscr{C}_{2\pi}^{\mathbb{C}} \longrightarrow \mathcal{T}_n^{\mathbb{C}}$, it is sufficient to show the identity

$$G\left(\mathrm{e}^{\mathrm{i}j\cdot}\right) = F_n\left(\mathrm{e}^{\mathrm{i}j\cdot}\right) \qquad \text{for all } j \in \mathbb{Z}. \tag{6.47}$$

To this end, we take a closer look at the operator G. First we note

$$\left(T_s\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \mathrm{e}^{\mathrm{i}j(x+s)} = \mathrm{e}^{\mathrm{i}jx}\mathrm{e}^{\mathrm{i}js}$$

and this implies

$$\left(LT_s\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \mathrm{e}^{\mathrm{i}js}\left(L\mathrm{e}^{\mathrm{i}j\cdot}\right)(x)$$

and, moreover,

$$\left(T_{-s}LT_s\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \mathrm{e}^{\mathrm{i}js}\left(L\mathrm{e}^{\mathrm{i}j\cdot}\right)(x-s). \tag{6.48}$$

Case 1: For $|j| \le n$, we have (since L is surjective)

$$(Lf)(x) = \mathrm{e}^{\mathrm{i}jx} \in \mathcal{T}_n^{\mathbb{C}}$$

for one $f \in \mathscr{C}_{2\pi}^{\mathbb{C}}$. Together with the projection property $L(Lf) = Lf$, this implies (for this particular f) the identity

$$(L(Lf))(x) = \left(L\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = (Lf)(x) = \mathrm{e}^{\mathrm{i}jx},$$

i.e., $\left(L\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \mathrm{e}^{\mathrm{i}jx}$. In combination with (6.48), we further obtain

$$\left(T_{-s}LT_s\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \mathrm{e}^{\mathrm{i}js}\mathrm{e}^{\mathrm{i}j(x-s)} = \mathrm{e}^{\mathrm{i}jx}$$

and so

$$\left(G\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}jx}\,\mathrm{d}s = \mathrm{e}^{\mathrm{i}jx} = \left(F_n\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x).$$

Case 2: For $|j| > n$, we have $\left(F_n\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = 0$. Moreover, the function $\mathrm{e}^{\mathrm{i}js}$ is orthogonal to the trigonometric polynomial $\left(L\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x-s) \in \mathcal{T}_n^{\mathbb{C}}$. From this and by (6.48), we obtain

$$\left(G\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}js}\left(L\left(\mathrm{e}^{\mathrm{i}j\cdot}\right)\right)(x-s)\,\mathrm{d}s = 0.$$

Altogether, the identity (6.47) holds, as stated. ■

Obviously, the result of Theorem 6.54 makes our situation worse. Indeed, we can formulate one more negative consequence from the Charshiladse-Losinski theorem.

Corollary 6.56. *Let $(L_n)_{n\in\mathbb{N}_0}$ be a sequence of continuous and surjective linear projection operators $L_n : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$. Then, there is a function $f \in \mathscr{C}_{2\pi}$ satisfying*

$$\|L_n f\|_\infty \longrightarrow \infty \quad \textit{for } n \to \infty,$$

whereby

$$\|L_n f - f\|_\infty \longrightarrow \infty \quad \textit{for } n \to \infty.$$

■

Corollary 6.56 can be proven by similar arguments as for Corollary 6.53.

We finally draw another negative conclusion from the Banach-Steinhaus theorem, which prohibits uniform convergence for sequences of interpolation polynomials. The following important result is due to Faber[10] [23].

Theorem 6.57. (Faber, 1914). *For any sequence $(I_n)_{n\in\mathbb{N}_0}$ of interpolation operators $I_n : \mathscr{C}[a,b] \longrightarrow \mathcal{P}_n$, there is a continuous function $f \in \mathscr{C}[a,b]$, for which the corresponding sequence $(I_n f)_{n\in\mathbb{N}_0}$ of interpolation polynomials $I_n f \in \mathcal{P}_n$ does not converge uniformly to f.* □

For a proof of the Faber theorem, we refer to Exercise 6.93.

6.4 The Jackson Theorems

In this section, we analyze the asymptotic behaviour of the minimal distances

$$\eta_\infty(f,\mathcal{T}_n) := \inf_{T\in\mathcal{T}_n} \|T-f\|_\infty \qquad \text{for } f \in \mathscr{C}_{2\pi}$$
$$\eta_\infty(f,\mathcal{P}_n) := \inf_{p\in\mathcal{P}_n} \|p-f\|_\infty \qquad \text{for } f \in \mathscr{C}[a,b]$$

for $n \to \infty$ with respect to the maximum norm $\|\cdot\|_\infty$. According to the Weierstrass theorems, Corollaries 6.12 and 6.17, we can rely on the convergence

$$\eta_\infty(f,\mathcal{T}_n) \longrightarrow 0 \quad \text{and} \quad \eta_\infty(f,\mathcal{P}_n) \longrightarrow 0 \quad \text{for } n \to \infty.$$

In this section, we quantify the asymptotic decay of the zero sequences $(\eta_\infty(f,\mathcal{T}_n))_{n\in\mathbb{N}_0}$ and $(\eta_\infty(f,\mathcal{P}_n))_{n\in\mathbb{N}_0}$ for $n \to \infty$.

We begin our analysis with the trigonometric case, i.e., with the asymptotic behaviour of $(\eta_\infty(f,\mathcal{T}_n))_{n\in\mathbb{N}_0}$. On this occasion, we first recall the convergence rates of the Fourier partial sums $F_n f$ for $f \in \mathscr{C}_{2\pi}$. By the estimate

$$\eta_\infty(f,\mathcal{T}_n) \le \|F_n f - f\|_\infty \qquad \text{for } n \in \mathbb{N}_0$$

we expect for $f \in \mathscr{C}^k_{2\pi}$, $k \ge 1$, at least the convergence rate $k-1$, according to Corollary 6.47. However, as it turns out, we gain even more. In fact, we will obtain the convergence rate k, i.e.,

$$\eta_\infty(f,\mathcal{T}_n) = \mathcal{O}(n^{-k}) \quad \text{for } n \to \infty \qquad \text{for } f \in \mathscr{C}^k_{2\pi}.$$

Note that this complies with the convergence behaviour of Fourier partial sums $F_n f$ with respect to the Euclidean norm $\|\cdot\|$. Indeed, in that case, we have, by Theorem 6.44, the asymptotic behaviour

$$\eta(f,\mathcal{T}_n) = o(n^{-k}) \quad \text{for } n \to \infty \qquad \text{for } f \in \mathscr{C}^k_{2\pi}.$$

For an intermediate conclusion, we note one important principle:

The smoother $f \in \mathscr{C}^k_{2\pi}$ is, i.e., the larger $k \in \mathbb{N}$, the faster the convergence of the minimal distances $\eta(f,\mathcal{T}_n)$ and $\eta_\infty(f,\mathcal{T}_n)$ to zero, for $n \to \infty$.

[10] Georg Faber (1877-1966), German mathematician

Remark 6.58. On this occasion, we recall Remark 6.45, where we had drawn a similar conclusion for the approximation by Fourier partial sums with respect to the Euclidean norm. As regards our above intermediate conclusion, we remark that the converse of that principle is covered by the classical *Bernstein theorems* (see e.g. [11]), albeit we decided to refrain from discussing Bernstein theorems in more details. □

In this section, we develop suitable conditions on $f \in \mathscr{C}_{2\pi} \setminus \mathscr{C}^1_{2\pi}$, under which the sequence $(F_n f)_{n\in\mathbb{N}_0}$ of Fourier partial sums converges uniformly to f. In this way, we give an answer to the question which we formulated at the outset of Section 6.3.4. But first we require some preparations. Let $\Pi_n : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ denote the *nonlinear* projection operator, which assigns $f \in \mathscr{C}_{2\pi}$ to its unique best approximation $\Pi_n f \in \mathcal{T}_n$ with respect to the maximum norm $\|\cdot\|_\infty$, so that

$$\eta_\infty(f, \mathcal{T}_n) = \|\Pi_n f - f\|_\infty \quad \text{for all } f \in \mathscr{C}_{2\pi}.$$

Then, we have the estimate

$$\begin{aligned} \|F_n f - f\|_\infty &= \|F_n f - \Pi_n f + \Pi_n f - f\|_\infty \\ &= \|F_n(f - \Pi_n f) + (\Pi_n f - f)\|_\infty \\ &= \|(I - F_n)(\Pi_n f - f)\|_\infty \\ &\le \|I - F_n\|_\infty \cdot \|\Pi_n f - f\|_\infty \\ &= \|I - F_n\|_\infty \cdot \eta_\infty(f, \mathcal{T}_n), \end{aligned} \tag{6.49}$$

where I denotes the identity on $\mathscr{C}_{2\pi}$. By Theorem 6.51 the sequence of operator norms $\|F_n\|_\infty = \lambda_n$ diverges *logarithmically*, so that

$$\|I - F_n\|_\infty \le \|I\|_\infty + \|F_n\|_\infty = \mathcal{O}(\log(n)) \quad \text{for } n \to \infty. \tag{6.50}$$

On the ground of this observation, the asymptotic analysis of the minimal distances $\eta_\infty(f, \mathcal{T}_n)$ is of primary interest: Namely, if we can show, for $f \in \mathscr{C}_{2\pi}$, that the sequence $(\eta_\infty(f, \mathcal{T}_n))_{n\in\mathbb{N}_0}$ converges to zero at least *algebraically*, so that

$$\log(n) \cdot \eta_\infty(f, \mathcal{T}_n) \longrightarrow 0 \quad \text{for } n \to \infty, \tag{6.51}$$

then the sequence $(F_n f)_{n\in\mathbb{N}_0}$ converges by (6.49) and (6.50) uniformly to f.

To this end, the following inequalities of Jackson[11] are indeed very useful.

We begin our asymptotic analysis of the minimal distances $\eta_\infty(f, \mathcal{T}_n)$ for *continuously differentiable* functions $f \in \mathscr{C}^1_{2\pi}$. Recall that in this case the uniform convergence of the Fourier partial sums $F_n f$ to f is already guaranteed by Corollary 6.46 and quantified by Corollary 6.47. Nevertheless, the following Jackson theorem is of fundamental importance for further investigations concerning convergence rates of the minimal distance $(\eta_\infty(f, \mathcal{T}_n))_{n\in\mathbb{N}_0}$.

[11] Dunham Jackson (1888-1946), US-American mathematician

Theorem 6.59. (Jackson 1). *For $f \in \mathscr{C}^1_{2\pi}$, we have*

$$\eta_\infty(f, \mathcal{T}_n) \le \frac{\pi}{2(n+1)} \|f'\|_\infty = \mathcal{O}(n^{-1}) \quad \text{for } n \to \infty. \tag{6.52}$$

Remark 6.60. The estimate of Jackson 1, Theorem 6.59, is *sharp*, i.e., there is a function $f \in \mathscr{C}^1_{2\pi} \setminus \mathcal{T}_n$ for which equality holds in (6.52). For more details, we refer to Exercise 6.95. □

Our proof for Theorem 6.59 is based on the following two lemmas.

Lemma 6.61. *We have*

$$\min_{a_1,\dots,a_n \in \mathbb{R}} \int_0^\pi \left| \xi - \sum_{j=1}^n a_j \sin(j\xi) \right| \mathrm{d}\xi = \frac{\pi^2}{2(n+1)}. \tag{6.53}$$

Lemma 6.62. *For $A_1, \dots, A_n \in \mathbb{R}$, let $L_n : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$ be a linear operator of the form*

$$(L_n f)(x) := \frac{a_0}{2} + \sum_{j=1}^n A_j \left[a_j \cos(jx) + b_j \sin(jx) \right] \qquad \text{for } f \in \mathscr{C}_{2\pi}, \tag{6.54}$$

where $a_0 = (f, 1)$, $a_j = (f, \cos(j\cdot))$ and $b_j = (f, \sin(j\cdot))$, for $1 \le j \le n$, are the Fourier coefficients of f in (6.29). Then we have, for $f \in \mathscr{C}^1_{2\pi}$, the error representation

$$(L_n f - f)(x) = \frac{1}{\pi} \int_{-\pi}^{\pi} \left[\frac{\xi}{2} + \sum_{j=1}^n \frac{(-1)^j}{j} A_j \sin(j\xi) \right] f'(x + \pi - \xi)\, \mathrm{d}\xi. \tag{6.55}$$

Now we can prove the statement of Jackson 1, Theorem 6.59.

Proof. **(Jackson 1).** For $f \in \mathscr{C}^1_{2\pi}$, we have for the minimal distance

$$\eta_\infty(f, \mathcal{T}_n) = \inf_{T \in \mathcal{T}_n} \|T - f\|_\infty$$

the estimate

$$\begin{aligned}
\eta_\infty(f, \mathcal{T}_n) &\le \|L_n f - f\|_\infty \\
&\le \|f'\|_\infty \cdot \frac{1}{\pi} \int_{-\pi}^{\pi} \left| \frac{\xi}{2} + \sum_{j=1}^n \frac{(-1)^j}{j} A_j \sin(j\xi) \right| \mathrm{d}\xi \\
&= \|f'\|_\infty \cdot \frac{1}{\pi} \int_0^{\pi} \left| \xi + \sum_{j=1}^n \frac{2(-1)^j}{j} A_j \sin(j\xi) \right| \mathrm{d}\xi \\
&= \|f'\|_\infty \cdot \frac{1}{\pi} \cdot \frac{\pi^2}{2(n+1)} \\
&= \|f'\|_\infty \cdot \frac{\pi}{2(n+1)},
\end{aligned}$$

where in the second line we use the error representation (6.55). Moreover, in the penultimate line we choose *optimal* coefficients $A_1, \dots, A_n$ according to (6.53). ■

Now let us prove the two lemmas.

Proof. **(Lemma 6.62).** Suppose $f \in \mathscr{C}^1_{2\pi}$. We use the notation

$$g(\xi) := \frac{\xi}{2} + \sum_{j=1}^{n} \frac{(-1)^j}{j} A_j \sin(j\xi)$$

for the first factor of the integrand in (6.55). This way we obtain

$$\begin{aligned}
&\frac{1}{\pi} \int_{-\pi}^{\pi} g(\xi) f'(x + \pi - \xi) \, \mathrm{d}\xi \\
&= -\frac{1}{\pi} \Big[g(\xi) f(x + \pi - \xi) \Big]_{\xi=-\pi}^{\xi=\pi} + \frac{1}{\pi} \int_{-\pi}^{\pi} g'(\xi) f(x + \pi - \xi) \, \mathrm{d}\xi \\
&= -\frac{1}{\pi} \frac{\pi}{2} f(x) - \frac{1}{\pi} \frac{\pi}{2} f(x + 2\pi) + \frac{1}{\pi} \int_{-\pi}^{\pi} g'(x + \pi - \sigma) f(\sigma) \, \mathrm{d}\sigma \\
&= -f(x) + \frac{1}{\pi} \int_{-\pi}^{\pi} g'(x + \pi - \sigma) f(\sigma) \, \mathrm{d}\sigma
\end{aligned}$$

after integration by parts from the error representation (6.55). Now we have

$$\begin{aligned}
&g'(x + \pi - \sigma) \\
&= \frac{1}{2} + \sum_{j=1}^{n} \frac{(-1)^j}{j} A_j \cdot j \cdot \cos(j(x + \pi - \sigma)) \\
&= \frac{1}{2} + \sum_{j=1}^{n} (-1)^j A_j \left[\cos(j(x + \pi)) \cos(j\sigma) + \sin(j(x + \pi)) \sin(j\sigma) \right] \\
&= \frac{1}{2} + \sum_{j=1}^{n} (-1)^j A_j \left[(-1)^j \left(\cos(jx) \cos(j\sigma) + \sin(jx) \sin(j\sigma) \right) \right] \\
&= \frac{1}{2} + \sum_{j=1}^{n} A_j \left[\cos(jx) \cos(j\sigma) + \sin(jx) \sin(j\sigma) \right]
\end{aligned}$$

and so

$$\begin{aligned}
\frac{1}{\pi} \int_{-\pi}^{\pi} g'(x + \pi - \sigma) f(\sigma) \, \mathrm{d}\sigma &= \frac{a_0}{2} + \sum_{j=1}^{n} A_j \left[a_j \cos(jx) + b_j \sin(jx) \right] \\
&= (L_n f)(x),
\end{aligned}$$

which already shows that the stated error representation holds. ■

Proof. **(Lemma 6.61).** For arbitrary $a_1, \ldots, a_n \in \mathbb{R}$, we have the estimate

$$\begin{aligned}
&\int_0^\pi \left| \xi - \sum_{j=1}^n a_j \sin(j\xi) \right| \mathrm{d}\xi \\
&\geq \left| \int_0^\pi \left[\xi - \sum_{j=1}^n a_j \sin(j\xi) \right] \operatorname{sgn}(\sin((n+1)\xi)) \, \mathrm{d}\xi \right| \qquad (6.56) \\
&= \left| \int_0^\pi \xi \cdot \operatorname{sgn}(\sin((n+1)\xi)) \, \mathrm{d}\xi \right| \qquad (6.57) \\
&= \left| \sum_{k=0}^n (-1)^k \int_{k\pi/(n+1)}^{(k+1)\pi/(n+1)} \xi \, \mathrm{d}\xi \right| \\
&= \left| \frac{1}{2} \frac{\pi^2}{(n+1)^2} \sum_{k=0}^n (-1)^k \left((k+1)^2 - k^2 \right) \right| \\
&= \left| \frac{\pi^2}{2(n+1)^2} \sum_{k=0}^n (-1)^k (2k+1) \right| \\
&= \left| \frac{\pi^2}{2(n+1)^2} \cdot (n+1) \right| = \frac{\pi^2}{2(n+1)},
\end{aligned}$$

where for the equality in (6.57) we use the identity

$$\int_0^\pi \sin(j\xi) \cdot \operatorname{sgn}(\sin((n+1)\xi)) \, \mathrm{d}\xi = 0 \qquad \text{for } j < n+1. \qquad (6.58)$$

We prove the statement (6.58) by Lemma 6.63.

For the solution of the optimization problem (6.53), we determine coefficients $a_1, \ldots, a_n \in \mathbb{R}$, such that equality holds in (6.56). In this case, the function

$$g(\xi) = \xi - \sum_{j=1}^n a_j \sin(j\xi)$$

must necessarily change signs at the points $\xi_k = k\pi/(n+1) \in (0, \pi)$, for $1 \leq k \leq n$. Indeed, this is because the function $\operatorname{sgn}(\sin((n+1)\xi))$ has sign changes on $(0, \pi)$ only at the points $\xi_1, \ldots, \xi_n$.

Note that this requirement yields n conditions on the sought coefficients $a_1, \ldots, a_n \in \mathbb{R}$, where these conditions are the interpolation conditions

$$\xi_k = \sum_{j=1}^n a_j \sin(j\xi_k) \qquad \text{for } 1 \leq k \leq n. \qquad (6.59)$$

But the interpolation problem (6.59) has a unique solution, since the trigonometric polynomials $\sin(j\cdot)$, $1 \leq j \leq n$, form a Haar system on $(0, \pi)$ (see Exercise 5.54). ∎

Finally, it remains to show the identity (6.58).

Lemma 6.63. *For $n \in \mathbb{N}$, we have the identity*

$$\int_0^\pi \sin(j\xi) \cdot \operatorname{sgn}(\sin((n+1)\xi))\,\mathrm{d}\xi = 0 \qquad \text{for } 1 \le j < n+1. \tag{6.60}$$

Proof. The integrand in (6.60) is an even function. Now we regard the integral in (6.60) on $[-\pi, \pi]$ (rather than on $[0, \pi]$). By using the identity

$$\sin(j\xi) = \frac{1}{2\mathrm{i}} \left(\mathrm{e}^{\mathrm{i}j\xi} - \mathrm{e}^{-\mathrm{i}j\xi}\right)$$

it is sufficient to show

$$I_j := \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}j\xi} \cdot \operatorname{sgn}(\sin((n+1)\xi))\,\mathrm{d}\xi = 0 \qquad \text{for } 1 \le |j| < n+1. \tag{6.61}$$

After the substitution $\xi = \sigma + \pi/(n+1)$ in (6.61) the representation

$$\begin{aligned} I_j &= \int_{-\pi-\pi/(n+1)}^{\pi-\pi/(n+1)} \mathrm{e}^{\mathrm{i}j(\sigma+\pi/(n+1))} \cdot \operatorname{sgn}(\sin((n+1)\sigma + \pi))\,\mathrm{d}\sigma \\ &= -\mathrm{e}^{\mathrm{i}j\pi/(n+1)} \int_{-\pi}^{\pi} \mathrm{e}^{\mathrm{i}j\sigma} \cdot \operatorname{sgn}(\sin((n+1)\sigma))\,\mathrm{d}\sigma \\ &= -\mathrm{e}^{\mathrm{i}j\pi/(n+1)} \cdot I_j \end{aligned}$$

holds. Since $-\mathrm{e}^{\mathrm{i}j\pi/(n+1)} \neq 1$, this implies $I_j = 0$ for $1 \le |j| < n+1$. ■

We wish to work with weaker conditions on f (i.e., weaker than $f \in \mathscr{C}^1_{2\pi}$ as in Jackson 1). In the next Jackson theorem, we only need *Lipschitz*[12] *continuity* for f.

Definition 6.64. *A function $f : [a,b] \longrightarrow \mathbb{R}$ is said to be* `Lipschitz continuous` *on $[a,b] \subset \mathbb{R}$, if there is a constant $L > 0$ satisfying*

$$|f(x) - f(y)| \le L|x - y| \qquad \text{for all } x, y \in \mathbb{R}.$$

In this case, L is called a `Lipschitz constant` *of f on $[a,b]$.* ○

Remark 6.65. For a compact interval $[a,b] \subset \mathbb{R}$, every function $f \in \mathscr{C}^1[a,b]$ is Lipschitz continuous on $[a,b]$. Indeed, this is because in this case, the mean value theorem applies, so that for any $x, y \in [a,b]$, we have the representation

$$f(x) - f(y) = f'(\xi) \cdot (x - y) \qquad \text{for some } \xi \in (a,b)$$

and this implies the estimate

$$|f(x) - f(y)| \le \|f'\|_\infty \cdot |x - y| \qquad \text{for all } x, y \in [a,b].$$

Therefore, $L = \|f'\|_\infty$ is a Lipschitz constant of f on $[a,b]$. □

[12] Rudolf Lipschitz (1832-1903), German mathematician

Theorem 6.66. (Jackson 2). *Let $f \in \mathscr{C}_{2\pi}$ be Lipschitz continuous on $[0, 2\pi]$ with Lipschitz constant $L > 0$. Then, we have*

$$\eta_\infty(f, \mathcal{T}_n) \leq \frac{\pi \cdot L}{2(n+1)} = \mathcal{O}(n^{-1}) \quad \text{for } n \to \infty.$$

Remark 6.67. The estimate of Jackson 2, Theorem 6.66, is *sharp*. For more details, we refer to Exercise 6.95. □

Proof. For $\delta > 0$, we consider the *local mean value* function

$$\varphi_\delta(x) = \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} f(\xi)\, \mathrm{d}\xi \qquad \text{for } x \in \mathbb{R} \tag{6.62}$$

of f on $(x - \delta, x + \delta)$. Then, we have

$$\varphi'_\delta(x) = \frac{f(x+\delta) - f(x-\delta)}{2\delta} \qquad \text{for all } x \in \mathbb{R},$$

and so φ_δ is in $\mathscr{C}^1_{2\pi}$. Moreover, φ_δ satisfies the uniform bound

$$|\varphi'_\delta(x)| \leq L \qquad \text{for all } x \in \mathbb{R},$$

i.e., $\|\varphi'_\delta\|_\infty \leq L$. By Jackson 1, Theorem 6.59, this implies

$$\eta_\infty(\varphi_\delta, \mathcal{T}_n) \leq \frac{\pi \cdot L}{2(n+1)}.$$

Moreover, we have

$$\begin{aligned} |\varphi_\delta(x) - f(x)| &= \frac{1}{2\delta} \left| \int_{x-\delta}^{x+\delta} (f(\xi) - f(x))\, \mathrm{d}\xi \right| \leq \frac{L}{2\delta} \int_{x-\delta}^{x+\delta} |\xi - x|\, \mathrm{d}\xi \\ &= \frac{L}{2\delta} \cdot \delta^2 = \frac{L}{2} \cdot \delta \longrightarrow 0 \quad \text{for } \delta \to 0. \end{aligned}$$

Now let $T^*(\varphi_\delta) \in \mathcal{T}_n$ be the best approximation to φ_δ from $\mathcal{T}_n$ with respect to $\|\cdot\|_\infty$, so that

$$\eta_\infty(\varphi_\delta, \mathcal{T}_n) = \|T^*(\varphi_\delta) - \varphi_\delta\|_\infty.$$

Then, we have

$$\begin{aligned} \eta_\infty(f, \mathcal{T}_n) &\leq \|T^*(\varphi_\delta) - f\|_\infty \\ &\leq \|T^*(\varphi_\delta) - \varphi_\delta\|_\infty + \|\varphi_\delta - f\|_\infty \\ &\leq \frac{\pi \cdot L}{2(n+1)} + \frac{L}{2} \cdot \delta, \end{aligned}$$

whereby for $\delta \searrow 0$, we obtain

$$\eta_\infty(f, \mathcal{T}_n) \leq \frac{\pi \cdot L}{2(n+1)}$$

as stated. ■

To obtain even weaker conditions on the target function $f \in \mathscr{C}_{2\pi}$, we now work with the *modulus of continuity*.

Definition 6.68. *For $[a,b] \subset \mathbb{R}$, let $f \in \mathscr{C}[a,b]$ and $\delta > 0$. Then,*

$$\omega(f,\delta) = \sup_{\substack{x, x+\sigma \in [a,b] \\ |\sigma| \le \delta}} |f(x+\sigma) - f(x)|$$

is called `modulus of continuity` *of f on $[a,b]$ with respect to δ.* ◯

Remark 6.69. Note that the modulus of continuity $\omega(f,\delta)$ quantifies the *local* distance between the function values of f uniformly on $[a,b]$. In fact, the smaller the modulus of continuity $\omega(f,\delta)$, the smaller is the *local* variation of f on $[a,b]$. For a compact interval $[a,b] \subset \mathbb{R}$, the modulus of continuity $\omega(f,\delta)$ of $f \in \mathscr{C}[a,b]$ is finite by

$$\omega(f,\delta) \le 2\|f\|_{\infty,[a,b]} \qquad \text{for all } b-a \ge \delta > 0,$$

and, moreover, we have the convergence

$$\omega(f,\delta) \longrightarrow 0 \quad \text{for } \delta \searrow 0.$$

For $f \in \mathscr{C}^1[a,b]$ and $x, x+\sigma \in [a,b]$, we have

$$f(x+\sigma) - f(x) = \sigma \cdot f'(\xi) \qquad \text{for some } \xi \in (x, x+\sigma)$$

by the mean value theorem, and so

$$\omega(f,\delta) \le \delta \cdot \|f'\|_\infty.$$

For a Lipschitz continuous function $f \in \mathscr{C}[a,b]$ with Lipschitz constant $L > 0$ we finally have

$$\omega(f,\delta) \le \delta \cdot L.$$

□

The following Jackson theorem gives an upper bound for the minimal distance $\eta_\infty(f, \mathcal{T}_n)$ by involving the modulus of continuity of $f \in \mathscr{C}_{2\pi}$.

Theorem 6.70. (Jackson 3). *For $f \in \mathscr{C}_{2\pi}$, we have*

$$\eta_\infty(f, \mathcal{T}_n) \le \frac{3}{2} \cdot \omega\left(f, \frac{\pi}{n+1}\right). \tag{6.63}$$

Remark 6.71. The estimate of Jackson 3, Theorem 6.70, is *not* sharp. For more details, we refer to Exercise 6.97. □

Proof. For the local mean value function $\varphi_\delta \in \mathscr{C}^1_{2\pi}$ of f on $(x-\delta, x+\delta)$ in (6.62), we can give a uniform bound on the pointwise error by

$$|\varphi_\delta(x) - f(x)| \le \frac{1}{2\delta} \left| \int_{x-\delta}^{x+\delta} (f(\xi) - f(x)) \, \mathrm{d}\xi \right| \le \frac{1}{2\delta} \cdot 2\delta \cdot \omega(f, \delta) = \omega(f, \delta).$$

Moreover, φ'_δ is uniformly bounded above by

$$\|\varphi'_\delta\|_\infty \le \frac{1}{2\delta} \cdot \omega(f, 2\delta).$$

Now let $T^*(\varphi_\delta) \in \mathcal{T}_n$ be the best approximation to φ_δ from $\mathcal{T}_n$ with respect to $\|\cdot\|_\infty$. Then, by Jackson 1, Theorem 6.59, this implies for $\delta > 0$ the estimate

$$\begin{aligned} \eta_\infty(f, \mathcal{T}_n) &\le \|T^*(\varphi_\delta) - f\|_\infty \\ &\le \|T^*(\varphi_\delta) - \varphi_\delta\|_\infty + \|\varphi_\delta - f\|_\infty \\ &\le \frac{\pi}{2(n+1)} \cdot \frac{1}{2\delta} \cdot \omega(f, 2\delta) + \omega(f, \delta) \\ &\le \omega(f, 2\delta) \left(\frac{\pi}{4\delta(n+1)} + 1 \right). \end{aligned}$$

Letting $\delta = \pi/(2(n+1))$, this gives the stated estimate in (6.63). ∎

Next, we analyze the asymptotic decay rate of the minimal distances $\eta_\infty(f, \mathcal{T}_n)$, for smoother target functions f. To be more precise, we prove asymptotic convergence rates for $f \in \mathscr{C}^k_{2\pi}$, $k \in \mathbb{N}$. Given our previous results, we can, for *smoother* $f \in \mathscr{C}^k_{2\pi}$, i.e., for larger k, expect *faster* convergence of the zero sequence $(\eta_\infty(f, \mathcal{T}_n))_{n \in \mathbb{N}_0}$. Our perception matches with the result of the following Jackson theorem.

Theorem 6.72. (Jackson 4). *For $f \in \mathscr{C}^k_{2\pi}$, $k \ge 1$, we have*

$$\eta_\infty(f, \mathcal{T}_n) \le \left(\frac{\pi}{2(n+1)} \right)^k \cdot \|f^{(k)}\|_\infty = \mathcal{O}\left(n^{-k}\right) \qquad \text{for } n \to \infty.$$

Our proof for Theorem 6.72 is based on two lemmas.

Lemma 6.73. *For $f \in \mathscr{C}^1_{2\pi}$ and $n \in \mathbb{N}$, the estimate*

$$\eta_\infty(f, \mathcal{T}_n) \le \frac{\pi}{2(n+1)} \cdot \eta_\infty(f', \mathcal{T}'_n),$$

holds, where the linear space

$$\mathcal{T}'_n := \operatorname{span}\{\cos(k\cdot), \sin(k\cdot) \mid 1 \le k \le n\} \qquad \text{for } n \in \mathbb{N}$$

consists of all trigonometric polynomials from $\mathcal{T}_n$ without the constants.

Remark 6.74. For $n \in \mathbb{N}$, we have

$$\mathcal{T}_n' = \{T' \in \mathscr{C}_{2\pi} \,|\, T \in \mathcal{T}_n\} \subset \mathcal{T}_n$$

and this explains our notation $\mathcal{T}_n'$. By $\mathcal{T}_n' \subset \mathcal{T}_n$, we find the estimate

$$\eta_\infty(f, \mathcal{T}_n) \le \eta_\infty(f, \mathcal{T}_n')$$

for all $f \in \mathscr{C}_{2\pi}$. □

Proof. **(Lemma 6.73).** Let $T^* \in \mathcal{T}_n'$ be best approximation to f' from $\mathcal{T}_n'$. For

$$T(x) := \int_0^x T^*(\xi)\,\mathrm{d}\xi \in \mathcal{T}_n$$

we have $T' = T^*$ and so

$$\|(T-f)'\|_\infty = \|T^* - f'\|_\infty = \eta_\infty(f', \mathcal{T}_n').$$

But this implies, by using Jackson 1, Theorem 6.59, the stated estimate:

$$\eta_\infty(f, \mathcal{T}_n) = \eta_\infty(T-f, \mathcal{T}_n) \le \frac{\pi}{2(n+1)} \cdot \|(T-f)'\|_\infty = \frac{\pi}{2(n+1)} \cdot \eta_\infty(f', \mathcal{T}_n').$$

■

Lemma 6.75. *Let $f \in \mathscr{C}^1_{2\pi}$ satisfy*

$$\int_{-\pi}^{\pi} f(x)\,\mathrm{d}x = 0. \tag{6.64}$$

Then we have, for any $n \in \mathbb{N}$, the two estimates

$$\eta_\infty(f, \mathcal{T}_n') \le \frac{\pi}{2(n+1)} \cdot \|f'\|_\infty \tag{6.65}$$

$$\eta_\infty(f, \mathcal{T}_n') \le \frac{\pi}{2(n+1)} \cdot \eta_\infty(f', \mathcal{T}_n'). \tag{6.66}$$

Proof. For the modified Fourier partial sum $L_n f$ in (6.54),

$$(L_n f)(x) = \frac{a_0}{2} + \sum_{k=1}^{n} A_k(a_k \cos(kx) + b_k \sin(kx)),$$

we have $a_0 \equiv a_0(f) = (f, 1) = 0$ by (6.64) and so $L_n f \in \mathcal{T}_n'$. Therefore, we have (6.65), since

$$\eta_\infty(f, \mathcal{T}_n') \le \|L_n f - f\|_\infty \le \|f'\|_\infty \cdot \frac{\pi}{2(n+1)}$$

holds for *optimal* coefficients $A_1, \ldots, A_n$ (like in the proof of Jackson 1).

To show (6.66), suppose that $T^* \in \mathcal{T}_n'$ is best approximation to f' from $\mathcal{T}_n'$. For

$$T(x) := \int_0^x T^*(\xi)\,\mathrm{d}\xi \in \mathcal{T}_n$$

we have $T' = T^*$. Moreover, for

$$S(x) := T(x) - \frac{1}{2\pi}\int_{-\pi}^{\pi} T(\xi)\,\mathrm{d}\xi = T(x) - \frac{a_0(T)}{2}$$

we have $a_0(S) = (S,1) = 0$. Therefore, $S \in \mathcal{T}_n'$ and $S' = T^*$. But this already implies the stated estimate (6.66) by

$$\eta_\infty(f,\mathcal{T}_n') = \eta_\infty(S-f,\mathcal{T}_n') \le \frac{\pi}{2(n+1)}\cdot\|S'-f'\|_\infty = \frac{\pi}{2(n+1)}\cdot\eta_\infty(f',\mathcal{T}_n').$$

■

Now we are in a position where we can prove Jackson 4, Theorem 6.72.

Proof. **(Jackson 4).** For $f \in \mathscr{C}_{2\pi}^1$ we have

$$\int_{-\pi}^{\pi} f'(\xi)\,\mathrm{d}\xi = f(\pi) - f(-\pi) = 0.$$

Now the estimate (6.66) in Lemma 6.75 implies

$$\eta_\infty(f',\mathcal{T}_n') \le \left(\frac{\pi}{2(n+1)}\right)^{k-2}\cdot\eta_\infty(f^{(k-1)},\mathcal{T}_n') \qquad \text{for } f \in \mathscr{C}_{2\pi}^{k-1}$$

by induction on $k \ge 2$. Moreover, by Lemma 6.73 and (6.65), we get

$$\begin{aligned}\eta_\infty(f,\mathcal{T}_n) &\le \frac{\pi}{2(n+1)}\cdot\eta_\infty(f',\mathcal{T}_n')\\ &\le \left(\frac{\pi}{2(n+1)}\right)^{k-1}\cdot\eta_\infty(f^{(k-1)},\mathcal{T}_n')\\ &\le \left(\frac{\pi}{2(n+1)}\right)^{k-1}\frac{\pi}{2(n+1)}\|f^{(k)}\|_\infty\\ &= \left(\frac{\pi}{2(n+1)}\right)^{k}\|f^{(k)}\|_\infty\end{aligned}$$

for $f \in \mathscr{C}_{2\pi}^k$, where $k \ge 1$. ■

Now we return to the discussion from the outset of this section concerning the uniform convergence of Fourier partial sums. In that discussion, we developed the error estimate (6.49),

$$\|F_n f - f\|_\infty \le \|I - F_n\|_\infty\cdot\eta_\infty(f,\mathcal{T}_n) \qquad \text{for } f \in \mathscr{C}_{2\pi}.$$

Moreover, we took further note on the application of the Jackson theorems. We summarize the discussion of this section by the *Dini*[13]*-Lipschitz theorem*, each of whose results follows directly from (6.51),

$$\log(n) \cdot \eta_\infty(f, \mathcal{T}_n) \longrightarrow 0 \quad \text{for } n \to \infty,$$

and one corresponding Jackson inequality in Theorems 6.66-6.72.

Theorem 6.76. (Dini-Lipschitz, 1872).
If $f \in \mathscr{C}_{2\pi}$ satisfies one of the following conditions, then the sequence $(F_n f)_{n\in\mathbb{N}_0}$ of Fourier partial sums $F_n f$ converges uniformly to f, i.e.,

$$\|F_n f - f\|_\infty \longrightarrow 0 \quad \textit{for } n \to \infty,$$

at the following convergence rates.

(a) If

$$\log(n) \cdot \omega(f, 1/n) = o(1) \quad \textit{for } n \to \infty,$$

then we have (by Jackson 3)

$$\|F_n f - f\|_\infty = o(1) \quad \textit{for } n \to \infty.$$

(b) If f is Lipschitz continuous, then we have (by Jackson 2)

$$\|F_n f - f\|_\infty = \mathcal{O}(\log(n)/n) \quad \textit{for } n \to \infty.$$

(c) If $f \in \mathscr{C}^k_{2\pi}$, for $k \geq 1$, then we have (by Jackson 4)

$$\|F_n f - f\|_\infty = \mathcal{O}(\log(n)/n^k) \quad \textit{for } n \to \infty.$$

□

Finally, we transfer the results of the theorems Jackson 1-4 concerning the approximation of $f \in \mathscr{C}_{2\pi}$ by trigonometric polynomials from $\mathcal{T}_n$ to the case of approximation of $f \in \mathscr{C}[-1, 1]$ by algebraic polynomials from $\mathcal{P}_n$.

Theorem 6.77. (Jackson 5). *For the minimal distances*

$$\eta_\infty(f, \mathcal{P}_n) = \inf_{p\in\mathcal{P}_n} \|p - f\|_{\infty,[-1,1]} \qquad \textit{for } f \in \mathscr{C}[-1, 1]$$

the following estimates hold.

(a) For $f \in \mathscr{C}[-1, 1]$, we have

$$\eta_\infty(f, \mathcal{P}_n) \leq \frac{3}{2} \cdot \omega\left(f, \frac{\pi}{n+1}\right).$$

[13] Ulisse Dini (1845-1918), Italian mathematician and politician

(b) If f is Lipschitz continuous with Lipschitz constant $L > 0$, then we have

$$\eta_\infty(f, \mathcal{P}_n) \le \frac{3\pi \cdot L}{2(n+1)}.$$

(c) For $f \in \mathscr{C}^k[-1,1]$, $k \ge 1$, we have

$$\begin{aligned}\eta_\infty(f, \mathcal{P}_n) &\le \left(\frac{\pi}{2}\right)^k \frac{1}{(n+1)n(n-1)\dots(n-(k-2))} \|f^{(k)}\|_\infty \\ &= \mathcal{O}(n^{-k}) \quad \text{for } n \to \infty.\end{aligned}$$

We split the proof of Jackson 5, Theorem 6.77, into several lemmas. The following lemma reveals the structural connection between the trigonometric and the algebraic case.

Lemma 6.78. *For $f \in \mathscr{C}[-1,1]$ and $g(\varphi) = f(\cos(\varphi)) \in \mathscr{C}_{2\pi}$, we have*

$$\eta_\infty(f, \mathcal{P}_n) = \eta_\infty(g, \mathcal{T}_n).$$

Proof. For $f \in \mathscr{C}[-1,1]$ the function $g \in \mathscr{C}_{2\pi}$ is even. Therefore, the unique best approximation $T^* \in \mathcal{T}_n$ to g is even, so that we have

$$T^*(\varphi) = p(\cos(\varphi)) \qquad \text{for } \varphi \in [0, 2\pi]$$

for some $p \in \mathcal{P}_n$. Moreover, we find the relation

$$\eta_\infty(g, \mathcal{T}_n) = \|T^* - g\|_\infty = \|p - f\|_\infty = \|p^* - f\|_\infty = \eta_\infty(f, \mathcal{P}_n),$$

where $p^* \in \mathcal{P}_n$ is the unique best approximation to f from $\mathcal{P}_n$. ∎

Lemma 6.79. *For $f \in \mathscr{C}[-1,1]$ and $g(\varphi) = f(\cos(\varphi)) \in \mathscr{C}_{2\pi}$, we have*

$$\omega(g, \delta) \le \omega(f, \delta) \qquad \text{for all } \delta > 0.$$

Proof. By the mean value theorem, we have

$$|\cos(\varphi + \varepsilon) - \cos(\varphi)| \le \varepsilon \qquad \text{for } \varepsilon > 0$$

which in turn implies

$$\begin{aligned}\omega(g, \delta) &= \sup_{|\varepsilon| \le \delta} |g(\varphi + \varepsilon) - g(\varphi)| = \sup_{|\varepsilon| \le \delta} |f(\cos(\varphi + \varepsilon)) - f(\cos(\varphi))| \\ &\le \sup_{|\sigma| \le \delta} |f(x + \sigma) - f(x)| = \omega(f, \delta).\end{aligned}$$

∎

Now we prove statements (a) and (b) of Theorem 6.77.

Proof. **(Jackson 5, parts (a),(b)).**

(a): From Jackson 3, Theorem 6.70, in combination with Lemma 6.78 and Lemma 6.79, we can conclude

$$\eta_\infty(f,\mathcal{P}_n) = \eta_\infty(g,\mathcal{T}_n) \le \frac{3}{2}\cdot\omega\left(g,\frac{\pi}{n+1}\right) \le \frac{3}{2}\cdot\omega\left(f,\frac{\pi}{n+1}\right)$$

for $f \in \mathscr{C}[-1,1]$.

(b): Statement (a) implies, for Lipschitz continuous $f \in \mathscr{C}[-1,1]$ with Lipschitz constant L the estimate

$$\eta_\infty(f,\mathcal{P}_n) \le \frac{3}{2}\cdot\omega\left(f,\frac{\pi}{n+1}\right) \le \frac{3}{2}\cdot\frac{\pi\cdot L}{n+1}.$$

■

To prove part (c) of Jackson 5, Theorem 6.77, we use the following lemma.

Lemma 6.80. *For $f \in \mathscr{C}^1[-1,1]$, we have*

$$\eta_\infty(f,\mathcal{P}_n) \le \frac{\pi}{2(n+1)}\eta_\infty(f',\mathcal{P}_{n-1}).$$

Proof. Let $p^* \in \mathcal{P}_{n-1}$ be best approximation to f' from $\mathcal{P}_{n-1}$. For

$$p(x) = \int_0^x p^*(\xi)\,\mathrm{d}\xi \in \mathcal{P}_n$$

we have $p' = p^*$ and so

$$\eta_\infty(f,\mathcal{P}_n) = \eta_\infty(p-f,\mathcal{P}_n) \le \frac{\pi}{2(n+1)}\|p'-f'\|_\infty = \frac{\pi}{2(n+1)}\eta_\infty(f',\mathcal{P}_{n-1})$$

holds by Jackson 1, Theorem 6.59, and Lemma 6.78. ■

Now we can prove statement (c) of Jackson 5, Theorem 6.77.

Proof. **(Jackson 5, part (c)).**
For $f \in \mathscr{C}^k[-1,1]$, we obtain from Lemma 6.80 the estimate

$$\eta_\infty(f,\mathcal{P}_n) \le \left(\frac{\pi}{2}\right)^k \frac{1}{(n+1)n(n-1)\dots(n+2-k)}\cdot\eta_\infty(f^{(k)},\mathcal{P}_{n-k})$$

by induction on $k \ge 1$. This already implies, by

$$\eta_\infty(f^{(k)},\mathcal{P}_{n-k}) \le \|f^{(k)}-0\|_\infty = \|f^{(k)}\|_\infty,$$

the stated estimate

$$\eta_\infty(f,\mathcal{P}_n) \le \left(\frac{\pi}{2}\right)^k \frac{1}{(n+1)n(n-1)\dots(n-(k-2))}\|f^{(k)}\|_\infty.$$

■

We close this chapter by giving a reformulation of the Dini-Lipschitz theorem for the algebraic case.

Theorem 6.81. (Dini-Lipschitz).
If $f \in \mathscr{C}[-1,1]$ satisfies one of the following conditions, then the sequence $(\Pi_n f)_{n\in\mathbb{N}_0}$ of `Chebyshev partial sums`

$$\Pi_n f = \sum_{k=0}^{n} \frac{(f,T_k)_w}{\|T_k\|_w^2} T_k$$

in (4.32) converges uniformly to f, i.e.,

$$\|\Pi_n f - f\|_\infty \longrightarrow 0 \quad \text{for } n \to \infty,$$

at the following convergence rates.

(a) If

$$\log(n) \cdot \omega(f, 1/n) = o(1) \quad \text{for } n \to \infty,$$

then we have by Jackson 5 (a)

$$\|\Pi_n f - f\|_\infty = o(1) \quad \text{for } n \to \infty.$$

(b) If f is Lipschitz continuous, then we have by Jackson 5 (b)

$$\|\Pi_n f - f\|_\infty = \mathcal{O}(\log(n)/n) \quad \text{for } n \to \infty.$$

(c) If $f \in \mathscr{C}^k[-1,1]$, for $k \geq 1$, then we have by Jackson 5 (c)

$$\|\Pi_n f - f\|_\infty = \mathcal{O}(\log(n)/n^k) \quad \text{for } n \to \infty.$$

□

For the proof of the Dini-Lipschitz theorem we refer to Exercise 6.99.

6.5 Exercises

Exercise 6.82. Prove the following results.

(a) Show that for a set of $n+1$ pairwise distinct interpolation points

$$a \le x_0 < \ldots < x_n \le b$$

where $n \in \mathbb{N}$, the corresponding interpolation operator $I_n : \mathscr{C}[a,b] \longrightarrow \mathcal{P}_n$ is not necessarily monotone. To this end, construct one counterexample for the case $n = 2$ with three interpolation points $a = x_0 < x_1 < x_2 = b$.

(b) Develop for the case $n = 1$ a necessary and sufficient condition for two interpolation points

$$a \le x_0 < x_1 \le b$$

under which the interpolation operator I_1 is monotone.

Exercise 6.83. For $n \in \mathbb{N}_0$, consider the Bernstein polynomials

$$\beta_k^{(n)}(x) = \binom{n}{k} x^k \, (1-x)^{n-k} \in \mathcal{P}_n \qquad \text{for } x \in [0,1] \text{ and } 0 \le k \le n.$$

(a) Show that the Bernstein polynomials $\beta_k^{(n)}$ are non-negative on the unit interval $[0,1]$, where they are a partition of unity (cf. Remark 6.7 (b),(c)).

(b) Determine the zeros (including their multiplicities) and the maximum of the Bernstein polynomial $\beta_k^{(n)}$ on $[0,1]$, for $0 \le k \le n$, $n \in \mathbb{N}_0$.

(c) Prove the recursion formula

$$\beta_k^{(n)}(x) = x \, \beta_{k-1}^{(n-1)}(x) + (1-x) \, \beta_k^{(n-1)}(x) \qquad \text{for } x \in [0,1],$$

for $n \in \mathbb{N}$ and $k = 0, \ldots, n$, with initial and boundary values

$$\beta_0^{(0)} \equiv 1, \quad \beta_{-1}^{(n-1)} \equiv 0, \quad \beta_n^{(n-1)} \equiv 0.$$

(d) Show that the Bernstein polynomials $\beta_0^{(n)}, \ldots, \beta_n^{(n)}$ of degree $n \in \mathbb{N}_0$ form a basis for the polynomial space $\mathcal{P}_n$ (cf. Remark 6.7 (a)).

Exercise 6.84. Consider the Bernstein operator $B_n : \mathscr{C}[0,1] \longrightarrow \mathcal{P}_n$,

$$(B_n f)(x) = \sum_{j=0}^{n} f(j/n) \beta_j^{(n)}(x) \qquad \text{for } f \in \mathscr{C}[0,1] \text{ and } n \in \mathbb{N}_0,$$

with the Bernstein polynomials $\beta_j^{(n)}(x) = \binom{n}{j} x^j (1-x)^{n-j}$, for $0 \le j \le n$.

Show that, for any $f \in \mathscr{C}[0,1]$, the sequence $((B_n f)')_{n \in \mathbb{N}_0}$ of derivatives of $B_n f$ converges uniformly on $[0,1]$ to f', i.e.,

$$\lim_{n \to \infty} \| B_n'(f) - f' \|_\infty = 0.$$

Exercise 6.85. Prove the following results.

(a) For a compact interval $[a,b] \subset \mathbb{R}$, let $f \in \mathscr{C}[a,b]$. Show that f vanishes identically on $[a,b]$, if and only if all *moments* of f on $[a,b]$ vanish, i.e., if and only if

$$m_n = \int_a^b x^n \, f(x) \, \mathrm{d}x = 0 \qquad \text{for all } n \in \mathbb{N}_0.$$

(b) Suppose $f \in \mathscr{C}_{2\pi}$. Show that f vanishes identically on $\mathbb{R}$, if and only if all *Fourier coefficients* of f vanish, i.e., if and only if

$$c_j = \frac{1}{2\pi} \int_0^{2\pi} f(x) \, \mathrm{e}^{-\mathrm{i}jx} \, \mathrm{d}x = 0 \qquad \text{for all } j \in \mathbb{Z}.$$

Exercise 6.86. Prove the following generalization of the Korovkin theorem.

Let $\Omega \subset \mathbb{R}^d$ be a compact domain, where $d \in \mathbb{N}$. Moreover, suppose for $s_1, \ldots, s_m \in \mathscr{C}(\Omega)$ that there are functions $a_1, \ldots, a_m \in \mathscr{C}(\Omega)$ satisfying

$$p_t(x) = \sum_{j=1}^{m} a_j(t) \, s_j(x) \geq 0 \qquad \text{for all } t, x \in \Omega,$$

where $p_t(x) = 0$, if and only if $t = x$. Then, for any sequence $(L_n)_{n\in\mathbb{N}}$ of linear positive operators $L_n : \mathscr{C}(\Omega) \longrightarrow \mathscr{C}(\Omega)$ satisfying

$$\lim_{n\to\infty} \|L_n s_j - s_j\|_\infty = 0 \qquad \text{for all } 1 \leq j \leq m$$

we have the convergence

$$\lim_{n\to\infty} \|L_n s - s\|_\infty = 0 \qquad \text{for all } s \in \mathscr{C}(\Omega).$$

Conclude from this the statement of the Korovkin theorem, Theorem 6.11.

Exercise 6.87. Consider for $n \in \mathbb{N}_0$ the operator $\Pi_n^* : \mathscr{C}_{2\pi} \longrightarrow \mathcal{T}_n$, which assigns $f \in \mathscr{C}_{2\pi}$ to its (strongly unique) best approximation $\Pi_n^* f$ from $\mathcal{T}_n$ with respect to $\|\cdot\|_\infty$, so that

$$\eta_\infty(f, \mathcal{T}_n) = \inf_{T\in\mathcal{T}_n} \|T - f\|_\infty = \|\Pi_n^* f - f\|_\infty.$$

Investigate Π_n^* for the following (possible) properties.

(a) projection property;
(b) surjectivity;
(c) linearity;
(d) continuity;
(e) boundedness.

Exercise 6.88. Let $(u_n)_{n\in\mathbb{Z}}$ be a system of elements in a Hilbert space $\mathcal{H}$. Prove the equivalence of the following two statements.

(a) The system $(u_n)_{n\in\mathbb{Z}}$ is a Riesz basis of $\mathcal{H}$.
(b) There is a linear, continuous and invertible operator $T : \mathcal{H} \longrightarrow \mathcal{H}$ and a complete orthonormal system $(e_n)_{n\in\mathbb{Z}}$ in $\mathcal{H}$ satisfying

$$Te_n = u_n \qquad \text{for all } n \in \mathbb{Z}.$$

Exercise 6.89. Prove the completeness of the orthonormal system

$$\{\mathrm{e}^{\mathrm{i}j\cdot} \mid j \in \mathbb{Z}\} \subset \mathscr{C}^{\mathbb{C}}_{2\pi}$$

in $\mathscr{C}^{\mathbb{C}}_{2\pi}$ with respect to the Euclidean norm $\|\cdot\|_{\mathbb{C}}$.

Hint: Corollary 6.41 and Remark 6.42.

Exercise 6.90. Consider the linear space $\mathscr{C}^{\mathbb{C}}_{2L}$ of complex-valued $2L$-periodic continuous functions, equipped with the inner product

$$(f, g) = \frac{1}{2L}\int_0^{2L} f(x)\cdot\overline{g(x)}\,\mathrm{d}x \qquad \text{for } f, g \in \mathscr{C}^{\mathbb{C}}_{2L}.$$

(a) Determine a complete orthonormal system $(e_j)_{j\in\mathbb{Z}}$ in $\mathscr{C}^{\mathbb{C}}_{2L}$.
(b) Develop the Fourier coefficients $c_j = (f, e_j)$ of $f \in \mathscr{C}^{\mathbb{C}}_{2L}$.
(c) Formulate the Parseval identity in $\mathscr{C}^{\mathbb{C}}_{2L}$ with respect to $(e_j)_{j\in\mathbb{Z}}$.

Exercise 6.91. Let $c_j(f)$ be the complex Fourier coefficients of $f \in \mathscr{C}_{2\pi}$. Show that the estimate

$$|c_j(f)| \le C(1 + |j|)^{-(k+1+\varepsilon)} \qquad \text{for all } j \in \mathbb{Z},$$

for some $C > 0$ and $\varepsilon > 0$, implies $f \in \mathscr{C}^k_{2\pi}$ (cf. Remark 6.45).

Hint: Analyze the (uniform) convergence of the Fourier partial sums

$$(F_n f)(x) = \sum_{j=-n}^{n} c_j(f)\mathrm{e}^{\mathrm{i}jx}$$

and their derivatives.

Exercise 6.92. Show for $f \in \mathscr{C}^1_{2\pi}$ the identity

$$F_n f' = (F_n f)' \qquad \text{for all } n \in \mathbb{N}$$

for the Fourier partial sums $F_n f'$ of the derivative $f' \in \mathscr{C}_{2\pi}$.

Exercise 6.93. Prove Faber's theorem, Theorem 6.57: For any sequence $(I_n)_{n\in\mathbb{N}_0}$ of interpolation operators $I_n : \mathscr{C}[a, b] \longrightarrow \mathcal{P}_n$, there is a continuous function $f \in \mathscr{C}[a, b]$, for which the corresponding sequence $(I_n f)_{n\in\mathbb{N}_0}$ of interpolation polynomials $I_n f \in \mathcal{P}_n$ does *not* converge uniformly to f.

Exercise 6.94. Let $[a,b] \subset \mathbb{R}$ be a compact interval. For the numerical integration of

$$I_a^b(f) = \int_a^b f(x)\,\mathrm{d}x \qquad \text{for } f \in \mathscr{C}[a,b]$$

we apply the *Newton-Cotes*[14] *quadrature*. For $n \in \mathbb{N}$, the n-th Newton-Cotes quadrature formula is defined as

$$Q_n(f) = (b-a)\sum_{j=0}^{n} \alpha_{j,n}\, f(x_{j,n})$$

at equidistant knots

$$x_{j,n} = a + j\,\frac{b-a}{n} \qquad \text{for } j = 0,\dots,n$$

and weights

$$\alpha_{j,n} = \frac{1}{b-a}\int_a^b L_{j,n}(x)\,\mathrm{d}x \qquad \text{for } j = 0,\dots,n,$$

where $\{L_{0,n},\dots,L_{n,n}\} \subset \mathcal{P}_n$ are the Lagrange basis functions for the knot set $X_n = \{x_{0,n},\dots,x_{n,n}\}$ (cf. the discussion on Lagrange bases in Section 2.3).

Show that there is a continuous function $f \in \mathscr{C}[a,b]$, for which the sequence of Newton-Cotes approximations $((Q_n f))_{n\in\mathbb{N}}$ diverges.

Hint: Apply the *Kuzmin*[15] *theorem*, according to which the sum of the weights' moduli $|\alpha_{j,n}|$ diverges, i.e.,

$$\sum_{j=0}^{n} |\alpha_{j,n}| \longrightarrow \infty \qquad \text{for } n \to \infty.$$

Exercise 6.95. Show that the estimate of Jackson 1, Theorem 6.59,

$$\eta_\infty(f,\mathcal{T}_n) \le \frac{\pi}{2(n+1)}\|f'\|_\infty \qquad \text{for } f \in \mathscr{C}^1_{2\pi}, \tag{6.67}$$

is *sharp*, i.e., there is a function $f \in \mathscr{C}^1_{2\pi}\setminus\mathcal{T}_n$ for which equality holds in (6.67).

Conclude from this that the estimate of Jackson 2, Theorem 6.66,

$$\eta_\infty(f,\mathcal{T}_n) \le \frac{\pi\cdot L}{2(n+1)} \qquad \text{for } f \text{ Lipschitz continuous with constant } L > 0$$

is also sharp.

[14] Roger Cotes (1682-1716), English mathematician
[15] Rodion Ossijewitsch Kuzmin (1891-1949), Russian mathematician

Exercise 6.96. Prove the *theorem of de La Vallée Poussin*[16] : Let $f \in \mathscr{C}_{2\pi}$ and $T_n \in \mathcal{T}_n$. If there exist $2n+2$ pairwise distinct points

$$0 \le x_0 < \ldots < x_{2n+1} < 2\pi,$$

such that $T_n - f$ has alternating signs on x_k, $k = 0, \ldots, 2n+1$, then we have

$$\eta_\infty(f, \mathcal{T}_n) \ge \min_{0 \le k \le 2n+1} |(T_n - f)(x_k)|.$$

Exercise 6.97. The estimate of Jackson 3, Theorem 6.70, is *not* sharp. Show that the estimate

$$\eta_\infty(f, \mathcal{T}_n) \le \omega\left(f, \frac{\pi}{n+1}\right) \qquad \text{for } f \in \mathscr{C}_{2\pi}$$

is sharp (under the assumptions and with the notations in Theorem 6.70).

Hint: Apply the theorem of de La Vallée Poussin from Exercise 6.96.

Exercise 6.98. Verify the following properties of the modulus of continuity

$$\omega(f, \delta) = \sup_{\substack{x, x+\sigma \in \mathbb{R} \\ |\sigma| \le \delta}} |f(x+\sigma) - f(x)|$$

of $f : \mathbb{R} \longrightarrow \mathbb{R}$ on $\mathbb{R}$ with respect to $\delta > 0$ (cf. Definition 6.68).

(a) $\omega(f, (n+\theta)\delta) \le n\omega(f, \delta) + \omega(f, \theta\delta)$ for all $\theta \in [0, 1)$ and $n \in \mathbb{N}$.
(b) $\omega(f, \delta) \le n\omega(f, \delta/n)$ for all $n \in \mathbb{N}$.

Exercise 6.99. Prove part (c) of the Dini-Lipschitz theorem, Theorem 6.81, in two steps as follows. First show that, for any $f \in \mathscr{C}^1[-1, 1]$, the sequence $(\mathit{\Pi}_n f)_{n \in \mathbb{N}_0}$ of Chebyshev partial sums

$$\mathit{\Pi}_n f = \sum_{j=0}^{n} \frac{(f, T_j)_w}{\|T_j\|_w^2} T_j \qquad \text{where } T_j = \cos(j \arccos(\cdot)) \in \mathcal{P}_j$$

converges *uniformly* on $[-1, 1]$ to f, i.e.,

$$\lim_{n \to \infty} \|\mathit{\Pi}_n f - f\|_\infty = 0.$$

From this conclude, for $f \in \mathscr{C}^k[-1, 1]$, $k \ge 1$, the convergence behaviour

$$\|\mathit{\Pi}_n f - f\|_\infty = o(n^{1-k}) \quad \text{for} \quad n \to \infty.$$

[16] Charles-Jean de La Vallée Poussin (1866-1962), Belgian mathematician

7 Basic Concepts of Signal Approximation

In this chapter, we study basic concepts of mathematical signal analysis. To this end, we first introduce the *continuous* Fourier transform $\mathcal{F}$,

$$(\mathcal{F}f)(\omega) = \int_{\mathbb{R}} f(x) \cdot \mathrm{e}^{-\mathrm{i}x\omega} \, \mathrm{d}\omega \qquad \text{for } f \in \mathrm{L}^1(\mathbb{R}), \tag{7.1}$$

as a linear integral transform on the Banach space $\mathrm{L}^1(\mathbb{R})$ of absolutely Lebesgue-integrable functions. We motivate the transfer from Fourier series of *periodic* functions $f \in \mathscr{C}_{2\pi}^{\mathbb{C}}$ to Fourier transforms of *non-periodic* functions $f \in \mathrm{L}^1(\mathbb{R})$. In particular, we provide a heuristic account to the Fourier transformation $\mathcal{F}f$ in (7.1), where we depart from Fourier partial sums $F_n f$, for $f \in \mathscr{C}_{2\pi}$. Then, we analyze the following relevant questions.

(1) Is the Fourier transform $\mathcal{F}$ invertible?
(2) Can $\mathcal{F}$ be transferred to the Hilbert space $\mathrm{L}^2(\mathbb{R})$?
(3) Can $\mathcal{F}$ be applied to multivariate functions $f \in \mathrm{L}^p(\mathbb{R}^d)$, for $p = 1, 2$?

We give positive answers to all questions (1)-(3). The answer to (1) leads us, for $f \in \mathrm{L}^1(\mathbb{R})$, with $\mathcal{F}f \in \mathrm{L}^1(\mathbb{R})$, to the Fourier inversion formula

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} (\mathcal{F}f)(\omega) \mathrm{e}^{\mathrm{i}x\omega} \, \mathrm{d}\omega \qquad \text{for almost every } x \in \mathbb{R}.$$

To analyze (2), we study the spectral properties of $\mathcal{F}$, where we identify the *Hermite functions* h_n in (4.55) as eigenfunctions of $\mathcal{F}$. As we show, the Hermite functions $(h_n)_{n \in \mathbb{N}_0}$ form a complete orthogonal system in the Hilbert space $\mathrm{L}^2(\mathbb{R})$. This result leads us to the *Plancherel theorem*, Theorem 7.30, providing the continuous extension of $\mathcal{F}$ to an automorphism on $\mathrm{L}^2(\mathbb{R})$. The basic properties of the Fourier operator $\mathcal{F}$ can be generalized from the univariate case to the multivariate case, and this gives an answer to (3).

Finally, we formulate and prove the celebrated *Shannon sampling theorem*, Theorem 7.34 (in Section 7.3), giving a fundamental result of mathematical signal processing. According to the Shannon sampling theorem, a *signal*, i.e., a *function* $f \in \mathrm{L}^2(\mathbb{R})$, with bounded frequency density can be reconstructed *exactly* from its samples (i.e., function values) on an infinite uniform grid at a sufficiently small sampling rate. Our proof of the Shannon sampling theorem serves to demonstrate the relevance and the significance of the introduced Fourier methods.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_7

The second half of this chapter is devoted to *wavelets*. *Wavelets* are popular and powerful tools of modern mathematical signal processing, in particular for the approximation of functions $f \in \mathrm{L}^2(\mathbb{R})$. A wavelet approximation to f is essentially based on a *multiresolution of* $\mathrm{L}^2(\mathbb{R})$, i.e., on a nested sequence

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset V_{j-1} \subset V_j \subset \cdots \subset \mathrm{L}^2(\mathbb{R}) \tag{7.2}$$

of closed *scale spaces* $V_j \subset \mathrm{L}^2(\mathbb{R})$. The nested sequence in (7.2) leads us to stable approximation methods, where f is represented on different frequency bands by orthogonal projectors $\Pi_j : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_j$. More precisely, for a fixed scaling function $\varphi \in \mathrm{L}^2(\mathbb{R})$, the scale spaces $V_j \subset \mathrm{L}^2(\mathbb{R})$ in (7.2) are generated by dilations and translations of basis functions $\varphi_k^j(x) := 2^{j/2}\varphi(2^j x - k)$, for $j, k \in \mathbb{Z}$, so that

$$V_j = \overline{\operatorname{span}\{\varphi_k^j : k \in \mathbb{Z}\}} \subset \mathrm{L}^2(\mathbb{R}) \qquad \text{for } j \in \mathbb{Z}.$$

Likewise, for a corresponding *wavelet* function $\psi \in \mathrm{L}^2(\mathbb{R})$, the orthogonal complement $W_j \subset V_{j+1}$ of V_j in V_{j+1},

$$V_{j+1} = W_j \oplus V_j \qquad \text{for } j \in \mathbb{Z},$$

is generated by basis functions $\psi_k^j(x) := 2^{j/2}\psi(2^j x - k)$, for $j, k \in \mathbb{Z}$, so that

$$W_j = \overline{\operatorname{span}\{\psi_k^j \,|\, k \in \mathbb{Z}\}} \qquad \text{for } j \in \mathbb{Z}.$$

The basic construction of wavelet approximations to $f \in \mathrm{L}^2(\mathbb{R})$ is based on *refinement equations* of the form

$$\varphi(x) = \sum_{k\in\mathbb{Z}} h_k \varphi(2x - k) \qquad \text{and} \qquad \psi(x) = \sum_{k\in\mathbb{Z}} g_k \varphi(2x - k),$$

for specific *coefficient masks* $(h_k)_{k\in\mathbb{Z}}, (g_k)_{k\in\mathbb{Z}} \subset \ell^2$.

The development of wavelet methods, dating back to the early 1980s, has since then gained enormous popularity in applications of information technology, especially in image and signal processing. Inspired by a wide range of applications in science and engineering, this has led to rapid progress concerning both computational methods and the mathematical theory of wavelets.

Therefore, it is by no means possible for us to give a complete overview over the multiple facets of wavelet methods. Instead, we have decided to present selected basic principles of wavelet approximation. To this end, we restrict the discussion of this chapter to the rather simple *Haar wavelet*

$$\psi(x) = \chi_{[0,1/2)}(x) - \chi_{[1/2,1)}(x) = \begin{cases} 1 & \text{for } x \in [0, 1/2), \\ -1 & \text{for } x \in [1/2, 1), \\ 0 & \text{otherwise,} \end{cases}$$

and its corresponding scaling function $\varphi = \chi_{[0,1)}$. For a more comprehensive account to the mathematical theory of wavelets, we recommend the classical textbooks [14, 18, 49] and, moreover, the more recent textbooks [29, 31, 69] for a more pronounced connection to Fourier analysis.

7.1 The Continuous Fourier Transform

In Section 4.3, we introduced Fourier partial sums $F_n f$ to approximate continuous periodic functions $f \in \mathscr{C}_{2\pi}$, or, $f \in \mathscr{C}_{2\pi}^{\mathbb{C}}$. Recall that we restricted ourselves (without loss of generality) to continuous functions f with period $T = 2\pi$ (according to Definition 2.32).

In the following discussion, we assume $f \equiv f_T : \mathbb{R} \longrightarrow \mathbb{C}$ to be a time-continuous *signal* (i.e., a *function*) with period T, for some $T > 0$. In this case, by following along the lines of our derivations in Section 4.3, we obtain for the complex n-th Fourier partial sum $F_n f_T$ of f_T the representation

$$(F_n f_T)(x) = \sum_{j=-n}^{n} c_j \mathrm{e}^{\mathrm{i}j\omega x} \tag{7.3}$$

with the frequency $\omega = 2\pi/T$ and with the complex Fourier coefficients

$$c_j = \frac{1}{T} \int_0^T f_T(\xi) \mathrm{e}^{-\mathrm{i}j\omega\xi} \, \mathrm{d}\xi = \frac{1}{T} \int_{-T/2}^{T/2} f_T(\xi) \mathrm{e}^{-\mathrm{i}j\omega\xi} \, \mathrm{d}\xi \tag{7.4}$$

for $j = -n, \ldots, n$, as in (4.23) and in (4.24). Technically speaking, the Fourier coefficient c_j gives the amplification factor for the Fourier mode $\mathrm{e}^{-\mathrm{i}j\omega\cdot}$ of the frequency $\omega_j = j \cdot \omega = j \cdot 2\pi/T$, for $j = -n, \ldots, n$, i.e., the Fourier coefficients c_j yield the amplitudes of the present fundamental Fourier modes $\mathrm{e}^{-\mathrm{i}j\omega\cdot}$.

Then, in Section 6.3 we analyzed the convergence of Fourier partial sums. According to Theorem 6.48, the Fourier series representation

$$f_T(x) = \sum_{j=-\infty}^{\infty} c_j \mathrm{e}^{\mathrm{i}j\omega x} \tag{7.5}$$

holds pointwise at all points $x \in \mathbb{R}$, where f_T is differentiable.

The bi-infinite sequence $(c_j)_{j\in\mathbb{Z}}$ of complex Fourier coefficients in (7.5) is called the *discrete Fourier spectrum* of the signal f_T. Therefore, the discrete Fourier spectrum of f_T is a sequence $(c_j)_{j\in\mathbb{Z}}$ of Fourier coefficients from which we can reconstruct f_T exactly via (7.5). If only finitely many Fourier coefficients c_j do not vanish, then only the frequencies of the (finitely many) corresponding Fourier modes $\mathrm{e}^{-\mathrm{i}j\omega\cdot}$ will appear in the representation (7.5). In this case, the *frequency spectrum* of f_T is bounded.

Now we derive an alternative representation for the Fourier series in (7.5). To this end, we first introduce the *mesh width*

$$\Delta\omega := \omega_{j+1} - \omega_j \equiv \frac{2\pi}{T} \qquad \text{for } j \in \mathbb{Z}$$

as the difference between two consecutive frequencies of the Fourier modes $\mathrm{e}^{-\mathrm{i}j\omega\cdot}$. Then, the Fourier series in (7.5) can be written as

$$f_T(x) = \sum_{j=-\infty}^{\infty} \frac{1}{2\pi} \int_{-T/2}^{T/2} f_T(\xi) \mathrm{e}^{\mathrm{i}\omega_j(x-\xi)} \,\mathrm{d}\xi \cdot \Delta\omega. \tag{7.6}$$

The Fourier series representation (7.6) of the T-periodic signal f_T leads us to the following questions.

- Is there a representation as in (7.6) for *non-periodic* signals $f : \mathbb{R} \longrightarrow \mathbb{C}$?
- If so, how would we represent the Fourier spectrum of f?

In the following analysis on these questions, we consider a non-periodic signal $f : \mathbb{R} \longrightarrow \mathbb{C}$ as a signal with *infinite* period, i.e., we consider the limit

$$f(x) = \lim_{T\to\infty} f_T(x) \qquad \text{for } x \in \mathbb{R}, \tag{7.7}$$

where the T-periodic signal f_T is assumed to coincide on $(-T/2, T/2)$ with f.

Moreover, we regard the function

$$g_T(\omega) := \int_{-T/2}^{T/2} f_T(\xi) \mathrm{e}^{-\mathrm{i}\omega\xi} \,\mathrm{d}\xi$$

of the frequency variable ω, whereby we obtain for f_T the representation

$$f_T(x) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} g_T(\omega_j)\, \mathrm{e}^{\mathrm{i}\omega_j x} \cdot \Delta\omega \tag{7.8}$$

from (7.6). We remark that the infinite series in (7.8) is a Riemannian sum on the knot sequence $\{w_j\}_{j\in\mathbb{Z}}$. Note that the mesh width $\Delta\omega$ of the sequence $\{w_j\}_{j\in\mathbb{Z}}$ is, for large enough $T > 0$, arbitrarily small. This observation leads us, via the above-mentioned limit in (7.7), to the function

$$g(\omega) := \lim_{T\to\infty} g_T(\omega) = \int_{-\infty}^{\infty} f(\xi) \mathrm{e}^{-\mathrm{i}\omega\xi} \,\mathrm{d}\xi \qquad \text{for } \omega \in \mathbb{R}. \tag{7.9}$$

To guarantee the well-definedness of the function g in (7.9), we mainly require the existence of the Fourier integral on the right hand side in (7.9) for all frequencies ω. To this end, we assume $f \in \mathrm{L}^1(\mathbb{R})$, i.e., we assume the function f to be *absolutely integrable*. In this case, the Fourier integral in (7.9) is, due to $|\mathrm{e}^{-\mathrm{i}\omega\cdot}| \equiv 1$, finite, for all frequencies ω. Recall that we work here and throughout this work with Lebesgue integration.

Definition 7.1. *For $f \in \mathrm{L}^1(\mathbb{R})$, the function*

$$(\mathcal{F}f)(\omega) = \hat{f}(\omega) := \int_{\mathbb{R}} f(x) \mathrm{e}^{-\mathrm{i}x\omega} \,\mathrm{d}x \qquad \textit{for } \omega \in \mathbb{R} \tag{7.10}$$

is called the `Fourier transform` *of f. The* `Fourier operator`*, which assigns $f \in \mathrm{L}^1(\mathbb{R})$ to its Fourier transform $\mathcal{F}f = \hat{f}$, is denoted as $\mathcal{F}$.* ◯

Note that the Fourier transform $\mathcal{F}$ is a linear integral transform which maps a function $f \equiv f(x)$ of the *spatial variable* x (or, a signal f of the *time variable*) to a function $\mathcal{F}f = \hat{f} \equiv \hat{f}(\omega)$ of the *frequency variable* ω. The application of the Fourier transform is (especially for signals) referred to as *time-frequency analysis*. Moreover, the function $\mathcal{F}f = \hat{f}$ is called the *continuous Fourier spectrum* of f. If we regard the Fourier integral in (7.10) as parameter integral of the frequency variable ω, then we will see that the Fourier transform $\hat{f} : \mathbb{R} \longrightarrow \mathbb{C}$ of $f \in \mathrm{L}^1(\mathbb{R})$ is a function that is uniformly continuous on $\mathbb{R}$ (see Exercise 7.56). In particular, we have $\hat{f} \in \mathscr{C}(\mathbb{R})$. Moreover, due to the estimate

$$|\hat{f}(\omega)| = \left| \int_{\mathbb{R}} f(x) \mathrm{e}^{-\mathrm{i}x\omega} \, \mathrm{d}x \right| \leq \int_{\mathbb{R}} |f(x)| \, \mathrm{d}x = \|f\|_{\mathrm{L}^1(\mathbb{R})}, \tag{7.11}$$

the function $\hat{f}$ is uniformly bounded on $\mathbb{R}$ by the L^1-norm $\|f\|_{\mathrm{L}^1(\mathbb{R})}$ of f.

We note the following fundamental properties of $\mathcal{F}$ (see Exercise 7.58).

Proposition 7.2. *The Fourier transform $\mathcal{F} : \mathrm{L}^1(\mathbb{R}) \longrightarrow \mathscr{C}(\mathbb{R})$ has the following properties, where we assume $f \in \mathrm{L}^1(\mathbb{R})$ for all statements (a)-(e).*

(a) For $f_{x_0} := f(\cdot - x_0)$, where $x_0 \in \mathbb{R}$, we have

$$(\mathcal{F}f_{x_0})(\omega) = \mathrm{e}^{-\mathrm{i}\omega x_0} (\mathcal{F}f)(\omega) \qquad \text{for all } \omega \in \mathbb{R}.$$

(b) For $f_\alpha := f(\alpha \cdot)$, where $\alpha \in \mathbb{R} \setminus \{0\}$, we have

$$(\mathcal{F}f_\alpha)(\omega) = \frac{1}{|\alpha|} (\mathcal{F}f)(\omega/\alpha) \qquad \text{for all } \omega \in \mathbb{R}. \tag{7.12}$$

(c) For the conjugate complex $\bar{f} \in \mathrm{L}^1(\mathbb{R})$, where $\bar{f}(x) = \overline{f(x)}$, we have

$$(\mathcal{F}\bar{f})(\omega) = \overline{(\mathcal{F}f)}(-\omega) \qquad \text{for all } \omega \in \mathbb{R}. \tag{7.13}$$

(d) For the Fourier transform of the derivative f' of f, we have

$$(\mathcal{F}f')(\omega) = \mathrm{i}\omega (\mathcal{F}f)(\omega) \qquad \text{for all } \omega \in \mathbb{R}$$

under the assumption $f \in \mathscr{C}^1(\mathbb{R}) \cap \mathrm{L}^1(\mathbb{R})$ with $f' \in \mathrm{L}^1(\mathbb{R})$.

(e) For the derivative of the Fourier transform $\mathcal{F}f$ of f, we have

$$\frac{\mathrm{d}}{\mathrm{d}\omega} (\mathcal{F}f)(\omega) = -\mathrm{i}(\mathcal{F}(xf))(\omega) \qquad \text{for all } \omega \in \mathbb{R}$$

under the assumption $xf \in \mathrm{L}^1(\mathbb{R})$. □

All properties in Proposition 7.2 can be shown by elementary calculations.

In the following discussion, we work with functions of compact *support*.

Definition 7.3. *For a continuous function* $f : \mathbb{R} \longrightarrow \mathbb{C}$*, we call the point set*

$$\operatorname{supp}(f) := \overline{\{x \in \mathbb{R} \,|\, f(x) \neq 0\}} \subset \mathbb{R}$$

`support` *of* f*. Therefore,* f *has* `compact support`*, if* $\operatorname{supp}(f)$ *is compact.* ◯

We denote by $\mathscr{C}_c(\mathbb{R})$ the linear space of all continuous functions with compact support. Recall that $\mathscr{C}_c(\mathbb{R})$ is *dense* in $\mathrm{L}^1(\mathbb{R})$, i.e., for any $f \in \mathrm{L}^1(\mathbb{R})$ and $\varepsilon > 0$ there is a function $g \in \mathscr{C}_c(\mathbb{R})$ satisfying $\|f - g\|_{\mathrm{L}^1(\mathbb{R})} < \varepsilon$.

According to the *Riemann*[1]*-Lebesgue lemma*, the Fourier transform $\hat{f}$ of $f \in \mathrm{L}^1(\mathbb{R})$ vanishes at infinity.

Lemma 7.4. (Riemann-Lebesgue).
The Fourier transform $\hat{f}$ *of* $f \in \mathrm{L}^1(\mathbb{R})$ *vanishes at infinity, i.e.,*

$$\hat{f}(\omega) \longrightarrow 0 \quad \text{for } |\omega| \to \infty.$$

Proof. Let g be a continuous function with compact support, i.e., $g \in \mathscr{C}_c(\mathbb{R})$.

Due to statement (a) in Proposition 7.2, the function

$$g_{-\pi/\omega} = g(\cdot + \pi/\omega) \in \mathscr{C}_c(\mathbb{R}) \subset \mathrm{L}^1(\mathbb{R}) \qquad \text{for } \omega \neq 0$$

has the Fourier transform

$$(\mathcal{F} g_{-\pi/\omega})(\omega) = \mathrm{e}^{\mathrm{i}\pi}(\mathcal{F}g)(\omega) = -(\mathcal{F}g)(\omega) \qquad \text{for } \omega \neq 0.$$

This implies the representation

$$2(\mathcal{F}g)(\omega) = (\mathcal{F}g)(\omega) - (\mathcal{F}g_{-\pi/\omega})(\omega) = \int_{\mathbb{R}} (g(x) - g(x + \pi/\omega))\mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}x,$$

whereby, in combination with the dominated convergence theorem, we get

$$|\hat{g}(\omega)| = |(\mathcal{F}g)(\omega)| \leq \frac{1}{2}\int_{\mathbb{R}} |g(x) - g(x + \pi/\omega)|\,\mathrm{d}x \longrightarrow 0 \quad \text{for } |\omega| \to \infty. \tag{7.14}$$

Now $\mathscr{C}_c(\mathbb{R})$ is dense in $\mathrm{L}^1(\mathbb{R})$, so that for any $f \in \mathrm{L}^1(\mathbb{R})$ and $\varepsilon > 0$ there is one $g \in \mathscr{C}_c(\mathbb{R})$ satisfying $\|f - g\|_{\mathrm{L}^1(\mathbb{R})} < \varepsilon$. From this, the statement follows from the estimate (7.11), whereby

$$|\hat{f}(\omega) - \hat{g}(\omega)| \leq \|f - g\|_{\mathrm{L}^1(\mathbb{R})} < \varepsilon \qquad \text{for all } \omega \in \mathbb{R},$$

in combination with the property (7.14). ■

[1] Bernhard Riemann (1826-1866), German mathematician

Remark 7.5. By the Riemann-Lebesgue lemma, the Fourier transform $\mathcal{F}$ is a linear mapping between the Banach space $(\mathrm{L}^1(\mathbb{R}), \|\cdot\|_{\mathrm{L}^1(\mathbb{R})})$ of all absolutely integrable functions and the Banach space $(\mathscr{C}_0(\mathbb{R}), \|\cdot\|_\infty)$ of all continuous functions that are vanishing at infinity, i.e.,

$$\mathcal{F} : \mathrm{L}^1(\mathbb{R}) \longrightarrow \mathscr{C}_0(\mathbb{R}).$$

□

In our following discussion, two questions are of fundamental importance:

- Is the Fourier transform $\mathcal{F}$ invertible?
- Can the Fourier transform $\mathcal{F}$ be transferred to the Hilbert space $\mathrm{L}^2(\mathbb{R})$?

To give an answer to these questions, we require only a few preparations. First, we prove the following result.

Proposition 7.6. *For $f, g \in \mathrm{L}^1(\mathbb{R})$ both functions $\hat{f}g$ and $f\hat{g}$ are integrable. Moreover, we have*

$$\int_{\mathbb{R}} \hat{f}(x)g(x)\,\mathrm{d}x = \int_{\mathbb{R}} f(\omega)\hat{g}(\omega)\,\mathrm{d}\omega. \tag{7.15}$$

Proof. Since the functions $\hat{f}$ and $\hat{g}$ are continuous and bounded, respectively, both functions $\hat{f}g$ and $f\hat{g}$ are integrable. By using the *Fubini*[2] *theorem*, we can conclude

$$\begin{aligned}\int_{\mathbb{R}} f(\omega)\hat{g}(\omega)\,\mathrm{d}\omega &= \int_{\mathbb{R}} f(\omega)\left(\int_{\mathbb{R}} g(x)\mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}x\right)\mathrm{d}\omega \\ &= \int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(\omega)\mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}\omega\right) g(x)\,\mathrm{d}x = \int_{\mathbb{R}} \hat{f}(x)g(x)\,\mathrm{d}x.\end{aligned}$$

■

Now let us discuss two important examples for Fourier transforms.

Example 7.7. For $\alpha > 0$, let $\sqcap_\alpha = \chi_{[-\alpha,\alpha]}$ be the indicator function of the compact interval $[-\alpha, \alpha] \subset \mathbb{R}$. Then,

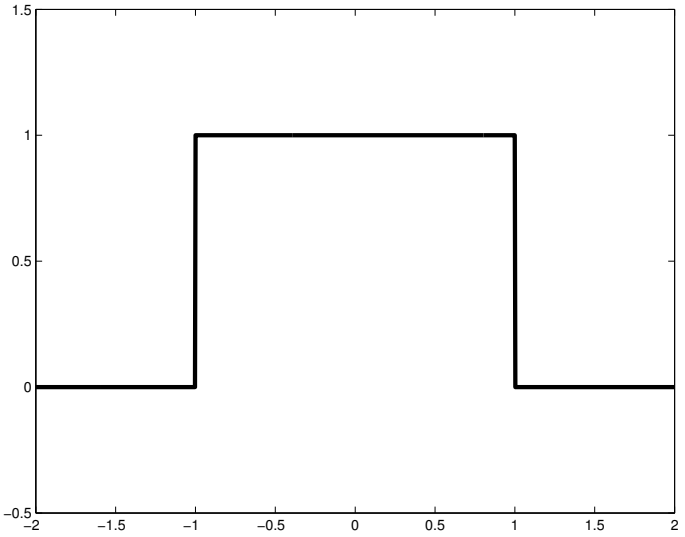
$$(\mathcal{F}\sqcap_1)(\omega) = \int_{-1}^{1} \mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}x = 2 \cdot \mathrm{sinc}(\omega) \qquad \text{for } \omega \in \mathbb{R}$$

is the Fourier transform of $\sqcap_1$, where the (continuous) function

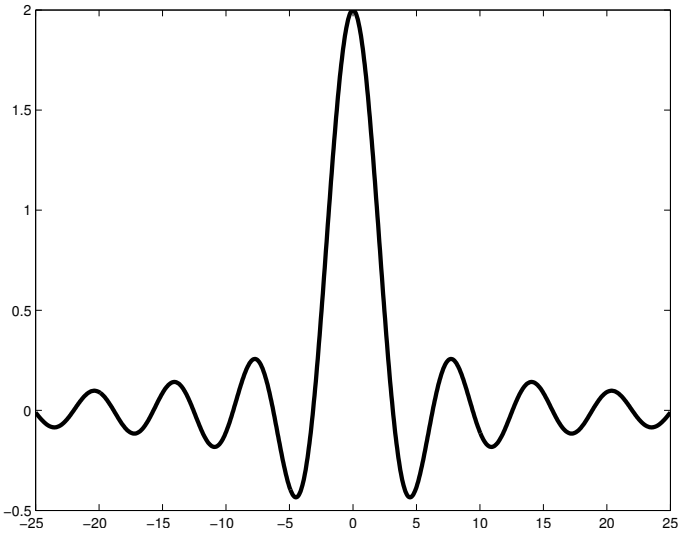
$$\mathrm{sinc}(\omega) := \begin{cases} \sin(\omega)/\omega & \text{for } \omega \neq 0 \\ 1 & \text{for } \omega = 0 \end{cases}$$

is called *sinus cardinalis* (or, *sinc function*) (see Figure 7.1).

[2] Guido Fubini (1879-1943), Italian mathematician



(a) The indicator function $\sqcap_1$



(b) The Fourier transform $\mathcal{F}\sqcap_1 = 2\text{sinc}$

Fig. 7.1. The sinc function yields the Fourier transform of the function $\sqcap_1$.

By the scaling property in (7.12), we find that $(\mathcal{F}\sqcap_\alpha)(\omega) = 2\alpha \cdot \operatorname{sinc}(\alpha\omega)$, for $\omega \in \mathbb{R}$, is the Fourier transform of $\sqcap_\alpha$. Note that the Fourier transform $\mathcal{F}\sqcap_\alpha$ of $\sqcap_\alpha \in \mathrm{L}^1(\mathbb{R})$ is *not* contained in $\mathrm{L}^1(\mathbb{R})$, since the sinc function is *not* absolutely integrable. ♢

Example 7.8. We compute the Fourier transform of the *Gauss function*

$$g_\alpha(x) = \mathrm{e}^{-\alpha x^2} \qquad \text{for } x \in \mathbb{R}$$

for $\alpha > 0$ by

$$\begin{aligned}
\widehat{g_\alpha}(\omega) &= \int_{\mathbb{R}} e^{-\alpha x^2} \mathrm{e}^{-\mathrm{i}x\omega}\, \mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{-\alpha(x^2 + \mathrm{i}x\omega/\alpha)}\, \mathrm{d}x \\
&= \int_{\mathbb{R}} \mathrm{e}^{-\alpha\left(x^2 + \frac{\mathrm{i}x\omega}{\alpha} + \left(\frac{\mathrm{i}\omega}{2\alpha}\right)^2\right)} \mathrm{e}^{\alpha\left(\frac{\mathrm{i}\omega}{2\alpha}\right)^2}\, \mathrm{d}x \\
&= \mathrm{e}^{-\omega^2/(4\alpha)} \int_{\mathbb{R}} \mathrm{e}^{-\alpha\left(x + \frac{\mathrm{i}\omega}{2\alpha}\right)^2}\, \mathrm{d}x \\
&= \sqrt{\frac{\pi}{\alpha}} \cdot \mathrm{e}^{-\omega^2/(4\alpha)},
\end{aligned}$$

where in the last line we used the well-known identity

$$\int_{\mathbb{R}} \mathrm{e}^{-\alpha(x+\mathrm{i}y)^2}\, \mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{-\alpha x^2}\, \mathrm{d}x = \sqrt{\frac{\pi}{\alpha}} \qquad \text{for } \alpha > 0.$$

In conclusion, we note the following observation:

The Fourier transform of a Gauss function is a Gauss function.

In particular, for $\alpha = 1/2$, we have

$$\widehat{g_{1/2}} = \sqrt{2\pi} \cdot g_{1/2}, \tag{7.16}$$

i.e., the Gauss function $g_{1/2}$ is an *eigenfunction* of $\mathcal{F}$ for the eigenvalue $\sqrt{2\pi}$. The identity in (7.16) immediately implies the representation

$$\mathrm{e}^{-x^2/2} = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \mathrm{e}^{-y^2/2} \cdot \mathrm{e}^{\mathrm{i}xy}\, \mathrm{d}y \qquad \text{for all } x \in \mathbb{R}, \tag{7.17}$$

where we used the symmetry $g_{1/2}(x) = g_{1/2}(-x)$, for $x \in \mathbb{R}$. ♢

Now we can determine the operator norm of $\mathcal{F}$.

Proposition 7.9. *The Fourier transform* $\mathcal{F} : \mathrm{L}^1(\mathbb{R}) \longrightarrow \mathscr{C}_0(\mathbb{R})$ *has operator norm one, i.e.,*

$$\|\mathcal{F}\|_{\mathrm{L}^1(\mathbb{R}) \to \mathscr{C}_0(\mathbb{R})} = 1.$$

Proof. For $f \in \mathrm{L}^1(\mathbb{R})$, the Fourier transform $\mathcal{F}f = \hat{f}$ is bounded, due to (7.11), where $\|\mathcal{F}f\|_\infty \le \|f\|_{\mathrm{L}^1(\mathbb{R})}$. This implies $\|\mathcal{F}\|_{\mathrm{L}^1(\mathbb{R})\to\mathscr{C}_0(\mathbb{R})} \le 1$. For the Gauss function $g_{1/2}(x) = \exp(-x^2/2)$ from Example 7.8, we obtain $\|g_{1/2}\|_{\mathrm{L}^1(\mathbb{R})} = \sqrt{2\pi}$, on the one hand, whereas, on the other hand, we have $\|\mathcal{F}g_{1/2}\|_\infty = \sqrt{2\pi}$, due to (7.16). This implies

$$\|\mathcal{F}\|_{\mathrm{L}^1(\mathbb{R})\to\mathscr{C}_0(\mathbb{R})} = \sup_{f\in\mathrm{L}^1(\mathbb{R})\setminus\{0\}} \frac{\|\mathcal{F}f\|_\infty}{\|f\|_{\mathrm{L}^1(\mathbb{R})}} = 1.$$

■

From the result of Proposition 7.9, we can draw the following conclusion.

Corollary 7.10. *Let $(f_n)_{n\in\mathbb{N}}$ be a convergent sequence in $\mathrm{L}^1(\mathbb{R})$ with limit $f \in \mathrm{L}^1(\mathbb{R})$. Then, the corresponding sequence $(\hat{f}_n)_{n\in\mathbb{N}}$ of Fourier transforms $\mathcal{F}f_n = \hat{f}_n \in \mathscr{C}_0(\mathbb{R})$ converges uniformly on $\mathbb{R}$ to $\hat{f}$.*

Proof. The statement follows immediately from the estimate

$$\|\hat{f}_n - \hat{f}\|_\infty = \|\mathcal{F}(f_n - f)\|_\infty \le \|\mathcal{F}\| \cdot \|f_n - f\|_{\mathrm{L}^1(\mathbb{R})} = \|f_n - f\|_{\mathrm{L}^1(\mathbb{R})},$$

where $\|\mathcal{F}\| = \|\mathcal{F}\|_{\mathrm{L}^1(\mathbb{R})\to\mathscr{C}_0(\mathbb{R})} = 1$, due to Proposition 7.9. ■

In the following discussion, the *convolution* in L^1 is of primary importance.

Definition 7.11. *For $f, g \in \mathrm{L}^1(\mathbb{R})$ the function*

$$(f * g)(x) := \int_{\mathbb{R}} f(x-y)g(y)\,\mathrm{d}y \qquad \text{for } x \in \mathbb{R} \tag{7.18}$$

is called `convolution product`*, in short,* `convolution`*, between f and g.* ◯

We note that the convolution between L^1-functions is well-defined, i.e., for $f, g \in \mathrm{L}^1(\mathbb{R})$, the integral in (7.18) is finite, for all $x \in \mathbb{R}$. Moreover, the convolution $f * g$ is in $\mathrm{L}^1(\mathbb{R})$. We take note of this important result as follows.

Proposition 7.12. *For $f, g \in \mathrm{L}^1(\mathbb{R})$, the convolution $f * g$ is absolutely integrable, and we have the estimate*

$$\|f * g\|_{\mathrm{L}^1(\mathbb{R})} \le \|f\|_{\mathrm{L}^1(\mathbb{R})} \cdot \|g\|_{\mathrm{L}^1(\mathbb{R})}. \tag{7.19}$$

Proof. For $f, g \in \mathrm{L}^1(\mathbb{R})$, we have the representation

$$\begin{aligned}\int_{\mathbb{R}} (f * g)(x)\,\mathrm{d}x &= \int_{\mathbb{R}}\int_{\mathbb{R}} f(x-y)g(y)\,\mathrm{d}y\,\mathrm{d}x \\ &= \int_{\mathbb{R}} g(y)\left(\int_{\mathbb{R}} f(x-y)\,\mathrm{d}x\right)\mathrm{d}y = \int_{\mathbb{R}} f(x)\,\mathrm{d}x \cdot \int_{\mathbb{R}} g(y)\,\mathrm{d}y,\end{aligned}$$

by using the Fubini theorem. Therefore, $f * g$ is integrable. From a similar representation for $|f * g|$, we get the estimate in (7.19). ■

Remark 7.13. Due to Proposition 7.12, the Banach space $\mathrm{L}^1(\mathbb{R})$ is closed under the convolution product $*$, i.e., we have $f * g \in \mathrm{L}^1(\mathbb{R})$ for $f, g \in \mathrm{L}^1(\mathbb{R})$. Moreover, for $f, g \in \mathrm{L}^1(\mathbb{R})$, we have the identity

$$(f * g)(x) = \int_{\mathbb{R}} f(x-y)g(y)\,\mathrm{d}y = \int_{\mathbb{R}} f(y)g(x-y)\,\mathrm{d}y = (g * f)(x)$$

for all $x \in \mathbb{R}$, i.e., the convolution $*$ is commutative on $\mathrm{L}^1(\mathbb{R})$.

Therefore, $\mathrm{L}^1(\mathbb{R})$ is a commutative Banach algebra. □

Due to Proposition 7.12 and Remark 7.13, we can apply the Fourier transform $\mathcal{F}$ to the convolution of two L^1-functions. As we show now, the Fourier transform $\mathcal{F}(f * g)$ of the convolution $f * g$, for $f, g \in \mathrm{L}^1(\mathbb{R})$, coincides with the algebraic product of their Fourier transforms $\mathcal{F}f$ and $\mathcal{F}g$.

Theorem 7.14. (Fourier convolution theorem).
For $f, g \in \mathrm{L}^1(\mathbb{R})$ we have

$$\mathcal{F}(f * g) = (\mathcal{F}f) \cdot (\mathcal{F}g).$$

Proof. With application of the Fubini theorem we immediately obtain by

$$\begin{aligned}(\mathcal{F}(f * g))(\omega) &= \int_{\mathbb{R}} (f * g)(x)\mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}x = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x-y)g(y)\,\mathrm{d}y \right) \mathrm{e}^{-\mathrm{i}x\omega}\,\mathrm{d}x \\ &= \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x-y)\mathrm{e}^{-\mathrm{i}(x-y)\omega}\,\mathrm{d}x \right) g(y)\mathrm{e}^{-\mathrm{i}y\omega}\,\mathrm{d}y \\ &= \int_{\mathbb{R}} (\mathcal{F}f)(\omega)g(y)\mathrm{e}^{-\mathrm{i}y\omega}\,\mathrm{d}y = (\mathcal{F}f)(\omega) \cdot (\mathcal{F}g)(\omega)\end{aligned}$$

the stated representation for all $\omega \in \mathbb{R}$. ■

Next, we specialize the Fourier convolution theorem to autocorrelations.

Definition 7.15. *For $f \in \mathrm{L}^1(\mathbb{R})$ the convolution product*

$$(f * f^*)(x) = \int_{\mathbb{R}} f(x-y)f^*(y)\,\mathrm{d}y = \int_{\mathbb{R}} f(x+y)f(y)\,\mathrm{d}y \qquad \text{for } x \in \mathbb{R}$$

is called `autocorrelation` *of f, where the function $f^* \in \mathrm{L}^1(\mathbb{R})$ is defined as $f^*(x) := f(-x)$ for all $x \in \mathbb{R}$.* ◯

From the Fourier convolution theorem, Theorem 7.14, in combination with statement (c) in Proposition 7.2, we immediately get the following result.

Corollary 7.16. *For real-valued $f \in \mathrm{L}^1(\mathbb{R})$, we have the representation*

$$\mathcal{F}(f * f^*)(\omega) = |(\mathcal{F}f)(\omega)|^2 \qquad \text{for all } \omega \in \mathbb{R}$$

for the Fourier transform of the autocorrelation of f. □

7.1.1 The Fourier Inversion Theorem

In this section, we prove the Fourier inversion formula on $\mathrm{L}^1(\mathbb{R})$, as we already motivated in the previous section. To this end, we derive a continuous version of the Fourier series representation in (7.5) by using the continuous Fourier spectrum $\hat{f} = \mathcal{F}f$, for $f \in \mathrm{L}^1(\mathbb{R})$.

In order to do so, we need only a few preparations.

Definition 7.17. *A sequence of functions $(\delta_k)_{k\in\mathbb{N}}$ is called a* `Dirac`[3] `sequence` *in $\mathrm{L}^1(\mathbb{R})$, if all of the following conditions are satisfied.*

(a) For all $k \in \mathbb{N}$, we have the positivity

$$\delta_k(x) \geq 0 \qquad \text{for almost every } x \in \mathbb{R}.$$

(b) For all $k \in \mathbb{N}$, we have the normalization

$$\int_{\mathbb{R}} \delta_k(x)\,\mathrm{d}x = 1.$$

(c) For all $r > 0$, we have

$$\lim_{k\to\infty} \int_{\mathbb{R}\setminus[-r,r]} \delta_k(x)\,\mathrm{d}x = 0.$$

○

If we interpret the functions $\delta_k \in \mathrm{L}^1(\mathbb{R})$ of a Dirac sequence as (non-negative) mass densities, then the total mass will be normalized to unity, due to property (b). Moreover, the total mass will at increasing $k \in \mathbb{N}$ be concentrated around zero. This observation motivates the following example.

Example 7.18. For the Gauss function $g_{1/2}(x) = \mathrm{e}^{-x^2/2}$ in Example 7.8, we have

$$\int_{\mathbb{R}} g_{1/2}(x)\,\mathrm{d}x = \int_{\mathbb{R}} \mathrm{e}^{-x^2/2}\,\mathrm{d}x = \sqrt{2\pi}.$$

Now we let $\delta_1(x) := \frac{1}{\sqrt{2\pi}} g_{1/2}(x)$ and, moreover, $\delta_k(x) := k\delta_1(kx)$, for $k > 1$, so that

$$\delta_k(x) = \frac{k}{\sqrt{2\pi}} \cdot \mathrm{e}^{-k^2x^2/2} \qquad \text{for } k \in \mathbb{N}. \tag{7.20}$$

By elementary calculations, we see that the Gauss sequence $(\delta_k)_{k\in\mathbb{N}}$ satisfies all conditions (a)-(c) in Definition 7.17, i.e., $(\delta_k)_{k\in\mathbb{N}}$ is a Dirac sequence. ♢

Next, we prove an important approximation theorem for L^1-functions, according to which any $f \in \mathrm{L}^1(\mathbb{R})$ can be approximated arbitrarily well by its convolutions $f * \delta_k$ with elements of a Dirac sequence $(\delta_k)_{k\in\mathbb{N}}$ in $\mathrm{L}^1(\mathbb{R})$.

[3] Paul Adrien Maurice Dirac (1902-1984), English physicist

Theorem 7.19. (Dirac approximation theorem).
Let $f \in \mathrm{L}^1(\mathbb{R})$ and $(\delta_k)_{k\in\mathbb{N}}$ be a Dirac sequence in $\mathrm{L}^1(\mathbb{R})$. Then, we have

$$\|f - f * \delta_k\|_{\mathrm{L}^1(\mathbb{R})} \longrightarrow 0 \quad \text{for } k \to \infty, \tag{7.21}$$

i.e., the sequence $(f * \delta_k)_{k\in\mathbb{N}}$ converges in $\mathrm{L}^1(\mathbb{R})$ to f.

Proof. Let g be a continuously differentiable function with compact support, i.e., $g \in \mathscr{C}_c^1(\mathbb{R})$. Then, the functions g and g' are bounded on $\mathbb{R}$, i.e., there is a constant $M > 0$ with $\max(\|g\|_\infty, \|g'\|_\infty) \le M$. We let $K := |\mathrm{supp}(g)| < \infty$ for the (finite) length $|\mathrm{supp}(g)|$ of the support interval $\mathrm{supp}(g) \subset \mathbb{R}$ of g.

We estimate the L^1-error $\|g - g * \delta_k\|_{\mathrm{L}^1(\mathbb{R})}$ from above by

$$\begin{aligned}
\|g - g * \delta_k\|_{\mathrm{L}^1(\mathbb{R})} &= \int_{\mathbb{R}} \left| \int_{\mathbb{R}} \delta_k(y)\,(g(x) - g(x-y))\,\mathrm{d}y \right| \mathrm{d}x \\
&\le \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \delta_k(y)\,|g(x) - g(x-y)|\,\mathrm{d}y \right) \mathrm{d}x, \\
&= \int_{\mathbb{R}} \delta_k(y) \left(\int_{\mathbb{R}} |g(x) - g(x-y)|\,\mathrm{d}x \right) \mathrm{d}y,
\end{aligned} \tag{7.22}$$

where we used the properties (a) and (b) in Definition 7.17. Note that the function $h_y := g - g(\cdot - y)$ satisfies, for any $y \in \mathbb{R}$, the estimate

$$|h_y(x)| \le \|g'\|_\infty \cdot |y| \le M \cdot |y| \qquad \text{for all } x \in \mathbb{R}. \tag{7.23}$$

Now we split the outer integral in (7.22) into a sum of two terms, which we estimate uniformly from above by (7.23), so that we have, for any $\rho > 0$,

$$\begin{aligned}
&\|g - g * \delta_k\|_{\mathrm{L}^1(\mathbb{R})} \\
&\le \int_{-\rho}^{\rho} \delta_k(y) \left(\int_{\mathbb{R}} |h_y(x)|\,\mathrm{d}x \right) \mathrm{d}y + \int_{\mathbb{R}\setminus(-\rho,\rho)} \delta_k(y) \left(\int_{\mathbb{R}} |h_y(x)|\,\mathrm{d}x \right) \mathrm{d}y \\
&\le 2 \cdot |\mathrm{supp}(g)| \cdot \|g'\|_\infty \cdot \rho + 2 \cdot |\mathrm{supp}(g)| \cdot \|g\|_\infty \int_{\mathbb{R}\setminus(-\rho,\rho)} \delta_k(y)\,\mathrm{d}y \\
&\le 4 \cdot K \cdot M \cdot \rho
\end{aligned}$$

for all $k \ge N \equiv N(\rho) \in \mathbb{N}$ satisfying

$$\int_{\mathbb{R}\setminus(-\rho,\rho)} \delta_k(y)\,\mathrm{d}y \le \rho,$$

by using property (c) in Definition 7.17. For $\varepsilon > 0$ we have $\|g - g*\delta_k\|_{\mathrm{L}^1(\mathbb{R})} < \varepsilon$, for all $k \ge N$, with assuming $\rho < \varepsilon/(4KM)$, Therefore, $g \in \mathscr{C}_c^1(\mathbb{R})$ can be approximated arbitrarily well in $\mathrm{L}^1(\mathbb{R})$ by convolutions $g * \delta_k$. Finally, $\mathscr{C}_c^1(\mathbb{R})$ is dense in $\mathrm{L}^1(\mathbb{R})$, which implies the stated L^1-convergence in (7.21) for $f \in \mathrm{L}^1(\mathbb{R})$. ∎

Now we turn to the Fourier inversion formula. At the outset of Section 7.1, we derived the representation (7.8) for periodic functions. We can transfer the inversion formula (7.8) from the discrete case to the continuous case. This motivates the following definition.

Definition 7.20. *For $g \in \mathrm{L}^1(\mathbb{R})$ we call the function*

$$(\mathcal{F}^{-1}g)(x) = \check{g}(x) := \frac{1}{2\pi}\int_{\mathbb{R}} g(\omega)\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega \qquad \text{for } x \in \mathbb{R} \tag{7.24}$$

`inverse Fourier transform` *of g. The inverse* `Fourier operator`*, which assigns $g \in \mathrm{L}^1(\mathbb{R})$ to its inverse Fourier transform $\check{g}$, is denoted as $\mathcal{F}^{-1}$.* ◯

Now we can prove the Fourier inversion formula

$$f = \mathcal{F}^{-1}\mathcal{F}f$$

under suitable assumptions on $f \in \mathrm{L}^1(\mathbb{R})$.

Theorem 7.21. (Fourier inversion formula).
For $f \in \mathrm{L}^1(\mathbb{R})$ satisfying $\hat{f} = \mathcal{F}f \in \mathrm{L}^1(\mathbb{R})$ the `Fourier inversion formula`

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\omega)\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega \qquad \text{for almost every } x \in \mathbb{R} \tag{7.25}$$

holds with equality at every point $x \in \mathbb{R}$, where f is continuous.

Proof. In the following proof, we utilize the Dirac sequence $(\delta_k)_{k\in\mathbb{N}}$ of Gauss functions from Example 7.18. For δ_k in (7.20), the identity (7.17) yields the representation

$$\delta_k(x) = \frac{k}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-y^2/2}\cdot \mathrm{e}^{\mathrm{i}kxy}\,\mathrm{d}y = \frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\omega^2/(2k^2)}\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega \tag{7.26}$$

for all $k \in \mathbb{N}$. This in turn implies

$$\begin{aligned}(f * \delta_k)(x) &= \int_{\mathbb{R}} f(y)\left(\frac{1}{2\pi}\int_{\mathbb{R}} \mathrm{e}^{-\omega^2/(2k^2)}\cdot \mathrm{e}^{\mathrm{i}(x-y)\omega}\,\mathrm{d}\omega\right)\mathrm{d}y \\ &= \frac{1}{2\pi}\int_{\mathbb{R}}\left(\int_{\mathbb{R}} f(y)\cdot \mathrm{e}^{-\mathrm{i}y\omega}\,\mathrm{d}y\right)\mathrm{e}^{-\omega^2/(2k^2)}\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\omega)\cdot \mathrm{e}^{-\omega^2/(2k^2)}\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega, \end{aligned} \tag{7.27}$$

where, for changing order of integration, we applied the dominated convergence theorem with the dominating function $|f(y)|\mathrm{e}^{-\omega^2}$.

For $k \to \infty$, the sequence of integrals in (7.27) converges to

$$\frac{1}{2\pi}\int_{\mathbb{R}} \hat{f}(\omega)\cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega,$$

where we use the assumption $\hat{f} \in L^1(\mathbb{R})$.

According to the Dirac approximation theorem, Theorem 7.19, the sequence of Dirac approximations $f * \delta_k$ converges in $L^1(\mathbb{R})$ to f, for $k \to \infty$. This already proves the stated Fourier inversion formula (7.25) in $L^1(\mathbb{R})$.

Finally, the parameter integral in (7.25) is a continuous function at x. Therefore, we have equality in (7.25), provided that f is continuous at x. ■

Remark 7.22. According to Remark 7.5, the Fourier transform maps any $f \in L^1(\mathbb{R})$ to a *continuous* function $\hat{f} \in \mathscr{C}_0(\mathbb{R})$. Therefore, by the Fourier inversion formula, Theorem 7.21, there exists for any $f \in L^1(\mathbb{R})$ satisfying $\hat{f} \in L^1(\mathbb{R})$ a *continuous* representative $\tilde{f} \in L^1(\mathbb{R})$, which coincides with f almost everywhere on $\mathbb{R}$ (i.e., $f \equiv \tilde{f}$ in the L^1-sense), and for which the Fourier inversion formula holds on $\mathbb{R}$. □

The Fourier inversion formula implies the injectivity of $\mathcal{F}$ on $L^1(\mathbb{R})$.

Corollary 7.23. *Suppose $\mathcal{F}f = 0$ for $f \in L^1(\mathbb{R})$. Then, $f = 0$ almost everywhere, i.e., the Fourier transform $\mathcal{F} : L^1(\mathbb{R}) \longrightarrow \mathscr{C}_0(\mathbb{R})$ is injective.* ■

In the following discussion, we will often apply the Fourier inversion formula to *continuous* functions $f \in L^1(\mathbb{R}) \cap \mathscr{C}(\mathbb{R})$. By the following result, we can in this case drop the assumption $\hat{f} \in L^1(\mathbb{R})$ (see Exercise 7.64).

Corollary 7.24. *For $f \in L^1(\mathbb{R}) \cap \mathscr{C}(\mathbb{R})$ the Fourier inversion formula*

$$f(x) = \lim_{\varepsilon \searrow 0} \left(\frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}x\omega} \mathrm{e}^{-\varepsilon|\omega|^2} \, \mathrm{d}\omega \right) \qquad \text{for all } x \in \mathbb{R} \tag{7.28}$$

holds. ■

7.2 The Fourier Transform on $L^2(\mathbb{R})$

In this section, we transfer the Fourier transform $\mathcal{F} : L^1(\mathbb{R}) \longrightarrow \mathscr{C}_0(\mathbb{R})$ from $L^1(\mathbb{R})$ to $L^2(\mathbb{R})$. We remark, however, that the Banach space $L^1(\mathbb{R})$ is *not* a subspace of the Hilbert space $L^2(\mathbb{R})$ (see Exercise 7.57). For this reason, we first consider the `Schwartz`[4] `space`

$$\mathcal{S}(\mathbb{R}) = \left\{ f \in \mathscr{C}^\infty(\mathbb{R}) \,\middle|\, x^k \cdot \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell} f(x) \text{ is bounded for all } k, \ell \in \mathbb{N}_0 \right\}$$

of all *rapidly decaying* $\mathscr{C}^\infty$ functions.

[4] Laurent Schwartz (1915-2002), French mathematician

Remark 7.25. Every function $f \in \mathcal{S}(\mathbb{R})$ and all of its derivatives $f^{(k)}$, for $k \in \mathbb{N}$, are rapidly decaying to zero around infinity, i.e., for any (complex-valued) polynomial $p \in \mathcal{P}^{\mathbb{C}}$ and for any $k \in \mathbb{N}_0$, we have

$$p(x) f^{(k)}(x) \longrightarrow 0 \quad \text{for } |x| \to \infty.$$

Therefore, all derivatives $f^{(k)}$ of $f \in \mathcal{S}(\mathbb{R})$, for $k \in \mathbb{N}$, are also contained in $\mathcal{S}(\mathbb{R})$. Obviously, we have the inclusion $\mathcal{S}(\mathbb{R}) \subset \mathrm{L}^1(\mathbb{R})$, and so $f \in \mathcal{S}(\mathbb{R})$ and all its derivatives $f^{(k)}$, for $k \in \mathbb{N}$, are absolutely integrable, i.e., we have $f^{(k)} \in \mathrm{L}^1(\mathbb{R})$ for all $k \in \mathbb{N}_0$. □

Typical examples of elements in the Schwartz space $\mathcal{S}(\mathbb{R})$ are $\mathscr{C}^\infty$ functions with compact support. Another example is the Gauss function g_α, for $\alpha > 0$, from Example 7.8. Before we give further examples of functions in the Schwartz space $\mathcal{S}(\mathbb{R})$, we first note a few observations.

According to Remark 7.25, every function $f \in \mathcal{S}(\mathbb{R})$ and all its derivatives $f^{(k)}$, for $k \in \mathbb{N}$, have a Fourier transform. Moreover, for $f \in \mathcal{S}(\mathbb{R})$ and $k, \ell \in \mathbb{N}_0$, we have the representations

$$\begin{aligned} \frac{\mathrm{d}^\ell}{\mathrm{d}\omega^\ell}(\mathcal{F}f)(\omega) &= (-\mathrm{i})^\ell (\mathcal{F}(x^\ell f))(\omega) \qquad \text{for all } \omega \in \mathbb{R} \\ (\mathcal{F}f^{(k)})(\omega) &= (\mathrm{i}\omega)^k (\mathcal{F}f)(\omega) \qquad \text{for all } \omega \in \mathbb{R}, \end{aligned}$$

as they directly follow (by induction) from Proposition 7.2 (d)-(e) (see Exercise 7.59). This yields the uniform estimate

$$\left| \omega^k \frac{\mathrm{d}^\ell}{\mathrm{d}\omega^\ell}(\mathcal{F}f)(\omega) \right| \leq \left\| \frac{\mathrm{d}^k}{\mathrm{d}x^k}(x^\ell f(x)) \right\|_{\mathrm{L}^1(\mathbb{R})} \qquad \text{for all } \omega \in \mathbb{R}. \tag{7.29}$$

i.e., all functions $\omega^k (\mathcal{F}f)^{(\ell)}(\omega)$, for $k, \ell \in \mathbb{N}_0$, are bounded. Therefore, we see that the Fourier transform $\mathcal{F}f$ of any $f \in \mathcal{S}(\mathbb{R})$ is also contained in $\mathcal{S}(\mathbb{R})$. By the Fourier inversion formula, Theorem 7.25, the Fourier transform $\mathcal{F}$ is bijective on $\mathcal{S}(\mathbb{R})$. We reformulate this important result as follows.

Theorem 7.26. *The Fourier transform $\mathcal{F} : \mathcal{S}(\mathbb{R}) \longrightarrow \mathcal{S}(\mathbb{R})$ is an automorphism on the Schwartz space $\mathcal{S}(\mathbb{R})$, i.e., $\mathcal{F}$ is linear and bijective on $\mathcal{S}(\mathbb{R})$.* ■

Now we make an important example for a family of functions that are contained in the Schwartz space $\mathcal{S}(\mathbb{R})$. To this end, we recall the Hermite polynomials H_n from Section 4.4.3 and their associated Hermite functions h_n from Exercise 4.42.

Example 7.27. The *Hermite functions*

$$h_n(x) = H_n(x) \cdot \mathrm{e}^{-x^2/2} \qquad \text{for } n \in \mathbb{N}_0 \tag{7.30}$$

are contained in the Schwartz space $\mathcal{S}(\mathbb{R})$. Indeed, this follows from the rapid decay of the Gauss function $g_{1/2}(x) = \exp(-x^2/2)$, cf. Example 7.8. ♢

The Schwartz space $\mathcal{S}(\mathbb{R})$ is obviously contained in any Banach space $\mathrm{L}^p(\mathbb{R})$, for $1 \leq p \leq \infty$. In particular, $\mathcal{S}(\mathbb{R})$ is a subspace of the Hilbert space $\mathrm{L}^2(\mathbb{R})$, i.e., $\mathcal{S}(\mathbb{R}) \subset \mathrm{L}^2(\mathbb{R})$. In the following discussion, we work with the L^2-inner product

$$(f, g) = \int_{\mathbb{R}} f(x)\overline{g(x)}\,\mathrm{d}x \qquad \text{for } f, g \in \mathrm{L}^2(\mathbb{R}). \tag{7.31}$$

Now we prove the completeness of the Hermite functions in $\mathrm{L}^2(\mathbb{R})$.

Proposition 7.28. *The Hermite functions $(h_n)_{n\in\mathbb{N}_0}$ in (7.30) are a complete orthogonal system in the Hilbert space $\mathrm{L}^2(\mathbb{R})$.*

Proof. The orthogonality of $(h_n)_{n\in\mathbb{N}_0}$ follows from the orthogonality of the Hermite polynomials in Theorem 4.28. According to (4.47), we have

$$(h_m, h_n) = 2^n n!\sqrt{\pi} \cdot \delta_{mn} \qquad \text{for all } m, n \in \mathbb{N}_0. \tag{7.32}$$

Now we show the completeness of the system $(h_n)_{n\in\mathbb{N}_0}$. To this end, we use the completeness criterion in Theorem 6.26, as follows.

Suppose that $f \in \mathrm{L}^2(\mathbb{R})$ satisfies $(f, h_n) = 0$ for all $n \in \mathbb{N}_0$. Then, we consider the function $g : \mathbb{C} \longrightarrow \mathbb{C}$, defined as

$$g(z) = \int_{\mathbb{R}} h_0(x) f(x) \mathrm{e}^{-\mathrm{i}xz}\,\mathrm{d}x \qquad \text{for } z \in \mathbb{C}.$$

Note that g is holomorphic on $\mathbb{C}$, and, moreover, we have

$$g^{(m)}(z) = (-\mathrm{i})^m \int_{\mathbb{R}} x^m h_0(x) f(x) \mathrm{e}^{-\mathrm{i}xz}\,\mathrm{d}x \qquad \text{for } m \in \mathbb{N}_0.$$

Therefore, $g^{(m)}(0)$ can be written as a linear combination of the inner products (f, h_k), for $k = 0, \ldots, m$, so that $g^{(m)}(0) = 0$ for all $m \in \mathbb{N}_0$. From this, we conclude $g \equiv 0$, since g is holomorphic, which in turn implies $\mathcal{F}(h_0 f) = 0$. By Corollary 7.23, we get $h_0 f = 0$ almost everywhere. In particular, we have $f = 0$ almost everywhere. Due to the completeness criterion, Theorem 6.26, the orthogonal system $(h_n)_{n\in\mathbb{N}_0}$ is complete in $\mathrm{L}^2(\mathbb{R})$. ∎

Theorem 7.29. *For any $n \in \mathbb{N}_0$, the Hermite function h_n in (7.30) is an eigenfunction of the Fourier transform for the eigenvalue $\sqrt{2\pi}(-\mathrm{i})^n$, i.e.,*

$$\widehat{h_n} = \sqrt{2\pi}(-\mathrm{i})^n h_n \qquad \textit{for all } n \in \mathbb{N}_0.$$

Proof. We prove the statement by induction on $n \in \mathbb{N}_0$.

Initial step: For $n = 0$ the statement holds for $h_0 = g_{1/2}$ by (7.16).

Induction hypothesis: Assume that the Hermite function h_n is an eigenfunction of the Fourier transform for the eigenvalue $\sqrt{2\pi}(-\mathrm{i})^n$, for $n \in \mathbb{N}_0$.

Induction step ($n \longrightarrow n+1$)*:* By partial integration, we obtain

$$\begin{aligned}
\widehat{h_{n+1}}(\omega) &= \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\omega} h_{n+1}(x)\,\mathrm{d}\omega \\
&= \int_{\mathbb{R}} \mathrm{e}^{-\mathrm{i}x\omega}(-1)^{n+1}\mathrm{e}^{x^2/2}\frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\mathrm{d}^n}{\mathrm{d}x^n}\mathrm{e}^{-x^2}\right)\mathrm{d}x \\
&= \lim_{R\to\infty}\left[\mathrm{e}^{-\mathrm{i}x\omega}(-1)^{n+1}\mathrm{e}^{x^2/2}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\mathrm{e}^{-x^2}\right]_{x=-R}^{x=R} \\
&\quad - \int_{\mathbb{R}}(-\mathrm{i}\omega + x)\mathrm{e}^{-\mathrm{i}x\omega}\mathrm{e}^{x^2/2}(-1)^{n+1}\frac{\mathrm{d}^n}{\mathrm{d}x^n}\mathrm{e}^{-x^2}\,\mathrm{d}x \\
&= \lim_{R\to\infty}\left[-\mathrm{e}^{-\mathrm{i}x\omega}h_n(x)\right]_{x=-R}^{x=R} + \int_{\mathbb{R}}(-\mathrm{i}\omega + x)\mathrm{e}^{-\mathrm{i}x\omega}h_n(x)\,\mathrm{d}x \\
&= -\mathrm{i}\omega\widehat{h_n}(\omega) + \widehat{xh_n}(\omega).
\end{aligned}$$

From the induction hypothesis and Proposition 7.2 (e), we conclude

$$\widehat{h_{n+1}}(\omega) = \sqrt{2\pi}(-\mathrm{i})^{n+1}\left(\omega h_n(\omega) - h_n'(\omega)\right). \tag{7.33}$$

Now the three-term recursion of the Hermite polynomials in (4.48) can be transferred to the Hermite functions, so that

$$h_{n+1}(x) = 2xh_n(x) - 2nh_{n-1}(x) \qquad \text{for } n \geq 0 \tag{7.34}$$

holds with the initial values $h_{-1} \equiv 0$ and $h_0(x) = \exp(-x^2/2)$. By using the recursion $H_n'(x) = 2nH_{n-1}(x)$, for $n \in \mathbb{N}$, from Corollary 4.30, we get

$$\begin{aligned}
h_n'(x) &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\mathrm{e}^{-x^2/2}\cdot H_n(x)\right) = -x\cdot\mathrm{e}^{-x^2/2}\cdot H_n(x) + \mathrm{e}^{-x^2/2}\cdot H_n'(x) \\
&= -xh_n(x) + \mathrm{e}^{-x^2/2}(2nH_{n-1}(x)) \\
&= 2nh_{n-1}(x) - xh_n(x).
\end{aligned} \tag{7.35}$$

Moreover, the representations in (7.34) and (7.35) imply the recursion

$$h_{n+1}(x) = xh_n(x) - h_n'(x) \qquad \text{for } n \geq 0 \tag{7.36}$$

(cf. Exercise 4.42). Therefore, $\widehat{h_{n+1}} = \sqrt{2\pi}(-\mathrm{i})^{n+1}h_{n+1}$ by (7.33) and (7.36). ∎

Given the completeness of the Hermite functions $(h_n)_{n\in\mathbb{N}_0}$ in $\mathrm{L}^2(\mathbb{R})$, according to Theorem 7.28, there is a unique extension of the Fourier transform $\mathcal{F} : \mathcal{S}(\mathbb{R}) \longrightarrow \mathcal{S}(\mathbb{R})$ to the Hilbert space $\mathrm{L}^2(\mathbb{R})$. Moreover, by the spectral property of the orthonormal system $(h_n)_{n\in\mathbb{N}_0}$ in $\mathrm{L}^2(\mathbb{R})$, as shown in Theorem 7.29, the Parseval identity (6.12) can also be extended to $\mathrm{L}^2(\mathbb{R})$. This important result is referred to as the *Plancherel*[5] *theorem.*

[5] Michel Plancherel (1885-1967), Swiss mathematician

Theorem 7.30. (Plancherel theorem).
The Fourier transform $\mathcal{F} : \mathcal{S}(\mathbb{R}) \longrightarrow \mathcal{S}(\mathbb{R})$ *can uniquely be extended to a bounded and bijective linear mapping on the Hilbert space* $\mathrm{L}^2(\mathbb{R})$. *The extended Fourier transform* $\mathcal{F} : \mathrm{L}^2(\mathbb{R}) \longrightarrow \mathrm{L}^2(\mathbb{R})$ *has the following properties.*

(a) The `Parseval identity`

$$(\mathcal{F}f, \mathcal{F}g) = 2\pi(f, g) \qquad \text{for all } f, g \in \mathrm{L}^2(\mathbb{R}),$$

holds, so that in particular

$$\|\mathcal{F}f\|_{\mathrm{L}^2(\mathbb{R})} = \sqrt{2\pi}\|f\|_{\mathrm{L}^2(\mathbb{R})} \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}).$$

(b) The `Fourier inversion formula` *holds on* $\mathrm{L}^2(\mathbb{R})$, *i.e.,*

$$\mathcal{F}^{-1}(\mathcal{F}f) = f \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}).$$

(c) For the operator norms of $\mathcal{F}$ *and* $\mathcal{F}^{-1}$ *on* $\mathrm{L}^2(\mathbb{R})$, *we have*

$$\|\mathcal{F}\|_{\mathrm{L}^2(\mathbb{R})\to\mathrm{L}^2(\mathbb{R})} = (2\pi)^{1/2}$$
$$\|\mathcal{F}^{-1}\|_{\mathrm{L}^2(\mathbb{R})\to\mathrm{L}^2(\mathbb{R})} = (2\pi)^{-1/2}.$$

■

We close this section by the following remark.

Remark 7.31. The Fourier operator $\mathcal{F} : \mathrm{L}^2(\mathbb{R}) \longrightarrow \mathrm{L}^2(\mathbb{R})$ is uniquely determined by the properties in Theorem 7.30. Moreover, we remark that the Fourier transform $\mathcal{F} : \mathrm{L}^1(\mathbb{R}) \longrightarrow \mathscr{C}_0(\mathbb{R})$ maps any $f \in \mathrm{L}^1(\mathbb{R})$ to a unique uniformly continous function $\mathcal{F}f \in \mathscr{C}_0(\mathbb{R})$. In contrast, the Fourier transform $\mathcal{F} : \mathrm{L}^2(\mathbb{R}) \longrightarrow \mathrm{L}^2(\mathbb{R})$ maps any $f \in \mathrm{L}^2(\mathbb{R})$ to a function $\mathcal{F}f \in \mathrm{L}^2(\mathbb{R})$ that is merely *almost everywhere* unique. □

7.3 The Shannon Sampling Theorem

This section is devoted to the *Shannon*[6] *sampling theorem*, which is a fundamental result in mathematical signal processing. According to the Shannon sampling theorem, any *signal* $f \in \mathrm{L}^2(\mathbb{R})$ with bounded frequency density can be reconstructed *exactly* from its samples (i.e., function values) on an infinite uniform grid $\{jd \,|\, j \in \mathbb{Z}\} \subset \mathbb{R}$ at a sufficiently small sampling rate $d > 0$.

We formulate the mathematical assumptions on f as follows.

Definition 7.32. *A function* $f \in \mathrm{L}^2(\mathbb{R})$ *is said to be* `band-limited`, *if its Fourier transform* $\mathcal{F}f$ *has compact support* $\mathrm{supp}(\mathcal{F}f)$, *i.e., if there is some constant* $L > 0$ *satisfying* $\mathrm{supp}(\mathcal{F}f) \subset [-L, L]$, *where the smallest constant* L *with this property is called the* `bandwidth` *of* f. ◯

[6] Claude Elwood Shannon (1916-2001), US-American mathematician

Remark 7.33. Every band-limited function f is analytic. This important result is due to the *Paley*[7]*-Wiener*[8] *theorem*. A detailed discussion concerning the analyticity of Fourier transforms can be found in [58, Section IX.3]. □

Theorem 7.34. (Shannon sampling theorem).
Let $f \in \mathrm{L}^2(\mathbb{R})$ be a band-limited function with bandwidth $L > 0$. Then, we have the reconstruction formula

$$f(x) = \sum_{j\in\mathbb{Z}} f(j\pi/L) \cdot \operatorname{sinc}(Lx - j\pi) \qquad \text{for all } x \in \mathbb{R}. \tag{7.37}$$

Proof. Without loss of generality, we assume $L = \pi$ for the bandwidth of f, since otherwise we can resort to the case $\hat{g}(\omega) = \hat{f}(\omega \cdot \pi/L)$.

For a fixed $x \in \mathbb{R}$, we work with the function $e_x(\omega) = \exp(\mathrm{i}x\omega)$. By $e_x \in \mathrm{L}^2[-\pi, \pi]$, the Fourier series representation

$$e_x(\omega) = \sum_{j\in\mathbb{Z}} c_j(e_x) \cdot \mathrm{e}^{\mathrm{i}j\omega}$$

holds in the L^2-sense. The Fourier coefficients $c_j(e_x)$ of e_x can be computed as

$$c_j(e_x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e_x(\omega) \cdot \mathrm{e}^{-\mathrm{i}j\omega}\,\mathrm{d}\omega = \operatorname{sinc}(\pi(x-j)) \qquad \text{for all } j \in \mathbb{Z}.$$

Now f has a continuous representative in L^2 which satisfies the representation

$$f(x) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}x\omega}\,\mathrm{d}\omega \tag{7.38}$$

$$= \sum_{j\in\mathbb{Z}} \operatorname{sinc}(\pi(x-j)) \cdot \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}j\omega}\,\mathrm{d}\omega \tag{7.39}$$

$$= \sum_{j\in\mathbb{Z}} f(j) \cdot \operatorname{sinc}(\pi(x-j)) \tag{7.40}$$

pointwise for all $x \in \mathbb{R}$. Note that we have applied the Fourier inversion formula of the Plancherel theorem, Theorem 7.30, to obtain (7.38) and (7.40).

Finally, we remark that the interchange of integration and summation in (7.39) is valid by the Parseval identity

$$\frac{1}{2\pi} \int_{-\pi}^{\pi} g(\omega)\overline{h(\omega)}\,\mathrm{d}\omega = \sum_{j\in\mathbb{Z}} c_j(g) \cdot \overline{c_j(h)} \qquad \text{for all } g, h \in \mathrm{L}^2[-\pi, \pi],$$

which completes our proof for the stated reconstruction formula in (7.37). ■

[7] Raymond Paley (1907-1933), English mathematician
[8] Norbert Wiener (1894-1964), US-American mathematician

Remark 7.35. By the Shannon sampling theorem, Theorem 7.34, any band-limited function $f \in \mathrm{L}^1(\mathbb{R}) \cap \mathscr{C}(\mathbb{R})$, or, $f \in \mathrm{L}^2(\mathbb{R})$ with bandwidth $L > 0$ can uniquely be reconstructed from its values on the uniform sampling grid $\{jd \,|\, j \in \mathbb{Z}\} \subset \mathbb{R}$ for all sampling rates $d \leq \pi/L$. Therefore, the *optimal* sampling rate is $d^* = \pi/L$, and this rate corresponds to half of the *smallest* wave length $2\pi/L$ that is present in the signal f. The optimal sampling rate $d^* = \pi/L$ is called the *Nyquist rate* (or, *Nyquist distance*). □

Remark 7.36. In the commonly used literature, various formulations of the Shannon sampling theorem are given for band-limited functions $f \in \mathrm{L}^1(\mathbb{R})$, rather than for $f \in \mathrm{L}^2(\mathbb{R})$. We remark that the representation in (7.37) does also hold for band-limited functions $f \in \mathrm{L}^1(\mathbb{R})$, or, to be more precise, the representation in (7.37) holds pointwise for a continuous representative of $f \in \mathrm{L}^1(\mathbb{R})$. In fact, this statement can be shown (for compact $\mathrm{supp}(\hat{f}) \subset \mathbb{R}$) by following along the lines of our proof for Theorem 7.34. □

Remark 7.37. The Shannon sampling theorem is, in its different variants, also connected with the names of *Nyquist*[9], *Whittaker*[10], and *Kotelnikov*[11]. In fact, Kotelnikov had formulated and published the sampling theorem already in 1933, although his work was widely unknown for a long time. Shannon formulated the sampling theorem in 1948, where he used this result as a starting point for his theory on maximal channel capacities. □

7.4 The Multivariate Fourier Transform

In this section, we introduce the Fourier transform for complex-valued functions $f \equiv f(x_1, \ldots, x_d)$ of d real variables. To this end, we can rely on basic concepts for the univariate case, $d = 1$, from the previous sections. Again, we first regard the Fourier transform on the Banach space of all absolutely integrable functions,

$$\mathrm{L}^1(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \longrightarrow \mathbb{C} \;\middle|\; \int_{\mathbb{R}^d} |f(x)| \,\mathrm{d}x < \infty \right\},$$

equipped with the L^1-norm

$$\|f\|_{\mathrm{L}^1(\mathbb{R}^d)} = \int_{\mathbb{R}^d} |f(x)| \,\mathrm{d}x \qquad \text{for } f \in \mathrm{L}^1(\mathbb{R}^d).$$

[9] Harry Nyquist (1889-1976), US-American electrical engineer
[10] Edmund Taylor Whittaker (1873-1956), British astronomer, mathematician
[11] Vladimir Kotelnikov (1908-2005), Russian pioneer of information theory

Definition 7.38. *For $f \in \mathrm{L}^1(\mathbb{R}^d)$, the function*

$$(\mathcal{F}_d f)(\omega) = \hat{f}(\omega) := \int_{\mathbb{R}^d} f(x) \mathrm{e}^{-\mathrm{i}\langle x,\omega\rangle} \,\mathrm{d}x \qquad \text{for } \omega \in \mathbb{R}^d \tag{7.41}$$

is called the `Fourier transform` *of f. The* `Fourier operator`*, which assigns $f \in \mathrm{L}^1(\mathbb{R}^d)$ to its d-variate Fourier transform $\mathcal{F}_d f = \hat{f}$, is denoted as $\mathcal{F}_d$. Likewise, for $g \in \mathrm{L}^1(\mathbb{R}^d)$ the function*

$$(\mathcal{F}_d^{-1} g)(x) = \check{g}(x) := (2\pi)^{-d} \int_{\mathbb{R}} g(\omega) \cdot \mathrm{e}^{\mathrm{i}\langle x,\omega\rangle} \,\mathrm{d}\omega \qquad \text{for } x \in \mathbb{R} \tag{7.42}$$

is called the `inverse Fourier transform` *of g. The operator, which assigns $g \in \mathrm{L}^1(\mathbb{R}^d)$ to its inverse Fourier transform $\check{g}$, is denoted as $\mathcal{F}_d^{-1}$.* ◯

By separation of the variables in the $\mathbb{R}^d$-inner product $\langle \cdot, \cdot \rangle$,

$$\langle x, \omega \rangle = x_1\omega_1 + \ldots + x_d\omega_d \qquad \text{for } x = (x_1, \ldots, x_d)^T, \omega = (\omega_1, \ldots, \omega_d)^T \in \mathbb{R}^d,$$

appearing in the Fourier transform's formulas (7.41) and (7.42) we can, via

$$\mathrm{e}^{\pm\mathrm{i}\langle x,\omega\rangle} = \mathrm{e}^{\pm\mathrm{i}x_1\omega_1} \cdot \ldots \cdot \mathrm{e}^{\pm\mathrm{i}x_d\omega_d},$$

generalize the results for the univariate case, $d = 1$, to the multivariate case, $d \geq 1$. In the following of this section, we merely quote results that are needed in Chapters 8 and 9. Of course, the Fourier inversion formula from Theorem 7.21 is of central importance.

Theorem 7.39. (Fourier inversion formula).
For $f \in \mathrm{L}^1(\mathbb{R}^d)$ with $\hat{f} = \mathcal{F}_d f \in \mathrm{L}^1(\mathbb{R}^d)$, the `Fourier inversion formula`

$$f(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}\langle x,\omega\rangle} \,\mathrm{d}\omega \qquad \text{for almost every } x \in \mathbb{R}^d \tag{7.43}$$

holds with equality at every point $x \in \mathbb{R}^d$, where f is continuous. □

As in Corollary 7.24, formula (7.43) holds also for $f \in \mathrm{L}^1(\mathbb{R}^d) \cap \mathscr{C}(\mathbb{R}^d)$.

Corollary 7.40. *For $f \in \mathrm{L}^1(\mathbb{R}^d) \cap \mathscr{C}(\mathbb{R}^d)$, the Fourier inversion formula*

$$f(x) = \lim_{\varepsilon \searrow 0} \left((2\pi)^{-d} \int_{\mathbb{R}^d} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}\langle x,\omega\rangle} \mathrm{e}^{-\varepsilon\|\omega\|_2^2} \,\mathrm{d}\omega \right) \qquad \text{for all } x \in \mathbb{R}^d \tag{7.44}$$

holds. □

An important example is the Fourier transform of the Gauss function.

Example 7.41. The d-variate Fourier transform of the *Gauss function*

$$g_\alpha(x) = \mathrm{e}^{-\alpha\|x\|_2^2} \qquad \text{for } x \in \mathbb{R}^d \text{ and } \alpha > 0$$

is

$$(\mathcal{F}_d g_\alpha)(\omega) = \left(\frac{\pi}{\alpha}\right)^{d/2} \mathrm{e}^{-\|\omega\|_2^2/(4\alpha)} \qquad \text{for } \omega \in \mathbb{R}^d.$$

♢

Moreover, we apply the Fourier transform $\mathcal{F}_d$ to convolutions.

Definition 7.42. *For $f, g \in \mathrm{L}^1(\mathbb{R}^d)$, the function*

$$(f * g)(x) := \int_{\mathbb{R}^d} f(x-y)g(y)\,\mathrm{d}y \qquad \textit{for } x \in \mathbb{R}^d \tag{7.45}$$

is called the `convolution product`*, in short,* `convolution`*, between f and g. Moreover, for $f \in \mathrm{L}^1(\mathbb{R}^d)$ the convolution product*

$$(f * f^*)(x) = \int_{\mathbb{R}^d} f(x-y)f^*(y)\,\mathrm{d}y = \int_{\mathbb{R}^d} f(x+y)f(y)\,\mathrm{d}y \qquad \textit{for } x \in \mathbb{R}^d$$

is called the `autocorrelation` *of f, where $f^*(x) := f(-x)$ for all $x \in \mathbb{R}^d$.* ◯

As for the univariate case, in Theorem 7.14 and Corollary 7.16, the Fourier convolution theorem holds for the multivariate Fourier transform.

Theorem 7.43. (Fourier convolution theorem).
For $f, g \in \mathrm{L}^1(\mathbb{R}^d)$, the identity

$$\mathcal{F}_d(f * g) = (\mathcal{F}_d f) \cdot (\mathcal{F}_d g).$$

holds. In particular, for real-valued $f \in \mathrm{L}^1(\mathbb{R})$, we have

$$\mathcal{F}_d(f * f^*)(\omega) = |(\mathcal{F}_d f)(\omega)|^2 \qquad \textit{for all } \omega \in \mathbb{R}^d$$

for the Fourier transform of the autocorrelation of f. □

By following along the lines of Section 7.2, we can transfer the multivariate Fourier transform $\mathcal{F}_d : \mathrm{L}^1(\mathbb{R}^d) \longrightarrow \mathscr{C}_0(\mathbb{R}^d)$ to the Hilbert space

$$\mathrm{L}^2(\mathbb{R}^d) = \left\{ f : \mathbb{R}^d \longrightarrow \mathbb{C} \;\middle|\; \int_{\mathbb{R}^d} |f(x)|^2\,\mathrm{d}x < \infty \right\}$$

of all square-integrable function, being equipped with the L^2-inner product

$$(f, g) = \int_{\mathbb{R}^d} f(x)\overline{g(x)}\,\mathrm{d}x \qquad \text{for } f, g \in \mathrm{L}^2(\mathbb{R}^d)$$

and the Euclidean norm $\|\cdot\|_{\mathrm{L}^2(\mathbb{R}^d)} = (\cdot,\cdot)^{1/2}$. To this end, we first introduce the Fourier transform $\mathcal{F}_d$ on the *Schwartz space*

$$\mathcal{S}(\mathbb{R}^d) = \left\{ f \in \mathscr{C}^\infty(\mathbb{R}^d) \,\middle|\, x^k \cdot \frac{\mathrm{d}^\ell}{\mathrm{d}x^\ell} f(x) \text{ is bounded for all } k, \ell \in \mathbb{N}_0^d \right\}$$

of all *rapidly decaying* $\mathscr{C}^\infty$ functions. As for the univariate case, Theorem 7.26, the Fourier transform $\mathcal{F}_d$ is bijective on $\mathcal{S}(\mathbb{R}^d)$.

Theorem 7.44. *The multivariate Fourier transform* $\mathcal{F}_d : \mathcal{S}(\mathbb{R}^d) \longrightarrow \mathcal{S}(\mathbb{R}^d)$ *is an automorphism on the Schwartz space* $\mathcal{S}(\mathbb{R}^d)$. □

This implies the Plancherel theorem, as in Theorem 7.30 for $d = 1$.

Theorem 7.45. (Plancherel theorem).
The Fourier transform $\mathcal{F}_d : \mathcal{S}(\mathbb{R}^d) \longrightarrow \mathcal{S}(\mathbb{R}^d)$ *can uniquely be extended to a bounded and bijective linear mapping on the Hilbert space* $\mathrm{L}^2(\mathbb{R}^d)$. *The extended Fourier transform* $\mathcal{F}_d : \mathrm{L}^2(\mathbb{R}^d) \longrightarrow \mathrm{L}^2(\mathbb{R}^d)$ *has the following properties.*

(a) The `Parseval identity`

$$(\mathcal{F}_d f, \mathcal{F}_d g) = (2\pi)^d (f, g) \qquad \text{for all } f, g \in \mathrm{L}^2(\mathbb{R}^d),$$

holds, so that in particular

$$\|\mathcal{F}_d f\|_{\mathrm{L}^2(\mathbb{R}^d)} = (2\pi)^{d/2} \|f\|_{\mathrm{L}^2(\mathbb{R}^d)} \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}^d).$$

(b) The `Fourier inversion formula`

$$\mathcal{F}_d^{-1}(\mathcal{F}_d f) = f \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}^d)$$

holds on $\mathrm{L}^2(\mathbb{R})$, *i.e.,*

$$f(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \hat{f}(\omega) \mathrm{e}^{\mathrm{i}\langle x, \omega \rangle} \, \mathrm{d}\omega \qquad \text{for almost every } x \in \mathbb{R}^d.$$

(c) For the operator norms of $\mathcal{F}_d$ *and* $\mathcal{F}_d^{-1}$ *on* $\mathrm{L}^2(\mathbb{R}^d)$, *we have*

$$\|\mathcal{F}_d\|_{\mathrm{L}^2(\mathbb{R}^d) \to \mathrm{L}^2(\mathbb{R}^d)} = (2\pi)^{d/2}$$
$$\|\mathcal{F}_d^{-1}\|_{\mathrm{L}^2(\mathbb{R}^d) \to \mathrm{L}^2(\mathbb{R}^d)} = (2\pi)^{-d/2}.$$

□

7.5 The Haar Wavelet

In this section, we turn to the construction and analysis of wavelet methods. Wavelets are important building blocks for multiresolution representations of signals $f \in \mathrm{L}^2(\mathbb{R})$. To this end, suitable *wavelet bases* of $\mathrm{L}^2(\mathbb{R})$ are utilized. A very simple-structured wavelet basis of $\mathrm{L}^2(\mathbb{R})$ is due to the work [32] of Alfréd Haar in 1910. In the following discussion, we explain important principles of wavelet methods by using the *Haar*[12] *wavelet*.

[12] Alfréd Haar (1885-1933), Hungarian mathematician

Let us first introduce a basic ingredient. For an interval $I \subset \mathbb{R}$, we denote by $\chi_I : \mathbb{R} \longrightarrow \mathbb{R}$,

$$\chi_I(x) := \begin{cases} 1 & \text{for } x \in I, \\ 0 & \text{otherwise,} \end{cases}$$

the *indicator function* of I. Now we can give a definition for the Haar wavelet.

Definition 7.46. *The function $\psi : \mathbb{R} \longrightarrow \mathbb{R}$, defined as*

$$\psi(x) = \chi_{[0,1/2)}(x) - \chi_{[1/2,1)}(x) = \begin{cases} 1 & \text{for } x \in [0,1/2), \\ -1 & \text{for } x \in [1/2,1), \\ 0 & \text{otherwise,} \end{cases}$$

is called `Haar wavelet`. ◯

In the following, we wish to construct a wavelet basis of $\mathrm{L}^2(\mathbb{R})$ by using the Haar wavelet ψ. To this end, we apply *dilations* (i.e., scalings) and *translations* (i.e., shifts) to the argument of ψ. To be more precise, we consider, for $j, k \in \mathbb{Z}$, the *wavelet functions*

$$\psi_k^j(x) := 2^{j/2}\psi(2^j x - k) \qquad \text{for } x \in \mathbb{R} \tag{7.46}$$

that are generated from the Haar wavelet ψ by multiplication of ψ with factor $2^{j/2}$, along with the application of dilations with 2^j and translations about k on the argument of ψ. In particular, for $j = k = 0$, we get the Haar wavelet $\psi = \psi_0^0$. Figure 7.2 shows the function graphs of ψ_k^j, for $j = -1, 0, 1$.

Let us note only a few elementary properties of the wavelet functions ψ_k^j.

Proposition 7.47. *For ψ_k^j in (7.46), the following statements hold.*

(a) The wavelet functions ψ_k^j have zero mean, i.e.,

$$\int_{-\infty}^{\infty} \psi_k^j(x)\,\mathrm{d}x = 0 \qquad \text{for all } j, k \in \mathbb{Z}.$$

(b) The wavelet functions ψ_k^j have unit L^2-norm, i.e.,

$$\|\psi_k^j\|_{\mathrm{L}^2(\mathbb{R})} = 1 \qquad \text{for all } j, k \in \mathbb{Z}.$$

(c) For any $j, k \in \mathbb{Z}$, the wavelet function ψ_k^j has compact support, where

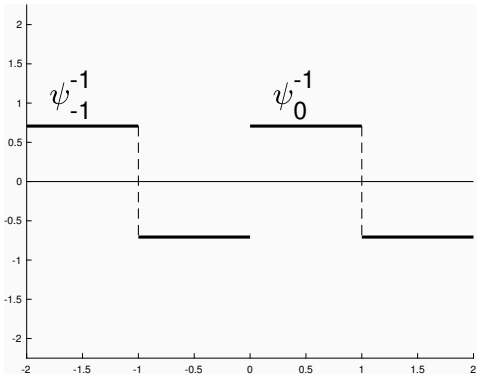
$$\operatorname{supp}(\psi_k^j) = [2^{-j}k, 2^{-j}(k+1)].$$

□

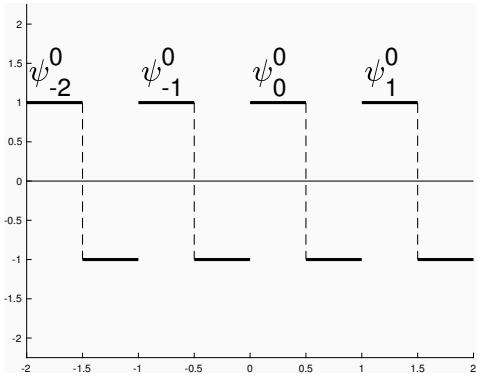
Proposition 7.47 (a)-(c) can be proven by elementary calculations.

Another important property is the orthonormality of the function system $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ with respect to the L^2-inner product $(\cdot,\cdot)$, defined as

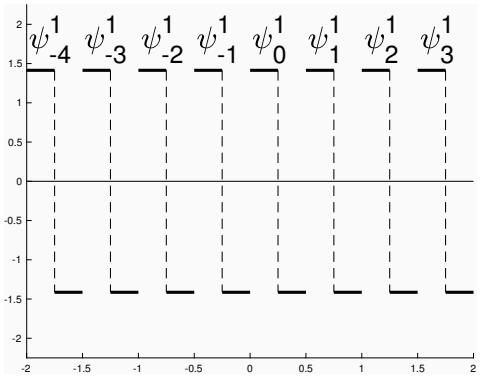
$$(f, g) := \int_{\mathbb{R}} f(x)g(x)\,\mathrm{d}x \qquad \text{for } f, g \in \mathrm{L}^2(\mathbb{R}).$$



ψ_k^{-1} for $k = -1, 0$



ψ_k^0, for $k = -2, -1, 0, 1$



ψ_k^1 for $k = -4, \ldots, 3$

Fig. 7.2. The Haar wavelet $\psi = \psi_0^0$ generates the functions $\psi_k^j = 2^{j/2}\psi(2^j \cdot -k)$.

Proposition 7.48. *The function system* $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ *is orthonormal in* $\mathrm{L}^2(\mathbb{R})$*, i.e.,*

$$(\psi_k^j, \psi_\ell^m) = \delta_{jm}\delta_{k\ell} \qquad \text{for all } j,k,\ell,m \in \mathbb{Z}.$$

Proof. According to Proposition 7.47 (b), any ψ_k^j has unit L^2-norm.

Now suppose that ψ_k^j and ψ_ℓ^m, are, for $j,k,\ell,m \in \mathbb{Z}$, distinct.

Case 1: If $j = m$, then $k \neq \ell$. In this case, the intersection of the support intervals of ψ_k^j and ψ_ℓ^m contains at most one point, according to Proposition 7.47 (c), so that $(\psi_k^j, \psi_\ell^m) = 0$.

Case 2: If $j \neq m$, then we assume $m > j$ (without loss of generality). In this case, we either have, for $\ell \neq 2^{m-j}k, \ldots, 2^{m-j}(k+1) - 1$,

$$\operatorname{supp}(\psi_k^j) \cap \operatorname{supp}(\psi_\ell^m) = \emptyset,$$

whereby $(\psi_k^j, \psi_\ell^m) = 0$, or we have, for $\ell = 2^{m-j}k, \ldots, 2^{m-j}(k+1) - 1$,

$$\operatorname{supp}(\psi_\ell^m) = [2^{-m}\ell, 2^{-m}(\ell+1)] \subset [2^{-j}k, 2^{-j}(k+1)] = \operatorname{supp}(\psi_k^j),$$

so that

$$(\psi_k^j, \psi_\ell^m) = \pm 2^{j/2} \int_{\operatorname{supp}(\psi_\ell^m)} \psi_\ell^m(x)\,\mathrm{d}x = 0.$$

This completes our proof. ∎

Next, we wish to construct a sequence of approximations to $f \in \mathrm{L}^2(\mathbb{R})$ on different scales, i.e., at different resolutions. To this end, we work with a decomposition of $\mathrm{L}^2(\mathbb{R})$ into "finer" and "coarser" closed subspaces. In the following construction of such closed subspaces, the relation between the Haar-Wavelet ψ and its `scaling function`

$$\varphi = \chi_{[0,1)}$$

plays an important role. For the functions

$$\varphi_k^j(x) := 2^{j/2}\varphi(2^j x - k) \qquad \text{for } j,k \in \mathbb{Z}, \tag{7.47}$$

generated by φ, we note the following elementary properties.

Proposition 7.49. *For φ_k^j in (7.47) the following statements hold.*

(a) We have the orthonormality relation

$$(\varphi_k^j, \varphi_\ell^j) = \delta_{k\ell} \qquad \text{for all } j,k,\ell \in \mathbb{Z}.$$

(b) For any $j,k \in \mathbb{Z}$, the function φ_k^j has compact support, where

$$\operatorname{supp}(\varphi_k^j) = [2^{-j}k, 2^{-j}(k+1)].$$

□

Proposition 7.49 can be proven by elementary calculations.

Now we turn to a very important property of φ, which in particular explains the naming *scaling function*.

Proposition 7.50. *The* `refinement equations`

$$\varphi_k^{j-1} = 2^{-1/2}(\varphi_{2k}^j + \varphi_{2k+1}^j) \qquad \text{for all } j,k \in \mathbb{Z} \tag{7.48}$$

$$\psi_k^{j-1} = 2^{-1/2}(\varphi_{2k}^j - \varphi_{2k+1}^j) \qquad \text{for all } j,k \in \mathbb{Z} \tag{7.49}$$

hold.

Proof. By the representation

$$\varphi(x) = \varphi(2x) + \varphi(2x-1) \qquad \text{for all } x \in \mathbb{R}$$

the refinement equation in (7.48) holds for $j = k = 0$. By linear transformation of the argument $x \mapsto 2^{j-1}x - k$, this implies the representation in (7.48). Likewise, the representation in (7.49) can be verified, by now from the identity

$$\psi(x) = \varphi(2x) - \varphi(2x-1) \qquad \text{for all } x \in \mathbb{R}.$$

■

By the refinement equations in (7.48) and (7.49), the *coarser* functions φ_k^{j-1} and ψ_k^{j-1} are represented by a unique linear combination of two *finer* functions, φ_{2k}^j and φ_{2k+1}^j, respectively. We collect all functions of *refinement level* $j \in \mathbb{Z}$ in the L^2-closure

$$V_j = \overline{\operatorname{span}\{\varphi_k^j : k \in \mathbb{Z}\}} \subset \mathrm{L}^2(\mathbb{R}) \qquad \text{for } j \in \mathbb{Z} \tag{7.50}$$

of all linear combinations of functions φ_k^j, for $k \in \mathbb{Z}$. For the properties of the *scale spaces* V_j, we note the following observation.

Proposition 7.51. *The spaces V_j in (7.50) have the following properties.*

(a) V_j is $2^{-j}\mathbb{Z}$-translation-invariant, i.e., $f \in V_j$ implies $f(\cdot - 2^{-j}k) \in V_j$.
(b) The inclusion $V_{j-1} \subset V_j$ holds.

Proof. Property (a) follows from the scale-invariance of the wavelet basis,

$$\varphi_\ell^j(x - 2^{-j}k) = 2^{j/2}\varphi(2^j(x - 2^{-j}k) - \ell) = 2^{j/2}\varphi(2^j x - (k + \ell)) = \varphi_{k+\ell}^j.$$

Property (b) follows directly from the refinement equation in (7.48). ■

Remark 7.52. According to property (b) in Proposition 7.51, the *coarser* scale space V_{j-1} (spanned by the *coarser* basis elements φ_k^{j-1}) is contained in the *finer* scale space V_j (spanned by the *finer* basis elements φ_k^j). Therefore, the scale spaces $(V_j)_{j\in\mathbb{Z}}$ in (7.50) are a nested sequence

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots \subset V_{j-1} \subset V_j \subset \cdots \subset \mathrm{L}^2(\mathbb{R}) \tag{7.51}$$

of subspaces in $\mathrm{L}^2(\mathbb{R})$. □

Now we study further properties of the nested sequence $(V_j)_{j\in\mathbb{Z}}$. To this end, we work with the *orthogonal projection operator* $\Pi_j : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_j$, for $j \in \mathbb{Z}$, which assigns every $f \in \mathrm{L}^2(\mathbb{R})$ to its unique best approximation $s_j^* = \Pi_j f$ in $\mathrm{L}^2(\mathbb{R})$. According to our discussion in Section 6.2, we have the series representation

$$\Pi_j f = \sum_{k\in\mathbb{Z}} (f, \varphi_k^j)\varphi_k^j \in V_j \qquad \text{for } f \in \mathrm{L}^2(\mathbb{R}) \tag{7.52}$$

for the *orthogonal projection* of f on V_j, as in (6.9). The following result describes the asymptotic behaviour of the approximations $(\Pi_j f)_{j\in\mathbb{Z}}$ to any function $f \in \mathrm{L}^2(\mathbb{R})$ with respect to $\|\cdot\| = \|\cdot\|_{\mathrm{L}^2(\mathbb{R})}$.

Proposition 7.53. *For the sequence $(\Pi_j f)_{j\in\mathbb{Z}}$ of orthogonal projections $\Pi_j f$ of $f \in \mathrm{L}^2(\mathbb{R})$ in (7.52) the following statements hold.*

(a) The sequence $(\Pi_j f)_{j\in\mathbb{Z}}$ converges for $j \to \infty$ w.r.t. $\|\cdot\|$ to f, i.e.,

$$\|\Pi_j f - f\| \longrightarrow 0 \qquad \text{for } j \to \infty.$$

(b) The sequence $(\Pi_j f)_{j\in\mathbb{Z}}$ converges for $j \to -\infty$ to zero, i.e.,

$$\|\Pi_j f\| \longrightarrow 0 \qquad \text{for } j \to -\infty.$$

Proof. Let $\varepsilon > 0$ and $f \in \mathrm{L}^2(\mathbb{R})$. Then there is, for a (sufficiently fine) dyadic decomposition of $\mathbb{R}$, a step function $T \in \mathrm{L}^2(\mathbb{R})$ with $\|T - f\| < \varepsilon/2$. Moreover, for the indicator functions $\chi_{I_k^j}$ of the dyadic intervals $I_k^j := [2^{-j}k, 2^{-j}(k+1))$, we have the reproduction property $\Pi_j \chi_{I_k^j} = \chi_{I_k^j}$, for all $k \in \mathbb{Z}$. Therefore, there is a level index $j_0 \in \mathbb{Z}$ with $T = \Pi_j T$ for all $j \geq j_0$. From this, we can conclude statement (a) by the estimate

$$\begin{aligned} \|\Pi_j f - f\| &\leq \|\Pi_j(f - T)\| + \|\Pi_j T - T\| + \|T - f\| \\ &\leq \|\Pi_j\| \cdot \|f - T\| + \|T - f\| < \varepsilon \qquad \text{for } j \geq j_0, \end{aligned}$$

where we use $\|\Pi_j\| = 1$ from Proposition 4.7.

To prove statement (b), we take on given $\varepsilon > 0$ a continuous function g with compact support $\operatorname{supp}(g) = [-R, R]$, for $R > 0$, such that $\|f - g\| < \varepsilon/2$. Now for $2^j \le R^{-1}$, we have

$$\begin{aligned}\Pi_j g &= 2^j \left[\left(\int_{-R}^{0} g(x)\,\mathrm{d}x \right) \chi_{I_{-1}^j} + \left(\int_{0}^{R} g(x)\,\mathrm{d}x \right) \chi_{I_0^j} \right] \\ &= 2^j (c_{-1} \chi_{I_{-1}^j} + c_0 \chi_{I_0^j}),\end{aligned}$$

where $c_{-1} = (g, \varphi_{-1}^j)$ and $c_0 = (g, \varphi_0^j)$. Then, we have $\|\Pi_j g\|^2 = 2^j (c_{-1}^2 + c_0^2)$ and, moreover, $\|\Pi_j g\| < \varepsilon/2$ for $j \equiv j(\varepsilon) \in \mathbb{Z}$ small enough. For this j, we finally get

$$\|\Pi_j f\| \le \|\Pi_j (f - g)\| + \|\Pi_j g\| \le \|f - g\| + \|\Pi_j g\| < \varepsilon$$

by the triangle inequality, and so (b) is also proven. ∎

Proposition 7.53 implies a fundamental property of the scale spaces V_j.

Theorem 7.54. *The system $(V_j)_{j \in Z}$ of scale spaces V_j in (7.50) forms a* `multiresolution analysis` *of $\mathrm{L}^2(\mathbb{R})$ by satisfying the following conditions.*

(a) The scale spaces in $(V_j)_{j \in \mathbb{Z}}$ are nested, so that the inclusions (7.51) hold.
(b) The system $(V_j)_{j \in Z}$ is complete in $\mathrm{L}^2(\mathbb{R})$, i.e., $\mathrm{L}^2(\mathbb{R}) = \overline{\bigcup_{j \in \mathbb{Z}} V_j}$.
(c) The system $(V_j)_{j \in Z}$ satisfies the separation $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$.

Proof. Property (a) holds according to Remark 7.52.

Property (b) follows from Proposition 7.53 (a) and Theorem 6.21.

To prove (c), let $f \in \mathrm{L}^2(\mathbb{R})$ be an element in $\bigcap_{j \in \mathbb{Z}} V_j$. Then, f must have the form

$$f(x) = \begin{cases} c_\ell & \text{for } x \in (-\infty, 0), \\ c_r & \text{for } x \in [0, \infty), \end{cases}$$

for some constants $c_\ell, c_r \in \mathbb{R}$. Since $f \in \mathrm{L}^2(\mathbb{R})$, we have $c_\ell = c_r = 0$ and so $f \equiv 0$. Hence, statement (c) is proven. ∎

In the following analysis, we consider the *orthogonal complement*

$$W_{j-1} = \{ w \in V_j \,|\, (w, v) = 0 \text{ for all } v \in V_{j-1} \} \subset V_j \qquad \text{for } j \in \mathbb{Z}$$

of V_{j-1} in V_j, where we use the notation

$$V_j = W_{j-1} \oplus V_{j-1} \tag{7.53}$$

for the orthogonality relation between W_{j-1} and V_{j-1}. In this way, the linear scale space V_j is by (7.53) decomposed into a *smooth* scale space V_{j-1}

containing the *low frequency* functions of V_j and a *rough* orthogonal complement space W_{j-1} containing the *high frequency* functions from V_j. A recursive decomposition of the scale spaces V_ℓ yields the representation

$$V_j = W_{j-1} \oplus W_{j-2} \oplus \cdots \oplus W_{j-\ell} \oplus V_{j-\ell} \qquad \text{for } \ell \in \mathbb{N}, \tag{7.54}$$

whereby the scale space V_j is being decomposed in a finite sequence of subspace with increasing smoothness. By Theorem 7.54, we get the decomposition

$$\mathrm{L}^2(\mathbb{R}) = \bigoplus_{j\in\mathbb{Z}} W_j, \tag{7.55}$$

i.e., $\mathrm{L}^2(\mathbb{R})$ is decomposed into the orthogonal subspaces W_j. The linear function spaces W_j are called `wavelet spaces`. The following result establishes a fundamental relation between the wavelet functions $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ of the Haar wavelets ψ and the wavelet spaces W_j.

Theorem 7.55. *The functions* $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ *form an* `orthonormal basis` *of* $\mathrm{L}^2(\mathbb{R})$*, i.e.,* $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ *is a complete orthonormal system in* $\mathrm{L}^2(\mathbb{R})$*.*

Proof. The orthonormality of the functions $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ is covered by Proposition 7.48. Therefore, it remains to prove the completeness of the orthonormal system $\{\psi_k^j\}_{j,k\in\mathbb{Z}}$ in $\mathrm{L}^2(\mathbb{R})$. Due to the decomposition in (7.55), it is sufficient to show that the wavelet space W_j is, for any refinement level $j \in \mathbb{Z}$, generated by the functions ψ_k^j, for $k \in \mathbb{Z}$, i.e.,

$$W_j = \overline{\operatorname{span}\{\psi_k^j \,|\, k \in \mathbb{Z}\}} \qquad \text{for } j \in \mathbb{Z}.$$

To this end, we first verify the orthogonality relation

$$(\psi_k^{j-1}, \varphi_\ell^{j-1}) = 0 \qquad \text{for all } k, \ell \in \mathbb{Z}. \tag{7.56}$$

We get (7.56) as follows. For $k \neq \ell$, we have $\operatorname{supp}(\psi_k^{j-1}) \cap \operatorname{supp}(\varphi_\ell^{j-1}) = \emptyset$, whereby $(\psi_k^{j-1}, \varphi_\ell^{j-1}) = 0$. For $k = \ell$, the orthogonality in (7.56) follows from Proposition 7.47 (a). Now by the orthogonality relation in (7.56), we have

$$\psi_k^{j-1} \in W_{j-1} \qquad \text{for all } k \in \mathbb{Z}.$$

The refinement equations (7.48) and (7.49) in Proposition 7.50 imply

$$\begin{aligned} \varphi_{2k}^j &= 2^{-1/2}\left(\varphi_k^{j-1} + \psi_k^{j-1}\right) \\ \varphi_{2k+1}^j &= 2^{-1/2}\left(\varphi_k^{j-1} - \psi_k^{j-1}\right). \end{aligned}$$

Therefore, any basis element $\{\varphi_k^j\}_{k\in\mathbb{Z}}$ of V_j can be represented as a unique linear combination of basis elements in $\{\varphi_k^{j-1}\}_{k\in\mathbb{Z}} \subset V_{j-1}$ and elements in $\{\psi_k^{j-1}\}_{k\in\mathbb{Z}}$, and so the statement follows from the decomposition (7.53). ∎

According to our more general discussion concerning complete orthogonal systems in Section 6.2, we obtain for all elements of the Hilbert space $\mathrm{L}^2(\mathbb{R})$ the representation

$$f = \sum_{j,k\in\mathbb{Z}} (f, \psi_k^j)\psi_k^j \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}). \tag{7.57}$$

This representation follows directly from Theorem 6.21 (b) and Theorem 7.55.

Now we organize the representation (7.57) for $f \in \mathrm{L}^2(\mathbb{R})$ on multiple wavelet scales. Our starting point for doing so is the multiresolution analysis of $\mathrm{L}^2(\mathbb{R})$ in Theorem 7.54. For simplification we suppose $\mathrm{supp}(f) \subset [0,1]$. We approximate f on the scale space V_j, for $j \in \mathbb{N}$, by the orthogonal projectors $\mathit{\Pi}_j : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_j$, given as

$$\mathit{\Pi}_j f = \sum_{k=0}^{N-1} c_k^j \varphi_k^j \in V_j \qquad \text{for } f \in \mathrm{L}^2(\mathbb{R}), \tag{7.58}$$

where $c_k^j := (f, \varphi_k^j)$, for $k = 0, \ldots, N-1$, and where we assume $N = 2^j$. The representation in (7.58) follows directly from (7.52), where the range of the summation index $k \in \{0, \ldots, N-1\}$ in (7.58) is due to

$$\mathrm{supp}(f) \subset [0,1] \qquad \text{and} \qquad \mathrm{supp}(\varphi_k^j) = [2^{-j}k, 2^{-j}(k+1)].$$

By (7.53), $\mathit{\Pi}_{j-1}^{\perp} = \mathit{\Pi}_j - \mathit{\Pi}_{j-1}$ is the *orthogonal projector* of $\mathrm{L}^2(\mathbb{R})$ onto W_{j-1}, so that the decomposition

$$\mathit{\Pi}_j f = \mathit{\Pi}_{j-1}^{\perp} f + \mathit{\Pi}_{j-1} f \qquad \text{for all } f \in \mathrm{L}^2(\mathbb{R}) \tag{7.59}$$

holds. The orthogonal projector $\mathit{\Pi}_{j-1}^{\perp} : \mathrm{L}^2(\mathbb{R}) \longrightarrow W_{j-1}$ is described by

$$\mathit{\Pi}_{j-1}^{\perp} f = \sum_{k=0}^{N/2-1} d_k^{j-1} \psi_k^{j-1} \qquad \text{for } f \in \mathrm{L}^2(\mathbb{R}), \tag{7.60}$$

where $d_k^{j-1} := (f, \psi_k^{j-1})$, for $k = 0, \ldots, N/2-1$.

By (7.58) and (7.60), the identity (7.59) can be written in the basis form

$$\sum_{k=0}^{N-1} c_k^j \varphi_k^j = \sum_{k=0}^{N/2-1} d_k^{j-1} \psi_k^{j-1} + \sum_{k=0}^{N/2-1} c_k^{j-1} \varphi_k^{j-1}. \tag{7.61}$$

With the recursive decomposition of the scale spaces in (7.54), for $\ell = j$,

$$V_j = W_{j-1} \oplus W_{j-2} \oplus \cdots \oplus W_0 \oplus V_0 \qquad \text{for } j \in \mathbb{N},$$

we can write the orthogonal projector $\mathit{\Pi}_j : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_j$ as a *telescoping sum*

$$\Pi_j f = \sum_{r=0}^{j-1} \Pi_r^{\perp} f + \Pi_0 f \qquad \text{for } f \in \mathrm{L}^2(\mathbb{R}), \tag{7.62}$$

whereby $\Pi_j f \in V_j$ is decomposed into a sum of functions $\Pi_r^{\perp} f \in W_r$, for $r = j-1, \ldots, 0$, and $\Pi_0 f \in V_0$ with increasing smoothness, i.e., from high frequency to low frequency terms. By (7.58) and (7.60), we can rewrite (7.62) in basis form as

$$\sum_{k=0}^{N-1} c_k^j \varphi_k^j = \sum_{r=0}^{j-1} \sum_{k=0}^{2^r-1} d_k^r \psi_k^r + c_0^0 \varphi_0^0. \tag{7.63}$$

In practice, however, we have only *discrete* samples of $f \in \mathrm{L}^2(\mathbb{R})$. Suppose the function values $f(2^{-j}k)$ are known for all $k = 0, \ldots, N-1$, where $N = 2^j$. Then, f is interpolated by the function

$$s = \sum_{k=0}^{N-1} f(2^{-j}k)\varphi(2^j \cdot -k) \in V_j$$

at the sample points. Indeed, by $\varphi(k) = \delta_{0k}$ we get

$$s(2^{-j}\ell) = f(2^{-j}\ell) \qquad \text{for } \ell = 0, \ldots, N-1.$$

For the approximation of f, we use the function values of the *finest* level,

$$c_k^j \approx 2^{-j/2} f(2^{-j}k) \qquad \text{for } k = 0, \ldots, N-1,$$

for the coefficients $c^j = (c_k^j)_{k=0}^{N-1} \in \mathbb{R}^N$ in (7.58).

Now we consider the representation of $\Pi_j f$ in (7.63). Our aim is to compute, from the input coefficients $c^j = (c_k^j)_{k=0}^{N-1} \in \mathbb{R}^N$, all *wavelet coefficients*

$$d = (c^0, d^0, (d^1)^T, \ldots, (d^{j-1})^T)^T \in \mathbb{R}^N \tag{7.64}$$

of the representation in (7.63), where

$$c^0 = (c_0^0) \in \mathbb{R}^1 \quad \text{and} \quad d^r = (d_k^r)_{k=0}^{2^r-1} \in \mathbb{R}^{2^r} \text{ for } r = 0, \ldots, j-1.$$

The linear mapping $T : \mathbb{R}^N \longrightarrow \mathbb{R}^N$, which maps any data vector $c^j \in \mathbb{R}^N$ to its corresponding wavelet coefficients $d \in \mathbb{R}^N$ in (7.64) is bijective, and referred to as `discrete wavelet analysis`. In the following discussion, we describe the discrete wavelet analysis in detail.

The computation of the wavelet coefficients d in (7.64) can be performed by recursive decompositions: At the first decomposition level, we compute $c^{j-1} = (c_k^{j-1})_{k=0}^{N/2-1}$ and $d^{j-1} = (d_k^{j-1})_{k=0}^{N/2-1}$ in (7.61). To this end, we apply the refinement equation in (7.48) to the representation in (7.61), whereby

$$\sum_{k=0}^{N/2-1} c_{2k}^j \varphi_{2k}^j + \sum_{k=0}^{N/2-1} c_{2k+1}^j \varphi_{2k+1}^j =$$
$$2^{-1/2} \left(\sum_{k=0}^{N/2-1} (c_{2k}^j + c_{2k+1}^j) \varphi_k^{j-1} + \sum_{k=0}^{N/2-1} (c_{2k}^j - c_{2k+1}^j) \psi_k^{j-1} \right).$$

By comparison of coefficients, we obtain the *decomposition equation*

$$\begin{bmatrix} H_j \\ G_j \end{bmatrix} c^j = \begin{bmatrix} c^{j-1} \\ d^{j-1} \end{bmatrix} \qquad \text{or} \qquad T_j \cdot c^j = \begin{bmatrix} c^{j-1} \\ d^{j-1} \end{bmatrix} \tag{7.65}$$

with the orthogonal *decomposition matrix* $T_j \in \mathbb{R}^{N\times N}$ containing the matrix blocks

$$H_j = 2^{-1/2} \begin{bmatrix} 1\ 1 & & \\ & \ddots & \\ & & 1\ 1 \end{bmatrix},\ G_j = 2^{-1/2} \begin{bmatrix} 1\ -1 & & \\ & \ddots & \\ & & 1\ -1 \end{bmatrix} \in \mathbb{R}^{N/2\times N}.$$

In the next level, the vector $c^{j-1} \in \mathbb{R}^{N/2}$ is decomposed into the vectors $c^{j-2} \in \mathbb{R}^{N/4}$ and $d^{j-2} \in \mathbb{R}^{N/4}$. The resulting recursion is called the `pyramid algorithm`. The decomposition scheme of the pyramid algorithm is represented as follows.

$$\begin{array}{ccccccccc} & & d^{j-1} & & d^{j-2} & & \dots & & d^0 \\ & \nearrow & & \nearrow & & \nearrow & & \nearrow & \\ c^j & \longrightarrow & c^{j-1} & \longrightarrow & c^{j-2} & \longrightarrow & \dots & \longrightarrow & c^0 \end{array}$$

We can describe the decompositions of the pyramid algorithm as linear mappings $T : \mathbb{R}^N \longrightarrow \mathbb{R}^N$, $c^j \longmapsto Tc^j = d$, whose matrix representation

$$T \cdot c^j = T_1 \cdot T_2 \cdot \ldots \cdot T_{j-1} \cdot T_j \cdot c^j = (c^0, d^0, (d^1)^T, \ldots, (d^{j-1})^T)^T = d$$

contains the decomposition matrices T_{j-r}, $r = 0, \ldots, j-1$, of the recursion levels. The orthogonal decomposition matrices are *block diagonal* of the form

$$T_{j-r} = \begin{bmatrix} \boxed{\begin{matrix} H_{j-r} \\ G_{j-r} \end{matrix}} & & \\ & \boxed{I_r} & \end{bmatrix} \in \mathbb{R}^{N\times N} \qquad \text{for } r = 0, \ldots, j-1 \tag{7.66}$$

with $H_{j-r}, G_{j-r} \in \mathbb{R}^{N/2^{r+1}\times N/2^r}$ and the identities $I_r \in \mathbb{R}^{N(1-2^{-r})\times N(1-2^{-r})}$.

Therefore, the orthogonal matrix

$$T = T_1 \cdot T_2 \cdot \ldots \cdot T_{j-1} \cdot T_j \in \mathbb{R}^{N\times N} \tag{7.67}$$

represents the discrete wavelet analysis.

For given wavelet coefficients d in (7.64), the coefficients $c^j = (c_k^j)_{k=0}^{N-1}$ can thereby be reconstructed from $\Pi_j f$ in (7.63). The linear mapping of this reconstruction is called `discrete wavelet synthesis`. The wavelet synthesis is represented by the inverse matrix

$$T^{-1} = T_j^{-1} \cdot T_{j-1}^{-1} \cdot \ldots \cdot T_2^{-1} \cdot T_1^{-1} = T_j^T \cdot T_{j-1}^T \cdot \ldots \cdot T_2^T \cdot T_1^T \in \mathbb{R}^{N\times N}$$

of T in (7.67), so that

$$c^j = T_j^T \cdot \ldots \cdot T_1^T \cdot d.$$

The discrete wavelet analysis and the discrete wavelet synthesis are associated with the terms *discrete wavelet transform* (wavelet analysis) and *inverse discrete wavelet transformation* (wavelet synthesis).

Due to the orthogonality of the matrices T_{j-r} in (7.66), the wavelet transform is *numerically stable*, since

$$\|d\|_2 = \|T_1 \cdot \ldots \cdot T_j \cdot c^j\|_2 = \|c^j\|_2.$$

Moreover, the complexity of the wavelet transform is only linear, since the j decomposition steps (for $r = 0, 1, \ldots, j-1$) require altogether

$$N + N/2 + \ldots + 2 = 2N - 2 = \mathcal{O}(N) \quad \text{for } N \to \infty$$

operations.

7.6 Exercises

Exercise 7.56. Show that the Fourier transform $\hat{f} : \mathbb{R} \longrightarrow \mathbb{C}$,

$$\hat{f}(\omega) = \int_{\mathbb{R}} f(x) \mathrm{e}^{-\mathrm{i}x\omega} \, \mathrm{d}x \qquad \text{for } \omega \in \mathbb{R},$$

of $f \in \mathrm{L}^1(\mathbb{R})$ is a uniformly continuous function on $\mathbb{R}$.

Exercise 7.57. Consider the Banach space $(\mathrm{L}^1(\mathbb{R}), \|\cdot\|_{\mathrm{L}^1(\mathbb{R})})$ and the Hilbert space $(\mathrm{L}^2(\mathbb{R}), \|\cdot\|_{\mathrm{L}^2(\mathbb{R})})$. Show that neither the inclusion $\mathrm{L}^1(\mathbb{R}) \subset \mathrm{L}^2(\mathbb{R})$ nor the inclusion $\mathrm{L}^2(\mathbb{R}) \subset \mathrm{L}^1(\mathbb{R})$ holds. Make a (non-trivial) example for a linear space $\mathcal{S}$ satisfying $\mathcal{S} \subset \mathrm{L}^1(\mathbb{R})$ and $\mathcal{S} \subset \mathrm{L}^2(\mathbb{R})$.

Exercise 7.58. Consider Proposition 7.2.

(a) Prove the properties (a)-(e) in Proposition 7.2.
(b) Give a *multivariate* formulation for each of the statements (a)-(e).

Exercise 7.59. Prove the following statements for the Fourier transform $\mathcal{F}$.

(a) For the Fourier transform of the k-th derivative $f^{(k)}$ of f, we have

$$(\mathcal{F} f^{(k)})(\omega) = (\mathrm{i}\omega)^k (\mathcal{F} f)(\omega) \qquad \text{for all } \omega \in \mathbb{R}$$

under the assumption $f^{(k)} \in \mathscr{C}(\mathbb{R}) \cap \mathrm{L}^1(\mathbb{R})$.

(b) For the k-th derivative of the Fourier transform $\mathcal{F} f$ of f, we have

$$\frac{\mathrm{d}^k}{\mathrm{d}\omega^k} (\mathcal{F} f)(\omega) = (-\mathrm{i})^k (\mathcal{F}(x^k f))(\omega) \qquad \text{for all } \omega \in \mathbb{R}$$

under the assumption $x^k f \in \mathrm{L}^1(\mathbb{R})$.

Exercise 7.60. Conclude from the results in Exercise 7.59 the statement: *"$f \in \mathrm{L}^1(\mathbb{R})$ is smooth, if and only if $\mathcal{F}f$ has rapid decay around infinity".* Be more precise on this and quantify the decay and the smoothness of f.

Exercise 7.61. Let $f \in \mathrm{L}^1(\mathbb{R}) \setminus \{0\}$ be a function with compact support. Prove the following statements for the Fourier transform $\mathcal{F}f = \hat{f}$ of f.

(a) $\hat{f}$ has arbitrarily many derivatives, i.e., $\hat{f} \in \mathscr{C}^\infty(\mathbb{R})$;
(b) $\hat{f}$ does *not* have compact support.

Exercise 7.62. Prove the estimate

$$\|f * g\|_\infty \leq \|f\|_{\mathrm{L}^1(\mathbb{R})} \cdot \|g\|_\infty \qquad \text{for all } f \in \mathrm{L}^1(\mathbb{R}), g \in \mathscr{C}_0(\mathbb{R}).$$

Exercise 7.63. Prove the convolution formula

$$\mathcal{F}_d f * \mathcal{F}_d g = (2\pi)^d \mathcal{F}_d(f \cdot g) \qquad \text{for all } f, g \in \mathrm{L}^1(\mathbb{R}^d)$$

in the frequency domain of the multivariate Fourier transform $\mathcal{F}_d$.

Exercise 7.64. Prove for $f \in \mathrm{L}^1(\mathbb{R}) \cap \mathscr{C}(\mathbb{R})$ the Fourier inversion formula

$$f(x) = \lim_{\varepsilon \searrow 0} \left(\frac{1}{2\pi} \int_{\mathbb{R}} \hat{f}(\omega) \cdot \mathrm{e}^{\mathrm{i}x\omega} \mathrm{e}^{-\varepsilon|\omega|^2} \, \mathrm{d}\omega \right) \qquad \text{for all } x \in \mathbb{R},$$

i.e., prove Corollary 7.24 as a conclusion from Theorem 7.21.

Hint: see [26, Chapter 7].

Exercise 7.65. Let $f : \mathbb{R} \longrightarrow \mathbb{R}$ be a Lebesgue-measurable function with $f(x) \neq 0$ for almost every $x \in \mathbb{R}$. Moreover, suppose that f satisfies the decay condition

$$|f(x)| \leq C \cdot \mathrm{e}^{-\tau|x|} \qquad \text{for all } x \in \mathbb{R}$$

for some $C, \tau > 0$. Show that the system $(x^n f(x))_{n \in \mathbb{N}_0}$ is *complete* in $\mathrm{L}^2(\mathbb{R})$, i.e.,

$$\overline{\operatorname{span}\{x^n f(x) \,|\, n \in \mathbb{N}_0\}} = \mathrm{L}^2(\mathbb{R}).$$

Hint: Proposition 7.28.

Exercise 7.66. Prove the statements of Proposition 7.47.

Exercise 7.67. Let $V_0 \subset V_1$ be closed subspaces of $\mathrm{L}^2(\mathbb{R})$. Moreover, let $\Pi_\ell : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_\ell$ be linear projectors of $\mathrm{L}^2(\mathbb{R})$ onto V_ℓ, for $\ell = 0, 1$.

(a) Show that the operator $P = \Pi_1 - \Pi_0 : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_1$ is a projector of $\mathrm{L}^2(\mathbb{R})$ onto V_1, if and only if $\Pi_0 \circ \Pi_1 = \Pi_0$.
(b) Make an example for two projectors $\Pi_\ell : \mathrm{L}^2(\mathbb{R}) \longrightarrow V_\ell$, for $\ell = 0, 1$, such that the condition $\Pi_0 \circ \Pi_1 = \Pi_0$ is violated.

Exercise 7.68. For $\psi \in \mathrm{L}^2(\mathbb{R})$, let $\{\psi(\cdot - k) \,|\, k \in \mathbb{Z}\}$ be a Riesz basis of

$$W_0 = \overline{\operatorname{span}\{\psi(\cdot - k) \,|\, k \in \mathbb{Z}\}}$$

with Riesz constants $0 < A \leq B < \infty$. Moreover, let

$$\psi_k^j := 2^{j/2}\psi(2^j \cdot - k) \qquad \text{for } j, k \in \mathbb{Z}.$$

(a) Show that $\{\psi_k^j \,|\, k \in \mathbb{Z}\}$ is a Riesz basis of

$$W_j = \overline{\operatorname{span}\{\psi_k^j \,|\, k \in \mathbb{Z}\}} \qquad \text{for } j \in \mathbb{Z}$$

with Riesz constants $0 < A \leq B < \infty$.

(b) Show that $\{\psi_k^j \,|\, j, k \in \mathbb{Z}\}$ is a Riesz basis of $\mathrm{L}^2(\mathbb{R})$ with Riesz constants $0 < A \leq B < \infty$, provided that

$$\mathrm{L}^2(\mathbb{R}) = \bigoplus_{j \in \mathbb{Z}} W_j.$$

8 Kernel-based Approximation

This chapter is devoted to interpolation and approximation of *multivariate* functions. Throughout this chapter, $f : \Omega \longrightarrow \mathbb{R}$ denotes a continuous function on a domain $\Omega \subset \mathbb{R}^d$, for $d > 1$. Moreover, $X = \{x_1, \ldots, x_n\} \subset \Omega$ is a set of pairwise distinct interpolation points where we assume that the function values of f at X are known. We collect these function values in a data vector

$$f_X = (f(x_1), \ldots, f(x_n))^T = (f_1, \ldots, f_n)^T \in \mathbb{R}^n. \tag{8.1}$$

Since we do not make any assumptions on the distribution of the points X in the domain Ω, the point set X is considered as *scattered*. We formulate the basic interpolation problem for scattered data sets (X, f_X) as follows.

Problem 8.1. On given interpolation points $X = \{x_1, \ldots, x_n\} \subset \Omega$, where $\Omega \subset \mathbb{R}^d$ for $d > 1$, and function values $f_X \in \mathbb{R}^n$ find an *interpolant* $s \in \mathscr{C}(\Omega)$ satisfying $s_X = f_X$, so that s satisfies the *interpolation conditions*

$$s(x_j) = f(x_j) \qquad \text{for all } 1 \leq j \leq n. \tag{8.2}$$

□

According to the Mairhuber-Curtis theorem, Theorem 5.25, there are no non-trivial Haar systems in the *truly* multivariate case, i.e., for multivariate parameter domains $\Omega \subset \mathbb{R}^d$, $d > 1$, containing at least one interior point. Therefore, the interpolation problem for the multivariate case, as formulated by Problem 8.1, is much harder than that for the univariate case.

To solve the posed multivariate interpolation problem, we construct families of basis functions that are generated by a *reproducing kernel* K of a Hilbert space $\mathcal{F}$. The construction of such kernels K requires suitable characterizations for *positive definite functions*, as we explain this in detail later in this chapter. To this end, we rely on fundamental results from functional analysis, which we develop here. For only a few standard results from functional analysis, we omit their proofs and rather refer to the textbook [33].

In the following discussion of this chapter, we show how positive definite kernels lead to *optimal* solutions of the interpolation problem, Problem 8.1. Moreover, we discuss other features and advantages of the proposed interpolation method, where aspects of numerical relevance, e.g. stability and update strategies, are included in our discussion. Finally, we briefly address basic aspects of kernel-based learning methods.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_8

8.1 Multivariate Lagrange Interpolation

8.1.1 Discussion of the Interpolation Problem

Before we develop concrete solutions to Problem 8.1, we discuss the interpolation problem in (8.2) from a more general viewpoint. To this end, suppose that for a continuous function and for pairwise distinct interpolation points $X = \{x_1, \ldots, x_n\} \subset \Omega \subset \mathbb{R}^d$, $d > 1$, a data vector f_X containing function values of the form (8.1) is given.

To solve the interpolation problem in (8.2), we fix a *suitable* (finite-dimensional) subspace $\mathcal{S} \subset \mathscr{C}(\Omega)$, from which we wish to determine an interpolant $s \in \mathcal{S}$ satisfying the interpolation conditions (8.2). To this end, we choose a set $\mathcal{B} = \{s_1, \ldots, s_n\} \subset \mathscr{C}(\Omega)$ of n linearly independent continuous functions $s_j : \Omega \longrightarrow \mathbb{R}$, $1 \leq j \leq n$, so that the finite-dimensional interpolation space

$$\mathcal{S} = \operatorname{span}\{s_1, \ldots, s_n\} \subset \mathscr{C}(\Omega)$$

consists of all linear combinations of functions in $\mathcal{B}$. In this approach, the sought interpolant $s \in \mathcal{S}$ is assumed to be of the form

$$s = \sum_{j=1}^{n} c_j s_j. \tag{8.3}$$

Now the solution of Problem 8.1 leads us to the linear system

$$V_{\mathcal{B},X} \cdot c = f_X$$

for (unknown) coefficients $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ of s in (8.3), where

$$V_{\mathcal{B},X} = (s_j(x_k))_{1 \leq j,k \leq n} \in \mathbb{R}^{n \times n}$$

is the *generalized Vandermonde-Matrix* with respect to the basis $\mathcal{B}$.

We wish to determine the basis $\mathcal{B}$, and so the interpolation space $\mathcal{S}$, such that the interpolation problem (8.2) has for *any* set of interpolation points X and function values f_X a unique solution s from $\mathcal{S}$, i.e., we require the regularity of $V_{\mathcal{B},X}$ for *any* set of interpolation points X.

As we recall from our discussion in Chapter 5, especially in Section 5.3, there is, according to the Mairhuber-Curtis theorem, Theorem 5.25, *no* nontrivial Haar space $\mathcal{S} \subset \mathscr{C}(\Omega)$ on domains $\Omega \subset \mathbb{R}^d$ containing bifurcations. The negative result of Mairhuber-Curtis is in particular critical for the case of multivariate domains. In other words, according to Mairhuber-Curtis, there is for $n \geq 2$ no Haar system $\{s_1, \ldots, s_n\}$, such that for *any* data vector f_X the interpolation problem $f_X = s_X$ assuming an interpolant $s \in \operatorname{span}\{s_1, \ldots, s_n\}$ is guaranteed to have a unique solution.

To further explain this dilemma, we refer to the characterization of Haar spaces in Theorem 5.23. According to Theorem 5.23, for the unique solution

of the interpolation problems (8.2) we need to work with a basis $\mathcal{B}$ whose elements do necessarily depend on the interpolation points X. To construct such *data-dependent* bases $\mathcal{B} = \{s_1, \ldots, s_n\}$, we choose the approach

$$s_j \equiv K(\cdot, x_j) \qquad \text{for } 1 \le j \le n, \tag{8.4}$$

so that the j-th basis function $s_j \in \mathcal{B}$ depends on the j-th interpolation point $x_j \in X$. In this approach, $K : \Omega \times \Omega \longrightarrow \mathbb{R}$ in (8.4) denotes a *suitable* continuous function, whose structural properties are discussed in the following section.

Note that our assumption in (8.4) leads us, for a fixed set of interpolation points $X = \{x_1, \ldots, x_n\} \subset \Omega$ to the finite-dimensional interpolation space

$$\mathcal{S}_X = \operatorname{span}\{K(\cdot, x_j) \,|\, x_j \in X\} \subset \mathscr{C}(\Omega),$$

from which we wish to choose an interpolant of the form

$$s = \sum_{j=1}^{n} c_j K(\cdot, x_j). \tag{8.5}$$

The solution of the interpolation problems $f_X = s_X$ is in this case given by the solution $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ of the linear equation system

$$A_{K,X} \cdot c = f\big|_X$$

with the *interpolation matrix* $A_{K,X} = (K(x_k, x_j))_{1 \le j,k \le n} \in \mathbb{R}^{n \times n}$.

8.1.2 Lagrange Interpolation by Positive Definite Functions

For the sake of unique interpolation, in Problem 8.1, and with assuming (8.5), the matrix $A_{K,X}$ must necessarily be regular. Indeed, this follows directly from Theorem 5.23. In the following discussion, we wish to construct continuous functions $K : \Omega \times \Omega \longrightarrow \mathbb{R}$, such that $A_{K,X}$ is symmetric positive definite for *all* finite sets X of interpolation points, in which case $A_{K,X}$ would be regular. Obviously, the matrix $A_{K,X}$ is symmetric, if the function K is symmetric, i.e., if $K(x, y) = K(y, x)$ for all $x, y \in \mathbb{R}^d$. The requirement for $A_{K,X}$ to be positive definite leads us to the notion of positive definite functions. Since we allow arbitrary parameter domains $\Omega \subset \mathbb{R}^d$, we will from now restrict ourselves (without loss of generality) to the case $\Omega = \mathbb{R}^d$.

Definition 8.2. *A continuous and symmetric function $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is said to be* `positive definite` *on $\mathbb{R}^d$, $K \in \mathbf{PD}_d$, if for any set of pairwise distinct interpolation points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $n \in \mathbb{N}$, the matrix*

$$A_{K,X} = (K(x_k, x_j))_{1 \le j,k \le n} \in \mathbb{R}^{n \times n}, \tag{8.6}$$

is symmetric and positive definite. ◯

We summarize our discussion as follows (cf. Theorem 5.23).

Theorem 8.3. *For $K \in \mathbf{PD}_d$, let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, for $n \in \mathbb{N}$, be a finite point set. Then, the following statements are true.*

(a) The matrix $A_{K,X}$ in (8.6) is positive definite.
(b) If $s \in \mathcal{S}_X$ vanishes on X, i.e., if $s_X = 0$, then $s \equiv 0$.
(c) The interpolation problem $s_X = f_X$ has a unique solution $s \in \mathcal{S}_X$ of the form (8.5), whose coefficient vector $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ is determined by the unique solution of the linear system $A_{K,X} \cdot c = f_X$.

■

By Theorem 8.3, the posed interpolation problem, in Problem 8.1, has for $K \in \mathbf{PD}_d$ a unique solution $s \in \mathcal{S}_X$ of the form (8.5). In this case, for any fixed set of interpolation points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ there is a unique `Lagrange basis` $\{\ell_1, \ldots, \ell_n\} \subset \mathcal{S}_X$, whose `Lagrange basis functions` ℓ_j, $1 \le j \le n$, are uniquely determined by the solution of the *cardinal* interpolation problem

$$\ell_j(x_k) = \delta_{jk} = \begin{cases} 1 & \text{for } j = k \\ 0 & \text{for } j \neq k \end{cases} \qquad \text{for all } 1 \le j, k \le n. \tag{8.7}$$

Therefore, the Lagrange basis functions are also often referred to as *cardinal interpolants*. We can represent the elements of the Lagrange basis $\{\ell_1, \ldots, \ell_n\}$ as follows.

Proposition 8.4. *Let $K \in \mathbf{PD}_d$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$. Then, the Lagrange basis $\{\ell_1, \ldots, \ell_n\} \subset \mathcal{S}_X$ for X is uniquely determined by the solution of the linear system*

$$A_{K,X} \cdot \ell(x) = R(x) \qquad \text{for } x \in \mathbb{R}^d, \tag{8.8}$$

where

$$\ell(x) = (\ell_1(x), \ldots, \ell_n(x))^T \in \mathbb{R}^n \text{ and } R(x) = (K(x, x_1), \ldots, K(x, x_n))^T \in \mathbb{R}^n.$$

The interpolant $s \in \mathcal{S}_X$ satisfying $s_X = f_X$ has the `Lagrange representation`

$$s(x) = \langle f_X, \ell(x) \rangle, \tag{8.9}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product on the Euclidean space $\mathbb{R}^n$.

Proof. For $x = x_j$, the right hand side $R(x_j)$ in (8.8) coincides with the j-th column of $A_{K,X}$, and so the j-th unit vector $e_j \in \mathbb{R}^n$ is the unique solution of the linear equation system (8.8), i.e.,

$$\ell(x_j) = e_j \in \mathbb{R}^n \qquad \text{for all } 1 \le j \le n.$$

In particular, ℓ_j satisfies the conditions (8.7) of cardinal interpolation. Moreover, any Lagrange basis function ℓ_j can, by using $\ell(x) = A_{K,X}^{-1} R(x)$, uniquely be represented as a linear combination

$$\ell_j(x) = e_j^T A_{K,X}^{-1} R(x) \qquad \text{for } 1 \le j \le n \tag{8.10}$$

of the basis functions $K(x, x_j)$ in $R(x)$, i.e., $\ell_j \in \mathcal{S}_X$ for $1 \le j \le n$. From (8.10), we obtain, in particular, the stated representation in (8.8).

Finally, the interpolant s in (8.5) can, by using

$$s(x) = \langle c, R(x) \rangle = \langle A_{K,X}^{-1} f_X, R(x) \rangle = \langle f_X, A_{K,X}^{-1} R(x) \rangle = \langle f_X, \ell(x) \rangle,$$

be represented as a unique linear combination in the Lagrange basis, where

$$s(x) = \sum_{j=1}^{n} f(x_j) \ell_j(x)$$

and so we find the Lagrange representation, as stated in (8.9). ■

8.1.3 Construction of Positive Definite Functions

In this section, we discuss the construction and characterization of positive definite functions. To this end, we use the continuous multivariate Fourier transform from Section 7.4.

But let us first note two simple observations. For $K \in \mathbf{PD}_d$ and $X = \{x\}$, for $x \in \mathbb{R}^d$, the matrix $A_{K,X} \in \mathbb{R}^{1\times 1}$ is positive definite, i.e., $K(x, x) > 0$. For $X = \{x, y\}$, with $x, y \in \mathbb{R}^d$, $x \neq y$, we have $\det(A_{K,X}) > 0$, whereby $K(x, y)^2 < K(x, x)K(y, y)$.

In our subsequent construction of positive definite functions we assume

$$K(x, y) := \Phi(x - y) \qquad \text{for } x, y \in \mathbb{R}^d \tag{8.11}$$

for an *even* continuous function $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$, i.e., $\Phi(x) = \Phi(-x)$ for all $x \in \mathbb{R}^d$. Important special cases for such Φ are *radially symmetric* functions.

Definition 8.5. *A continuous function $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is* `radially symmetric` *on $\mathbb{R}^d$, with respect to the Euclidean norm $\|\cdot\|_2$, in short, Φ is* `radially symmetric`*, if there exists a continuous function $\phi : [0, \infty) \longrightarrow \mathbb{R}$ satisfying $\Phi(x) = \phi(\|x\|_2)$ for all $x \in \mathbb{R}^d$.* ◯

Obviously, every radially symmetric function $\Phi = \phi(\|\cdot\|_2)$ is even. In the following discussion, we call Φ or ϕ *positive definite*, respectively, in short, $\Phi \in \mathbf{PD}_d$ or $\phi \in \mathbf{PD}_d$, if and only if $K \in \mathbf{PD}_d$.

We summarize our observations for $K \in \mathbf{PD}_d$ in (8.11) as follows.

Remark 8.6. Let $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ be even and positive definite, i.e., $\Phi \in \mathbf{PD}_d$. Then, the following statements hold.

(a) $\Phi(0) > 0$;
(b) $|\Phi(x)| < \Phi(0)$ for all $x \in \mathbb{R}^d \setminus \{0\}$.

From now, we assume the normalization $\Phi(0) = 1$. This is without loss of generality, since for $\Phi \in \mathbf{PD}_d$, we have $\alpha\,\Phi \in \mathbf{PD}_d$ for any $\alpha > 0$. □

Now let us discuss the construction of positive definite functions. This is done by using the continuous *Fourier transform*

$$\hat{f}(\omega) := \int_{\mathbb{R}^d} f(x) \mathrm{e}^{-\mathrm{i}\langle x,\omega\rangle}\,\mathrm{d}x \qquad \text{for } f \in \mathrm{L}^1(\mathbb{R}^d).$$

The following fundamental result is due to Bochner[1] who studied in [8] positive (semi-)definite functions of one variable. We can make use of the *Bochner theorem* in [8] to prove suitable characterizations for multivariate positive definite functions.

Theorem 8.7. (Bochner, 1932).
Suppose that $\Phi \in \mathscr{C}(\mathbb{R}^d) \cap \mathrm{L}^1(\mathbb{R}^d)$ is an even function. If the Fourier transform $\hat{\Phi}$ of Φ is positive on $\mathbb{R}^d$, $\hat{\Phi} > 0$, then Φ is positive definite on $\mathbb{R}^d$, $\Phi \in \mathbf{PD}_d$.

Proof. For $\Phi \in \mathscr{C}(\mathbb{R}^d) \cap \mathrm{L}^1(\mathbb{R}^d)$, the Fourier inversion formula

$$\Phi(x) = (2\pi)^{-d} \int_{\mathbb{R}^d} \hat{\Phi}(\omega) \mathrm{e}^{\mathrm{i}\langle x,\omega\rangle}\,\mathrm{d}\omega$$

holds (see Corollary 7.24). Moreover, $\hat{\Phi}$ is continuous on $\mathbb{R}^d$ (cf. our discussion in Section 7.1). If $\hat{\Phi} > 0$ on $\mathbb{R}^d$, then the quadratic form

$$c^T A_{K,X} c = \sum_{j,k=1}^{n} c_j c_k \Phi(x_j - x_k) = (2\pi)^{-d} \int_{\mathbb{R}^d} \left| \sum_{j=1}^{n} c_j \mathrm{e}^{\mathrm{i}\langle x_j,\omega\rangle} \right|^2 \hat{\Phi}(\omega)\,\mathrm{d}\omega$$

is non-negative for any pair (c, X) of a vector $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ and a point set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, i.e., $c^T A_{K,X} c \geq 0$. If $c^T A_{K,X} c = 0$, then the *symbol function*

$$S(\omega) \equiv S_{c,X}(\omega) = \sum_{j=1}^{n} c_j e^{i\langle x_j,\omega\rangle} \qquad \text{for } \omega \in \mathbb{R}^d$$

must vanish identically on $\mathbb{R}^d$, due to the positivity of $\hat{\Phi}$ on $\mathbb{R}^d$. By the linear independence of the functions $\mathrm{e}^{\mathrm{i}\langle x_j,\cdot\rangle}$, we can conclude $c = 0$ from $S \equiv 0$ (see Exercise 8.61). Therefore, we have $c^T A_{K,X} c > 0$ for all $c \in \mathbb{R}^n \setminus \{0\}$ and $X \subset \mathbb{R}^d$ with $|X| = n \in \mathbb{N}$. ■

[1] Salomon Bochner (1899-1982), mathematician

Remark 8.8. We could also work with weaker assumptions on $\hat{\Phi} \in \mathscr{C}(\mathbb{R}^d)$ in Theorem 8.7, if we merely require non-negativity $\hat{\Phi} \geq 0$, with $\hat{\Phi} \not\equiv 0$, for $\hat{\Phi}$. But the somewhat stronger requirements for $\hat{\Phi}$ in Theorem 8.7 are sufficient for our purposes and, in fact, quite convenient in our following discussion. □

By using Bochner's characterization of Theorem 8.7, we can make three examples for positive definite *radially symmetric* functions Φ.

Example 8.9. The `Gauss function`

$$\Phi(x) = \mathrm{e}^{-\|x\|_2^2} \qquad \text{for } x \in \mathbb{R}^d$$

is for any $d \geq 1$ positive definite on $\mathbb{R}^d$, $\Phi \in \mathbf{PD}_d$, by Example 7.41, where

$$\hat{\Phi}(\omega) = \pi^{d/2}\mathrm{e}^{-\|\omega\|_2^2/4} > 0,$$

and so $K(x, y) = \exp(-\|x - y\|_2^2) \in \mathbf{PD}_d$, according to Theorem 8.7. ♢

Example 8.10. The `inverse multiquadric`

$$\Phi(x) = \left(1 + \|x\|_2^2\right)^{-\beta/2} \qquad \text{for } \beta > d/2$$

is positive definite on $\mathbb{R}^d$ for all $d \in \mathbb{N}$. The Fourier transform of Φ is given as

$$\hat{\Phi}(\omega) = (2\pi)^{-d/2} \cdot \frac{2^{1-\beta}}{\Gamma(\beta)} \|\omega\|_2^{\beta - d/2} K_{d/2-\beta}(\|\omega\|_2), \tag{8.12}$$

where

$$K_\nu(z) = \int_0^\infty \mathrm{e}^{-z\cosh(x)} \cosh(\nu x)\, \mathrm{d}x \qquad \text{for } z \in \mathbb{C} \text{ with } |\arg(z)| < \pi/2$$

is the *modified Bessel function of the third kind* of order $\nu \in \mathbb{C}$. We decided to omit the rather technical details concerning the Fourier transform $\hat{\Phi}$ in (8.12) and its positivity, but rather refer to [72, Theorem 6.13]. ♢

Example 8.11. The `radial characteristic functions`

$$\Phi(x) = (1 - \|x\|_2)_+^\beta = \begin{cases} (1 - \|x\|_2)^\beta & \text{for } \|x\|_2 < 1 \\ 0 & \text{for } \|x\|_2 \geq 1 \end{cases}$$

of Askey [2] are for $d \geq 2$ positive definite on $\mathbb{R}^d$, provided that $\beta \geq (d+1)/2$. In this case, the Fourier transform $\hat{\Phi}$ of Φ can (up to some positive constant) be represented as

$$\hat{\Phi}(s) = s^{-(d/2+\beta+1)} \int_0^s (s - t)^\beta t^{d/2} J_{(d-2)/2}(t)\, \mathrm{d}t > 0 \tag{8.13}$$

for $s = \|\omega\|_2$, where

$$J_\nu(z) = \sum_{j=0}^{\infty} \frac{(-1)^j (z/2)^{\nu+2j}}{j!\Gamma(\nu+j+1)} \qquad \text{for } z \in \mathbb{C} \setminus \{0\}$$

is the *Bessel function of the first kind* of order $\nu \in \mathbb{C}$. Again, we decided to omit the technical details concerning the Fourier transform $\hat{\Phi}$ in (8.13). Further details on the construction and characterization of these early examples for compactly supported radial positive definite functions are in [37]. ♢

Now that we have provided three explicit examples for positive definite (radial) functions, we remark that the characterization of Bochner's theorem allows us to construct even larger classes of positive definite functions. This is done by using convolutions. Recall that for any pair $f, g \in \mathrm{L}^1(\mathbb{R}^d)$ of functions, the Fourier transform maps the *convolution product* $f * g \in \mathrm{L}^1(\mathbb{R}^d)$,

$$(f * g)(x) = \int_{\mathbb{R}^d} f(x-y) g(y)\, dy \qquad \text{for } f, g \in \mathrm{L}^1(\mathbb{R}^d)$$

to the product of their Fourier transforms, i.e.,

$$\widehat{f * g} = \hat{f} \cdot \hat{g} \qquad \text{for } f, g \in \mathrm{L}^1(\mathbb{R}^d)$$

by the Fourier convolution theorem, Theorem 7.14.

For $g(x) = f^*(x) = f(-x)$, we get the non-negative autocorrelation

$$\widehat{f * f^*} = \hat{f} \cdot \overline{\hat{f}} = |\hat{f}|^2 \qquad \text{for } f \in \mathrm{L}^1(\mathbb{R}^d).$$

This gives a simple method for constructing positive definite functions.

Corollary 8.12. *For any function $\Psi \in L^1(\mathbb{R}^d) \setminus \{0\}$, its autocorrelation*

$$\Phi(x) = (\Psi * \Psi^*)(x) = \int_{\mathbb{R}^d} \Psi(x-y)\Psi(-y)\, dy$$

is positive definite, $\Phi \in \mathbf{PD}_d$.

Proof. For $\Psi \in \mathrm{L}^1(\mathbb{R}^d)\setminus\{0\}$, we have $\Phi \in \mathrm{L}^1(\mathbb{R}^d)\setminus\{0\}$, and so $\hat{\Phi} \in \mathscr{C}(\mathbb{R}^d)\setminus\{0\}$. Moreover, the Fourier transform $\hat{\Phi} = |\hat{\Psi}|^2$ of the autocorrelation $\Phi = \Psi * \Psi^*$ is, due to the Fourier convolution theorem, Theorem 7.43, non-negative, so that $\Phi \in \mathbf{PD}_d$, due to Remark 8.8. ■

The practical value of the construction resulting from Corollary 8.12 is, however, rather limited. This is because the autocorrelations $\Psi * \Psi^*$ are rather awkward to evaluate. To avoid numerical integration, one would prefer to work with explicit (preferably simple) analytic expressions for positive definite functions $\Phi = \Psi * \Psi^*$.

We remark that the basic idea of Corollary 8.12 has led to the construction of *compactly supported* positive definite (radial) functions, dating back to

earlier Göttingen works of Schaback & Wendland [62] (in 1993), Wu [74] (in 1994), and Wendland [71] (in 1995). In their constructions, explicit formulas were given for autocorrelations $\Phi = \Psi * \Psi^*$, whose generators $\Psi(x) = \psi(\|x\|_2)$, $x \in \mathbb{R}^d$, are specific *radially symmetric* and compactly supported functions $\psi : [0, \infty) \longrightarrow \mathbb{R}$. This has provided a large family of continuous, radially symmetric, and compactly supported functions $\Phi = \Psi * \Psi^*$, as they were later popularized by Wendland [71], who used the radial characteristic functions of Example 8.11 for Ψ to obtain piecewise polynomial positive definite compactly supported radial functions of minimal degree. For further details concerning the construction of compactly supported positive definite radial functions, we refer to the survey [61] of Schaback.

8.2 Native Reproducing Kernel Hilbert Spaces

The discussion of this section is devoted to *reproducing kernel Hilbert spaces* $\mathcal{F}$ which are generated by positive definite functions $K \in \mathbf{PD}_d$. In particular, for any fixed $K \in \mathbf{PD}_d$, the positive definite function K is shown to be the *reproducing kernel* of its associated Hilbert space $\mathcal{F} \equiv \mathcal{F}_K$, whose structure is entirely determined by the properties of K. Therefore, $\mathcal{F}$ is also referred to as the *native reproducing kernel Hilbert space* of K, in short, *native space*.

To introduce $\mathcal{F}$, we first define, for a fixed positive definite $K \in \mathbf{PD}_d$, the *reconstruction space*

$$\mathcal{S} = \{s \in \mathcal{S}_X \,|\, X \subset \mathbb{R}^d, |X| < \infty\} \tag{8.14}$$

containing all (potential) interpolants of the form

$$s(x) = \sum_{j=1}^{n} c_j K(x, x_j) \tag{8.15}$$

for some $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$.

Note that any $s \in \mathcal{S}$ in (8.15) can be rewritten as

$$s(x) \equiv s_\lambda(x) := \lambda^y K(x, y) \qquad \text{for } \lambda = \sum_{j=1}^{n} c_j \delta_{x_j} \tag{8.16}$$

where δ_x is the *Dirac*[2] *point evaluation functional*, defined by $\delta_x(f) = f(x)$, and λ^y in (8.16) denotes action of the linear functional λ on variable y.

[2] Paul Adrien Maurice Dirac (1902-1984), English physicist

8.2.1 Topology of the Reconstruction Space and Duality

Now we consider the linear space

$$\mathcal{L} = \left\{ \lambda = \sum_{j=1}^{n} c_j \delta_{x_j} \,\middle|\, c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n, X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d, n \in \mathbb{N} \right\}$$

containing all *finite* linear combinations of δ-functionals. We equip $\mathcal{L}$ with the inner product

$$(\lambda, \mu)_K := \lambda^x \mu^y K(x, y) = \sum_{j=1}^{n_\lambda} \sum_{k=1}^{n_\mu} c_j d_k K(x_j, y_k) \qquad \text{for } \lambda, \mu \in \mathcal{L}, \tag{8.17}$$

for $K \in \mathbf{PD}_d$, where

$$\lambda = \sum_{j=1}^{n_\lambda} c_j \delta_{x_j} \in \mathcal{L} \quad \text{and} \quad \mu = \sum_{k=1}^{n_\mu} d_k \delta_{y_k} \in \mathcal{L}.$$

By $\|\cdot\|_K := (\cdot,\cdot)_K^{1/2}$, $\mathcal{L}$ is a Euclidean space. Likewise, via the duality relation in (8.16), we can equip $\mathcal{S}$ with the inner product

$$(s_\lambda, s_\mu)_K := (\lambda, \mu)_K \qquad \text{for } s_\lambda, s_\mu \in \mathcal{S} \tag{8.18}$$

and the norm $\|\cdot\|_K = (\cdot,\cdot)_K^{1/2}$. Note that the normed linear spaces $\mathcal{S}$ and $\mathcal{L}$ are isometric isomorphic, $\mathcal{S} \cong \mathcal{L}$, via the linear bijection $\lambda \longmapsto s_\lambda$ and by the norm isometry

$$\|\lambda\|_K = \|s_\lambda\|_K \qquad \text{for all } \lambda \in \mathcal{L}. \tag{8.19}$$

Before we study the topology of the spaces $\mathcal{L}$ and $\mathcal{S}$ in more detail, we first discuss a few concrete examples for inner products and norms of elements in $\mathcal{L}$ and $\mathcal{S}$.

Example 8.13. For any pair of point evaluation functionals $\delta_{z_1}, \delta_{z_2} \in \mathcal{L}$, with $z_1, z_2 \in \mathbb{R}^d$, their inner product is given by

$$(\delta_{z_1}, \delta_{z_2})_K = \delta_{z_1}^x \delta_{z_2}^y K(x, y) = K(z_1, z_2) = \Phi(z_1 - z_2).$$

Moreover, for the norm of any $\delta_z \in \mathcal{L}$, $z \in \mathbb{R}^d$, we obtain

$$\|\delta_z\|_K^2 = (\delta_z, \delta_z)_K = \delta_z^x \delta_z^y K(x, y) = K(z, z) = \Phi(0) = 1,$$

with using the normalization $\Phi(0) = 1$, as introduced in Remark 8.6. Likewise, we have

$$(K(\cdot, z_1), K(\cdot, z_2))_K = K(z_1, z_2) = \Phi(z_1 - z_2) \tag{8.20}$$

for all $z_1, z_2 \in \mathbb{R}^d$ and

$$\|K(\cdot, z)\|_K = \|\delta_z\|_K = 1 \qquad \text{for all } z \in \mathbb{R}^d.$$

$\diamondsuit$

To extend this first elementary example, we regard, for a fixed point set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, the linear bijection operator $G : \mathbb{R}^n \longrightarrow \mathcal{S}_X$, defined as

$$G(c) = \sum_{j=1}^{n} c_j K(\cdot, x_j) = \langle c, R(x) \rangle \qquad \text{for } c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n. \tag{8.21}$$

Proposition 8.14. *For any $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, we have*

$$(G(c), G(d))_K = \langle c, d \rangle_{A_{K,X}} \qquad \textit{for all } c, d \in \mathbb{R}^n,$$

where

$$\langle c, d \rangle_{A_{K,X}} := c^T A_{K,X} d \qquad \textit{for } c, d \in \mathbb{R}^n$$

denotes the inner product generated by the positive definite matrix $A_{K,X}$. In particular, G is an isometry by

$$\|G(c)\|_K = \|c\|_{A_{K,X}} \qquad \textit{for all } c \in \mathbb{R}^n,$$

where $\| \cdot \|_{A_{K,X}} := \langle \cdot, \cdot \rangle_{A_{K,X}}^{1/2}$.

Proof. By (8.20), we have

$$(G(c), G(d))_K = \sum_{j,k=1}^{n} c_j d_k (K(\cdot, x_j), K(\cdot, x_k))_K = c^T A_{K,X} d = \langle c, d \rangle_{A_{K,X}}$$

for all $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ and $d = (d_1, \ldots, d_n)^T \in \mathbb{R}^n$. ■

The result of Proposition 8.14 leads us to the dual operator of G.

Proposition 8.15. *For any finite point set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, the dual operator $G^* : \mathcal{S}_X \longrightarrow \mathbb{R}^n$ of G in (8.21), characterized by the relation*

$$(G(c), s)_K = \langle c, G^*(s) \rangle \qquad \textit{for } c \in \mathbb{R}^n \textit{ and } s \in \mathcal{S}_X, \tag{8.22}$$

is given as

$$G^*(s) = s_X \qquad \textit{for } s \in \mathcal{S}_X.$$

Proof. Note that for any $s \in \mathcal{S}_X$, there is a unique $d \in \mathbb{R}^n$ satisfying $G(d) = s$, so that we have

$$(G(c), s)_K = (G(c), G(d))_K = \langle c, d \rangle_{A_{K,X}} = \langle c, A_{K,X} d \rangle = \langle c, s_X \rangle$$

for all $c \in \mathbb{R}^n$, in which case the assertion follows directly from (8.22). ■

Next, we compute inner products and norms for the Lagrange basis functions $\ell_1, \ldots, \ell_n$ of $\mathcal{S}_X$. The following proposition yields an important result concerning our subsequent stability analysis of the interpolation method.

Proposition 8.16. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, the inner products between the Lagrange basis functions $\ell_j \in \mathcal{S}_X$ satisfying (8.7) are given as*

$$(\ell_j, \ell_k)_K = a_{jk}^{-1} \qquad \text{for all } 1 \le j, k \le n,$$

where $A_{K,X}^{-1} = (a_{jk}^{-1})_{1 \le j,k \le n} \in \mathbb{R}^{n \times n}$. In particular, the norm of $\ell_j \in \mathcal{S}_X$ is

$$\|\ell_j\|_K^2 = a_{jj}^{-1} \qquad \text{for all } 1 \le j \le n.$$

Proof. The representation of the Lagrange basis functions ℓ_j in (8.10) yields

$$(\ell_j, \ell_k)_K = e_j^T A_{K,X}^{-1} A_{K,X} A_{K,X}^{-1} e_k = e_j^T A_{K,X}^{-1} e_k = a_{jk}^{-1}$$

for all $1 \le j, k \le n$. ■

From Example 8.13 and Proposition 8.16, we see that the matrices

$$A_{K,X} = ((\delta_{x_j}, \delta_{x_k})_K)_{1 \le j,k \le n} \in \mathbb{R}^{n \times n}$$
$$A_{K,X}^{-1} = ((\ell_j, \ell_k)_K)_{1 \le j,k \le n} \in \mathbb{R}^{n \times n}$$

are *Gramian*, i.e., the entries of the symmetric positive definite matrices $A_{K,X}$ and $A_{K,X}^{-1}$ are represented by inner products, respectively.

8.2.2 Construction of the Native Hilbert Space

In this section, we introduce the native Hilbert space $\mathcal{F} \equiv \mathcal{F}_K$ of $K \in \mathbf{PD}_d$. To this end, we perform a *completion* of the Euclidean space $\mathcal{S}$. On this occasion, we recall the general concept of completion for normed linear spaces from functional analysis. But we decided to omit the proofs, where we rather refer to the more general discussion in [33, Appendix B]. The following result can, for instance, be found in [33, Corollary 16.11].

Theorem 8.17. (Completion of normed linear spaces). *Let $\mathcal{S}$ be a normed linear space. Then, $\mathcal{S}$ is isometric isomorphic to a dense subspace of a Banach space $\mathcal{F}$, which is, up to norm isomorphy, unique. The Banach space $\mathcal{F}$ is called* `completion` *of $\mathcal{S}$, in short, $\mathcal{F} = \overline{\mathcal{S}}$.* □

The concept of completion can, obviously, be applied to Euclidean spaces: For any Euclidean space $\mathcal{S}$ in (8.14), there is a unique (up to norm isomorphy) Hilbert space $\mathcal{F}$, which is the completion of $\mathcal{S}$ with respect to the Euclidean norm $\|\cdot\|_K$, i.e., $\mathcal{F} = \overline{\mathcal{S}}$. Likewise, for the dual space $\mathcal{L}$ of $\mathcal{S}$, there is a unique (up to norm isomorphy) Hilbert space $\mathcal{D}$ satisfying $\mathcal{D} = \overline{\mathcal{L}}$.

By the norm isomorphy in (8.19) and by the continuity of the norm $\|\cdot\|_K$, we obtain another important result, where we extend the linear bijection $\lambda \longmapsto s_\lambda$ between $\mathcal{L}$ and $\mathcal{S}$ to $\mathcal{D}$ and $\mathcal{F}$.

Proposition 8.18. *The Hilbert spaces $\mathcal{D}$ and $\mathcal{F}$ are isometric isomorphic,*

$$\mathcal{D} \cong \mathcal{F},$$

via the linear bijection $\lambda \longmapsto s_\lambda$ and by the norm isometry

$$\|\lambda\|_K = \|s_\lambda\|_K \qquad \text{for all } \lambda \in \mathcal{D}.$$

■

Remark 8.19. Any functional $\mu \in \mathcal{D}$ is continuous on the Hilbert space $\mathcal{F}$ by the Cauchy-Schwarz inequality

$$|\mu(s_\lambda)| = |\mu^x \lambda^y K(x,y)| = |(\mu,\lambda)_K| \leq \|\mu\|_K \cdot \|\lambda\|_K = \|\mu\|_K \cdot \|s_\lambda\|_K.$$

In particular, any point evaluation functional $\delta_x \in \mathcal{L}$, $x \in \mathbb{R}^d$, is continuous on $\mathcal{F}$, since we have

$$|\delta_x(f)| \leq \|\delta_x\|_K \cdot \|f\|_K = \|f\|_K \qquad \text{for all } f \in \mathcal{F},$$

where $\|\delta_x\|_K = 1$ (see Example 8.13). □

Resorting to functional analysis, we see that $\mathcal{F}$ is a reproducing kernel Hilbert space. But let us first recall some facts about *reproducing kernels* [1].

Definition 8.20. *Let $\mathcal{H}$ denote a Hilbert space of functions $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, with inner product $(\cdot,\cdot)_\mathcal{H}$. Then, a function $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ is said to be a* `reproducing kernel` *for $\mathcal{H}$, if $K(\cdot,x) \in \mathcal{H}$, for all $x \in \mathbb{R}^d$, and*

$$(K(\cdot,x), f)_\mathcal{H} = f(x) \qquad \text{for all } f \in \mathcal{H} \text{ and all } x \in \mathbb{R}^d.$$

○

Next, we prove an important result concerning the characterization of reproducing kernel Hilbert spaces. To this end, we rely on the *representation theorem of Fréchet*[3]*-Riesz*[4], giving another standard result from functional analysis, which can be found, for instance, in [33, Section 8.3].

Theorem 8.21. (Fréchet-Riesz representation theorem). *Let $\mathcal{H}$ be a Hilbert space. Then, there is for any bounded linear functional $\varphi : \mathcal{H} \longrightarrow \mathbb{R}$ a unique* `representer` *$u_\varphi \in \mathcal{H}$, satisfying*

$$\varphi(u) = (u_\varphi, u)_\mathcal{H} \qquad \text{for all } u \in \mathcal{H}.$$

The mapping $\varphi \longmapsto u_\varphi$ is linear, bijective and isometric from $\mathcal{H}'$ onto $\mathcal{H}$. □

[3] Maurice René Fréchet (1878-1973), French mathematician
[4] Frigyes Riesz (1880-1956), Hungarian mathematician

Theorem 8.22. *A Hilbert space $\mathcal{H}$ of functions $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ has a reproducing kernel, if and only if all point evaluation functionals $\delta_x : \mathcal{H} \longrightarrow \mathbb{R}$, for $x \in \mathbb{R}^d$, are continuous on $\mathcal{H}$.*

Proof. Suppose K is a reproducing kernel for $\mathcal{H}$. Then, by the estimate

$$|\delta_x(f)| = |f(x)| = |(K(\cdot, x), f)_{\mathcal{H}}| \leq \|K(\cdot, x)\|_{\mathcal{H}} \cdot \|f\|_{\mathcal{H}} \qquad \text{for } x \in \mathbb{R}^d$$

any point evaluation functional δ_x is bounded, and so continuous, on $\mathcal{H}$.

As for the converse, suppose that all point evaluation functionals δ_x are continuous on $\mathcal{H}$. Then, due to the Fréchet-Riesz representation theorem, Theorem 8.21, there is, for any $x \in \mathbb{R}^d$, a unique function $k_x \in \mathcal{H}$ satisfying

$$f(x) = \delta_x(f) = (k_x, f)_{\mathcal{H}} \qquad \text{for all } f \in \mathcal{H},$$

and so the function $K(\cdot, x) := k_x$ is a reproducing kernel for $\mathcal{H}$. ■

Remark 8.23. A reproducing kernel K for $\mathcal{H}$ is unique. Indeed, if $\tilde{K}$ is another reproducing kernel for $\mathcal{H}$, then we have

$$(\tilde{K}(\cdot, x), f)_{\mathcal{H}} = f(x) \qquad \text{for all } f \in \mathcal{H} \text{ and all } x \in \mathbb{R}^d.$$

For $k_x := K(\cdot, x)$ and $\tilde{k}_x := \tilde{K}(\cdot, x)$ this implies

$$(k_x - \tilde{k}_x, f)_{\mathcal{H}} = 0 \qquad \text{for all } f \in \mathcal{H} \text{ and all } x \in \mathbb{R}^d,$$

and so $k_x \equiv \tilde{k}_x$, i.e., $K \equiv \tilde{K}$. □

8.2.3 The Madych-Nelson Theorem

Now we are in a position where we can show that the positive definite function $K \in \mathbf{PD}_d$ is the (unique) reproducing kernel for the Hilbert space $\mathcal{F} \equiv \mathcal{F}_K$. To this end, we rely on the seminal works [45, 46, 47] of Madych and Nelson.

Theorem 8.24. (Madych-Nelson, 1983).
For any dual functional $\lambda \in \mathcal{D}$, we have the representation

$$\lambda(f) = (\lambda^y K(\cdot, y), f)_K \qquad \textit{for all } f \in \mathcal{F}. \tag{8.23}$$

Proof. For $\lambda \in \mathcal{L}$ and $s_\mu = \mu^y K(\cdot, y) \in \mathcal{S}$ the representation

$$(\lambda^y K(\cdot, y), s_\mu)_K = (s_\lambda, s_\mu)_K = (\lambda, \mu)_K = \lambda^x \mu^y K(x, y) = \lambda(s_\mu) \tag{8.24}$$

holds, with the inner products $(\cdot, \cdot)_K : \mathcal{L} \times \mathcal{L} \longrightarrow \mathbb{R}$ and $(\cdot, \cdot)_K : \mathcal{S} \times \mathcal{S} \longrightarrow \mathbb{R}$ in (8.17) and in (8.18). By continuous extension of the representation (8.24) from $\mathcal{L}$ to $\mathcal{D}$ and from $\mathcal{S}$ to $\mathcal{F}$ we already obtain the statement in (8.23). ■

From the result of Theorem 8.24, we note the following observation.

Remark 8.25. Any dual functional $\lambda \in \mathcal{D}$ is, according to (8.23) and in the sense of the Fréchet-Riesz representation theorem, Theorem 8.21, uniquely represented by the element $s_\lambda = \lambda^y K(\cdot, y) \in \mathcal{F}$. □

Now we can formulate the central result of this section.

Corollary 8.26. *Every positive definite function $K \in \mathbf{PD}_d$ is the unique reproducing kernel of the Hilbert space $\mathcal{F} \equiv \mathcal{F}_K$ generated by K.*

Proof. On the one hand, we have, for $\delta_x \in \mathcal{L}$, $x \in \mathbb{R}^d$, the representation

$$\delta_x^y K(\cdot, y) = K(\cdot, x) \in \mathcal{F} \qquad \text{for all } x \in \mathbb{R}^d.$$

On the other hand, by letting $\lambda = \delta_x \in \mathcal{L}$ in (8.23), we obtain

$$(K(\cdot, x), f)_K = f(x) \qquad \text{for all } f \in \mathcal{F} \text{ and all } x \in \mathbb{R}^d.$$

Therefore, K is the reproducing kernel of $\mathcal{F}$ according to Definition 8.20. ■

Another useful consequence of the Madych-Nelson theorem is as follows.

Corollary 8.27. *Every function $f \in \mathcal{F}$ is continuous on $\mathbb{R}^d$, $\mathcal{F} \subset \mathscr{C}(\mathbb{R}^d)$.*

Proof. Recall that we assume continuity for $K \in \mathbf{PD}_d$. Therefore, by

$$|f(x) - f(y)| = |(K(\cdot, x) - K(\cdot, y), f)_K| \le \|f\|_K \cdot \|K(\cdot, x) - K(\cdot, y)\|_K,$$

for $f \in \mathcal{F}$, and by

$$\begin{aligned} &\|K(\cdot, x) - K(\cdot, y)\|_K^2 \\ &= (K(\cdot, x), K(\cdot, x))_K - 2(K(\cdot, x), K(\cdot, y))_K + (K(\cdot, y), K(\cdot, y))_K \\ &= K(x, x) - 2K(x, y) + K(y, y) \end{aligned}$$

(cf. Example 8.13) we see that every $f \in \mathcal{F}$ is a continuous function. ■

8.3 Optimality of the Interpolation Method

In this section, we prove further results that directly follow from the Madych-Nelson theorem, Theorem 8.24. As we show, the proposed Lagrange interpolation method is *optimal* in two different senses.

8.3.1 Orthogonality and Best Approximation

The first optimality property is based on the Pythagoras theorem.

Corollary 8.28. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $\mathcal{F}$ can be decomposed as*

$$\mathcal{F} = \mathcal{S}_X \oplus \{f \in \mathcal{F} \,|\, f_X = 0\}\,, \tag{8.25}$$

where $\mathcal{S}_X^\perp = \{f \in \mathcal{F} \,|\, f_X = 0\}$ is the orthogonal complement of $\mathcal{S}_X$ in $\mathcal{F}$.

For $f \in \mathcal{F}$ and the unique interpolant $s \in \mathcal{S}_X$ to f on X satisfying $s_X = f_X$, the `Pythagoras theorem` *holds, i.e.,*

$$\|f\|_K^2 = \|s\|_K^2 + \|f - s\|_K^2. \tag{8.26}$$

Proof. For $f \in \mathcal{F}$, let $s \in \mathcal{S}_X$ be the unique interpolant to f from $\mathcal{S}_X$ satisfying $s_X = f_X$. Then, s can, according to (8.16), be represented as $s = \lambda^y K(\cdot, y)$, for some dual functional $\lambda \in \mathcal{L}$ of the form $\lambda = \sum_{j=1}^n c_j \delta_{x_j}$.

By the Madych-Nelson theorem, Theorem 8.24, we have

$$(s, g)_K = 0 \quad \text{for all } g \in \mathcal{F} \text{ with } \lambda(g) = 0,$$

i.e., s is perpendicular to the (algebraic) kernel of λ, and so the implication

$$g_X = 0 \quad \Longrightarrow \quad g \perp \mathcal{S}_X$$

holds. But this implies $f - s \perp \mathcal{S}_X$, or $f - s \in \mathcal{S}_X^\perp$, since $(f - s)_X = 0$. Therefore, the stated decomposition with the direct sum in (8.25) holds by

$$f = s + (f - s) \in \mathcal{S}_X \oplus \mathcal{S}_X^\perp$$

and, moreover,

$$\|f\|_K^2 = \|f - s + s\|_K^2 = \|f - s\|_K^2 + 2(f - s, s)_K + \|s\|_K^2 = \|f - s\|_K^2 + \|s\|_K^2,$$

whereby the Pythagoras theorem (8.26) is also proven. ∎

By the result of Corollary 8.28, we can identify the unique interpolant $s^* \in \mathcal{S}_X$ to f on X as the orthogonal projection of f onto $\mathcal{S}_X$. Therefore, the interpolant s^* is, according to Remark 4.2, the unique best approximation to f from $\mathcal{S}_X$ with respect to $(\mathcal{F}, \|\cdot\|_K)$. Note that the projection operator $\Pi_{\mathcal{S}_X} : \mathcal{F} \longrightarrow \mathcal{S}_X$, $f \longmapsto s^*$, satisfies the property

$$f - \Pi_{\mathcal{S}_X} f = (I - \Pi_{\mathcal{S}_X})(f) \perp \mathcal{S}_X \qquad \text{for all } f \in \mathcal{F},$$

so that the operator $I - \Pi_{\mathcal{S}_X} : \mathcal{F} \longrightarrow \mathcal{S}_X^\perp$ maps, according to the decomposition in (8.25), onto the orthogonal complement $\mathcal{S}_X^\perp \subset \mathcal{F}$ of $\mathcal{S}_X$ in $\mathcal{F}$. The linear operator $I - \Pi_{\mathcal{S}_X}$ is also a projection operator (cf. our general discussion on orthogonal projections in Section 4.2).

We summarize our observations as follows.

Corollary 8.29. *For $f \in \mathcal{F}$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ the unique interpolant $s^* \in \mathcal{S}_X$ to f on X, $s^*_X = f_X$, satisfies the following properties.*

(a) s^* is the unique orthogonal projection of $f \in \mathcal{F}$ onto $\mathcal{S}_X$.
(b) s^* is the unique best approximation to $f \in \mathcal{F}$ from $\mathcal{S}_X$ w.r.t. $\|\cdot\|_K$.

■

Moreover, Corollary 8.28 implies that the interpolant has *minimal variation.*

Corollary 8.30. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $f_X \in \mathbb{R}^n$, the interpolant $s \in \mathcal{S}_X$ satisfying $s_X = f_X$ is the unique minimizer of the energy functional $\|\cdot\|_K$ among all interpolants from $\mathcal{F}$ to the data f_X, i.e.,*

$$\|s\|_K \leq \|g\|_K \qquad \text{for all } g \in \mathcal{F} \text{ with } g_X = f_X.$$

The interpolant s is uniquely determined by this variational property. ■

Now we analyze to stability of the proposed interpolation method. To this end, we compute the norm of the interpolation operator $I_X : \mathcal{F} \longrightarrow \mathcal{S}_X$, which maps every $f \in \mathcal{F}$ to its unique interpolant $s \in \mathcal{S}_X$ satisfying $f_X = s_X$. On this occasion, we recall the definition for the norm of linear operators, which is in particular for $I_X : \mathcal{F} \longrightarrow \mathcal{S}$ with respect to $\|\cdot\|_K$ given as

$$\|I_X\|_K = \sup_{f \in \mathcal{F} \setminus \{0\}} \frac{\|I_X f\|_K}{\|f\|_K}.$$

Theorem 8.31. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, the native space norm $\|I_X\|_K$ of the interpolation operator $I_X : \mathcal{F} \longrightarrow \mathcal{S}_X$ is one, i.e.,*

$$\|I_X\|_K = 1.$$

Proof. The variational property in Corollary 8.30 implies

$$\|I_X f\|_K \leq \|f\|_K \qquad \text{for all } f \in \mathcal{F}, \tag{8.27}$$

and so $\|I_X\|_K \leq 1$. Due to the projection property $I_X s = s$, for all $s \in \mathcal{S}_X$, equality in (8.27) is attained at any $s \in \mathcal{S}_X$, i.e.,

$$\|I_X s\|_K = \|s\|_K \qquad \text{for all } f \in \mathcal{S}_X,$$

and therefore we have $\|I_X\|_K = 1$. ■

The above result allows us to draw the following conclusion.

Remark 8.32. By the stability property (8.27) in Theorem 8.31, the proposed interpolation method has *minimal* condition number w.r.t. $\|\cdot\|_K$. □

8.3.2 Norm Minimality of the Pointwise Error Functional

The second optimality property of the proposed interpolation method is concerning the *pointwise* error

$$\varepsilon_x(f) = f(x) - s(x) \qquad \text{for } x \in \mathbb{R}^d \tag{8.28}$$

between $f \in \mathcal{F}$ and the interpolant $s \in \mathcal{S}_X$ to f on X satisfying $s_X = f_X$.

By the Lagrange representation of s in (8.9), the pointwise *error functional* $\varepsilon_x : \mathcal{F} \longrightarrow [0, \infty)$ can be written as a linear combination of δ-functionals,

$$\varepsilon_x = \delta_x - \sum_{j=1}^{n} \ell_j(x)\delta_{x_j} = \delta_x - \ell(x)^T \delta_X \in \mathcal{L}, \tag{8.29}$$

where $\delta_X := (\delta_{x_1}, \ldots, \delta_{x_n})^T$. Moreover, we use the notation

$$\ell(x)^T R(x) = \sum_{j=1}^{n} \ell_j(x) K(x, x_j) = \sum_{j=1}^{n} \ell_j(x) \delta_{x_j}^y K(x, y) = \left(\delta_x, \ell^T(x)\delta_X\right)_K .$$

The pointwise error $\varepsilon_x(f)$ in (8.28) is bounded above as follows.

Corollary 8.33. *For $f \in \mathcal{F}$ and $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, let $s \in \mathcal{S}_X$ be the unique interpolant to f on X satisfying $s_X = f_X$. Then, for the pointwise error $\varepsilon_x(f)$ in (8.28) the estimate*

$$|\varepsilon_x(f)| \le \|\varepsilon_x\|_K \cdot \|f\|_K \tag{8.30}$$

holds, where the norm $\|\varepsilon_x\|_K$ of the error functional can be written as

$$\|\varepsilon_x\|_K^2 = 1 - \ell(x)^T A_{K,X} \ell(x) = 1 - \|\ell(x)\|_{A_{K,X}}^2, \tag{8.31}$$

by using the positive definite matrix $A_{K,X}$ in (8.6), so that

$$0 \le \|\varepsilon_x\|_K \le 1 \qquad \text{for all } x \in \mathbb{R}^d. \tag{8.32}$$

The error estimate in (8.30) is sharp*, where equality holds for the function*

$$f_x = \varepsilon_x^y K(\cdot, y) \in \mathcal{F}. \tag{8.33}$$

Proof. By the Madych-Nelson theorem, Theorem 8.24, we have

$$\varepsilon_x(f) = (\varepsilon_x^y K(\cdot, y), f)_K \qquad \text{for all } f \in \mathcal{F}, \tag{8.34}$$

so that (8.30) follows directly from (8.34) and the Cauchy-Schwarz inequality.

We compute the norm of the error functional ε_x in (8.29) by

$$\begin{aligned} \|\varepsilon_x\|_K^2 &= (\varepsilon_x, \varepsilon_x)_K = (\delta_x - \ell(x)^T\delta_X, \delta_x - \ell(x)^T\delta_X)_K \\ &= 1 - 2\ell(x)^T R(x) + \ell(x)^T A_{K,X}\ell(x) = 1 - \ell(x)^T A_{K,X}\ell(x), \end{aligned}$$

(cf. Example 8.13), where we use the representation in (8.8). The upper bound for $\|\varepsilon_x\|_K$ in (8.32) follows from the positive definiteness of $A_{K,X}$.

Finally, for the function f_x in (8.33) equality holds in (8.30), since we get

$$|\varepsilon_x(f_x)| = |(\varepsilon_x^y K(\cdot, y), f_x)_K| = (f_x, f_x)_K = (\varepsilon_x, \varepsilon_x)_K = \|\varepsilon_x\|_K \cdot \|f_x\|_K$$

from the Madych-Nelson theorem, and so the estimate in (8.30) is sharp. ■

Finally, we show the pointwise optimality of the interpolation method. To this end, we regard *quasi-interpolants* of the form

$$s_\ell = \ell^T f_X = \sum_{j=1}^n \ell_j f(x_j) \qquad \text{for } \ell = (\ell_1, \ldots, \ell_n)^T \in \mathbb{R}^n$$

along with their associated pointwise error functionals

$$\varepsilon_x^{(\ell)} = \delta_x - \sum_{j=1}^n \ell_j \delta_{x_j} = \delta_x - \ell^T \delta_X \in \mathcal{L} \qquad \text{for } x \in \mathbb{R}^d. \tag{8.35}$$

For the norm $\|\varepsilon_x^{(\ell)}\|_K$ we have, like in (8.31), the representation

$$\|\varepsilon_x^{(\ell)}\|_K^2 = 1 - 2\ell^T R(x) + \ell^T A_{K,X} \ell.$$

Now let us minimize the norm $\|\varepsilon_x^{(\ell)}\|_K$ under variation of the coefficients $\ell \in \mathbb{R}^n$. This leads us directly to the unconstrained optimization problem

$$\|\varepsilon_x^{(\ell)}\|_K^2 = 1 - 2\ell^T R(x) + \ell^T A_{K,X} \ell \longrightarrow \min_{\ell \in \mathbb{R}^n} ! \tag{8.36}$$

whose unique solution is the solution to the linear system $A_{K,X}\ell = R(x)$. But this already implies the pointwise optimality, that we state as follows.

Corollary 8.34. *Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and $x \in \mathbb{R}^d$. Then, the pointwise error functional ε_x in (8.29) is norm-minimal among all error functionals of the form (8.35), where*

$$\|\varepsilon_x\|_K < \|\varepsilon_x^{(\ell)}\|_K \qquad \text{for all } \ell \in \mathbb{R}^n \text{ with } A_{K,X}\ell \neq R(x),$$

i.e., ε_x is the unique solution to the optimization problem (8.36). ■

8.4 Orthonormal Systems, Convergence, and Updates

In this section, we discuss important numerical aspects of the interpolation method. First, we construct countable systems $\{u_j\}_{j \in \mathbb{N}} \subset \mathcal{S}$ of *orthonormal bases* in $\mathcal{S} \subset \mathcal{F}$, in short, *orthonormal systems*. On this occasion, we recall our discussion in Sections 4.2 and 6.2, where we have already explained important advantages of orthonormal systems. In particular, orthonormal systems and their associated orthogonal projection operators $\Pi : \mathcal{F} \longrightarrow \mathcal{S}$ lead us to *efficient* and *numerically stable* approximation methods.

8.4.1 Construction of Orthonormal Systems

Our following construction of orthonormal systems in $\mathcal{S} \subset \mathcal{F}$ relies on a familiar result from linear algebra, the *spectral theorem* for symmetric matrices.

Proposition 8.35. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, let*

$$A_{K,X} = Q^T D Q$$

be the eigendecomposition of the symmetric positive definite kernel matrix $A_{K,X} \in \mathbb{R}^{n \times n}$ in (8.6), with an orthogonal factor $Q \in \mathbb{R}^{n \times n}$ and a diagonal matrix $D = \operatorname{diag}(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{n \times n}$, whose elements $\sigma_1 \geq \ldots \geq \sigma_n > 0$ are the positive eigenvalues of $A_{K,X}$. Then, the functions

$$u_j(x) = e_j^T D^{-1/2} Q \cdot R(x) \qquad \text{for } 1 \leq j \leq n \tag{8.37}$$

form an orthonormal basis of $\mathcal{S}_X$, where $D^{-1/2} = \operatorname{diag}(\sigma_1^{-1/2}, \ldots, \sigma_n^{-1/2})$.

Proof. By the representation in (8.37), for $R(x) = (K(x, x_j))_{1 \leq j \leq n} \in \mathbb{R}^n$, any u_j is being expressed as a linear combination of the basis functions $\{K(\cdot, x_j)\}_{j=1}^n \subset \mathcal{S}_X$, and so $u_j \in \mathcal{S}_X$. By Proposition 8.14, we obtain the orthonormality relation

$$(u_j, u_k)_K = e_j^T D^{-1/2} Q A_{K,X} Q^T D^{-1/2} e_k = \langle e_j, e_k \rangle = \delta_{jk}$$

for all $1 \leq j, k \leq n$. ∎

Now we develop for $s, \tilde{s} \in \mathcal{S}_X$ useful representations of their inner products $(s, \tilde{s})_K$ and norms $\|s\|_K$. To this end, we work with the inner product

$$\langle c, d \rangle_{A_{K,X}^{-1}} = c^T A_{K,X}^{-1} d \qquad \text{for } c, d \in \mathbb{R}^n,$$

which is generated by the positive definite inverse $A_{K,X}^{-1}$ of $A_{K,X}$.

Proposition 8.36. *For $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, we have the representations*

$$(s, \tilde{s})_K = \langle s_X, \tilde{s}_X \rangle_{A_{K,X}^{-1}} \tag{8.38}$$

$$\|s\|_K = \|s_X\|_{A_{K,X}^{-1}} \tag{8.39}$$

for all $s, \tilde{s} \in \mathcal{S}_X$.

Proof. For $s, \tilde{s} \in \mathcal{S}_X$ we have the Lagrange representations

$$s(x) = \langle s_X, \ell(x) \rangle = \sum_{j=1}^n s(x_j) \ell_j(x) \quad \text{and} \quad \tilde{s}(x) = \langle \tilde{s}_X, \ell(x) \rangle = \sum_{k=1}^n \tilde{s}(x_k) \ell_k(x)$$

according to (8.9) in Proposition 8.4. From this and Proposition 8.16, we get

$$(s, \tilde{s})_K = \sum_{j,k=1}^n s(x_j) \tilde{s}(x_k) (\ell_j, \ell_k)_K = \sum_{j,k=1}^n s(x_j) \tilde{s}(x_k) a_{jk}^{-1} = \langle s_X, \tilde{s}_X \rangle_{A_{K,X}^{-1}},$$

and so (8.38) holds. For $s = \tilde{s}$ in (8.38), we have (8.39). ∎

8.4.2 On the Convergence of the Interpolation Method

In this section, we develop rather elementary convergence results for the proposed kernel-based interpolation method. We use the following notations. By X we denote a finite set of pairwise distinct interpolation points, where we further assume that X is contained in a compact domain $\Omega \subset \mathbb{R}^d$, i.e., $X \subset \Omega$. Moreover, we denote by $s_{f,X} \in \mathcal{S}_X$ the unique interpolant to $f : \Omega \longrightarrow \mathbb{R}$ on X, where we assume that the function f is contained in the linear subspace

$$\mathcal{F}_\Omega := \overline{\operatorname{span}\{K(\cdot, y) \,|\, y \in \Omega\}} \subset \mathcal{F}$$

i.e., $f \in \mathcal{F}_\Omega$. Finally,

$$h_{X,\Omega} := \sup_{y \in \Omega} \min_{x \in X} \|y - x\|_2 \tag{8.40}$$

is the `fill distance` of the interpolation points X in the compact set Ω.

In the following discussion, we analyze for a nested sequence

$$X_1 \subset X_2 \subset X_3 \subset \ldots \subset X_n \subset \ldots \subset \Omega \tag{8.41}$$

of (finite) point sets $X_n \subset \Omega$, for $n \in \mathbb{N}$, the asymptotic behaviour of the minimal distances

$$\eta_K(f, \mathcal{S}_{X_n}) := \|s_{f,X_n} - f\|_K = \inf_{s \in \mathcal{S}_{X_n}} \|s - f\|_K \qquad \text{for } f \in \mathcal{F}_\Omega \tag{8.42}$$

for $n \to \infty$. Moreover, we work with the (reasonable) assumption

$$h_{X_n,\Omega} \searrow 0 \quad \text{for } n \to \infty \tag{8.43}$$

concerning the asymptotic geometric distribution of the interpolation points X_n. Under this assumption, we already obtain our first convergence result.

Theorem 8.37. *Let $(X_n)_{n\in\mathbb{N}} \subset \Omega$ be a nested sequence of interpolation points, as in (8.41). Moreover, suppose that the associated fill distances $h_{X_n,\Omega}$ have asymptotic decay $h_{X_n,\Omega} \searrow 0$, as in (8.43). Then, we have, for any $f \in \mathcal{F}_\Omega$, the convergence*

$$\eta_K(f, \mathcal{S}_{X_n}) = \|s_{f,X_n} - f\|_K \longrightarrow 0 \quad \text{for } n \to \infty.$$

Proof. Suppose $y \in \Omega$. Then, according to our assumption in (8.43) there is a sequence $(x_n)_{n\in\mathbb{N}} \subset \Omega$ of interpolation points $x_n \in X_n$ satisfying

$$\|y - x_n\|_2 \le h_{X_n,\Omega} \longrightarrow 0 \quad \text{for } n \to \infty.$$

Moreover, we have

$$\eta_K^2(K(\cdot, y), \mathcal{S}_{X_n}) \le \|K(\cdot, x_n) - K(\cdot, y)\|_K^2 = 2 - 2K(y, x_n) \longrightarrow 0$$

for $n \to \infty$, due to the continuity of K and the normalization $K(w, w) = 1$.

Now for $Y = \{y_1, \ldots, y_N\} \subset \Omega$ and $c = (c_1, \ldots, c_N)^T \in \mathbb{R}^N$ we consider the function

$$f_{c,Y} = \sum_{j=1}^{N} c_j K(\cdot, y_j) \in \mathcal{S}_Y \subset \mathcal{F}_\Omega.$$

For any $y_j \in Y$, $1 \le j \le N$, we take a sequence $(x_n^{(j)})_{n\in\mathbb{N}} \subset \Omega$ of interpolation points $x_n^{(j)} \in X_n$ satisfying $\|y_j - x_n^{(j)}\|_2 \le h_{X_n,\Omega}$. Moreover, we consider the functions

$$s_{c,n} = \sum_{j=1}^{N} c_j K(\cdot, x_n^{(j)}) \in \mathcal{S}_{X_n} \qquad \text{for } n \in \mathbb{N}.$$

Then, we have

$$\begin{aligned}\eta_K(f_{c,Y}, \mathcal{S}_{X_n}) \le \|s_{c,n} - f_{c,Y}\|_K &= \left\| \sum_{j=1}^{N} c_j \left(K(\cdot, x_n^{(j)}) - K(\cdot, y_j) \right) \right\|_K \\ &\le \sum_{j=1}^{N} |c_j| \cdot \|K(\cdot, x_n^{(j)}) - K(\cdot, y_j)\|_K \longrightarrow 0 \quad \text{for } n \to \infty.\end{aligned}$$

This proves the convergence for the dense subset

$$\mathcal{S}_\Omega := \{f_{c,Y} \in \mathcal{S}_Y \,|\, |Y| < \infty\} \subset \mathcal{F}_\Omega.$$

By continuous extension, we finally obtain the stated convergence on $\mathcal{F}_\Omega$. ■

We remark that the proven convergence in Theorem 8.37 can be arbitrarily slow. Indeed, for any monotonically decreasing zero sequence $(\eta_n)_{n\in\mathbb{N}}$ of non-negative real numbers, i.e., $\eta_n \searrow 0$ for $n \to \infty$, there is a nested sequence of point sets $(X_n)_{n\in\mathbb{N}} \subset \Omega$, as in (8.41), and a function $f \in \mathcal{F}_\Omega$ satisfying

$$\eta_K(f, \mathcal{S}_{X_n}) \ge \eta_n \qquad \text{for all } n \in \mathbb{N}.$$

For the proof of this statement, we refer to Exercise 8.64.

Nevertheless, we can prove convergence rates for norms that are weaker than the native space norm $\|\cdot\|_K$. To make a prototypical case, we restrict ourselves to the maximum norm $\|\cdot\|_\infty$ (cf. Exercise 8.62). On this occasion, recall that any function $f \in \mathcal{F}$ is continuous, according to Corollary 8.27. In particular, we have $\mathcal{F}_\Omega \subset \mathscr{C}(\Omega)$, and so $\|\cdot\|_\infty$ is well-defined on $\mathcal{F}_\Omega$.

For our next convergence result we require the following lemma.

Lemma 8.38. *Let $K(x, y) = \Phi(x - y)$ be positive definite, $K \in \mathbf{PD}_d$, where $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is even and Lipschitz continuous with Lipschitz constant $L > 0$. Then, we have, for any $f \in \mathcal{F}_\Omega$, the estimate*

$$|f(x) - f(y)|^2 \le 2L\|x - y\|_2 \cdot \|f\|_K^2 \qquad \textit{for all } x, y \in \Omega.$$

Proof. Suppose $f \in \mathcal{F}_\Omega$ satisfies $\|f\|_K = 1$ (without loss of generality). Then,

$$\begin{aligned} |f(x) - f(y)|^2 &= |(f, \Phi(\cdot - x) - \Phi(\cdot - y))_K|^2 \leq \|\Phi(\cdot - x) - \Phi(\cdot - y)\|_K^2 \\ &= 2\Phi(0) - 2\Phi(x - y) \leq 2L\|x - y\|_2, \end{aligned}$$

where we use the reproduction property of K in $\mathcal{F}_\Omega$. ∎

Lemma 8.38 immediately implies the following error estimate.

Theorem 8.39. *Let $K(x, y) = \Phi(x-y)$ be positive definite, $K \in \mathbf{PD}_d$, where $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is even and Lipschitz continuous with Lipschitz constant $L > 0$. Moreover, let $X \subset \Omega$ be a finite subset of $\Omega \subset \mathbb{R}^d$. Then, we have, for any $f \in \mathcal{F}_\Omega$, the error estimate*

$$\|s_{f,X} - f\|_\infty \leq \sqrt{2Lh_{X,\Omega}} \cdot \|f\|_K.$$

Proof. Suppose $y \in \Omega$. Then, there is some $x \in X$ satisfying $\|y-x\|_2 \leq h_{X,\Omega}$. Then, from Lemma 8.38, and by $(s_{f,X} - f)(x) = 0$, we can conclude

$$|(s_{f,X} - f)(y)|^2 \leq 2Lh_{X,\Omega} \cdot \|f\|_K^2 \qquad \text{for all } y \in \Omega,$$

where we use the estimate $\|s_{f,X} - f\|_K \leq \|f\|_K$. ∎

From Theorem 8.39 we finally obtain our next convergence result.

Corollary 8.40. *Let $K(x, y) = \Phi(x - y)$ be positive definite, $K \in \mathbf{PD}_d$, where $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is even and Lipschitz continuous with Lipschitz constant $L > 0$. Moreover, let $(X_n)_{n\in\mathbb{N}} \subset \Omega$ be a nested point set of interpolation points, as in (8.41). Finally, assume for the associated fill distances $h_{X_n,\Omega}$ the asymptotic decay $h_{X_n,\Omega} \searrow 0$, as in (8.43). Then, we have, for any $f \in \mathcal{F}_\Omega$, the uniform convergence*

$$\|s_{f,X_n} - f\|_\infty = \mathcal{O}\left(h_{X_n,\Omega}^{1/2}\right) \qquad \text{for } n \to \infty$$

at convergence rate $1/2$. ∎

We remark that we can, under more restrictive assumptions on $\Phi \in \mathbf{PD}_d$, prove convergence rates that are even higher than those in Corollary 8.40. For a prototypical case, we refer to Exercise 8.66.

8.4.3 Update Strategies

Now we develop *update strategies* for the proposed interpolation method. To further explain this task, let us regard a set $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ of $n \in \mathbb{N}$ pairwise distinct points. In this case, one *update step* is initiated by adding a new point $x_{n+1} \in \mathbb{R}^d \setminus X_n$ to X_n. Typically, the insertion of x_{n+1} is motivated by the purpose to improve the quality of the approximation to

$f \in \mathcal{F}$ (according to our discussion in Section 8.4.2). When adding a new point x_{n+1}, this leads by

$$X_{n+1} := X_n \cup \{x_{n+1}\} \qquad \text{for } n \in \mathbb{N} \tag{8.44}$$

to an *updated* set of interpolation points X_{n+1}. Note that the update of X_n in (8.44) yields one additional interpolation condition

$$s(x_{n+1}) = f(x_{n+1})$$

in Problem 8.1, and so this requires an update for the interpolation method.

But we essentially wish to use the data of the interpolant $s_n \in \mathcal{S}_{X_n}$ of f on X_n, $s_{X_n} = f_{X_n}$, to *efficiently* compute the new data for the interpolant $s_{n+1} \in \mathcal{S}_{X_{n+1}}$ of f on X_{n+1}, $s_{X_{n+1}} = f_{X_{n+1}}$. Any method which performs such an efficient update for the relevant data is called an *update strategy*.

By iteration on the update step, we obtain, starting with an initial point set $X_1 = \{x_1\}$, for some $x_1 \in \mathbb{R}^d$, a nested sequence

$$X_1 \subset X_2 \subset X_3 \subset \ldots \subset X_n \subset \mathbb{R}^d \tag{8.45}$$

of subsets X_k, containing $|X_k| = k$ interpolation points each. Moreover, two subsequent point sets $X_k \subset X_{k+1}$ differ only about the point x_{k+1}, so that

$$\{x_{k+1}\} = X_{k+1} \setminus X_k \qquad \text{for } 1 \le k \le n-1.$$

Now we discuss the performance of selected update strategies. We begin with updates on the Lagrange bases. On this occasion, we introduce another orthonormal system for $\mathcal{S} \subset \mathcal{F}$.

Theorem 8.41. *Let $(X_m)_{m=1}^n$ be a nested sequence of point sets of the form (8.45). Moreover, let $\ell^{(m)} = \{\ell_1^{(m)}, \ldots, \ell_m^{(m)}\} \subset \mathcal{S}_{X_m}$ be their corresponding Lagrange bases, for $1 \le m \le n$, satisfying*

$$\ell_j^{(m)}(x_k) = \delta_{jk} \qquad \text{for } 1 \le j,k \le m.$$

Then, the sequence

$$\ell_1^{(1)}, \ldots, \ell_n^{(n)}$$

of the leading Lagrange basis functions forms an orthogonal system in $\mathcal{S}_{X_n}$, where

$$(\ell_j^{(j)}, \ell_k^{(k)})_K = \delta_{jk} \cdot a_{kk}^{-1} \qquad \text{for } 1 \le j,k \le n$$

with the diagonal entries a_{kk}^{-1} of the inverse A_{K,X_k}^{-1} of A_{K,X_k}, $1 \le k \le n$.

Proof. We distinguish two cases.

Case 1: For $j = k$ the statement follows from Proposition 8.16.

Case 2: Suppose $j \ne k$.

Assuming $j < k$ (without loss of generality), we have

$$\ell_k^{(k)}(x_j) = 0 \qquad \text{for all } x_j \in X_j \subset X_k,$$

i.e., $\ell_k^{(k)} \perp \mathcal{S}_{X_j}$. In particular, $(\ell_k^{(k)}, \ell_j^{(j)})_K = 0$. ∎

Next, we develop update strategies for the *Cholesky*[5] *decomposition* of the symmetric positive definite interpolation matrix $A_{K,X}$ in (8.6). To this end, we describe one update step, starting with $X_n = \{x_1, \ldots, x_n\}$. In the following discussion, it is convenient to use the abbreviation $A_n := A_{K,X_n}$.

Now our aim is to *efficiently* compute the coefficients

$$c^{(n+1)} = (c_1^{(n+1)}, \ldots, c_{n+1}^{(n+1)})^T \in \mathbb{R}^{n+1}$$

of the interpolant

$$s_{n+1} = \sum_{j=1}^{n+1} c_j^{(n+1)} K(\cdot, x_j) \in \mathcal{S}_{X_{n+1}}$$

to f on X_{n+1} via the solution to the linear system

$$A_{n+1} c^{(n+1)} = f_{X_{n+1}}$$

from the coefficients $c^{(n)} \in \mathbb{R}^n$ of the previous interpolant $s_n \in \mathcal{S}_{X_n}$ to f on X_n. On this occasion, we recall the Cholesky decomposition for symmetric positive definite matrices, which should be familiar from numerical mathematics (see e.g. [57, Theorem 3.6]). But let us first introduce *lower unitriangular matrices.*

Definition 8.42. *A* `lower unitriangular matrix` $L \in \mathbb{R}^{n \times n}$ *has the form*

$$L = \begin{bmatrix} 1 & & & \\ l_{21} & 1 & & \\ l_{31} & l_{32} & 1 & \\ \vdots & & \ddots & \ddots \\ l_{n1} & \cdots & \cdots & l_{n,n-1} \; 1 \end{bmatrix}$$

i.e., we have $l_{jj} = 1$ for the diagonal entries of L, $1 \le j \le n$, and vanishing entries above the diagonal, i.e., $l_{jk} = 0$ for all $1 \le j < k \le n$. ◯

Theorem 8.43. *Every symmetric positive definite matrix A has a unique factorization of the form*

$$A = LDL^T \tag{8.46}$$

with a lower unitriangular matrix L and a diagonal $D = \operatorname{diag}(d_1, \ldots, d_n)$ with positive diagonal entries $d_1, \ldots, d_n > 0$. □

For a diagonal matrix $D = \operatorname{diag}(d_1, \ldots, d_n)$ with positive diagonal entries, we let $D^{1/2} := \operatorname{diag}(\sqrt{d_1}, \ldots, \sqrt{d_n})$, so that $D^{1/2} \cdot D^{1/2} = D$. Now we can introduce the Cholesky decomposition.

[5] André-Louis Cholesky (1875-1918), French mathematician

Definition 8.44. *For a symmetric positive definite matrix A in (8.46), the unique factorization*

$$A = \bar{L}\bar{L}^T$$

with factor $\bar{L} := L \cdot D^{1/2}$, is called the `Cholesky decomposition` *of A.* ◯

Now we can describe the Cholesky update. We start with the Cholesky decomposition

$$A_n = \bar{L}_n \bar{L}_n^T \tag{8.47}$$

of $A_n = A_{K,X_n}$. When adding one interpolation point $x_{n+1} \in \mathbb{R}^d \setminus X_n$ to X_n, we wish to determine the Cholesky decomposition of $A_{n+1} := A_{K,X_{n+1}}$ for the interpolation points $X_{n+1} = X_n \cup \{x_{n+1}\}$. To this end, we can use the Cholesky decomposition of A_n in (8.47). In the following discussion, we let $\bar{L}_n := L_n \cdot D_n^{1/2}$ for $n \in \mathbb{N}$.

Theorem 8.45. *For $X_n = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, let $A_n = A_{K,X_n}$ be the interpolation matrix in (8.6), whose Cholesky decomposition is as in (8.47). Then, for $A_{n+1} = A_{K,X_{n+1}}$, $X_{n+1} = X_n \cup \{x_{n+1}\}$, the Cholesky decomposition*

$$A_{n+1} = \bar{L}_{n+1}\bar{L}_{n+1}^T \tag{8.48}$$

is given by the Cholesky factor

$$\bar{L}_{n+1} = \left[\begin{array}{c|c} \bar{L}_n & 0 \\ \hline S_n^T D_n^{-1/2} & \left(1 - S_n^T D_n^{-1} S_n\right)^{1/2} \end{array}\right] \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{8.49}$$

where $S_n \in \mathbb{R}^n$ is the unique solution of the triangular system $L_n S_n = R_n$, for $R_n := R(x_{n+1}) = (K(x_1, x_{n+1}), \ldots, K(x_n, x_{n+1}))^T \in \mathbb{R}^n$.

Proof. The matrix A_{n+1} has the form

$$A_{n+1} = \begin{bmatrix} A_n & R_n \\ R_n^T & 1 \end{bmatrix}$$

and, moreover, the decomposition

$$A_{n+1} = \left[\begin{array}{c|c} L_n & 0 \\ \hline S_n^T D_n^{-1} & 1 \end{array}\right] \cdot \left[\begin{array}{c|c} D_n & 0 \\ \hline 0 & 1 - S_n^T D_n^{-1} S_n \end{array}\right] \cdot \left[\begin{array}{c|c} L_n^T & D_n^{-1} S_n \\ \hline 0 & 1 \end{array}\right] \tag{8.50}$$

holds, as we can verify directly by multiplying the factors.

Now note that the three matrix factors on the right hand side in (8.50) have the required form of the unique decomposition $A_{n+1} = L_{n+1} D_{n+1} L_{n+1}^T$ for A_{n+1}, according to Theorem 8.43. Therefore, we have in particular

$$L_{n+1} = \left[\begin{array}{c|c} L_n & 0 \\ \hline S_n^T D_n^{-1} & 1 \end{array}\right] \in \mathbb{R}^{(n+1)\times(n+1)}$$

and $D_{n+1} = \operatorname{diag}(d_1, \ldots, d_n, 1 - S_n^T D_n^{-1} S_n) \in \mathbb{R}^{(n+1)\times(n+1)}$.

But this immediately yields the Cholesky decomposition in (8.48) with the Cholesky factor $\bar{L}_{n+1} = L_{n+1} \cdot D_{n+1}^{1/2}$, for which we can verify the stated form in (8.49) by multiplying the factors. ■

Now let us discuss the computational complexity of the Cholesky update. Essentially, we need to determine the vector S_n in (8.49), which can be computed by forward substitution as the solution of the triangular system in $\mathcal{O}(n^2)$ steps. This allows us to compute the required entries in the last row of the Cholesky factor $\bar{L}_{n+1}$ in (8.49) in $\mathcal{O}(n)$ steps. Altogether, we only require at most $\mathcal{O}(n^2)$ steps for the Cholesky update. In contrast, a complete Cholesky decomposition of A_{n+1} *without* using the Cholesky factor $\bar{L}_n$ of A_n costs $\mathcal{O}(n^3)$ floating point operations (flops).

We compute the coefficients $c^{(n+1)} = (c_1^{(n+1)}, \ldots, c^{(n+1)})^T \in \mathbb{R}^{n+1}$ of the interpolant

$$s_{n+1} = \sum_{j=1}^{n+1} c_j^{(n+1)} K(\cdot, x_j) \in \mathcal{S}_{X_{n+1}}$$

to f on X_{n+1} via the solution of the linear equation system

$$A_{n+1} c^{(n+1)} = f_{X_{n+1}} \tag{8.51}$$

efficiently as follows. To this end, we assume the coefficients $c^{(n)} \in \mathbb{R}^n$ of the previous interpolant $s_n \in \mathcal{S}_{X_n}$ to f on X_n to be known. Moreover, we employ the Cholesky decomposition $A_{n+1} = \bar{L}_{n+1} \bar{L}_{n+1}^T$ of A_{n+1} to compute the solution $c^{(n+1)} \in \mathbb{R}^{n+1}$ of (8.51) in two steps. This is done as follows.

(a) Solve the system $\bar{L}_{n+1} d^{(n+1)} = f_{X_{n+1}}$ by forward substitution.
(b) Solve the system $\bar{L}_{n+1}^T c^{(n+1)} = d^{(n+1)}$ by backward substitution.

Computational methods for solving (a) and (b) should be familiar from numerical mathematics. The numerical solution of the triangular systems in (a) and (b) require $\mathcal{O}(n^2)$ flops each. But we can entirely avoid the computational costs in (a). To this end, we take a closer look at the two systems in (a) and (b).

The system in (a), $\bar{L}_{n+1} d^{(n+1)} = f_{X_{n+1}}$, has the form

$$\left[\begin{array}{c|c} \bar{L}_n & 0 \\ \hline S_n^T D_n^{-1/2} & \left(1 - S_n^T D_n^{-1} S_n\right)^{1/2} \end{array}\right] \cdot \left[\begin{array}{c} d^{(n)} \\ \hline d_{n+1}^{(n+1)} \end{array}\right] = \left[\begin{array}{c} f_{X_n} \\ \hline f(x_{n+1}) \end{array}\right].$$

Note that we have already determined the solution $d^{(n)} \in \mathbb{R}^n$ of the triangular system $\bar{L}_n d^{(n)} = f_{X_n}$ with the computation of the interpolant s_n, whereby we obtain the last coefficient in $d^{(n+1)}$ by

$$d_{n+1}^{(n+1)} = \frac{f(x_{n+1}) - S_n^T D_n^{-1/2} d^{(n)}}{\left(1 - S_n^T D_n^{-1} S_n\right)^{1/2}}. \tag{8.52}$$

But we can avoid the computation of the entry $d_{n+1}^{(n+1)}$ in (8.52). To see this, we consider the system in (b), $\bar{L}_{n+1}^T c^{(n+1)} = d^{(n+1)}$, which has the form

$$\left[\begin{array}{c|c} \bar{L}_n^T & D_n^{-1/2} S_n \\ \hline 0 & \left(1 - S_n^T D_n^{-1} S_n\right)^{1/2} \end{array}\right] \cdot c^{(n+1)} = \left[\begin{array}{c} d^{(n)} \\ \hline d_{n+1}^{(n+1)} \end{array}\right].$$

For the last coefficient in $c^{(n+1)}$, we have the representation

$$c_{n+1}^{(n+1)} = \frac{d_{n+1}^{(n+1)}}{\left(1 - S_n^T D_n^{-1} S_n\right)^{1/2}} = \frac{f(x_{n+1}) - S_n^T D_n^{-1/2} d^{(n)}}{1 - S_n^T D_n^{-1} S_n}.$$

For the computation of the remaining n coefficients in $c^{(n+1)}$, we apply backward substitution. But in this case, the entry $d_{n+1}^{(n+1)}$ in (8.52) is not needed. Therefore, we require for the substitution in (a) *no* computational costs at all, while the backward substitution in (b) costs altogether $\mathcal{O}(n^2)$ flops.

8.5 Stability of the Reconstruction Scheme

In this section, we analyze the numerical stability of the kernel-based interpolation method. To this end, we first prove basic stability results, before we discuss the conditioning of the interpolation problem. The investigations of this section are motivated by the wavelet theory on time-frequency analysis, where the concept of *Riesz stability* plays an important role.

8.5.1 Riesz Bases and Riesz Stability

For the special case of kernel-based interpolation from *finite* data, we can characterize Riesz bases in a rather straightforward manner: For a *finite* set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ of pairwise distinct interpolation points, the basis functions $\mathcal{B}_X = \{K(\cdot, x_j)\}_{j=1}^n \subset \mathcal{S}_X$ are (obviously) a *Riesz basis* of $\mathcal{S}_X$, where we have the *Riesz stability estimate*

$$\sigma_{\min}(A_{K,X}) \|c\|_2^2 \leq \left\| \sum_{j=1}^n c_j K(\cdot, x_j) \right\|_K^2 \leq \sigma_{\max}(A_{K,X}) \|c\|_2^2 \tag{8.53}$$

for all $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$, whose *Riesz constants* are determined by the smallest eigenvalue $\sigma_{\min}(A_{K,X})$ and the largest eigenvalue $\sigma_{\max}(A_{K,X})$ of $A_{K,X}$. Indeed, according to Proposition 8.14, we have for $G : \mathbb{R}^n \longrightarrow \mathcal{S}_X$ in (8.21),

$$G(c) = \sum_{j=1}^n c_j K(\cdot, x_j),$$

the representation

$$\|G(c)\|_K^2 = \|c\|_{A_{K,X}}^2 = c^T A_{K,X} c \qquad \text{for all } c \in \mathbb{R}^n.$$

Therefore, the stated Riesz stability estimate in (8.53) holds by the *Courant*[6]*-Fischer*[7] *theorem*, which should be familiar from linear algebra. In fact, according to the Courant-Fischer theorem, the minimal eigenvalue $\sigma_{\min}(A)$ and the maximal eigenvalue $\sigma_{\max}(A)$ of a symmetric matrix A can be represented by the minimal and the maximal *Rayleigh*[8] *quotient*, respectively, i.e.,

$$\sigma_{\min}(A) = \min_{c \in \mathbb{R}^n \setminus \{0\}} \frac{\langle c, Ac \rangle}{\langle c, c \rangle} \quad \text{and} \quad \sigma_{\max}(A) = \max_{c \in \mathbb{R}^n \setminus \{0\}} \frac{\langle c, Ac \rangle}{\langle c, c \rangle}.$$

By Theorem 6.31, any Riesz basis $\mathcal{B}$ has a unique *dual Riesz basis* $\tilde{\mathcal{B}}$. Now let us determine the dual Riesz basis of $\mathcal{B}_X = \{K(\cdot, x_j)\}_{j=1}^n \subset \mathcal{S}_X$. To this end, we rely on the results from Section 6.2.2. By Theorem 6.31, we can identify the Lagrange basis of $\mathcal{S}_X$ as dual to $\mathcal{B}_X$, i.e., $\tilde{\mathcal{B}}_X = \{\ell_1, \ldots, \ell_n\} \subset \mathcal{S}_X$.

Theorem 8.46. *For any point set $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, the Lagrange basis $\tilde{\mathcal{B}}_X = \{\ell_j\}_{j=1}^n$ is the unique dual Riesz basis of $\mathcal{B}_X = \{K(\cdot, x_j)\}_{j=1}^n$. In particular, the orthonormality relation*

$$(K(\cdot, x_j), \ell_k)_K = \delta_{jk}, \tag{8.54}$$

holds, for all $1 \le j, k \le n$. Moreover, the stability estimates

$$\sigma_{\max}^{-1}(A_{K,X}) \|f_X\|_2^2 \le \left\| \sum_{j=1}^n f(x_j) \ell_j \right\|_K^2 \le \sigma_{\min}^{-1}(A_{K,X}) \|f_X\|_2^2, \tag{8.55}$$

hold, for all $f_X \in \mathbb{R}^n$, and, we have

$$\sigma_{\min}(A_{K,X}) \|s\|_K^2 \le \|s_X\|_2^2 \le \sigma_{\max}(A_{K,X}) \|s\|_K^2 \tag{8.56}$$

for all $s \in \mathcal{S}_X$.

Proof. The orthonormality relation in (8.54) follows from the reproduction property of the kernel K, whereby

$$(K(\cdot, x_j), \ell_k)_K = \ell_k(x_j) = \delta_{jk} \qquad \text{for all } 1 \le j, k \le n.$$

Due to Theorem 6.31, the Lagrange basis $\tilde{\mathcal{B}}_X = \{\ell_j\}_{j=1}^n \subset \mathcal{S}_X$ is the uniquely determined dual Riesz basis of $\mathcal{B}_X = \{K(\cdot, x_j)\}_{j=1}^n \subset \mathcal{S}_X$.

Moreover, by Proposition 8.36, the representation

$$\left\| \sum_{j=1}^n f(x_j) \ell_j \right\|_K^2 = \|f_X\|_{A_{K,X}^{-1}}^2 = f_X^T A_{K,X}^{-1} f_X \qquad \text{for all } f_X \in \mathbb{R}^n$$

[6] Richard Courant (1888-1972), German-US American mathematician
[7] Ernst Sigismund Fischer (1875-1954), Austrian mathematician
[8] John William Strutt, 3. Baron Rayleigh (1842-1919), English physicist

holds. According to the Courant-Fischer theorem, the Rayleigh estimates

$$\sigma_{\min}(A_{K,X}^{-1})\|f_X\|_2^2 \leq f_X^T A_{K,X}^{-1} f_X \leq \sigma_{\max}(A_{K,X}^{-1})\|f_X\|_2^2$$

hold for all $f_X \in \mathbb{R}^n$. This implies the stability estimate in (8.55), where

$$\sigma_{\max}^{-1}(A_{K,X}) = \sigma_{\min}(A_{K,X}^{-1}) \quad \text{and} \quad \sigma_{\min}^{-1}(A_{K,X}) = \sigma_{\max}(A_{K,X}^{-1}).$$

Letting $f = s \in \mathcal{S}_X$ in (8.55), we finally get

$$\sigma_{\max}^{-1}(A_{K,X})\|s_X\|_2^2 \leq \|s\|_K^2 = \left\| \sum_{j=1}^n s(x_j)\ell_j \right\|_K^2 \leq \sigma_{\min}^{-1}(A_{K,X})\|s_X\|_2^2$$

for all $s \in \mathcal{S}_X$, so that the stated estimates in (8.56) hold. ■

From the Riesz duality relation between the bases $\mathcal{B}_X = \{K(\cdot, x_j)\}_{j=1}^n$ and $\tilde{\mathcal{B}}_X = \{\ell_j\}_{j=1}^n$, in combination with Theorem 6.31, in particular with (6.22), we can conclude another important result.

Corollary 8.47. *For $f \in \mathcal{S}_X$, the representations*

$$f = \sum_{j=1}^n (f, K(\cdot, x_j))_K \, \ell_j = \sum_{j=1}^n (f, \ell_j)_K \, K(\cdot, x_j) \tag{8.57}$$

hold. ■

Remark 8.48. We can also verify the representations in (8.57) for

$$f = \sum_{j=1}^n c_j K(\cdot, x_j) = \sum_{j=1}^n f(x_j)\ell_j \in \mathcal{S}_X$$

directly: On the one hand, we get

$$c_j = \langle e_j, c \rangle = e_j^T A_{K,X}^{-1} f_X = f_X^T A_{K,X}^{-1} e_j = (f, \ell_j)_K$$

from Proposition 8.16. On the other hand, we have $(f, K(\cdot, x_j))_K = f(x_j)$ by the reproduction property of the kenel K, for all $1 \leq j \leq n$. □

8.5.2 Conditioning of the Interpolation Problem

In this section, we analyze the *conditioning* of the interpolation problem, Problem 8.1. Thereby, we quantify the sensitivity of the interpolation problem with respect to perturbations of the input data. We restrict ourselves to the interpolation problem for continuous functions $f \in \mathscr{C}(\Omega)$ on a fixed *compact* domain $\Omega \subset \mathbb{R}^d$, i.e., we only allow interpolation point sets X in Ω, $X \subset \Omega$. In practice, this requirement does usually not lead to severe restrictions.

For our subsequent analysis, we equip $\mathscr{C}(\Omega)$ with the maximum norm $\|\cdot\|_\infty$. Moreover, for any set of interpolation points $X = \{x_1, \ldots, x_n\} \subset \Omega$, we denote by $I_X : \mathscr{C}(\Omega) \longrightarrow \mathcal{S}_X$ the interpolation operator for X, which assigns every function $f \in \mathscr{C}(\Omega)$ to its unique interpolant $s \in \mathcal{S}_X$ satisfying $s_X = f_X$.

Definition 8.49. *For $X = \{x_1, \ldots, x_n\} \subset \Omega$, the* `condition number` *of the interpolation problem, Problem 8.1, is the smallest constant $\kappa_\infty \equiv \kappa_{\infty,X}$ satisfying*

$$\|I_X f\|_\infty \leq \kappa_\infty \cdot \|f\|_\infty \qquad \text{for all } f \in \mathscr{C}(\Omega),$$

i.e., κ_∞ is the operator norm $\|I_X\|_\infty$ of I_X on $\mathscr{C}(\Omega)$ w.r.t. $\|\cdot\|_\infty$. ◯

The operator norm $\|I_X\|_\infty = \kappa_\infty$ can be computed as follows.

Theorem 8.50. *For $X = \{x_1, \ldots, x_n\} \subset \Omega$, the norm $\|I_X\|_\infty$ of the interpolation operator $I_X : \mathscr{C}(\Omega) \longrightarrow \mathcal{S}_X$ is given by the* `Lebesgue constant`

$$\Lambda_\infty := \max_{x \in \Omega} \sum_{j=1}^{n} |\ell_j(x)| = \max_{x \in \Omega} \|\ell(x)\|_1, \tag{8.58}$$

i.e., $\|I_X\|_\infty = \Lambda_\infty$.

Proof. For any $f \in \mathscr{C}(\Omega)$, let $s = I_X(f) \in \mathcal{S}_X \subset \mathscr{C}(\Omega)$ denote the unique interpolant to f on X satisfying $f_X = s_X$. Using the Lagrange representation of s in (8.9), we obtain the estimate

$$\|I_X f\|_\infty = \|s\|_\infty \leq \max_{x \in \Omega} \sum_{j=1}^{n} |\ell_j(x)| \cdot |f(x_j)| \leq \Lambda_\infty \cdot \|f\|_\infty,$$

and therefore $\|I_X\|_\infty \leq \Lambda_\infty$.

In order to see that $\|I_X\|_\infty \geq \Lambda_\infty$ holds, suppose that the maximum of Λ_∞ in (8.58) is attained at $x^* \in \Omega$. Moreover, let $g \in \mathscr{C}(\Omega)$ be a function with unit norm $\|g\|_{L_\infty(\Omega)} = 1$, that is satisfying the interpolation conditions $g(x_j) = \operatorname{sgn}(\ell_j(x^*))$, for all $1 \leq j \leq n$. Then, we have

$$\|I_X g\|_\infty \geq (I_X g)(x^*) = \sum_{j=1}^{n} \ell_j(x^*) g(x_j) = \sum_{j=1}^{n} |\ell_j(x^*)| = \Lambda_\infty$$

and so $\|I_X g\|_\infty \geq \Lambda_\infty$, which implies $\|I_X\|_\infty \geq \Lambda_\infty$.

Altogether, the stated identity $\|I_X\|_\infty = \Lambda_\infty$ holds. ■

We can compute bounds on the Lebesgue constant Λ_∞ as follows.

Proposition 8.51. *For $X = \{x_1, \ldots, x_n\} \subset \Omega$, we have the estimates*

$$1 \leq \Lambda_\infty \leq \sum_{j=1}^{n} \sqrt{a_{jj}^{-1}} \leq n \cdot \sqrt{\sigma_{\max}(A_{K,X}^{-1})} \tag{8.59}$$

for the Lebesgue constant Λ_∞, where $a_{jj}^{-1} > 0$ is, for $1 \leq j \leq n$, the j-th diagonal entry in the inverse $A_{K,X}^{-1}$ of $A_{K,X}$.

Proof. We first prove the upper bound in (8.59). To this end, we assume that the maximum in (8.58) is attained at $x^* \in \Omega$. Then, from Example 8.13 and Proposition 8.16, we get the first upper bound in (8.59) by

$$\begin{aligned}\Lambda_\infty &= \sum_{j=1}^{n} |\ell_j(x^*)| = \sum_{j=1}^{n} |\delta_{x^*}(\ell_j)| \\ &\leq \sum_{j=1}^{n} \|\delta_{x^*}\|_K \cdot \|\ell_j\|_K = \sum_{j=1}^{n} \|\ell_j\|_K = \sum_{j=1}^{n} \sqrt{a_{jj}^{-1}}.\end{aligned}$$

But this immediately implies the second upper bound in (8.59) by

$$a_{jj}^{-1} \leq \sigma_{\max}(A_{K,X}^{-1}) \qquad \text{for all } 1 \leq j \leq n.$$

The lower bound in (8.59) holds by $\|\ell(x_j)\|_1 = 1$, for $1 \leq j \leq n$. ∎

We remark that the estimates for Λ_∞ in (8.59) are only rather rough. Optimal bounds for the spectral condition number of $A_{K,X}$ can be found in the recent work [22] of Diederichs.

8.6 Kernel-based Learning Methods

This section is devoted to one particular variant of linear regression. For the description of the basic method, we can directly link with our previous investigations in Sections 2.1 and 2.2. By the introduction of *kernel-based learning methods*, we provide an alternative method for data fitting by Lagrange interpolation. Kernel-based learning is particularly relevant, if the input data (X, f_X) are very large and decontaminated from noise. In such application scenarios, we wish to reduce, for a suitably chosen linear subspace $\mathcal{R} \subset \mathcal{S}$, the *empiric ℓ^2-data error*

$$\eta_X(f, s) = \frac{1}{N} \|s_X - f_X\|_2^2 \tag{8.60}$$

under variation of $s \in \mathcal{R}$. To this end, we construct an approximation s^* to f, $s^* \approx f$, which, in addition, satisfies specific smoothness requirements. We measure the smoothness of s^* by the *native* energy functional $J : \mathcal{S} \longrightarrow \mathbb{R}$,

$$J(s) := \|s\|_K^2 \qquad \text{for } s \in \mathcal{S}. \tag{8.61}$$

To make a compromise between the data error in (8.60) and the smoothness in (8.61), we consider in the following of this section the minimization of the cost functional $J_\alpha : \mathcal{S} \longrightarrow \mathbb{R}$, defined as

$$J_\alpha(s) = \eta_X(f, s) + \alpha J(s) \qquad \text{for } \alpha > 0. \tag{8.62}$$

The term $\alpha J(s)$ in (8.62) is called the *regularization term*, which penalizes *non-smooth* elements $s \in \mathcal{R}$, that are admissible for the optimization problem. Moreover, the *regularization parameter* $\alpha > 0$ is used to balance between the data error $\eta_X(f, s)$ and the smoothness $J(s)$ of s.

Therefore, we can view the approximation method of this section as a regularization method (see Section 2.2). According to the jargon of approximation theory, the proposed method of this section is also referred to as *penalized least squares approximation* (see, e.g. [30]).

8.6.1 Problem Formulation and Characterization of Solutions

To explain the basic approximation problem, let $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ be a finite point set. Moreover, suppose that $Y = \{y_1, \ldots, y_n\}$ is a subset of X, $Y \subset X$, whose size $|Y| = n$ is much smaller than the size $|X| = N$ of X, i.e., $n \ll N$. Then, our aim is to reconstruct an unknown function $f \in \mathcal{F}$ from its values $f_X \in \mathbb{R}^N$ by solving the following *unconstrained optimization problem*.

Problem 8.52. Let $\alpha \geq 0$. Determine from given data f_X and $Y \subset X$, an approximation $s_\alpha \in \mathcal{S}_Y$ to f satisfying

$$\frac{1}{N}\|(f - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 = \min_{s \in \mathcal{S}_Y} \left(\frac{1}{N}\|(f - s)_X\|_2^2 + \alpha\|s\|_K^2 \right). \tag{8.63}$$

We denote the optimization problem in (8.63) as (P_α). □

Before we discuss the well-posedness of problem (P_α), let us first make a few comments. For $\alpha = 0$, the optimization problem (P_0) obviously coincides with the basic problem of linear least squares approximation [7, 41]. For very large values of $\alpha > 0$, the smoothness term $\alpha\|s\|_K^2$ in (8.63) dominates the data error. In fact, we expect that any sequence $\{s_\alpha\}_\alpha$ of solutions s_α to (P_α) converges for $\alpha \to \infty$ to zero, which is the unique minimum of $J(s)$ on $\mathcal{S}_Y$.

Now we show that the optimization problem (P_α) has, for any $\alpha > 0$, a unique solution. To this end, we choose for the data error the representation

$$\eta_X(f, s) = \frac{1}{N}\|f_X - A_{X,Y}c\|_2^2 \qquad \text{for } s \in \mathcal{S}_Y, \tag{8.64}$$

with

$$A_{X,Y} = (K(x_k, y_j))_{1 \leq k \leq N; 1 \leq j \leq n} \in \mathbb{R}^{N \times n},$$

and the coefficient vector $c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n$ of

$$s = \sum_{j=1}^{n} c_j K(\cdot, y_j) \in \mathcal{S}_Y. \tag{8.65}$$

Therefore, by using the representation

$$J(s) = \|s\|_K^2 = c^T A_{K,Y} c \qquad \text{for } s \in \mathcal{S},$$

we can express the cost functional $J_\alpha : \mathcal{S} \longrightarrow \mathbb{R}$ in (8.63) for (P_α) as

$$J_\alpha(s) := \eta_X(f, s) + \alpha J(s) = \frac{1}{N}\|f_X - A_{X,Y} c\|_2^2 + \alpha c^T A_{K,Y} c. \tag{8.66}$$

Now we prove the existence and uniqueness for the solution of (P_α).

Theorem 8.53. *Let $\alpha \geq 0$. Then, the penalized least squares problem (P_α) has a unique solution $s_\alpha \in \mathcal{S}_Y$ of the form (8.65), where the coefficients $c_\alpha \in \mathbb{R}^n$ of s_α are uniquely determined by the solution of the normal equation*

$$\left[\frac{1}{N} A_{X,Y}^T A_{X,Y} + \alpha A_{K,Y}\right] c_\alpha = \frac{1}{N} A_{X,Y}^T f_X. \tag{8.67}$$

Proof. For any solution s_α of (P_α), the corresponding coefficient vector $c_\alpha \in \mathbb{R}^n$ minimizes the cost functional J_α in (8.66). Now the gradient of J_α does necessarily vanish at c_α, whereby the representation through the stated normal equation in (8.67) follows. Note that the coefficient matrix of the normal equation in (8.67) is, for any $\alpha \geq 0$, symmetric positive definite. Therefore, (P_α) has a unique solution. ■

An alternative characterization for the unique solution s_α of (P_α) follows from our previous results on Euclidean approximation (see Section 4.1).

Theorem 8.54. *For $\alpha \geq 0$, the solution $s_\alpha \equiv s_\alpha(f) \in \mathcal{S}_Y$ of (P_α) satisfies the condition*

$$\frac{1}{N}\langle (f - s_\alpha)_X, s_X\rangle = \alpha (s_\alpha, s)_K \qquad \textit{for all } s \in \mathcal{S}_Y. \tag{8.68}$$

Proof. We equip $\mathcal{F} \times \mathcal{F}$ with a positive semi-definite symmetric bilinear form,

$$[(f, g), (\tilde{f}, \tilde{g})]_\alpha := \frac{1}{N}\langle f_X, \tilde{f}_X\rangle + \alpha (g, \tilde{g})_K \qquad \text{for } f, g, \tilde{f}, \tilde{g} \in \mathcal{F},$$

yielding the semi-norm

$$|(f, g)|_\alpha^2 = \frac{1}{N}\|f_X\|_2^2 + \alpha \|g\|_K^2 \qquad \text{for } f, g \in \mathcal{F}.$$

Now the solution $s_\alpha \in S_Y$ of (P_α) corresponds to the best approximation $(s_\alpha^*, s_\alpha^*) \in S_Y \times S_Y$ to $(f, 0)$ with respect to $(\mathcal{F} \times \mathcal{F}, |\cdot|_\alpha)$, i.e.,

$$|(f,0)-(s_\alpha^*,s_\alpha^*)|_\alpha^2 = \inf_{s\in\mathcal{S}_Y} |(f,0)-(s,s)|_\alpha^2.$$

According to Remark 4.2, the best approximation s_α^* is unique and, moreover, characterized by the orthogonality condition

$$[(f,0)-(s_\alpha^*,s_\alpha^*),(s,s)]_\alpha = 0 \qquad \text{for all } s\in\mathcal{S}_Y,$$

which is for $s_\alpha^* = s_\alpha$ equivalent to the condition in (8.68). ■

The characterizations in the Theorems 8.53 and 8.54 are obviously equivalent. Indeed, if we replace $s\in\mathcal{S}_Y$ in (8.68) by the standard basis functions $K(\cdot,y_k)\in\mathcal{S}_Y$, for $1\le k\le n$, then, the condition in (8.68) can be expressed as

$$\frac{1}{N}\langle (f-s_\alpha)_X, R(y_k)\rangle = \alpha(s_\alpha, K(\cdot,y_k))_K \qquad \text{for all } 1\le k\le n, \tag{8.69}$$

where

$$R^T(y_k) = (K(x_1,y_k),\dots,K(x_N,y_k)) = e_k^T A_{X,Y}^T.$$

For s_α in (8.65) with corresponding coefficients $c_\alpha\in\mathbb{R}^n$, we get

$$(s_\alpha)_X = A_{X,Y}c_\alpha \in \mathbb{R}^N,$$

and so we obtain the normal equation in (8.69): On the one hand, the left hand side in (8.69) can be written as

$$\begin{aligned}\frac{1}{N}\langle (f-s_\alpha)_X, R(y_k)\rangle &= \frac{1}{N}\left[R^T(y_k)f_X - R^T(y_k)A_{X,Y}c_\alpha\right]\\ &= \frac{1}{N}\left[e_k^T A_{X,Y}^T f_X - e_k^T A_{X,Y}^T A_{X,Y}c_\alpha\right].\end{aligned}$$

On the other hand, the right hand side in (8.69) can be written as

$$\alpha(s_\alpha,K(\cdot,y_k))_K = \alpha s_\alpha(y_k) = \alpha e_k^T A_{K,Y}c_\alpha,$$

where we used the reproduction property

$$(s_\alpha,K(\cdot,y_k))_K = s_\alpha(y_k)$$

of the kernel K.

8.6.2 Stability, Sensitivity, Error Bounds, and Convergence

Next, we analyze the stability of the proposed regression method. To this end, we first bound the minimum of the cost functional in (8.63) as follows.

Theorem 8.55. *For any $\alpha\ge 0$, the solution $s_\alpha\equiv s_\alpha(f)\in\mathcal{S}_Y$ of (P_α) satisfies the stability estimate*

$$\frac{1}{N}\|(s_\alpha-f)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 \le (1+\alpha)\|f\|_K^2.$$

Proof. Let $s_f \in \mathcal{S}_Y$ denote the (unique) interpolant to f at Y satisfying $(s_f - f)_Y = 0$. Recall $\|s_f\|_K \le \|f\|_K$ from Corollary 8.30. Then, we have

$$\begin{aligned}
\frac{1}{N}\|(s_\alpha - f)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 &= \frac{1}{N}\sum_{k=1}^{N} |s_\alpha(x_k) - f(x_k)|^2 + \alpha\|s_\alpha\|_K^2 \\
&\le \frac{1}{N}\sum_{k=1}^{N} |s_f(x_k) - f(x_k)|^2 + \alpha\|s_f\|_K^2 \\
&\le \frac{1}{N}\sum_{x\in X\setminus Y} \|\varepsilon_x\|_K^2 \cdot \|f\|_K^2 + \alpha\|f\|_K^2 \\
&= \left(\frac{1}{N}\sum_{x\in X\setminus Y} \|\varepsilon_x\|_K^2 + \alpha\right)\|f\|_K^2 \\
&\le \left(\frac{N-n}{N} + \alpha\right)\|f\|_K^2 \le (1+\alpha)\|f\|_K^2,
\end{aligned}$$

where we use the pointwise error estimate in (8.30) along with the uniform estimate $\|\varepsilon_x\|_K \le 1$ in (8.32). ∎

Next, we analyze the sensitivity of problem (P_α) under variation of the smoothing parameter $\alpha \ge 0$. To this end, we first observe that the solution $s_\alpha \equiv s_\alpha(f)$ of problem (P_α) coincides with that of the target function s_0, i.e., $s_\alpha(s_0) = s_\alpha(f)$.

Lemma 8.56. *For any $\alpha \ge 0$, the solution $s_\alpha \equiv s_\alpha(f)$ of (P_α) satisfies the following properties.*

(a) The Pythagoras theorem, i.e.,

$$\|(f - s_\alpha)_X\|_2^2 = \|(f - s_0)_X\|_2^2 + \|(s_0 - s_\alpha)_X\|_2^2.$$

(b) The best approximation property $s_\alpha(s_0) = s_\alpha(f)$, i.e.,

$$\frac{1}{N}\|(s_0 - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 = \min_{s\in\mathcal{S}_Y}\left(\frac{1}{N}\|(s_0 - s)_X\|_2^2 + \alpha\|s\|_K^2\right).$$

Proof. Recall that the solution $s_\alpha(f)$ to (P_α) is characterized by the conditions (8.68) in Theorem 8.54, where for $\alpha = 0$, we obtain the characterization

$$\frac{1}{N}\langle (f - s_0)_X, s_X\rangle = 0 \qquad \text{for all } s \in \mathcal{S}_Y. \tag{8.70}$$

For $s \in \mathcal{S}_Y$, this implies the relation

$$\begin{aligned}
\|(f - s)_X\|_2^2 &= \langle (f - s_0 + s_0 - s)_X, (f - s_0 + s_0 - s)_X\rangle \\
&= \|(f - s_0)_X\|_2^2 + 2\langle (f - s_0)_X, (s_0 - s)_X\rangle + \|(s_0 - s)_X\|_2^2 \\
&= \|(f - s_0)_X\|_2^2 + \|(s_0 - s)_X\|_2^2,
\end{aligned}$$

and so, for $s = s_\alpha$, we get property (a).

To verify property (b), we subtract the representation in (8.70) from that in (8.68), whereby with

$$\frac{1}{N}\langle (s_0 - s_\alpha)_X, s_X \rangle = \alpha(s_\alpha, s)_K \qquad \text{for all } s \in \mathcal{S}_Y \tag{8.71}$$

we get the characterization (8.68) for the unique solution $s_\alpha(s_0)$ of (P_α), so that statement (b) follows from Theorem 8.54. ■

Next, we analyze the convergence of $\{s_\alpha\}_\alpha$ for $\alpha \searrow 0$. To this end, we first prove one stability estimate for s_α, along with one error bound for $s_\alpha - s_0$.

Theorem 8.57. *Let $f \in \mathcal{F}$ and $\alpha \geq 0$. Then, the solution $s_\alpha \equiv s_\alpha(f)$ to problem (P_α) has the following properties.*

(a) s_α satisfies the stability estimate

$$\|s_\alpha\|_K \leq \|s_0\|_K.$$

(b) s_α satisfies the error estimate

$$\frac{1}{N}\|(s_\alpha - s_0)_X\|_2^2 \leq \alpha \|s_0\|_K^2.$$

Proof. Letting $s = s_0 - s_\alpha$ in (8.71) we get

$$\frac{1}{N}\|(s_0 - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 = \alpha(s_\alpha, s_0)_K. \tag{8.72}$$

By using the Cauchy-Schwarz inequality, this implies

$$\frac{1}{N}\|(s_0 - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2 \leq \alpha\|s_\alpha\|_K \cdot \|s_0\|_K,$$

and so the statements in (a) and (b) hold. ■

Finally, we prove the convergence of s_α to s_0, for $\alpha \searrow 0$.

Theorem 8.58. *The solution s_α of (P_α) converges to the solution s_0 of (P_0), for $\alpha \searrow 0$, at the following convergence rates.*

(a) With respect to the norm $\|\cdot\|_K$, we have the convergence

$$\|s_\alpha - s_0\|_K^2 = \mathcal{O}(\alpha) \qquad \text{for } \alpha \searrow 0.$$

(b) With respect to the data error, we have the convergence

$$\frac{1}{N}\|(s_\alpha - s_0)_X\|_2^2 = o(\alpha) \qquad \text{for } \alpha \searrow 0.$$

Proof. To prove (a), first note that

$$\|s\|_X := \|s_X\|_2 \qquad \text{for } s \in \mathcal{S}_Y$$

is a norm on $\mathcal{S}_Y$. To see the definiteness of $\|\cdot\|_X$ on $\mathcal{S}_Y$, note that $\|s\|_X = 0$ implies $s_X = 0$, in particular $s_Y = 0$, since $Y \subset X$, in which case $s = 0$.

Moreover, since $\mathcal{S}_Y$ has finite dimension, the norms $\|\cdot\|_X$ and $\|\cdot\|_K$ are equivalent on $\mathcal{S}_Y$, so that there exists some constant $C > 0$ satisfying

$$\|s\|_K \le C\|s\|_X \qquad \text{for all } s \in \mathcal{S}_Y.$$

This, in combination with property (b) in Theorem 8.57, implies (a) by

$$\|s_\alpha - s_0\|_K^2 \le C^2\|(s_\alpha - s_0)_X\|_2^2 \le C^2 N\alpha\|s_0\|_K^2.$$

To prove (b), we recall the relation (8.72) to obtain

$$(s_\alpha, s_0)_K = \frac{1}{\alpha}\left[\frac{1}{N}\|(s_0 - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2\right] \qquad \text{for } \alpha > 0.$$

This in turn implies the identity

$$\|s_\alpha - s_0\|_K^2 = \|s_0\|_K^2 - \|s_\alpha\|_K^2 - \frac{2}{\alpha N}\|(s_0 - s_\alpha)_X\|_2^2 \tag{8.73}$$

by

$$\begin{aligned}\|s_\alpha - s_0\|_K^2 &= \|s_\alpha\|_K^2 - 2(s_\alpha, s_0)_K + \|s_0\|_K^2 \\ &= \|s_\alpha\|_K^2 + \|s_0\|_K^2 - \frac{2}{\alpha}\left[\frac{1}{N}\|(s_0 - s_\alpha)_X\|_2^2 + \alpha\|s_\alpha\|_K^2\right] \\ &= \|s_0\|_K^2 - \|s_\alpha\|_K^2 - \frac{2}{\alpha N}\|(s_0 - s_\alpha)_X\|_2^2.\end{aligned}$$

To complete our proof for (b), note that, by statement (a), the left hand side in (8.73) tends to zero, for $\alpha \searrow 0$, and so does the right hand side in (8.73) tend to zero. By the stability estimate in Theorem 8.57 (a), we get

$$0 \le \|s_0\|_K - \|s_\alpha\|_K \le \|s_\alpha - s_0\|_K \longrightarrow 0 \quad \text{for } \alpha \searrow 0,$$

so that $\|s_\alpha\|_K \longrightarrow \|s_0\|_K$ for $\alpha \searrow 0$. Therefore,

$$\frac{2}{\alpha N}\|(s_0 - s_\alpha)_X\|_2^2 \longrightarrow 0 \quad \text{for } \alpha \searrow 0,$$

which completes our proof for (b). ∎

8.7 Exercises

Exercise 8.59. Let $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$ be a continuous symmetric function, for $d > 1$. Moreover, suppose that for some $n \in \mathbb{N}$ all symmetric matrices of the form

$$A_{K,X} = (K(x_k, x_j))_{1 \le j,k \le n} \in \mathbb{R}^{n \times n},$$

for sets $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ of n pairwise distinct points, are regular. Show that *all* symmetric matrices $A_{K,X} \in \mathbb{R}^{n \times n}$ are positive definite, as soon as there is *one* point set $Y = \{y_1, \ldots, y_n\} \subset \mathbb{R}^d$ for which the matrix $A_{K,Y} \in \mathbb{R}^{n \times n}$ is symmetric positive definite.

Hint: Proof of the Mairhuber-Curtis theorem, Theorem 5.25.

Exercise 8.60. Let $\mathcal{F}$ be a Hilbert space of functions $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ with reproducing kernel $K : \mathbb{R}^d \times \mathbb{R}^d \longrightarrow \mathbb{R}$, $K \in \mathbf{PD}_d$. Moreover, for a set of interpolation points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, let $I_X : \mathcal{F} \longrightarrow \mathcal{S}_X$ denote the interpolation operator which assigns every function $f \in \mathcal{F}$ to its unique interpolant $s \in \mathcal{S}_X$ from $\mathcal{S}_X = \operatorname{span}\{K(\cdot, x_j) \,|\, 1 \le j \le n\}$ satisfying $s_X = f_X$.

Prove the following statements.

(a) If the interpolation method is *translation-invariant*, i.e., if we have for any finite set of interpolation points X the *translation invariance*

$$(I_X f)(x) = (I_{X+x_0} f_{x_0})(x + x_0) \qquad \text{for all } f \in \mathcal{F} \text{ and all } x_0 \in \mathbb{R}^d,$$

where $X + x_0 := \{x_1 + x_0, \ldots, x_n + x_0\} \subset \mathbb{R}^d$ and $f_{x_0} := f(\cdot - x_0)$, then K has necessarily the form $K(x, y) = \Phi(x - y)$, where $\Phi \in \mathbf{PD}_d$.

(b) If the interpolation method is translation-invariant and *rotation-invariant*, i.e., if for any finite set of interpolation points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ and for any rotation matrix $Q \in \mathbb{R}^{d \times d}$ the identity

$$(I_X f)(x) = (I_{QX} f_Q)(Qx) \qquad \text{for all } f \in \mathcal{F}$$

holds, where $QX := \{Qx_1, \ldots, Qx_n\} \subset \mathbb{R}^d$ and $f_Q := f(Q^T \cdot)$, then K has necessarily the form $K(x, y) = \phi(\|x - y\|_2)$, where $\phi \in \mathbf{PD}_d$.

Exercise 8.61. Let $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a finite set of $n \in \mathbb{N}$ points. Show that the functions $\mathrm{e}^{\mathrm{i}\langle x_j, \cdot \rangle}$, for $1 \le j \le n$, are linearly independent on $\mathbb{R}^d$, if and only if the points in X are pairwise distinct.

Hint: First prove the assertion for the univariate case, $d = 1$. To this end, consider, for pairwise distinct points $X = \{x_1, \ldots, x_n\} \subset \mathbb{R}$ the linear combination

$$S_{c,X}(\omega) = c_1 \mathrm{e}^{\mathrm{i}\langle x_1, \omega \rangle} + \ldots + c_n \mathrm{e}^{\mathrm{i}\langle x_n, \omega \rangle} \qquad \text{for } c = (c_1, \ldots, c_n)^T \in \mathbb{R}^n.$$

Then, evaluate the function $S_{c,X}$ and its derivatives $S_{c,X}^{(k)}$, for $1 \le k < n$, at $\omega = 0$, to show the implication

$$S_{c,X} \equiv 0 \quad \Longrightarrow \quad c = 0$$

by using the n linear conditions $S_{c,X}^{(k)}(0) = 0$, for $0 \leq k < n$. Finally, to prove the assertion for the multivariate case, $d > 1$, use the separation of the components in $\mathrm{e}^{\mathrm{i}\langle x_j, \omega \rangle}$, for $\omega = (\omega_1, \ldots, \omega_d)^T \in \mathbb{R}^d$ and $1 \leq j \leq n$.

Exercise 8.62. Let $K \in \mathbf{PD}_d$. Show that the native space norm $\|\cdot\|_K$ of the Hilbert space $\mathcal{F} \equiv \mathcal{F}_K$ is *stronger* than the maximum norm $\|\cdot\|_\infty$, i.e., if a sequence $(f_n)_{n\in\mathbb{N}}$ of functions in $\mathcal{F}$ converges w.r.t. $\|\cdot\|_K$ to $f \in \mathcal{F}$, so that $\|f_n - f\|_K \longrightarrow 0$ for $n \to \infty$, then $(f_n)_{n\in\mathbb{N}}$ does also converge w.r.t. the maximum norm $\|\cdot\|_\infty$ to f, so that $\|f_n - f\|_\infty \longrightarrow 0$ for $n \to \infty$.

Exercise 8.63. Let $\mathcal{H}$ be a Hilbert space of functions with reproducing kernel $K \in \mathbf{PD}_d$. Show that $\mathcal{H}$ is the native Hilbert space of K, i.e., $\mathcal{F}_K = \mathcal{H}$.

Hint: First, show the inclusion $\mathcal{F}_K \subset \mathcal{H}$. Then, consider the direct sum

$$\mathcal{H} = \mathcal{F}_K \oplus \mathcal{G}$$

to show $\mathcal{G} = \{0\}$, by contradiction.

Exercise 8.64. Let $(\eta_n)_{n\in\mathbb{N}}$ be a monotonically decreasing zero sequence of non-negative real numbers, i.e., $\eta_n \searrow 0$ for $n \to \infty$. Show that there is a nested sequence of point sets $(X_n)_{n\in\mathbb{N}} \subset \Omega$, as in (8.41), and a function $f \in \mathcal{F}_\Omega$ satisfying

$$\eta_K(f, \mathcal{S}_{X_n}) \geq \eta_n \qquad \text{for all } n \in \mathbb{N}.$$

Exercise 8.65. Let $K(x, y) = \Phi(x-y)$ be positive definite, $K \in \mathbf{PD}_d$, where $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is even and satisfies, for $\alpha > 0$, the growth condition

$$|\Phi(0) - \Phi(x)| \leq C\|x\|_2^\alpha \qquad \text{for all } x \in B_r(0), \tag{8.74}$$

around zero, for some $r > 0$ and some $C > 0$. Show that in this case, every $f \in \mathcal{F} \equiv \mathcal{F}_K$ is *globally* Hölder continuous with Hölder exponent $\alpha/2$, i.e.,

$$|f(x) - f(y)| \leq C\|x - y\|_2^{\alpha/2} \qquad \text{for all } x, y \in \mathbb{R}^d.$$

Conclude that no positive definite kernel function $K \in \mathbf{PD}_d$ satisfies the growth condition in (8.74) for $\alpha > 2$.

Exercise 8.66. Let $K(x, y) = \Phi(x-y)$ be positive definite, $K \in \mathbf{PD}_d$, where $\Phi : \mathbb{R}^d \longrightarrow \mathbb{R}$ is even and satisfies, for $\alpha > 0$, the growth condition

$$|\Phi(0) - \Phi(x)| \leq C\|x\|_2^\alpha \qquad \text{for all } x \in B_r(0),$$

around zero, for some $r > 0$ and $C > 0$. Moreover, for compact $\Omega \subset \mathbb{R}^d$, let $(X_n)_{n\in\mathbb{N}}$ be a nested sequence of subsets $X_n \subset \Omega$, as in (8.41), whose monotonically decreasing fill distances $h_{X_n,\Omega}$ are a zero sequence, i.e., $h_{X_n,\Omega} \searrow 0$ for $n \to \infty$. Show for $f \in \mathcal{F}_\Omega$ the uniform convergence

$$\|s_{f,X_n} - f\|_\infty = \mathcal{O}\left(h_{X_n,\Omega}^{\alpha/2}\right) \quad \text{for } n \to \infty.$$

Determine from this result the convergence rate for the special case of the Gauss kernel in Example 8.9.

Exercise 8.67. Show that the diagonal entry

$$\left(1 - S_n^T D_n^{-1} S_n\right)^{1/2}$$

of the Cholesky factor $\bar{L}_{n+1}$ in (8.49) is positive. To this end, first show the representation

$$1 - S_n^T D_n^{-1} S_n = \|\varepsilon_{x_{n+1},X_n}\|_K^2,$$

where $\varepsilon_{x_{n+1},X_n}$ is the error functional in (8.29) at $x_{n+1} \in X_{n+1} \setminus X_n$ with respect to the set of interpolation points X_n.

9 Computerized Tomography

Computerized tomography (CT) refers to a popular medical imaging method in diagnostic radiology, where large data samples are taken from a human body to generate slices of images to visualize the interior structure, e.g. of organs, muscles, brain tissue, or bones. But computerized tomography is also used in other relevant applications areas, e.g. in non-destructive evaluation of materials.

In the data acquisition of computerized tomography, a CT scan is being generated by a large set of X-ray beams with known intensity, where the X-ray beams pass through a medium (e.g. a human body) whose interior structure is to be recovered. Each CT datum is generated by one X-ray beam which is travelling along a straight line segment from an emitter to a detector.

If we identify the image domain with a convex set in the plane $\Omega \subset \mathbb{R}^2$, then (for each X-ray beam) the emitter is located at some position $\mathbf{x}_{\mathrm{E}} \in \Omega$, whereas the detector is located at another position $\mathbf{x}_{\mathrm{D}} \in \Omega$. Therefore, the X-ray beam passes through Ω along the straight line segment $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}] \subset \Omega$, from $\mathbf{x}_{\mathrm{E}}$ to $\mathbf{x}_{\mathrm{D}}$ (see Figure 9.1).

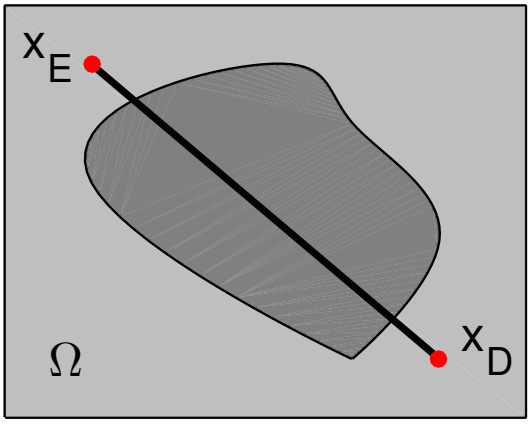


Fig. 9.1. X-ray beam travelling from emitter $\mathbf{x}_{\mathrm{E}}$ to detector $\mathbf{x}_{\mathrm{D}}$ along $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}] \subset \Omega$.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7_9

In the acquisition of one CT datum, the *initial intensity* $I_{\mathrm{E}} = I(\mathbf{x}_{\mathrm{E}})$ of the X-ray beam is being controlled at the emitter, whereas the *final intensity* $I_{\mathrm{D}} = I(\mathbf{x}_{\mathrm{D}})$ is being measured at the detector. Therefore, the difference $\Delta I = I_{\mathrm{E}} - I_{\mathrm{D}}$ gives the loss of intensity. Now the datum ΔI depends on the interior structure (i.e., on the material properties) of the medium along the straight line segment $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}]$. To be more precise, ΔI quantifies the medium's *absorption* of energy on $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}]$.

Now we explain, how the CT data ΔI are interpreted mathematically. By the law of `Lambert`[1]`-Beer`[2] [6]

$$\frac{\mathrm{d}I(\mathbf{x})}{\mathrm{d}\mathbf{x}} = -f(\mathbf{x})I(\mathbf{x}) \tag{9.1}$$

the rate of change for the X-ray intensity $I(\mathbf{x})$ at $\mathbf{x}$ is quantified by the factor $f(\mathbf{x})$, where $f(\mathbf{x})$ is referred to as the *attenuation-coefficient function*. Therefore, the attenuation-coefficient function $f(\mathbf{x})$ yields the energy absorption on the computational domain Ω, and so $f(\mathbf{x})$ represents an important material property of the scanned medium.

In the following of this chapter, we are interested in the reconstruction of $f(\mathbf{x})$. To this end, we further study the differential equation (9.1). By integrating (9.1) along the straight line segment $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}]$, we determine the loss of intensity (or, the loss of energy) of the X-ray beam on $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}]$ by

$$\int_{\mathbf{x}_{\mathrm{E}}}^{\mathbf{x}_{\mathrm{D}}} \frac{\mathrm{d}I(\mathbf{x})}{I(\mathbf{x})} = -\int_{\mathbf{x}_{\mathrm{E}}}^{\mathbf{x}_{\mathrm{D}}} f(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{9.2}$$

Now we can rewrite (9.2) as

$$\log\left(\frac{I(\mathbf{x}_{\mathrm{E}})}{I(\mathbf{x}_{\mathrm{D}})}\right) = \int_{\mathbf{x}_{\mathrm{E}}}^{\mathbf{x}_{\mathrm{D}}} f(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{9.3}$$

The intensity $I_{\mathrm{E}} = I(\mathbf{x}_{\mathrm{E}})$ at the emitter and the intensity $I_{\mathrm{D}} = I(\mathbf{x}_{\mathrm{D}})$ at the detector can be controlled or be measured. This measurement yields the *line integral* of the attenuation-coefficient function $f(\mathbf{x})$ along $[\mathbf{x}_{\mathrm{E}}, \mathbf{x}_{\mathrm{D}}]$,

$$\int_{\mathbf{x}_{\mathrm{E}}}^{\mathbf{x}_{\mathrm{D}}} f(\mathbf{x})\mathrm{d}\mathbf{x}. \tag{9.4}$$

In this chapter, we first explain how the attenuation-coefficient function $f(\mathbf{x})$ can be reconstructed *exactly* from the set of line integrals in (9.4). This leads us to a rather comprehensive mathematical discussion, from the problem formulation to the analytical solution. Then, we develop and analyze numerical algorithms to solve the reconstruction problem for $f(\mathbf{x})$ in relevant application scenarios. Numerical examples are presented for illustration.

[1] Johann Heinrich Lambert (1728-1777), mathematician, physicist, philosopher
[2] August Beer (1825-1863), German mathematician, chemist and physicist

9.1 The Radon Transform

9.1.1 Representation of Lines in the Plane

We represent any straight line $\ell \subset \mathbb{R}^2$ in the Euclidean plane by using polar coordinates. To this end, we consider the orthogonal projection $\mathbf{x}_\ell \in \ell$ of the origin $\mathbf{0} \in \mathbb{R}^2$ onto ℓ. Therefore, we can characterize $\mathbf{x}_\ell \in \ell$ as the unique best approximation to $\mathbf{0}$ from ℓ with respect to the Euclidean norm $\|\cdot\|_2$. Moreover, we consider the (unique) angle $\theta \in [0, \pi)$, for which the unit vector $\mathbf{n}_\theta = (\cos(\theta), \sin(\theta))$ is perpendicular to ℓ. Then, $\mathbf{x}_\ell \in \ell$ can be represented by

$$\mathbf{x}_\ell = (t\cos(\theta), t\sin(\theta)) \in \ell \subset \mathbb{R}^2$$

for some pair $(t, \theta) \in \mathbb{R} \times [0, \pi)$ of polar coordinates. For any straight line $\ell \subset \mathbb{R}^2$, the so constructed polar coordinates $(t, \theta) \in \mathbb{R} \times [0, \pi)$ are *unique*.

As for the converse, for any pair of polar coordinates $(t, \theta) \in \mathbb{R} \times [0, \pi)$ there is a *unique* straight line $\ell \equiv \ell_{t,\theta} \subset \mathbb{R}^2$, which is represented in this way by (t, θ). We introduce this representation as follows (see Figure 9.2).

Definition 9.1. *For any coordinate pair $(t, \theta) \in \mathbb{R} \times [0, \pi)$, we denote by $\ell_{t,\theta} \subset \mathbb{R}^2$ the unique straight line which passes through $\mathbf{x}_\ell = (t\cos(\theta), t\sin(\theta))$ and is perpendicular to the unit vector $\mathbf{n}_\theta = (\cos(\theta), \sin(\theta))$.* ◯

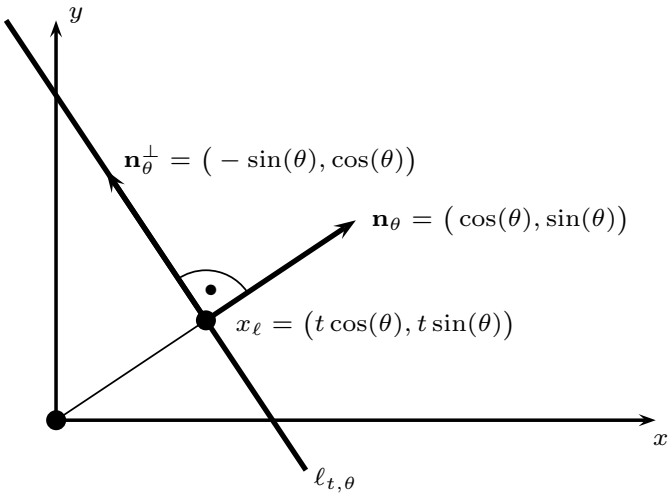


Fig. 9.2. Representation of straight line $\ell_{t,\theta} \subset \mathbb{R}^2$ by coordinates $(t, \theta) \in \mathbb{R} \times [0, \pi)$.

For the parameterization of a straight line $\ell_{t,\theta}$, for $(t, \theta) \in \mathbb{R} \times [0, \pi)$, we use the standard *point-vector representation*, whereby any point $(x, y) \in \ell_{t,\theta}$ in $\ell_{t,\theta}$ is uniquely represented as a linear combination of the form

$$(x, y) = t \cdot \mathbf{n}_\theta + s \cdot \mathbf{n}_\theta^\perp \tag{9.5}$$

with the curve parameter $s \in \mathbb{R}$ and the spanning unit vector

$$\mathbf{n}_\theta^\perp = (-\sin(\theta), \cos(\theta)),$$

which is perpendicular to $\mathbf{n}_\theta$, i.e., $\mathbf{n}_\theta^\perp \perp \mathbf{n}_\theta$ (see Figure 9.2). We can describe the relation between (t, s) and (x, y) in (9.5) via the linear system

$$\begin{aligned} x \equiv x(t, s) &= \cos(\theta)t - \sin(\theta)s \\ y \equiv y(t, s) &= \sin(\theta)t + \cos(\theta)s \end{aligned}$$

or

$$\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} t \\ s \end{bmatrix} = Q_\theta \cdot \begin{bmatrix} t \\ s \end{bmatrix} \tag{9.6}$$

with the rotation matrix $Q_\theta \in \mathbb{R}^{2\times 2}$. The inverse of the orthogonal matrix Q_θ is given by the rotation matrix $Q_{-\theta} = Q_\theta^T$, whereby the representation

$$\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} t \\ s \end{bmatrix} = Q_\theta^T \cdot \begin{bmatrix} x \\ y \end{bmatrix} \tag{9.7}$$

follows immediately from (9.6). Moreover, (9.6), or (9.7), yields the relation

$$t^2 + s^2 = x^2 + y^2, \tag{9.8}$$

which will be useful later in this chapter.

9.1.2 Formulation of the Reconstruction Problem

The basic reconstruction problem of computerized tomography, as sketched at the outset of this chapter, can be stated mathematically as follows.

Problem 9.2. Reconstruct a function $f \equiv f(x, y)$ from *line integrals*

$$\int_{\ell_{t,\theta}} f(x, y)\, \mathrm{d}x\, \mathrm{d}y, \tag{9.9}$$

that are assumed to be known for *all* straight lines $\ell_{t,\theta}$, $(t, \theta) \in \mathbb{R} \times [0, \pi)$. □

We remark that the CT reconstruction problem, Problem 9.2, cannot be solved for *all* bivariate functions f. But under suitable conditions on f we can reconstruct the function f *exactly* from its *Radon*[3] *data*

$$\left\{ \int_{\ell_{t,\theta}} f(x, y)\, \mathrm{d}x\, \mathrm{d}y \;\middle|\; (t, \theta) \in \mathbb{R} \times [0, \pi) \right\}. \tag{9.10}$$

[3] Johann Radon (1887-1956), Austrian mathematician

For any function $f \in \mathrm{L}^1(\mathbb{R}^2)$, the line integral in (9.9) is, for any coordinate pair $(t, \theta) \in \mathbb{R} \times [0, \pi)$, defined as

$$\int_{\ell_{t,\theta}} f(x, y)\, \mathrm{d}x\, \mathrm{d}y = \int_{\mathbb{R}} f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta))\, \mathrm{d}s, \quad (9.11)$$

where we use the coordinate transform in (9.6) with the arc length element

$$\|(\dot{x}(s), \dot{y}(s))\|_2\, \mathrm{d}s = \sqrt{(-\sin(\theta))^2 + (\cos(\theta))^2}\, \mathrm{d}s = \mathrm{d}s$$

on $\ell_{t,\theta}$. This finally leads us to the *Radon transform*.

Definition 9.3. *For $f \equiv f(x, y) \in \mathrm{L}^1(\mathbb{R}^2)$, the function*

$$\mathcal{R}f(t, \theta) = \int_{\mathbb{R}} f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta))\, ds \quad \text{for } t \in \mathbb{R}, \theta \in [0, \pi)$$

is called the `Radon transform` *of f.* ◯

Remark 9.4. The Radon transform $\mathcal{R}$ is well-defined on $\mathrm{L}^1(\mathbb{R}^2)$, where we have $\mathcal{R}f \in \mathrm{L}^1(\mathbb{R} \times [0, \pi))$ for $f \in \mathrm{L}^1(\mathbb{R}^2)$ (see Exercise 9.33). However, there are functions $f \in \mathrm{L}^1(\mathbb{R}^2)$, whose Radon transform $\mathcal{R}f \in \mathrm{L}^1(\mathbb{R} \times [0, \pi))$ is *not* finite in $(t, \theta) \in \mathbb{R} \times [0, \pi)$ (see Exercise 9.34). □

Note that the Radon transform $\mathcal{R}$ is a linear integral transform which maps a bivariate function $f \equiv f(x, y)$ in Cartesian coordinates (x, y) to a bivariate function $\mathcal{R}f(t, \theta)$ in polar coordinates (t, θ). This observation allows us to reformulate the reconstruction of f, Problem 9.2, more concise as follows. On this occasion, we implicitly accommodate the requirement $f \in \mathrm{L}^1(\mathbb{R}^2)$ to the list of our conditions on f.

Problem 9.5. Determine the inversion of the Radon transform $\mathcal{R}$. □

Before we turn to the solution of Problem 9.5, we first give some elementary examples of Radon transforms. We begin with the *indicator function* (i.e., the *characteristic function*) of the disk $\mathcal{B}_r = \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\|_2 \le r\}$, $r \neq 0$.

Example 9.6. For the indicator function $\chi_{\mathcal{B}_r}$ of the disk $\mathcal{B}_r$,

$$f(x, y) = \chi_{\mathcal{B}_r}(x, y) := \begin{cases} 1 & \text{for } x^2 + y^2 \le r^2, \\ 0 & \text{for } x^2 + y^2 > r^2, \end{cases}$$

we compute the Radon transform $\mathcal{R}f$ as follows. We first apply the variable transformation (9.6), which, in combination with the relation (9.8), gives

$$f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta)) = \begin{cases} 1 & \text{for } t^2 + s^2 \le r^2, \\ 0 & \text{for } t^2 + s^2 > r^2. \end{cases}$$

Note that $\mathcal{R}f(t,\theta) = 0$, if and only if the straight line $\ell_{t,\theta}$ does not intersect with the interior of the disk $\mathcal{B}_r$, i.e, if and only if $|t| \geq r$. Otherwise, i.e., for $|t| < r$, we obtain by

$$\mathcal{R}f(t,\theta) = \int_{\ell_{t,\theta}} f(x,y)\,\mathrm{d}(x,y) = \int_{-\sqrt{r^2-t^2}}^{\sqrt{r^2-t^2}} 1\,\mathrm{d}s = 2\sqrt{r^2-t^2}$$

the length of the straight line segment $\ell_{t,\theta} \cap \operatorname{supp}(f) = \ell_{t,\theta} \cap \mathcal{B}_r$. ♢

Example 9.7. We compute the Radon transform of the cone function

$$f(x,y) = \begin{cases} 1 - \sqrt{x^2+y^2} & \text{for } x^2+y^2 \leq 1, \\ 0 & \text{for } x^2+y^2 > 1, \end{cases}$$

or, by transformation (9.6) and on the relation (9.8),

$$f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta)) = \begin{cases} 1 - \sqrt{t^2+s^2} & \text{for } t^2+s^2 \leq 1, \\ 0 & \text{for } t^2+s^2 > 1. \end{cases}$$

In this case, we get $\mathcal{R}f(t,\theta) = 0$ for $|t| \geq 1$ and

$$\begin{aligned} \mathcal{R}f(t,\theta) &= \int_{\ell_{t,\theta}} f(x,y)\,\mathrm{d}(x,y) = \int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} \left(1 - \sqrt{t^2+s^2}\right)\,\mathrm{d}s \\ &= \sqrt{1-t^2} - \frac{t^2}{2}\log\left(\frac{1+\sqrt{1-t^2}}{1-\sqrt{1-t^2}}\right) \end{aligned}$$

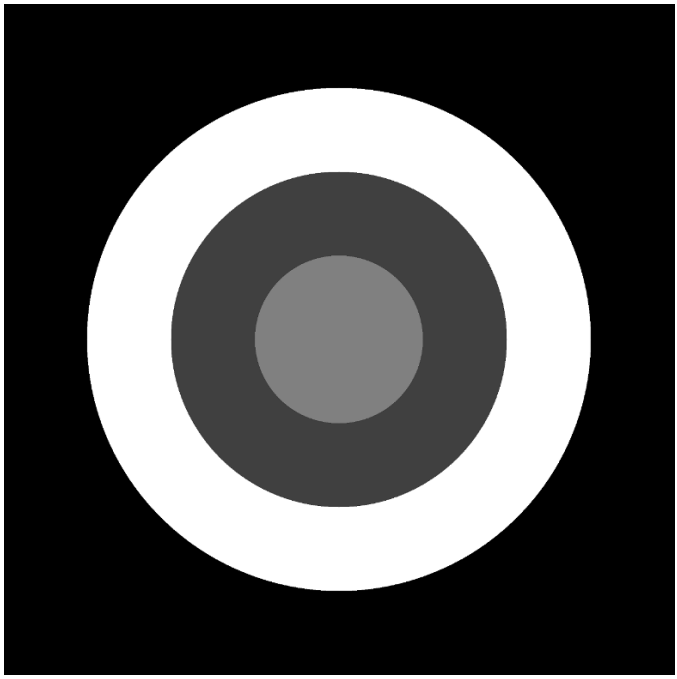
for $|t| < 1$. ♢

Remark 9.8. For any radially symmetric function $f(\cdot) = f(\|\cdot\|_2)$, the Radon transform $\mathcal{R}f(t,\theta)$ does only depend on $t \in \mathbb{R}$, but not on the angle $\theta \in [0,\pi)$. Indeed, in this case, we have the identity

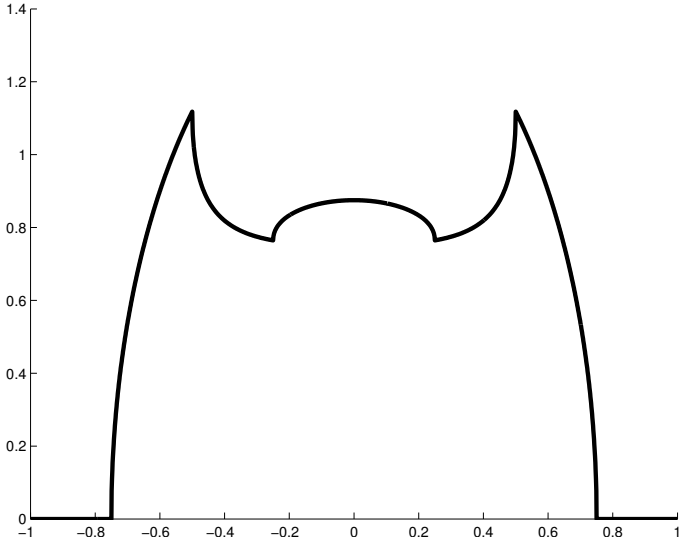
$$\begin{aligned} \mathcal{R}f(t,\theta) &= \int_{\ell_{t,\theta}} f(\|\mathbf{x}\|_2)\,\mathrm{d}\mathbf{x} = \int_{\ell_{t,0}} f(\|Q_\theta \mathbf{x}\|_2)\,\mathrm{d}\mathbf{x} = \int_{\ell_{t,0}} f(\|\mathbf{x}\|_2)\,\mathrm{d}\mathbf{x} \\ &= \mathcal{R}f(t,0) \end{aligned}$$

by application of the variable transform with the rotation matrix Q_θ in (9.6). This observation is consistent with our Examples 9.6 and 9.7. □

Now we construct another simple example from elementary functions. In medical image reconstruction, the term *phantom* is often used to denote test images whose Radon transforms can be computed analytically. The phantom *bull's eye* is only one such example for a popular test case.



(a) The phantom *bull's eye*



(b) The Radon transform of *bull's eye*

Fig. 9.3. *Bull's eye* and its Radon transform (see Example 9.9).

Example 9.9. The phantom *bull's eye* is given by the linear combination

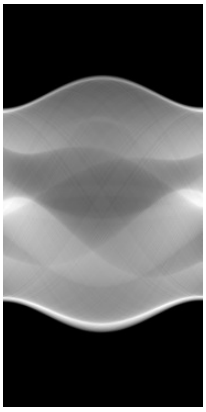
$$f(x,y) = \chi_{\mathcal{B}_{3/4}}(x,y) - \frac{3}{4}\chi_{\mathcal{B}_{1/2}}(x,y) + \frac{1}{4}\chi_{\mathcal{B}_{1/4}}(x,y) \tag{9.12}$$

of three indicator functions χ_{B_r} of the disks $\mathcal{B}_r$, for $r = 3/4, 1/2, 1/4$. To compute $\mathcal{R}f$, we apply the linearity of operator $\mathcal{R}$, whereby

$$\mathcal{R}f(t,\theta) = (\mathcal{R}\chi_{\mathcal{B}_{3/4}})(t,\theta) - \frac{3}{4}(\mathcal{R}\chi_{\mathcal{B}_{1/2}})(t,\theta) + \frac{1}{4}(\mathcal{R}\chi_{\mathcal{B}_{1/4}})(t,\theta). \tag{9.13}$$

Due to the radial symmetry of f (or, of χ_{B_r}), the Radon transform $\mathcal{R}f(t,\theta)$ does depend on t, but not on θ (cf. Remark 9.8). Now we can use the result of Example 9.6 to represent the Radon transform $\mathcal{R}f$ in (9.13) by linear combination of the Radon transforms $\mathcal{R}\chi_{\mathcal{B}_r}$, for $r = 3/4, 1/2, 1/4$. The phantom f and its Radon transform $\mathcal{R}f$ are shown in Figure 9.3. ♢

(a) Shepp-Logan-Phantom f



(b) Radon transform $\mathcal{R}f$

Fig. 9.4. The Shepp-Logan phantom and its sinogram.

For further illustration, we finally consider the *Shepp-Logan phantom* [66], a popular test case from medical imaging. The Shepp-Logan phantom f is a superposition of ten different ellipses to sketch a cross section of the human brain, see Figure 9.4 (a). In fact, the Shepp-Logan phantom is a very popular test case for numerical simulations, where the Radon transform $\mathcal{R}f$ of f can be computed analytically. Figure 9.4 (b) shows the Radon transform $\mathcal{R}f$ displayed in the *rectangular* coordinate system $\mathbb{R} \times [0, \pi)$. Such a representation of $\mathcal{R}f$ is called *sinogram*. In computerized tomography, the Shepp-Logan phantom (along with other popular test cases) is often used to evaluate the performance of numerical algorithms to reconstruct f from $\mathcal{R}f$.

9.2 The Filtered Back Projection

Now we turn to the inversion of the Radon transform, i.e., we wish to solve Problem 9.5. To this end, we first note some preliminary observations. Suppose we wish to reconstruct $f \equiv f(x, y)$ from given Radon data (9.10) only at one point (x, y). In this case, only those values of the line integrals (9.9) are relevant, whose *Radon lines* $\ell_{t,\theta}$ contain the point (x, y). Indeed, for all other straight lines $\ell_{t,\theta}$, which do *not* contain (x, y), the value $f(x, y)$ does *not* take influence on the line integral $\mathcal{R}f(t, \theta)$.

For this reason, we first wish to find out which Radon lines $\ell_{t,\theta}$ do contain the point (x, y). For any fixed angle $\theta \in [0, \pi)$, we can immediately work this out by using the relation (9.5). In fact, in this case, we necessarily require

$$t = x\cos(\theta) + y\sin(\theta),$$

see (9.7), and so this condition on t is also sufficient. Therefore, only the straight lines

$$\ell_{x\cos(\theta)+y\sin(\theta),\theta} \qquad \text{for } \theta \in [0, \pi)$$

contain the point (x, y). This observation leads us to the following definition for the back projection operator.

Definition 9.10. *For $h \in \mathrm{L}^1(\mathbb{R} \times [0, \pi))$, the function*

$$\mathcal{B}h(x, y) = \frac{1}{\pi}\int_0^\pi h(x\cos(\theta) + y\sin(\theta), \theta)\,\mathrm{d}\theta \qquad \text{for } (x, y) \in \mathbb{R}^2$$

is called the `back projection` *of h.* ◯

Remark 9.11. The back projection is a linear integral transform which maps a bivariate function $h \equiv h(t, \theta)$ in polar coordinates (t, θ) to a bivariate function $\mathcal{B}h(x, y)$ in Cartesian coordinates (x, y).

Moreover, the back projection $\mathcal{B}$ is (up to a positive factor) the *adjoint* operator of the Radon transform $\mathcal{R}f$. For more details on this, we refer to Exercise 9.39. □

Remark 9.12. The back projection $\mathcal{B}$ is *not* the inverse of the Radon transform $\mathcal{R}$. To see this, we make a simple counterexample. We consider the indicator function $f := \chi_{\mathcal{B}_1} \in \mathrm{L}^1(\mathbb{R}^2)$ of the unit ball $\mathcal{B}_1 = \{\mathbf{x} \in \mathbb{R}^2 \,|\, \|\mathbf{x}\|_2 \le 1\}$, whose (non-negative) Radon transform

$$\mathcal{R}f(t, \theta) = \begin{cases} 2\sqrt{1 - t^2} & \text{for } |t| \le 1, \\ 0 & \text{for } |t| > 1 \end{cases}$$

is computed in Example 9.6. Now we evaluate the back projection $\mathcal{B}(\mathcal{R}f)$ of $\mathcal{R}f$ at $(1 + \varepsilon, 0)$. For $\varepsilon \in (0, \sqrt{2} - 1)$, we have $1 + \varepsilon \in (1, \sqrt{2})$ and, moreover, $|(1 + \varepsilon)\cos(\theta)| < 1$ for $\theta \in [\pi/4, 3\pi/4]$. Therefore, we obtain

$$
\begin{aligned}
(\mathcal{B}(\mathcal{R}f))(1+\varepsilon,0) &= \frac{1}{\pi}\int_0^{\pi} \mathcal{R}f((1+\varepsilon)\cos(\theta),\theta)\,\mathrm{d}\theta \\
&\geq \frac{1}{\pi}\int_{\pi/4}^{3\pi/4} \mathcal{R}f((1+\varepsilon)\cos(\theta),\theta)\,\mathrm{d}\theta \\
&= \frac{2}{\pi}\int_{\pi/4}^{3\pi/4} \sqrt{1-(1+\varepsilon)^2\cos^2(\theta)}\,\mathrm{d}\theta > 0,
\end{aligned}
$$

i.e., we have $(\mathcal{B}(\mathcal{R}f))(1+\varepsilon,0) > 0$ for all $\varepsilon \in (0, \sqrt{2}-1)$.

Likewise, by the radial symmetry of f, we get for $\varphi \in (0, 2\pi)$

$$
(\mathcal{B}(\mathcal{R}f))((1+\varepsilon)\cos(\varphi),(1+\varepsilon)\sin(\varphi)) > 0 \qquad \text{for all } \varepsilon \in (0,\sqrt{2}-1),
$$

see Exercise 9.37, i.e., $\mathcal{B}(\mathcal{R}f)$ is positive on the open annulus

$$
\mathcal{R}_1^{\sqrt{2}} = \left\{ \mathbf{x} \in \mathbb{R}^2 \,\middle|\, 1 < \|\mathbf{x}\|_2 < \sqrt{2} \right\} \subset \mathbb{R}^2.
$$

However, we have $f \equiv 0$ on $\mathcal{R}_1^{\sqrt{2}}$, so that f is not reconstructed by the back projection $\mathcal{B}(\mathcal{R}f)$ of its Radon transform $\mathcal{R}f$, i.e., $f \neq \mathcal{B}(\mathcal{R}f)$. □

Figure 9.5 shows another counterexample by graphical illustration: In this case, the back projection $\mathcal{B}$ is applied to the Radon transform $\mathcal{R}f$ of the Shepp-Logan phantom f (cf. Figure 9.4). Observe that the *sharp* edges of phantom f are *blurred* by the back projection $\mathcal{B}$. In relevant applications, in particular for clinical diagnostics, such *smoothing effects* are clearly undesired. In the following discussion, we show how we can avoid such undesired effects by the application of *filters*.

(a) Shepp-Logan phantom f (b) back projection $\mathcal{B}(\mathcal{R}f)$.

Fig. 9.5. The Shepp-Logan phantom f and its back projection $\mathcal{B}(\mathcal{R}f)$.

Now we turn to the inversion of the Radon transform. To this end, we work with the *continuous* Fourier transform $\mathcal{F}$, which we apply to bivariate functions $f \equiv f(x,y)$ in Cartesian coordinates as usual, i.e., by using the *bivariate* Fourier transform $\mathcal{F} \equiv \mathcal{F}_2$. But for functions $h \equiv h(t,\theta)$ in polar coordinates we apply the *univariate* Fourier transform $\mathcal{F} \equiv \mathcal{F}_1$ to variable $t \in \mathbb{R}$, i.e., we keep the angle $\theta \in [0,\pi)$ fixed.

Definition 9.13. *For a function $f \equiv f(x,y) \in \mathrm{L}^1(\mathbb{R}^2)$ in Cartesian coordinates the* `Fourier transform` *$\mathcal{F}_2 f$ of f is defined as*

$$(\mathcal{F}_2 f)(X,Y) = \int_{\mathbb{R}^2} f(x,y)\mathrm{e}^{-\mathrm{i}(xX+yY)}\,\mathrm{d}(x,y).$$

For a function $h \equiv h(t,\theta)$ in polar coordinates satisfying $h(\cdot,\theta) \in \mathrm{L}^1(\mathbb{R})$, for all $\theta \in [0,\pi)$, the univariate `Fourier transform` *$\mathcal{F}_1 h$ of h is defined as*

$$(\mathcal{F}_1 h)(S,\theta) = \int_{\mathbb{R}} h(t,\theta)\mathrm{e}^{-\mathrm{i}St}\,\mathrm{d}t \qquad \text{for } \theta \in [0,\pi).$$

○

The following result will lead us directly to the inversion of the Radon transform. In fact, the *Fourier slice theorem* (also often referred to as *central slice theorem*) is an important result in Fourier analysis.

Theorem 9.14. (Fourier slice theorem). *For $f \in \mathrm{L}^1(\mathbb{R}^2)$, we have*

$$\mathcal{F}_2 f(S\cos(\theta), S\sin(\theta)) = \mathcal{F}_1(\mathcal{R}f)(S,\theta) \quad \text{for all } S \in \mathbb{R}, \theta \in [0,\pi). \tag{9.14}$$

Proof. For $f \equiv f(x,y) \in \mathrm{L}^1(\mathbb{R}^2)$, we consider the Fourier transform

$$\mathcal{F}_2 f(S\cos(\theta), S\sin(\theta)) = \int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)\mathrm{e}^{-iS(x\cos(\theta)+y\sin(\theta))}\,\mathrm{d}x\,\mathrm{d}y \tag{9.15}$$

at $(S,\theta) \in \mathbb{R} \times [0,\pi)$. By the variable transformation (9.6), the right hand side in (9.15) can be represented as

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta))\mathrm{e}^{-\mathrm{i}St}\,\mathrm{d}s\,\mathrm{d}t,$$

or, as

$$\int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(t\cos(\theta) - s\sin(\theta), t\sin(\theta) + s\cos(\theta))\,\mathrm{d}s \right) \mathrm{e}^{-\mathrm{i}St}\,\mathrm{d}t.$$

Note that the inner integral coincides with the Radon transform $\mathcal{R}f(t,\theta)$. But this already implies the stated identity

$$\mathcal{F}_2 f(S\cos(\theta), S\sin(\theta)) = \int_{\mathbb{R}} \mathcal{R}f(t,\theta)\mathrm{e}^{-\mathrm{i}St}\,\mathrm{d}t = \mathcal{F}_1(\mathcal{R}f)(S,\theta).$$

■

Theorem 9.15. (Filtered back projection formula).
For $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathscr{C}(\mathbb{R}^2)$, *the* `filtered back projection formula`

$$f(x,y) = \frac{1}{2}\mathcal{B}\left(\mathcal{F}_1^{-1}\left[|S|\mathcal{F}_1(\mathcal{R}f)(S,\theta)\right]\right)(x,y) \qquad \textit{for all } (x,y) \in \mathbb{R}^2 \tag{9.16}$$

holds.

Proof. For $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathscr{C}(\mathbb{R}^2)$, we have the Fourier inversion formula

$$f(x,y) = \mathcal{F}_2^{-1}(\mathcal{F}_2 f)(x,y) = \frac{1}{4\pi^2}\int_{\mathbb{R}}\int_{\mathbb{R}} (\mathcal{F}_2 f)(X,Y)\mathrm{e}^{\mathrm{i}(xX+yY)}\,\mathrm{d}X\,\mathrm{d}Y.$$

Changing variables from Cartesian coordinates (X,Y) to polar coordinates,

$$(X,Y) = (S\cos(\theta), S\sin(\theta)) \qquad \text{for } S \in \mathbb{R} \text{ and } \theta \in [0,\pi),$$

and by $\mathrm{d}X\,\mathrm{d}Y = |S|\,\mathrm{d}S\,\mathrm{d}\theta$ we get the representation

$$f(x,y) = \frac{1}{4\pi^2}\int_0^\pi \int_{\mathbb{R}} \mathcal{F}_2 f(S\cos(\theta), S\sin(\theta))\mathrm{e}^{\mathrm{i}S(x\cos(\theta)+y\sin(\theta))}|S|\,\mathrm{d}S\,\mathrm{d}\theta.$$

By the representation (9.14) in the Fourier slice theorem, this yields

$$\begin{aligned} f(x,y) &= \frac{1}{4\pi^2}\int_0^\pi \int_{\mathbb{R}} \mathcal{F}_1(\mathcal{R}f)(S,\theta)\mathrm{e}^{\mathrm{i}S(x\cos(\theta)+y\sin(\theta))}|S|\,\mathrm{d}S\,\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_0^\pi \mathcal{F}_1^{-1}\left[|S|\mathcal{F}_1(\mathcal{R}f)(S,\theta)\right](x\cos(\theta)+y\sin(\theta))\,\mathrm{d}\theta \\ &= \frac{1}{2}\mathcal{B}\left(\mathcal{F}_1^{-1}\left[|S|\mathcal{F}_1(\mathcal{R}f)(S,\theta)\right]\right)(x,y). \end{aligned}$$

■

Therefore, the reconstruction problem, Problem 9.5, is solved *analytically*. But the application of the formula (9.16) leads to critical *numerical* problems.

Remark 9.16. The filtered back projection (FBP) formula (9.16) is numerically unstable. We can explain this as follows. In the FBP formula (9.16) the Fourier transform $\mathcal{F}_1(\mathcal{R}f)$ of the Radon transform $\mathcal{R}f$ is being multiplied by the factor $|S|$. According to the jargon of signal processing, we say that $\mathcal{F}_1(\mathcal{R}f)$ is being *filtered* by $|S|$, which, on this occasion, explains the naming *filtered* back projection. Now the multiplication by the filter $|S|$ in (9.16) is very critical for *high* frequencies S, i.e., for S with large magnitude $|S|$. In fact, by the FBP formula (9.16) the high-frequency components in $\mathcal{R}f$ are amplified by the factor $|S|$. This is particularly critical for *noisy* Radon data, since the high-frequency noise level of the recorded signals $\mathcal{R}f$ is in this case exaggerated by application of the filter $|S|$.

Conclusion: The filtered back projection formula (9.16) is highly sensitive with respect to perturbations of the Radon data $\mathcal{R}f$ by noise. For this reason, the FBP formula (9.16) is entirely useless for practical purposes. □

9.3 Construction of Low-Pass Filters

To stabilize the filtered back projection, we replace the filter $|S|$ in the FPB formula (9.16) by a specific *low-pass filter*. In the general context of Fourier analysis, a low-pass filter is a function $F \equiv F(S)$ of the frequency variable S, which maps high-frequency parts of a signal to zero. To this end, we usually require compact support for F, so that $\operatorname{supp}(F) \subseteq [-L, L]$ for a fixed bandwidth $L > 0$, i.e., so that $F(S) = 0$ for all frequencies S with $|S| > L$.

In the particular context of the FBP formula (9.16), we require sufficient approximation quality for the low-pass filter F within the frequency band $[-L, L]$, i.e.,

$$F(S) \approx |S| \qquad \text{on } [-L, L].$$

To be more concrete on this, we explain our requirements for F as follows.

Definition 9.17. *Let $L > 0$. Moreover, suppose that $W \in \mathrm{L}^\infty(\mathbb{R})$ is an even function with compact support $\operatorname{supp}(W) \subseteq [-1, 1]$ satisfying $W(0) = 1$. A* `low-pass filter` *for the stabilization of (9.16) is a function $F : \mathbb{R} \longrightarrow \mathbb{R}$ of the form*

$$F(S) = |S| \cdot W(S/L) \qquad \text{for } S \in \mathbb{R},$$

where L denotes the `bandwidth` *and W is the* `window` *of $F \equiv F_{L,W}$.* ◯

Now let us make a few examples for commonly used low-pass filters. In the following discussion,

$$\sqcap_L(S) \equiv \chi_{[-L,L]}(S) = \begin{cases} 1 & \text{for } |S| \le L, \\ 0 & \text{for } |S| > L, \end{cases}$$

is, for $L > 0$, the indicator function of the interval $[-L, L]$, and we let $\sqcap := \sqcap_1$.

Example 9.18. The `Ram-Lak filter` F_{RL} is given by the window

$$W_{\mathrm{RL}}(S) = \sqcap(S),$$

so that

$$F_{\mathrm{RL}}(S) = |S| \cdot \sqcap_L(S) = \begin{cases} |S| & \text{for } |S| \le L, \\ 0 & \text{for } |S| > L. \end{cases}$$

The Ram-Lak filter is shown in Figure 9.6 (a). ♢

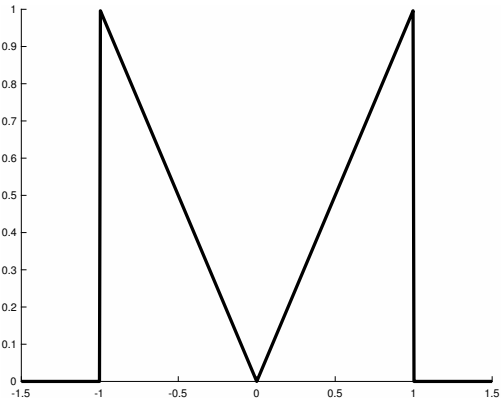
Example 9.19. The `Shepp-Logan filter` F_{SL} is given by the window

$$W_{\mathrm{SL}}(S) = \operatorname{sinc}(\pi S/2) \cdot \sqcap(S),$$

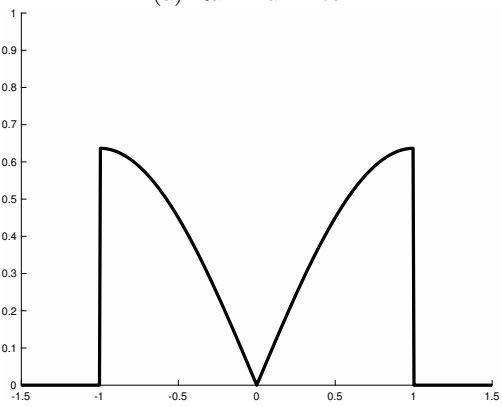
so that

$$F_{\mathrm{SL}}(S) = |S| \cdot \frac{\sin(\pi S/(2L))}{\pi S/(2L)} \cdot \sqcap_L(S) = \begin{cases} \frac{2L}{\pi} \cdot |\sin(\pi S/(2L))| & \text{for } |S| \le L, \\ 0 & \text{for } |S| > L. \end{cases}$$

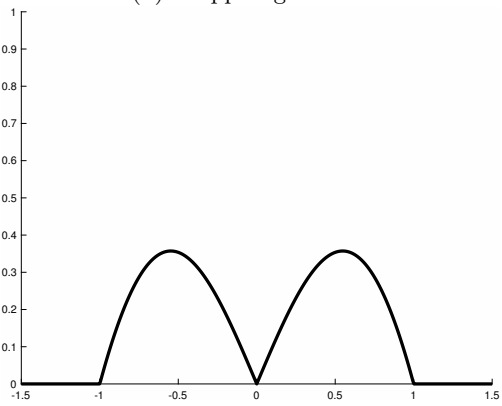
The Shepp-Logan filter is shown in Figure 9.6 (b). ♢



(a) Ram-Lak filter

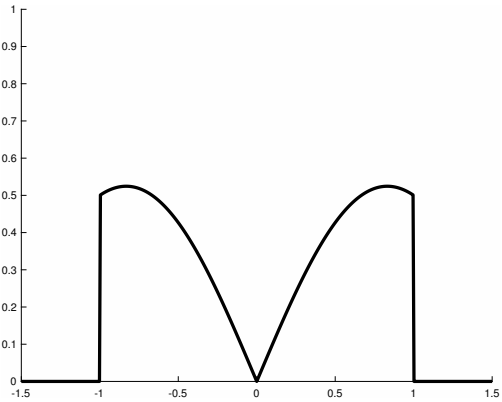


(b) Shepp-Logan filter

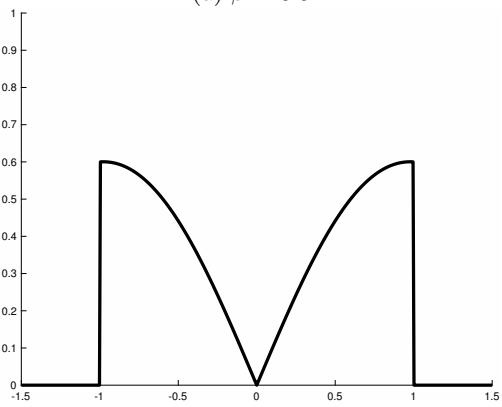


(c) cosine filter

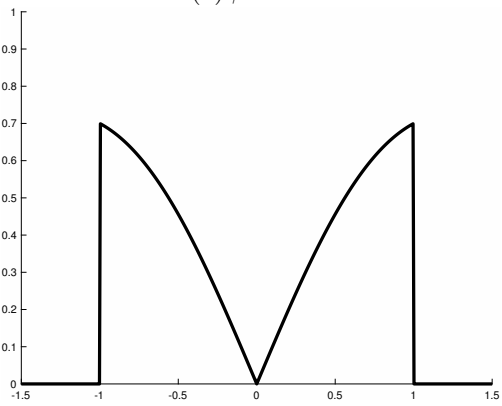
Fig. 9.6. Three commonly used low-pass filters (see Examples 9.18-9.20).



(a) $\beta = 0.5$

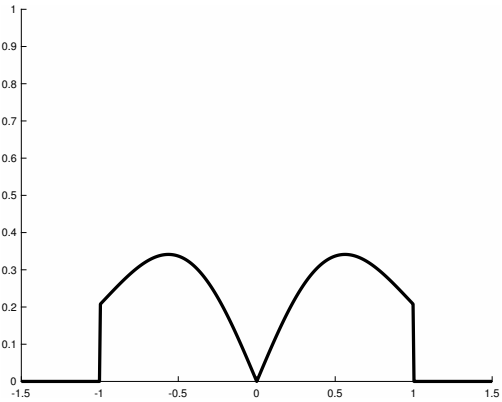


(b) $\beta = 0.6$

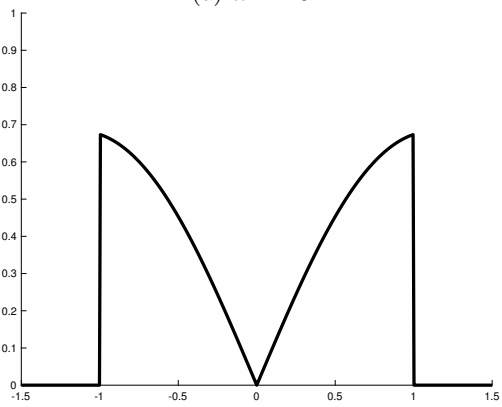


(c) $\beta = 0.7$

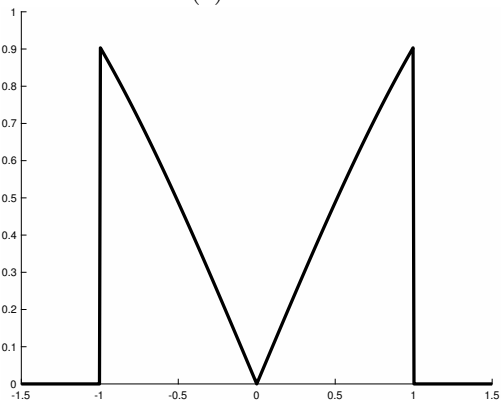
Fig. 9.7. The Hamming filter F_β for $\beta \in \{0.5, 0.6, 0.7\}$ (see Example 9.21).



(a) $\alpha = 2.5$



(b) $\alpha = 5.0$



(c) $\alpha = 10.0$

Fig. 9.8. The Gauss filter F_α for $\alpha \in \{2.5, 5.0, 10.0\}$ (see Example 9.22).

Example 9.20. The `cosine filter` F_{CF} is given by the window

$$W_{\mathrm{CF}}(S) = \cos(\pi S/2) \cdot \sqcap(S),$$

so that

$$F_{\mathrm{CF}}(S) = |S| \cdot \cos(\pi S/(2L)) \cdot \sqcap_L(S) = \begin{cases} |S| \cdot \cos(\pi S/(2L)) & \text{for } |S| \le L, \\ 0 & \text{for } |S| > L. \end{cases}$$

The cosine filter is shown in Figure 9.6 (c). ♢

Example 9.21. The `Hamming filter` F_β is given by the window

$$W_\beta(S) = (\beta + (1-\beta)\cos(\pi S)) \cdot \sqcap(S) \qquad \text{for } \beta \in [1/2, 1].$$

Note that the Hamming filter F_β is a combination of the Ram-Lak filter F_{RL} and the cosine filter F_{CF}. The Hamming filter F_β is shown in Figure 9.7, for $\beta \in \{0.5, 0.6, 0.7\}$. ♢

Example 9.22. The `Gauss filter` F_α is given by the window

$$W_\alpha(S) = \exp\left(-(\pi S/\alpha)^2\right) \cdot \sqcap(S) \qquad \text{for } \alpha > 1.$$

The Gauss filter F_α is shown in Figure 9.8, for $\alpha \in \{2.5, 5.0, 10.0\}$. ♢

If we replace the filter $|S|$ in (9.16) by a low-pass filter $F \equiv F(S)$, then the resulting reconstruction of f,

$$f_F(x,y) := \frac{1}{2}\mathcal{B}\left(\mathcal{F}_1^{-1}\left[F(S) \cdot \mathcal{F}_1(\mathcal{R}f)(S,\theta)\right]\right)(x,y), \tag{9.17}$$

will no longer be *exact*, i.e., the function f_F in (9.17) yields an *approximate* reconstruction of f, $f \approx f_F$. We analyze the approximation behaviour of f_F later in this chapter. But we develop a suitable representation for f_F in (9.17) already now.

Note that any low-pass filter F is absolutely integrable, i.e., $F \in \mathrm{L}^1(\mathbb{R})$. In particular, any low-pass filter F has, in contrast to the filter $|S|$, an inverse Fourier transform $\mathcal{F}_1^{-1}F$. We use this observation to simplify the representation of f_F in (9.17) by

$$f_F(x,y) = \frac{1}{2}\mathcal{B}\left((\mathcal{F}_1^{-1}F * \mathcal{R}f)(S,\theta)\right)(x,y). \tag{9.18}$$

To further simplify the representation in (9.18), we first prove a useful relation between $\mathcal{R}$ and $\mathcal{B}$. This relation involves the convolution product $*$ that we apply to bivariate functions in Cartesian coordinates and to bivariate functions in polar coordinates, according to the following definition.

Definition 9.23. *For $f \equiv f(x,y) \in \mathrm{L}^1(\mathbb{R}^2)$ and $g \equiv g(x,y) \in \mathrm{L}^1(\mathbb{R}^2)$, the* `convolution` *$f * g$ between the functions f and g is defined as*

$$(f * g)(X,Y) = \int_{\mathbb{R}} \int_{\mathbb{R}} f(X-x, Y-y) g(x,y) \,\mathrm{d}x \,\mathrm{d}y \qquad \text{for } X, Y \in \mathbb{R}.$$

For $\theta \in [0,\pi)$ and functions $g(\cdot,\theta), h(\cdot,\theta) \in \mathrm{L}^1(\mathbb{R})$, the `convolution` *$g * h$ between g and h is defined as*

$$(g * h)(T,\theta) = \int_{\mathbb{R}} g(T-t,\theta) h(t,\theta) \,\mathrm{d}t \qquad \text{for } T \in \mathbb{R}.$$

○

Theorem 9.24. *For $h \in \mathrm{L}^1(\mathbb{R} \times [0,\pi))$ and $f \in \mathrm{L}^1(\mathbb{R}^2)$, we have the relation*

$$\mathcal{B}\left(h * \mathcal{R}f\right)(X,Y) = \left(\mathcal{B}h * f\right)(X,Y) \qquad \text{for all } (X,Y) \in \mathbb{R}^2. \tag{9.19}$$

Proof. For the right hand side in (9.19), we obtain the representation

$$\begin{aligned}
&\left(\mathcal{B}h * f\right)(X,Y) \\
&= \int_{\mathbb{R}} \int_{\mathbb{R}} (\mathcal{B}h)(X-x, Y-y) f(x,y) \,\mathrm{d}x \,\mathrm{d}y \\
&= \frac{1}{\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \left[\int_0^{\pi} h((X-x)\cos(\theta) + (Y-y)\sin(\theta), \theta) \,\mathrm{d}\theta \right] f(x,y) \,\mathrm{d}x \,\mathrm{d}y.
\end{aligned}$$

By variable transformation on (x,y) by (9.6) and $\mathrm{d}x \,\mathrm{d}y = \mathrm{d}s \,\mathrm{d}t$, we obtain

$$\begin{aligned}
\left(\mathcal{B}h * f\right)(X,Y) &= \frac{1}{\pi} \int_0^{\pi} \int_{\mathbb{R}} h(X\cos(\theta) + Y\sin(\theta) - t, \theta)(\mathcal{R}f)(t,\theta) \,\mathrm{d}t \,\mathrm{d}\theta \\
&= \frac{1}{\pi} \int_0^{\pi} (h * \mathcal{R}f)(X\cos(\theta) + Y\sin(\theta), \theta) \,\mathrm{d}\theta \\
&= \mathcal{B}(h * \mathcal{R}f)(X,Y)
\end{aligned}$$

for all $(X,Y) \in \mathbb{R}^2$. ■

Theorem 9.24 and (9.18) provide a very useful representation for f_F, where we use the inverse Fourier transform $\mathcal{F}_1^{-1}F$ of the filter F by

$$(\mathcal{F}_1^{-1}F)(t,\theta) := (\mathcal{F}^{-1}F)(t) \qquad \text{for } t \in \mathbb{R} \text{ and } \theta \in [0,\pi)$$

as a bivariate function.

Corollary 9.25. *Let $f \in \mathrm{L}^1(\mathbb{R}^2)$. Moreover, let F be a filter satisfying $\mathcal{F}_1^{-1}F \in \mathrm{L}^1(\mathbb{R} \times [0,\pi))$. Then, the representation*

$$f_F(x,y) = \frac{1}{2}\left(\mathcal{B}(\mathcal{F}_1^{-1}F) * f\right)(x,y) = \left(K_F * f\right)(x,y), \tag{9.20}$$

holds, where

$$K_F(x,y) := \frac{1}{2}\mathcal{B}\left(\mathcal{F}_1^{-1}F\right)(x,y)$$

denotes the convolution kernel *of the low-pass filter F.* ■

Remark 9.26. The statement of Corollary 9.25 does also hold without the assumption $\mathcal{F}_1^{-1}F \in \mathrm{L}^1(\mathbb{R}\times[0,\pi))$, see [5]. Therefore, in Section 9.4, we apply Corollary 9.25 without any assumptions on the low-pass filter F. □

9.4 Error Estimates and Convergence Rates

To evaluate the quality of low-pass filters F, we analyze the *intrinsic* L^2-error

$$\|f - f_F\|_{\mathrm{L}^2(\mathbb{R}^2)} \tag{9.21}$$

that is incurred by the utilization of F. To this end, we consider for $\alpha > 0$ the `Sobolev`[4] `space`

$$\mathrm{H}^\alpha(\mathbb{R}^2) = \left\{ g \in \mathrm{L}^2(\mathbb{R}^2) \,|\, \|g\|_\alpha < \infty \right\} \subset \mathrm{L}^2(\mathbb{R}^2),$$

equipped with the norm $\|\cdot\|_\alpha$, where

$$\|g\|_\alpha^2 = \frac{1}{4\pi^2} \int_{\mathbb{R}} \int_{\mathbb{R}} \left(1 + x^2 + y^2\right)^\alpha |\mathcal{F}g(x,y)|^2 \,\mathrm{d}x\,\mathrm{d}y \qquad \text{for } g \in \mathrm{H}^\alpha(\mathbb{R}^2),$$

where we apply the bivariate Fourier transform, i.e., $\mathcal{F} = \mathcal{F}_2$.

We remark that estimates for the L^2-error in (9.21) and for the L^p-error were proven by Madych [44] in 1990. Moreover, pointwise error estimates and L^∞-estimates were studied by Munshi et al. [52, 53, 54] in 1991-1993. However, we do not further pursue their techniques here. Instead of this, we work with a more recent account by Beckmann [4, 5]. The following result of Beckmann [5] leads us, without any detour, to useful L^2-error estimates under rather weak assumptions on f and F in (9.21).

Theorem 9.27. *For $\alpha > 0$, let $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathrm{H}^\alpha(\mathbb{R}^2)$ and $W \in \mathrm{L}^\infty(\mathbb{R})$. Then, we have for the L^2-error (9.21) of the reconstruction $f_F = f * K_F$ in (9.20) the estimate*

$$\|f - f_F\|_{\mathrm{L}^2(\mathbb{R}^2)} \le \left(\Phi_{\alpha,W}^{1/2}(L) + L^{-\alpha} \right) \|f\|_\alpha, \tag{9.22}$$

where

$$\Phi_{\alpha,W}(L) := \sup_{S\in[-1,1]} \frac{(1 - W(S))^2}{(1 + L^2S^2)^\alpha} \quad \text{for } L > 0. \tag{9.23}$$

Proof. By $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathrm{H}^\alpha(\mathbb{R}^2)$, for $\alpha > 0$, we have $f \in \mathrm{L}^2(\mathbb{R}^2)$. Moreover, we have $f_F \in \mathrm{L}^2(\mathbb{R}^2)$, as shown in [5]. By application of the Fourier convolution theorem on $\mathrm{L}^2(\mathbb{R}^2)$, Theorem 7.43, in combination with the Plancherel theorem, Theorem 7.45, we get the representation

[4] Sergei Lvovich Sobolev (1908-1989), Russian mathematician

$$\begin{aligned}\|f - f * K_F\|^2_{\mathrm{L}^2(\mathbb{R}^2)} &= \frac{1}{4\pi^2}\,\|\mathcal{F}f - \mathcal{F}f \cdot \mathcal{F}K_F\|^2_{\mathrm{L}^2(\mathbb{R}^2;\mathbb{C})} \\ &= \frac{1}{4\pi^2}\,\|\mathcal{F}f - W_L \cdot \mathcal{F}f\|^2_{\mathrm{L}^2(\mathbb{R}^2;\mathbb{C})}, \end{aligned} \tag{9.24}$$

where for the scaled window $W_L(S) := W(S/L)$, $S \in \mathbb{R}$, we used the identity

$$W_L(\|(x,y)\|_2) = \mathcal{F}K_F(x,y) \qquad \text{for almost every } (x,y) \in \mathbb{R}^2 \tag{9.25}$$

(see Exercise 9.44). Since $\operatorname{supp}(W_L) \subset [-L, L]$, we can split the square error in (9.24) into a sum of two integrals,

$$\begin{aligned}&\frac{1}{4\pi^2}\,\|\mathcal{F}f - W_L \cdot \mathcal{F}f\|^2_{\mathrm{L}^2(\mathbb{R}^2;\mathbb{C})} \\ &= \frac{1}{4\pi^2}\int_{\|(x,y)\|_2 \le L} |(\mathcal{F}f - W_L \cdot \mathcal{F}f)(x,y)|^2\,\mathrm{d}(x,y) \qquad (9.26) \\ &\quad + \frac{1}{4\pi^2}\int_{\|(x,y)\|_2 > L} |\mathcal{F}f(x,y)|^2\,\mathrm{d}(x,y). \qquad (9.27)\end{aligned}$$

By $f \in \mathrm{H}^\alpha(\mathbb{R}^2)$, we estimate the integral in (9.27) from above by

$$\begin{aligned}&\frac{1}{4\pi^2}\int_{\|(x,y)\|_2 > L} |\mathcal{F}f(x,y)|^2\,\mathrm{d}(x,y) \\ &\le \frac{1}{4\pi^2}\int_{\|(x,y)\|_2 > L} \left(1 + x^2 + y^2\right)^\alpha L^{-2\alpha}\,|\mathcal{F}f(x,y)|^2\,\mathrm{d}(x,y) \\ &\le L^{-2\alpha}\,\|f\|^2_\alpha, \end{aligned} \tag{9.28}$$

whereas for the integral in (9.26), we obtain the estimate

$$\begin{aligned}&\frac{1}{4\pi^2}\int_{\|(x,y)\|_2 \le L} |(\mathcal{F}f - W_L \cdot \mathcal{F}f)(x,y)|^2\,\mathrm{d}(x,y) \\ &= \frac{1}{4\pi^2}\int_{\|(x,y)\|_2 \le L} \frac{|1 - W_L(\|(x,y)\|_2)|^2}{(1 + x^2 + y^2)^\alpha}\,\left(1 + x^2 + y^2\right)^\alpha\,|\mathcal{F}f(x,y)|^2\,\mathrm{d}(x,y) \\ &\le \left(\sup_{S \in [-L,L]} \frac{(1 - W_L(S))^2}{(1 + S^2)^\alpha}\right) \frac{1}{4\pi^2}\int_{\mathbb{R}}\int_{\mathbb{R}} \left(1 + x^2 + y^2\right)^\alpha\,|\mathcal{F}f(x,y)|^2\,\mathrm{d}x\,\mathrm{d}y \\ &= \left(\sup_{S \in [-1,1]} \frac{(1 - W(S))^2}{(1 + L^2S^2)^\alpha}\right) \|f\|^2_\alpha \\ &= \Phi_{\alpha,W}(L) \cdot \|f\|^2_\alpha. \end{aligned} \tag{9.29}$$

Finally, the sum of the two upper bounds in (9.29) and in (9.28) yields the stated error estimate in (9.22). ■

Remark 9.28. For the Ram-Lak filter from Example 9.18, we have $W \equiv 1$ on $[-1,1]$, and so $\Phi_{\alpha,W} \equiv 0$. In this case, Theorem 9.27 yields the error estimate

$$\|f - f_F\|_{\mathrm{L}^2(\mathbb{R}^2)} \leq L^{-\alpha}\|f\|_\alpha = \mathcal{O}\left(L^{-\alpha}\right) \quad \text{for } L \to \infty.$$

This further implies L^2-convergence, i.e., $f_F \longrightarrow f$ for $L \to \infty$, for the reconstruction method f_F at convergence rate α. □

In our subsequent analysis concerning arbitrary low-pass filters F, we use the following result from [4] to prove L^2-convergence $f_F \longrightarrow f$, for $L \to \infty$, i.e.,

$$\|f - f_F\|_{\mathrm{L}^2(\mathbb{R}^2)} \longrightarrow 0 \quad \text{for } L \to \infty.$$

Theorem 9.29. *Let $W \in \mathscr{C}[-1,1]$ satisfy $W(0) = 1$. Then, we have, for any $\alpha > 0$, the convergence*

$$\Phi_{\alpha,W}(L) = \max_{S\in[0,1]} \frac{(1 - W(S))^2}{(1 + L^2S^2)^\alpha} \longrightarrow 0 \quad \textit{for } L \to \infty. \tag{9.30}$$

Proof. Let $S^*_{\alpha,W,L} \in [0,1]$ be the *smallest* maximum on $[0,1]$ for the function

$$\Phi_{\alpha,W,L}(S) := \frac{(1 - W(S))^2}{(1 + L^2S^2)^\alpha} \quad \text{for } S \in [0,1].$$

Case 1: Suppose $S^*_{\alpha,W,L}$ is *uniformly* bounded away from zero, i.e., we have $S^*_{\alpha,W,L} \geq c > 0$ for all $L > 0$, for some $c \equiv c_{\alpha,W} > 0$. Then,

$$0 \leq \Phi_{\alpha,W,L}\big(S^*_{\alpha,W,L}\big) = \frac{\big(1 - W(S^*_{\alpha,W,L})\big)^2}{\big(1 + L^2(S^*_{\alpha,W,L})^2\big)^\alpha} \leq \frac{\|1 - W\|^2_{\infty,[-1,1]}}{(1 + L^2c^2)^\alpha} \longrightarrow 0$$

holds for $L \to \infty$.

Case 2: Suppose $S^*_{\alpha,W,L} \longrightarrow 0$ for $L \to \infty$. Then, we have

$$0 \leq \Phi_{\alpha,W,L}\big(S^*_{\alpha,W,L}\big) = \frac{\big(1 - W(S^*_{\alpha,W,L})\big)^2}{\big(1 + L^2(S^*_{\alpha,W,L})^2\big)^\alpha} \leq \big(1 - W(S^*_{\alpha,W,L})\big)^2 \longrightarrow 0,$$

for $L \to \infty$, by the continuity of W and with $W(0) = 1$. ■

Now the convergence of the reconstruction method f_F follows directly from Theorems 9.27 and 9.29.

Corollary 9.30. *For $\alpha > 0$, let $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathrm{H}^\alpha(\mathbb{R}^2)$. Moreover, let W be a continuous window on $[0,1]$ satisfying $W(0) = 1$. Then, the convergence*

$$\|f - f_F\|_{\mathrm{L}^2(\mathbb{R}^2)} \longrightarrow 0 \quad \text{for } L \to \infty \tag{9.31}$$

holds. ■

We remark that the assumptions in Corollary 9.30,

$$W \in \mathscr{C}([0,1]) \qquad \text{and} \qquad W(0) = 1$$

are satisfied by all windows W of the low-pass filters F in Examples 9.18-9.22. A more detailed discussion on error estimates and convergence rates for FBP reconstruction methods f_F can be found in the work [4] of Beckmann.

9.5 Implementation of the Reconstruction Method

In this section, we explain how to implement the FBP reconstruction method efficiently. The starting point for our discussion is the representation (9.18), whereby, for a fixed low-pass filter F, the corresponding reconstruction f_F, is given as

$$f_F(x,y) = \frac{1}{2}\mathcal{B}\left((\mathcal{F}_1^{-1}F * \mathcal{R}f)(S,\theta)\right)(x,y). \tag{9.32}$$

Obviously, we can only acquire and process *finitely* many Radon data $\mathcal{R}f(t,\theta)$ in practice. In the acquisition of Radon data, the X-ray beams are usually generated, such that the resulting Radon lines $\ell_{t,\theta} \subset \Omega$ are distributed in the image domain $\Omega \subset \mathbb{R}^2$ on *regular* geometries.

9.5.1 Parallel Beam Geometry

A commonly used method of data acquisition is referred to as *parallel beam geometry*. In this method, the Radon lines $\ell_{t,\theta}$ are collected in subsets of parallel lines. We can explain this more precisely as follows. For a uniform discretization of the angular variable $\theta \in [0,\pi)$ with N distinct angles

$$\theta_k := k\pi/N \qquad \text{for } k = 0,\ldots,N-1$$

and for a fixed sampling rate $d > 0$ with

$$t_j := j \cdot d \qquad \text{for } j = -M,\ldots,M,$$

a constant number of $2M+1$ Radon data is recorded per angle $\theta_k \in [0,\pi)$, along the parallel Radon lines $\{\ell_{t_j,\theta_k} \,|\, j = -M,\ldots,M\}$. Therefore, the resulting discretization of the Radon transform consists of $N \times (2M+1)$ Radon data

$$\{\mathcal{R}f(t_j,\theta_k) \,|\, j = -M,\ldots,M \text{ and } k = 0,\ldots,N-1\}. \tag{9.33}$$

Figure 9.9 shows 110 Radon lines $\ell_{t_j,\theta_k} \cap [-1,1]^2$ on parallel beam geometry for $N = 10$ angles θ_k and $2M+1 = 11$ Radon lines per angle.

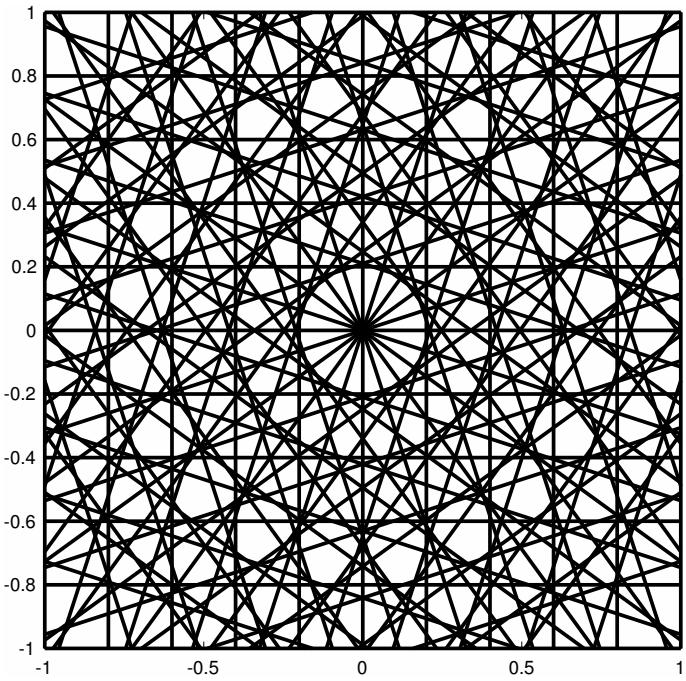


Fig. 9.9. *Parallel beam geometry.* Regular distribution of 110 Radon lines ℓ_{t_j,θ_k}, for $N = 10$ angles θ_k, $2M+1 = 11$ Radon lines per angle, at sampling rate $d = 0.2$.

9.5.2 Inverse Fourier transform of the low-pass filters

For our implementation of the FBP reconstruction method f_F in (9.32), we need the inverse Fourier transforms of the chosen low-pass filter F. Note that any low-pass filter F is, according to Definition 9.17, an even function. Therefore, the inverse Fourier transform $\mathcal{F}^{-1}F$ of F is an inverse cosine transform. This observation will simplify our following computations for $\mathcal{F}^{-1}F$.

We start with the Ram-Lak filter from Example 9.18.

Proposition 9.31. *The inverse Fourier transform of the Ram-Lak filter*

$$F_{\mathrm{RL}}(S) = |S| \cdot \sqcap(S) \qquad \text{for } S \in \mathbb{R}$$

is given by

$$\mathcal{F}^{-1}F_{\mathrm{RL}}(t) = \frac{1}{\pi}\left[\frac{(Lt)\cdot\sin(Lt)}{t^2} - \frac{2\cdot\sin^2(Lt/2)}{t^2}\right] \qquad \text{for } t \in \mathbb{R}. \tag{9.34}$$

The evaluation of $\mathcal{F}^{-1}F_{\mathrm{RL}}$ at $t_j = jd$, with sampling rate $d = \pi/L > 0$, yields

$$\mathcal{F}^{-1}F_{\mathrm{RL}}(\pi j/L) = \begin{cases} L^2/(2\pi) & \text{for } j = 0, \\ 0 & \text{for } j \neq 0 \text{ even}, \\ -2L^2/(\pi^3 \cdot j^2) & \text{for } j \text{ odd}. \end{cases} \tag{9.35}$$

Proof. The inverse Fourier transform $\mathcal{F}^{-1}F_{\mathrm{RL}}$ of the even function F_{RL} is given by the inverse cosine transform

$$\mathcal{F}^{-1}F_{\mathrm{RL}}(t) = \frac{1}{\pi}\int_0^L S \cdot \cos(tS)\,\mathrm{d}S.$$

Now we can compute the representation in (9.34) by elementary calculations,

$$\begin{aligned} \mathcal{F}^{-1}F_{\mathrm{RL}}(t) &= \frac{1}{\pi}\left[\frac{\cos(tS) + (tS)\cdot\sin(tS)}{t^2}\right]_{S=0}^{S=L} \\ &= \frac{1}{\pi}\left[\frac{\cos(Lt) + (Lt)\cdot\sin(Lt) - 1}{t^2}\right] \\ &= \frac{1}{\pi}\left[\frac{(Lt)\cdot\sin(Lt)}{t^2} - \frac{2\cdot\sin^2(Lt/2)}{t^2}\right], \end{aligned}$$

where we use the trigonometric identity $\cos(\theta) = 1 - 2\cdot\sin^2(\theta/2)$.

For the evaluation of $\mathcal{F}^{-1}F_{\mathrm{RL}}$ at $t = \pi j/L$, we obtain

$$\begin{aligned} \mathcal{F}^{-1}F_{\mathrm{RL}}(\pi j/L) &= \frac{1}{\pi}\left(\frac{(\pi j)\cdot\sin(\pi j)}{(\pi j/L)^2} - \frac{2\cdot\sin^2(\pi j/2)}{(\pi j/L)^2}\right) \\ &= \frac{L^2}{2\pi}\left(\frac{2\cdot\sin(\pi j)}{\pi j} - \left[\frac{\sin(\pi j/2)}{(\pi j/2)}\right]^2\right) \end{aligned}$$

and this already yields the stated representation in (9.35). ∎

Next, we consider the Shepp-Logan filter from Example 9.19.

Proposition 9.32. *The Shepp-Logan filter*

$$F_{\mathrm{SL}}(S) = \begin{cases} \frac{2L}{\pi}\cdot|\sin(\pi S/(2L))| & \text{for } |S| \leq L, \\ 0 & \text{for } |S| > L, \end{cases}$$

has the inverse Fourier transform

$$\mathcal{F}^{-1}F_{\mathrm{SL}}(t) = \frac{L}{\pi^2}\left(\frac{\cos(Lt - \pi/2) - 1}{t - \pi/(2L)} - \frac{\cos(Lt + \pi/2) - 1}{t + \pi/(2L)}\right) \quad \text{for } t \in \mathbb{R}.$$

The evaluation of $\mathcal{F}^{-1}F_{\mathrm{SL}}$ at $t_j = jd$, with sampling rate $d = \pi/L > 0$, yields

$$\mathcal{F}^{-1}F_{\mathrm{SL}}(\pi j/L) = \frac{4L^2}{\pi^3(1 - 4j^2)}. \tag{9.36}$$

Proof. We compute the inverse Fourier transform $\mathcal{F}^{-1}F_{\mathrm{SL}}$ by

$$\begin{aligned}\mathcal{F}^{-1}F_{\mathrm{SL}}(t) &= \frac{1}{\pi}\int_0^L \frac{2L}{\pi}\cdot\sin(\pi S/(2L))\cdot\cos(tS)\,\mathrm{d}S \\ &= \frac{L}{\pi^2}\left[\frac{\cos((t-\pi/(2L))S)}{t-\pi/(2L)} - \frac{\cos((t+\pi/(2L))S)}{t+\pi/(2L)}\right]_{S=0}^{S=L},\end{aligned}$$

where we used the trigonometric addition formula

$$2\sin\left(\frac{x-y}{2}\right)\cos\left(\frac{x+y}{2}\right) = \sin(x)-\sin(y)$$

for $x = (t+\pi/(2L))S$ and $y = (t-\pi/(2L))S$.

For the evaluation of $\mathcal{F}^{-1}F_{\mathrm{SL}}$ at $t = \pi j/L$, we obtain

$$\begin{aligned}\mathcal{F}^{-1}F_{\mathrm{SL}}(\pi j/L) &= \frac{L}{\pi^2}\left(\left[\frac{\cos(\pi j-\pi/2)}{\pi j/L-\pi/(2L)} - \frac{\cos(\pi j+\pi/2)}{\pi j/L+\pi/(2L)}\right]\right. \\ &\qquad \left. - \left[\frac{1}{\pi j/L-\pi/(2L)} - \frac{1}{\pi j/L+\pi/(2L)}\right]\right) \\ &= \frac{L}{\pi^2}\left(\frac{1}{\pi j/L+\pi/(2L)} - \frac{1}{\pi j/L-\pi/(2L)}\right) \\ &= \frac{4L^2}{\pi^3(1-4j^2)}.\end{aligned}$$

This completes our proof for the stated representation in (9.36). ∎

For the inverse Fourier transforms of the remaining filters F from Examples 9.20-9.22 we refer to Exercise 9.43.

9.5.3 Discretization of the Convolution

Next, we discretize the convolution operator $*$ in (9.32). Our purpose for doing so is to approximate, for any angle $\theta \equiv \theta_k \in [0,\pi)$, $k = 0,\dots,N-1$, the convolution product

$$(\mathcal{F}_1^{-1}F * \mathcal{R}f)(S,\theta) = \int_{\mathbb{R}} (\mathcal{F}_1^{-1}F)(S-t)\cdot\mathcal{R}f(t,\theta)\,\mathrm{d}t \tag{9.37}$$

between the functions

$$u(t) = \mathcal{F}_1^{-1}F(t) \quad \text{and} \quad v(t) = \mathcal{R}f(t,\theta) \qquad \text{for } t\in\mathbb{R}$$

from discrete data

$$u_j = \mathcal{F}_1^{-1}F(t_j) \quad \text{and} \quad v_j = \mathcal{R}f(t_j,\theta_k) \qquad \text{for } j\in\mathbb{Z},$$

which we acquire at $t_j = j \cdot d$ with sampling rate $d = \pi/L$. To this end, we replace the integral in (9.37), by application of the composite rectangle rule, with the (infinite) sum

$$(\mathcal{F}_1^{-1} F * \mathcal{R}f)(t_m, \theta_k) \approx \frac{\pi}{L} \sum_{j \in \mathbb{Z}} u_{m-j} \cdot v_j \qquad \text{for } m \in \mathbb{Z}, \tag{9.38}$$

i.e., we evaluate the convolution $u * v$ at $S = t_m = \pi m/L$, $m \in \mathbb{Z}$, numerically.

For the convergence of the sum in (9.38), we require $(u_j)_{j\in\mathbb{Z}}, (v_j)_{j\in\mathbb{Z}} \in \ell^1$. But in relevant application scenarios the situation is much easier. Indeed, we can assume compact support for the target attenuation-coefficient function f. In this case, the Radon transform $v = \mathcal{R}f(\cdot, \theta)$ also has compact support, for all $\theta \in [0, \pi)$, and so only *finitely* many Radon data in $(v_j)_{j\in\mathbb{Z}}$ do not vanish, so that the sum in (9.38) is *finite*.

According to our discussion concerning *parallel beam geometry*, we assume for the Radon data $\{\mathcal{R}f(t_j, \theta_k)\}_{j\in\mathbb{Z}}$, for any angle $\theta_k = k\pi/N \in [0, \pi)$, the form

$$v_j = \mathcal{R}f(t_j, \theta_k) \qquad \text{for } j = -M, \ldots, M.$$

But we choose $M \in \mathbb{N}$ large enough, so that $v_j = 0$ for all $|j| > M$. In this way, we can represent the series in (9.38) as a finite sum. This finally yields by

$$(\mathcal{F}_1^{-1} F * \mathcal{R}f)(t_m, \theta_k) \approx \frac{\pi}{L} \sum_{j=-M}^{M} u_{m-j} \cdot v_j \qquad \text{for } m \in \mathbb{Z} \tag{9.39}$$

a suitable discretization for the convolution product $(\mathcal{F}_1^{-1} F * \mathcal{R}f)$.

9.5.4 Discretization of the Back Projection

Finally, we turn to the discretization of the back projection. Recall that, according to Definition 9.10, the back projection of $h \in \mathrm{L}^1(\mathbb{R} \times [0, \pi))$ at (x, y) is defined as

$$\mathcal{B}h(x, y) = \frac{1}{\pi} \int_0^\pi h(x \cos(\theta) + y \sin(\theta), \theta)\, d\theta \qquad \text{for } (x, y) \in \mathbb{R}^2. \tag{9.40}$$

Further recall that for the FBP reconstruction method f_F in (9.32), the back projection $\mathcal{B}$ is applied to the function

$$h(S, \theta) = (\mathcal{F}_1^{-1} F * \mathcal{R}f)(S, \theta). \tag{9.41}$$

To numerically compute the integral in (9.40) at (x, y), we apply the composite rectangle rule, whereby

$$\mathcal{B}h(x, y) \approx \frac{1}{N} \sum_{k=0}^{N-1} h(x \cos(\theta_k) + y \sin(\theta_k), \theta_k). \tag{9.42}$$

This, however, leads us to the following problem.

To approximate $\mathcal{B}h(x,y)$ in (9.42) over the Cartesian grid of pixel points (x,y) we need, for any angle θ_k, $k=0,\dots,N-1$, the values $h(t,\theta_k)$ at

$$t = x\cos(\theta_k) + y\sin(\theta_k). \tag{9.43}$$

In the previous section, we have shown how to evaluate h in (9.41) at polar coordinates (t_m,θ_k) numerically from input data of the form

$$h(t_m,\theta_k) = (\mathcal{F}_1^{-1}F * \mathcal{R}f)(t_m,\theta_k) \qquad \text{for } m\in\mathbb{Z}. \tag{9.44}$$

Now note that t in (9.43) is *not* necessarily contained in the set $\{t_m\}_{m\in\mathbb{Z}}$. But we can determine the value $h(t,\theta_k)$ at $t = x\cos(\theta_k)+y\sin(\theta_k)$ from the data in (9.44) by interpolation, where we recommend the following methods.

Piecewise constant interpolation: In this method, the value $h(t,\theta_k)$ at $t\in[t_m,t_{m+1})$ is approximated by

$$h(t,\theta_k) \approx \mathcal{I}_0 h(t,\theta_k) := \begin{cases} h(t_m,\theta_k) & \text{for } t-t_m \le t_{m+1}-t, \\ h(t_{m+1},\theta_k) & \text{for } t-t_m > t_{m+1}-t. \end{cases}$$

Note that the resulting interpolant $\mathcal{I}_0 h(\cdot,\theta_k)$ is piecewise constant.

Interpolation by linear splines: In this method, the value $h(t,\theta_k)$ at $t\in[t_m,t_{m+1})$ is approximated by

$$h(t,\theta_k) \approx \mathcal{I}_1 h(t,\theta_k) := \frac{L}{\pi}\left[(t-t_m)h(t_{m+1},\theta_k) + (t_{m+1}-t)h(t_m,\theta_k)\right]$$

The *spline interpolant* $\mathcal{I}_1 h(\cdot,\theta_k)$ is globally continuous and piecewise linear.

We summarize the proposed FBP reconstruction method in Algorithm 10.

9.5.5 Numerical Reconstruction of the Shepp-Logan Phantom

We have implemented the FBP reconstruction method, Algorithm 10. For the purpose of illustration, we apply the FBP reconstruction method to the *Shepp-Logan phantom* [66] (see Figure 9.4 (a)). To this end, we use the Shepp-Logan filter F_{SL} from Example 9.19 (see Figure 9.6 (b)). The inverse Fourier transform $\mathcal{F}^{-1}F_{\mathrm{SL}}$ is given in Proposition 9.32, along with the functions values

$$(\mathcal{F}^{-1}F_{\mathrm{SL}})(\pi j/L) = \frac{4L^2}{\pi^3(1-4j^2)} \qquad \text{for } j\in\mathbb{Z},$$

which we use to compute the required convolutions in line 9 of Algorithm 10. To compute the back projection (line 16), we apply linear spline interpolation, i.e., we choose $\mathcal{I}=\mathcal{I}_1$ in line 13.

For our numerical experiments we used parameter values as in Table 9.1. Figure 9.10 shows the resulting FBP reconstructions of the Shepp-Logan phantoms on a grid of 512×512 pixels.

Algorithm 10 Reconstruction by filtered back projection

function FILTERED BACK PROJECTION($\mathcal{R}f$)
  Input: Radon data $\mathcal{R}f \equiv \mathcal{R}f(t_j, \theta_k)$, $k = 0, \dots, N-1$, $j = -M, \dots, M$;
    evaluation points $\{(x_n, y_m) \in \mathbb{R}^2 \,|\, (n,m) \in I_x \times I_y\}$ for (finite)
    index sets $I_x \times I_y \subset \mathbb{N} \times \mathbb{N}$.

  choose low-pass filter F with window W_F and bandwidth $L > 0$;
  for $k = 0, \dots, N-1$ **do**
    for $i \in I$ **do** ▷ for (finite) index set $I \subset \mathbb{Z}$
      let ▷ compute convolution product (9.39)

$$h_{ik} := \frac{\pi}{L} \sum_{j=-M}^{M} \mathcal{F}_1^{-1} F((i-j)\pi/L) \cdot \mathcal{R}f(t_j, \theta_k)$$

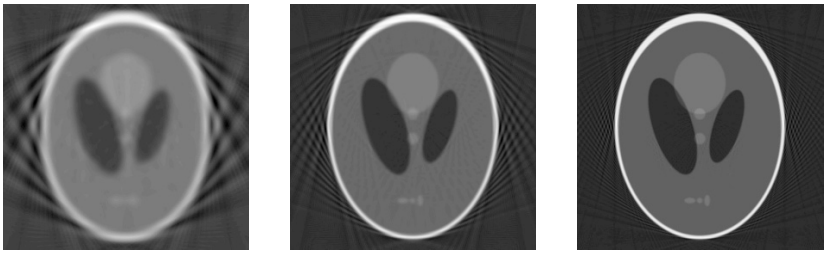
    end for
  end for

  choose interpolation method $\mathcal{I}$ ▷ e.g. linear splines $\mathcal{I}_1$
  for $n \in I_x$ **do**
    for $m \in I_y$ **do**
      let ▷ compute back projection (9.42)

$$f_{nm} := \frac{1}{2N} \sum_{k=0}^{N-1} \mathcal{I}h(x_n \cos(\theta_k) + y_m \sin(\theta_k), \theta_k).$$

    end for
  end for

  Output: reconstruction $\{f_{nm}\}_{(n,m) \in I_x \times I_y}$ with values $f_{nm} \approx f_F(x_n, y_m)$.
end function



(a) 2460 Radon lines (b) 15150 Radon lines (c) 60300 Radon lines

Fig. 9.10. Reconstruction of the Shepp-Logan phantom by filtered back projection, Algorithm 10, using the parameters in Table 9.1.

Table 9.1. Reconstruction of the Shepp-Logan phantom by filtered back projection, Algorithm 10. The following values were used for the bandwidth L, the sampling rate d, the number N of angles θ_k, at $2M+1$ parallel Radon lines ℓ_{t_j,θ_k} per angle θ_k. The resulting reconstructions on a grid of 512×512 pixels are shown in Figure 9.10.

parameter M	bandwidth $L = \pi \cdot M$	sampling rate $d = \pi/L$	# angles $N = 3 \cdot M$	# Radon lines $N \times (2M+1)$
20	20π	0.05	60	2460
50	50π	0.02	150	15150
100	30π	0.01	300	60300

9.6 Exercises

Exercise 9.33. Prove for $f \in \mathrm{L}^1(\mathbb{R}^2)$ the estimate

$$\|\mathcal{R}f(\cdot,\theta)\|_{\mathrm{L}^1(\mathbb{R})} \leq \|f\|_{\mathrm{L}^1(\mathbb{R}^2)} \qquad \text{for all } \theta \in [0,\pi)$$

to conclude

$$\mathcal{R}f \in \mathrm{L}^1(\mathbb{R} \times [0,\pi)) \qquad \text{for all } f \in \mathrm{L}^1(\mathbb{R}^2),$$

i.e., for $f \in \mathrm{L}^1(\mathbb{R}^2)$, we have

$$(\mathcal{R}f)(t,\theta) < \infty \qquad \text{for almost every } (t,\theta) \in \mathbb{R} \times [0,\pi).$$

Exercise 9.34. Consider the function $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$, defined as

$$f(\mathbf{x}) = \begin{cases} \|\mathbf{x}\|_2^{-3/2} & \text{for } \|\mathbf{x}\|_2 \leq 1 \\ 0 & \text{for } \|\mathbf{x}\|_2 > 1 \end{cases} \qquad \mathbf{x} = (x,y) \in \mathbb{R}^2.$$

Show that $(\mathcal{R}f)(0,0)$ is *not* finite, although $f \in \mathrm{L}^1(\mathbb{R}^2)$.

Exercise 9.35. Show that the Radon transform $\mathcal{R}f$ of $f \in \mathrm{L}^1(\mathbb{R}^2)$ has compact support, if f has compact support.

Does the converse of this statement hold? I.e., does $f \in \mathrm{L}^1(\mathbb{R}^2)$ necessarily have compact support, if $\mathrm{supp}(\mathcal{R}f)$ is compact?

Exercise 9.36. Recall the rotation matrix $Q_\theta \in \mathbb{R}^{2\times 2}$ in (9.6) and the unit vector $\mathbf{n}_\theta = (\cos(\theta), \sin(\theta))^T \in \mathbb{R}^2$, for $\theta \in [0,\pi)$, respectively.
Prove the following properties for the Radon transform $\mathcal{R}f$ of $f \in \mathrm{L}^1(\mathbb{R}^2)$.

(a) For $f_\theta(\mathbf{x}) = f(Q_\theta \mathbf{x})$, the identity

$$(\mathcal{R}f)(t,\theta+\varphi) = (\mathcal{R}f_\theta)(t,\varphi)$$

holds for all $t \in \mathbb{R}$ and all $\theta, \varphi \in [0,\pi)$.

(b) For $f_{\mathbf{x}_0}(\mathbf{x}) = f(\mathbf{x} + \mathbf{x}_0)$, where $\mathbf{x}_0 \in \mathbb{R}^2$, the identity

$$(\mathcal{R} f_{\mathbf{x}_0})(t, \theta) = (\mathcal{R} f)(t + \mathbf{n}_\theta^T \mathbf{x}_0, \theta)$$

holds for all $t \in \mathbb{R}$ and all $\theta \in [0, \pi)$.

Exercise 9.37. Show that for a radially symmetric function $f \in \mathrm{L}^1(\mathbb{R}^2)$, the backward projection $\mathcal{B}(\mathcal{R} f)$ of $\mathcal{R} f$ is radially symmetric.

Now consider the indicator function $f = \chi_{\mathcal{B}_1}$ of the unit ball $\mathcal{B}_1$ and its Radon transform $\mathcal{R} f$ from Example 9.6. Show that the backward projection $\mathcal{B}(\mathcal{R} f)$ of $\mathcal{R} f$ is positive on the open annulus

$$\mathcal{R}_1^{\sqrt{2}} = \left\{ \mathbf{x} \in \mathbb{R}^2 \,\middle|\, 1 < \|\mathbf{x}\|_2 < \sqrt{2} \right\} \subset \mathbb{R}^2.$$

Hint: Remark 9.12.

Exercise 9.38. Prove the *Radon convolution theorem*,

$$\mathcal{R}(f * g) = (\mathcal{R} f) * (\mathcal{R} g) \qquad \text{for } f, g \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathscr{C}(\mathbb{R}^2)$$

Hint: Use the Fourier slice theorem, Theorem 9.14.

Exercise 9.39. Show that the backward projection $\mathcal{B}$ is (up to factor π) the *adjoint* operator of the Radon transform $\mathcal{R}$. To this end, prove the relation

$$(\mathcal{R} f, g)_{\mathrm{L}^2(\mathbb{R} \times [0,\pi))} = \pi (f, \mathcal{B} g)_{\mathrm{L}^2(\mathbb{R}^2)}$$

for $g \in \mathrm{L}^2(\mathbb{R} \times [0, \pi))$ and $f \in \mathrm{L}^1(\mathbb{R}^2) \cap \mathrm{L}^2(\mathbb{R}^2)$ satisfying $\mathcal{R} f \in \mathrm{L}^2(\mathbb{R} \times [0, \pi))$.

Exercise 9.40. In this exercise, we consider a *spline filter* of first order, which is a low-pass filter $F : \mathbb{R} \longrightarrow \mathbb{R}$ of the form

$$F(S) = |S| \cdot \wedge(S) \cdot \sqcap(S)$$

(cf. Definition 9.17) with the *linear B-spline* $\wedge : \mathbb{R} \longrightarrow \mathbb{R}$, defined as

$$\wedge(S) = (1 - |S|)_+ = \begin{cases} 1 - |S| & \text{for } |S| \leq 1, \\ 0 & \text{for } |S| > 1. \end{cases}$$

(a) Show the representation

$$(\mathcal{F}_1^{-1} F)(x) = \frac{2}{\pi} \left[\frac{\sin^2(x/2) + \operatorname{sinc}(x) - 1}{x^2} \right] \qquad \text{for } x \in \mathbb{R}$$

for the inverse Fourier transform $\mathcal{F}_1^{-1} F$ of F.

(b) Use the result in (a) to compute $(\mathcal{F}_1^{-1} F)(\pi n)$ for $n \in \mathbb{Z}$.

Exercise 9.41. A spline filter F_k of order $k \in \mathbb{N}_0$ has the form

$$F_k(S) = |S| \cdot \wedge_k(S) \cdot \sqcap(S) \qquad \text{for } k \in \mathbb{N}_0, \tag{9.45}$$

where the B-spline $\wedge_k$ is defined by the recursion

$$\wedge_k(S) := (\wedge_{k-1} * \sqcap)(S/\alpha_k) \qquad \text{for } k \in \mathbb{N} \tag{9.46}$$

for the initial value $\wedge_0 := \sqcap$ and where, moreover, the positive scaling factor $\alpha_k > 0$ in (9.46) is chosen, such that $\operatorname{supp}(\wedge_k) = [-1, 1]$.

In this exercise, we construct a spline filter of second order.

(a) Show that the initial value $\wedge_0$ yields the Ram-Lak filter, i.e., $F_0 \equiv F_{\mathrm{RL}}$.
(b) Show that the scaling factor $\alpha_k > 0$ in (9.46) is, for any $k \in \mathbb{N}$, uniquely determined by the requirement $\operatorname{supp}(\wedge_k) = [-1, 1]$.
(c) Show that $\wedge_1$ generates by F_1 the spline filter from Exercise 9.40. Determine the scaling factor α_1 of F_1.
(d) Compute the second order spline filter F_2. To this end, determine the B-spline $\wedge_2$ in (9.46), along with its scaling factor α_2.

Exercise 9.42. Develop a construction scheme for higher order spline filters F_k of the form (9.45), where $k \geq 3$. To this end, apply the recursion in (9.46) and determine the scaling factors α_k, for $k \geq 3$.

Exercise 9.43. Compute the inverse Fourier transform $\mathcal{F}^{-1}F$ of the

(a) cosine filter $F = F_{\mathrm{CF}}$ from Example 9.20;
(b) Hamming filter $F = F_\beta$ from Example 9.21;
(c) Gauss filter $F = F_\alpha$ from Example 9.22.

Compute for each of the filters F in (a)-(c) the values

$$(\mathcal{F}^{-1}F)(\pi j/L) \qquad \text{for } j \in \mathbb{Z}.$$

Hint: Proposition 9.31 and Proposition 9.32.

Exercise 9.44. Let $F \equiv F_{L,W}$ be a low-pass filter with bandwidth $L > 0$ and window function $W : \mathbb{R} \longrightarrow \mathbb{R}$, according to Definition 9.17. Moreover, let

$$K_F(x, y) = \frac{1}{2}\mathcal{B}\left(\mathcal{F}_1^{-1}F\right)(x, y) \qquad \text{for } (x, y) \in \mathbb{R}^2$$

be the convolution kernel of F.

Prove for the scaled window $W_L(S) = W(S/L)$, $S \in \mathbb{R}$, the identity

$$W_L(\|(x, y)\|_2) = \mathcal{F}K_F(x, y) \tag{9.47}$$

In which sense does the identity in (9.47) hold?

Hint: Elaborate the details in the proof of [5, Proposition 4.1].

Exercise 9.45. Implement the reconstruction method of the filtered back projection (FBP), Algorithm 10. Apply the FBP method to the phantom *bull's eye* (see Example 9.9 and Figure 9.3). To this end, use the bandwidth $L = \pi \cdot M$, the sampling rate $d = \pi/L$, and $N = 3 \cdot M$ angles θ_k, with $2M+1$ parallel Radon lines ℓ_{t_j,θ_k} per angle θ_k, for $M = 10, 20, 50$.

For verification, the reconstructions with 512×512 pixels are displayed in Figure 9.11, where we used the Shepp-Logan filter F_{SL} from Example 9.19 (see Figure 9.6 (b)) in combination with linear spline interpolation.

(a) 630 Radon lines (b) 2460 Radon lines (c) 15150 Radon lines

Fig. 9.11. Reconstruction of *bull's eye* from Example 9.9 (see Figure 9.3).

References

1. N. Aronszajn: Theory of reproducing kernels. *Transactions of the AMS* **68**, 1950, 337–404.
2. R. Askey: Radial characteristic functions. TSR # 1262, Univ. Wisconsin, 1973.
3. S. Banach, S. Mazur: Zur Theorie der linearen Dimension. *Studia Mathematica* **4**, 1933, 100–112.
4. M. Beckmann, A. Iske: Error estimates and convergence rates for filtered back projection. *Mathematics of Computation*, published electronically on April 30, 2018, `https://doi.org/10.1090/mcom/3343`.
5. M. Beckmann, A. Iske: Approximation of bivariate functions from fractional Sobolev spaces by filtered back projection. HBAM 2017-05, U. Hamburg, 2017.
6. A. Beer: Bestimmung der Absorption des rothen Lichts in farbigen Flüssigkeiten. *Annalen der Physik und Chemie* **86**, 1852, 78–88.
7. Å. Bjørck: *Numerical Methods for Least Squares Problems*. SIAM, 1996.
8. S. Bochner: *Vorlesungen über Fouriersche Integrale*. Akademische Verlagsgesellschaft, Leipzig, 1932.
9. D. Braess: *Nonlinear Approximation Theory*. Springer, Berlin, 1986.
10. M.D. Buhmann: *Radial Basis Functions*. Cambridge University Press, Cambridge, UK, 2003.
11. E.W. Cheney: *Introduction to Approximation Theory*. Second edition, McGraw Hill, New York, NY, U.S.A., 1982.
12. W. Cheney, W. Light: *A Course in Approximation Theory*. Graduate Studies in Mathematics, vol. 101, AMS, Providence, RI, U.S.A., 2000.
13. O. Christensen: *An Introduction to Frames and Riesz Bases*. Second expanded edition, Birkhäuser, 2016.
14. C.K. Chui: *Wavelets: A Mathematical Tool for Signal Analysis*. Monographs on Mathematical Modeling and Computation. SIAM, 1997.
15. C.W. Clenshaw: A note on the summation of Chebyshev series. *Mathematics of Computation* **9**(51), 1955, 118–120.
16. J.W. Cooley, J.W. Tukey. An algorithm for the machine calculation of complex Fourier series. *Mathematics of Computation* **19**, 1965, 297–301.
17. P.C. Curtis Jr.: N-parameter families and best approximation. *Pacific Journal of Mathematics* **9**, 1959, 1013–1027.
18. I. Daubechies: *Ten Lectures on Wavelets*. SIAM, Philadelphia, 1992.
19. P.J. Davis: *Interpolation and Approximation*. 2nd edition, Dover, NY, 1975.
20. C. de Boor: *A Practical Guide to Splines*. Revised edition, Applied Mathematical Sciences, vol. 27, Springer, New York, 2001.
21. R.A. DeVore: Nonlinear approximation. *Acta Numerica*, 1998, 51–150.

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7

22. B. Diederichs, A. Iske: Improved estimates for condition numbers of radial basis function interpolation matrices. *J. Approximation Theory*, published electronically on October 16, 2017, `https://doi.org/10.1016/j.jat.2017.10.004`.
23. G. Faber: Über die interpolatorische Darstellung stetiger Funktionen. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **23**, 1914, 192–210.
24. G.E. Fasshauer: *Meshfree Approximation Methods with Matlab*. World Scientific, Singapore, 2007.
25. G.E. Fasshauer, M. McCourt: *Kernel-based Approximation Methods using Matlab*. World Scientific, Singapore, 2015.
26. G.B. Folland: *Fourier Analysis and its Applications*. Brooks/Cole, Pacific Grove, CA, U.S.A., 1992.
27. B. Fornberg, N. Flyer: *A Primer on Radial Basis Functions with Applications to the Geosciences*. SIAM, Philadelphia, 2015.
28. W. Gander, M.J. Gander, F. Kwok: *Scientific Computing – An Introduction using Maple and MATLAB*. Texts in CSE, volume 11, Springer, 2014.
29. C. Gasquet, P. Witomski: *Fourier Analysis and Applications*. Springer Science+Business Media, New York, 1999.
30. M. v. Golitschek: Penalized least squares approximation problems. *Jaen Journal on Approximation Theory* **1**(1), 2009, 83–96.
31. J. Gomes, L. Velho: *From Fourier Analysis to Wavelets*. Springer, 2015.
32. A. Haar: Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen 69*, 1910, 331–371.
33. M. Haase: *Functional Analysis: An Elementary Introduction*. American Mathematical Society, Providence, RI, U.S.A., 2014.
34. P.C. Hansen, J.G. Nagy, D.P. O'Leary: *Deblurring Images: Matrices, Spectra, and Filtering*. Fundamentals of Algorithms. SIAM, Philadelphia, 2006.
35. E. Hewitt, K.A. Ross: *Abstract Harmonic Analysis I*. Springer, Berlin, 1963.
36. K. Höllig, J. Hörner: *Approximation and Modeling with B-Splines*. SIAM, Philadelphia, 2013.
37. A. Iske: *Charakterisierung bedingt positiv definiter Funktionen für multivariate Interpolationsmethoden mit radialen Basisfunktionen*. Dissertation, Universität Göttingen, 1994.
38. A. Iske: *Multiresolution Methods in Scattered Data Modelling*. Lecture Notes in Computational Science and Engineering, vol. 37, Springer, Berlin, 2004.
39. J.L.W.V. Jensen: Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica* **30**, 1906, 175–193.
40. P. Jordan, J. von Neumann: On inner products in linear, metric spaces. *Annals of Mathematics* **36**(3), 1935, 719–723.
41. C.L. Lawson, R.J. Hanson: *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, NJ, U.S.A., 1974.
42. P.D. Lax: *Functional Analysis*. Wiley-Interscience, New York, U.S.A., 2002.
43. G.G. Lorentz, M. v. Golitschek, Y. Makovoz: *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften, Band 304, Springer, 2011.
44. W.R. Madych: Summability and approximate reconstruction from Radon transform data. In: *Integral Geometry and Tomography*, E. Grinberg and T. Quinto (eds.), AMS, Providence, RI, U.S.A., 1990, 189–219.
45. W.R. Madych, S.A. Nelson: *Multivariate Interpolation: A Variational Theory*. Technical Report, Iowa State University, 1983.
46. W.R. Madych, S.A. Nelson: Multivariate interpolation and conditionally positive definite functions. *Approx. Theory Appl.* **4**, 1988, 77–89.

47. W.R. Madych, S.A. Nelson: Multivariate interpolation and conditionally positive definite functions II. *Mathematics of Computation* **54**, 1990, 211–230.
48. J. Mairhuber: On Haar's theorem concerning Chebysheff problems having unique solutions. *Proc. Am. Math. Soc.* **7**, 1956, 609–615.
49. S. Mallat: *A Wavelet Tour of Signal Processing*. Academic Press, 1998.
50. G. Meinardus: *Approximation of Functions: Theory and Numerical Methods*. Springer, Berlin, 1967.
51. V. Michel: *Lectures on Constructive Approximation*. Birkhäuser, NY, 2013.
52. P. Munshi: Error analysis of tomographic filters I: theory. *NDT & E Int.* **25**, 1992, 191–194.
53. P. Munshi, R.K.S. Rathore, K.S. Ram, M.S. Kalra: Error estimates for tomographic inversion. *Inverse Problems* **7**, 1991, 399–408.
54. P. Munshi, R.K.S. Rathore, K.S. Ram, M.S. Kalra: Error analysis of tomographic filters II: results. *NDT & E Int.* **26**, 1993, 235–240.
55. J.J. O'Connor, E.F. Robertson: *MacTutor History of Mathematics archive*. `http://www-history.mcs.st-andrews.ac.uk`.
56. M.J.D. Powell: *Approximation Theory and Methods*. Cambridge University Press, Cambridge, UK, 1981.
57. A. Quarteroni, R. Sacco, F. Saleri: *Numerical Mathematics*. Springer, New York, 2000.
58. M. Reed, B. Simon: *Fourier Analysis, Self-Adjointness*. In: *Methods of Modern Mathematical Physics II*, Academic Press, New York, 1975.
59. E.Y. Remez: Sur le calcul effectiv des polynômes d'approximation des Tschebyscheff. *Compt. Rend. Acad. Sc.* **199**, 1934, 337.
60. E.Y. Remez: Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation. *Compt. Rend. Acad. Sc.* **198**, 1934, 2063.
61. R. Schaback: Creating surfaces from scattered data using radial basis functions. In: *Mathematical Methods for Curves and Surfaces*, M. Dæhlen, T. Lyche, and L.L. Schumaker (eds.), Vanderbilt University Press, Nashville, 1995, 477–496.
62. R. Schaback, H. Wendland: *Special Cases of Compactly Supported Radial Basis Functions*. Technical Report, Universität Göttingen, 1993.
63. R. Schaback, H. Wendland: *Numerische Mathematik*. Springer, Berlin, 2005.
64. L.L. Schumaker: *Spline Functions: Basic Theory*. Third Edition, Cambridge University Press, Cambridge, UK, 2007.
65. L.L. Schumaker: *Spline Functions: Computational Methods*. SIAM, 2015.
66. L.A. Shepp, B.F. Logan: The Fourier reconstruction of a head section. *IEEE Trans. Nucl. Sci.* **21**, 1974, 21–43.
67. G. Szegő: *Orthogonal Polynomials*. AMS, Providence, RI, U.S.A., 1939.
68. L.N. Trefethen: *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, 2013.
69. D.F. Walnut: *An Introduction to Wavelet Analysis*. Birkhäuser Basel, 2004.
70. G.A. Watson: *Approximation Theory and Numerical Methods*. John Wiley & Sons, Chichester, 1980.
71. H. Wendland: Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. *Advances in Comp. Math.* **4**, 1995, 389–396.
72. H. Wendland: *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.
73. Wikipedia. The free encyclopedia. `https://en.wikipedia.org/wiki/`
74. Z. Wu: Multivariate compactly supported positive definite radial functions. *Advances in Comp. Math.* **4**, 1995, 283–292.

Subject Index

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7

Name Index

© Springer Nature Switzerland AG 2018

A. Iske, *Approximation Theory and Algorithms for Data Analysis*, Texts in Applied Mathematics 68, https://doi.org/10.1007/978-3-030-05228-7