Kalle Åström
Lars-Erik Persson
Sergei D. Silvestrov *Editors*

Analysis for Science, Engineering and Beyond

The Tribute Workshop in Honour of Gunnar Sparr held in Lund, May 8-9, 2008

Springer

Springer Proceedings in Mathematics

Volume 6

For further volumes:
http://www.springer.com/series/8806

Springer Proceedings in Mathematics

The book series will feature volumes of selected contributions from workshops and conferences in all areas of current research activity in mathematics. Besides an overall evaluation, at the hands of the publisher, of the interest, scientific quality, and timeliness of each proposal, every individual contribution will be refereed to standards comparable to those of leading mathematics journals. It is hoped that this series will thus propose to the research community well-edited and authoritative reports on newest developments in the most interesting and promising areas of mathematical research today.

Kalle Åström
Lars-Erik Persson
Sergei D. Silvestrov
Editors

Analysis for Science, Engineering and Beyond

The Tribute Workshop in Honour of Gunnar Sparr held in Lund, May 8-9, 2008

Springer

Editors

Kalle Åström
Lund University
Centre for Mathematical Sciences
PO Box 118
221 00 Lund
Sweden
kalle@maths.lth.se

Sergei D. Silvestrov
Lund University
Centre for Mathematical Sciences
PO Box 118
221 00 Lund
Sweden
sergei.silvestrov@math.lth.se

and

Division of Applied Mathematics,
The School of Education, Culture
and Communication
Mälardalen University
PO BOX 883,
72123 Västerås, Sweden
sergei.silvestrov@mdh.se

Lars-Erik Persson
Luleå University of Technology
Dept. Mathematics
971 87 Luleå
Sweden

and

Narvik University College
PO Box 385
N 8505, Narvik, Norway
larserik@sm.luth.se

ISSN 2190-5614
ISBN 978-3-642-20235-3 e-ISBN 978-3-642-20236-0
DOI 10.1007/978-3-642-20236-0
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011942602

Mathematical Subject Classification (2010): 46-06, 62H35, 65D18, 68U10, 94A08, 26D10, 42A18, 42C40, 60G18, 46B70, 46L51, 46L52, 47A57, 92-08, 97B40, 97B10, 01A

Cover design: deblik, Berlin

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

This volume is dedicated to Professor Gunnar Sparr and his more than 40 years of excellent service and important research contributions to mathematics and its applications and to mathematics education.

Preface

This book is dedicated in celebration of Professor Gunnar Sparr's 65th anniversary and more than 40 years of exceptional service to mathematics and its applications in engineering and technology, to mathematics and engineering education, and to interdisciplinary, industrial and international cooperation. There are very few scientists who during their career manage to make successful and significant, lasting contributions simultaneously in the theoretical development of their subject, in its applications to other subjects, the development of new technology important for industries and society at large, and the development of education reaching far beyond their own subject, inspiring several generations of scientists and engineers. These characteristics of Gunnar Sparr – his deep, versatile and living expansion of the interplay between theoretical and applied aspects in mathematical research and education – is one of the main themes reflected throughout this book.

The individual chapters highlight some of Professor Sparr's broad areas of interest throughout the years and at the same time show several opportunities for further research both in mathematics and in applications in other sciences, engineering and beyond.

The volume starts with the comprehensive biographical review chapter about Gunnar Sparr and his contributions written by Lars-Erik Persson. This chapter describes the wide range of activities of Professor Gunnar Sparr in mathematics, engineering mathematics, and in the mathematical and engineering education. Scientifically he has made very important and long lasting contributions particularly in interpolation theory and mathematical computer vision. In particular, in the first field his contributions have been crucial for developing the modern theory concerning interpolation between three or more Banach spaces. In the latter field, his contributions are further strengthened by his being the founder of a large and internationally recognized research group. The chapter also describes the great impact of Gunnar Sparr on industrial mathematics in the region and Sweden, manifested by numerous collaboration projects, and even the founding of a successful start-up company. It also describes his role in the Swedish educational system, by being the driving force in the forming of a successful MSc-program

in Engineering Mathematics, with the mathematical sciences providing the main engineering tools. These remarkable contributions of Gunnar Sparr in mathematics and industrial mathematics as well as his outstanding role as an entrepreneur and ambassador for mathematics have also essentially influenced most of the chapters in this book.

Chapter 2, written by Georg Lindgren and Gunnar Sparr, starts with a historical account of engineering mathematics research at Lund Institute of Technology, where not only the mathematical sciences have played a role. The chapter then discusses important aspects of engineering mathematics education and the implementation of the highly successful education programme in Engineering Mathematics, which started in Lund in 2002. As verified by the employments of those who have finished, the education has been very successful in preparing students for mathematical work in a wide range of branches, as well as for doctoral studies in a wide range of subjects.

The other chapters in the volume are concerned with selected topics in contemporary mathematics and engineering mathematics.

Chapter 3, by Jan Koenderink, is concerned with edge detection, one of the most fundamental operations in image processing. In spite of the fundamental nature and numerous studies of edge detection, it remains yet unclear what precisely is meant by "edge"-apart from being what edge detectors detect, or by an "edge detector"-except for being that which detects edges. Though many edge detection algorithms are in common use, they "find" slightly different entities and it remains unclear how one may compare their effectiveness in detecting edges, although this is commonly done. In this chapter, a principled theory of edge detection is offered that is based on the structure of the 2-jet of the image at a certain scale. In this theory there is no such thing as an "edge detector"; edges are defined in terms of the 2-jet as a single object.

Chapter 4, by Johan Karlsson, Anders Ericsson and Kalle Åström, is devoted to a new method of shape modeling by optimising description length using gradients and parameterisation invariance. In statistical shape modelling, a dense correspondence between the shapes in the training set must be established. In recent years there has been a great deal of work on the automatic construction of Shape Models, and in particular, the minimum description length is shown to be useful in locating a dense correspondence between shapes. In this chapter, the gradient of the description length is derived, and the minimum description length is optimised using steepest descent, yielding faster and better models. To characterise shape properties that are invariant to similarity transformations, it is first necessary to normalize with respect to the similarity transformations from the annotated configurations. In this chapter, it is proposed to align shapes using the minimum description length criterion, and it is concluded that there is improvement in generalization in comparison to the normally used Procrustes analysis of minimizing the sum of squared distances between the corresponding landmarks under similarity transformations. Moreover, the novel theory to prevent the commonly occurring problem of clustering under correspondence optimization is presented. The problem is solved by calculating the covariance matrix of the shapes using a scalar product that is invariant to mutual

reparameterisations, and an algorithm for implementing the ideas is proposed that yields stability and model quality improvements with respect to several existing state of the art algorithms.

The next Chapter 5, by Anders Erikson and Kalle Åström, is about the thin-plate spline widely used in a number of areas such as image warping, shape analysis and scattered data interpolation. This natural interpolating function in two dimensions has a very intuitive interpretation as well as an elegant mathematical formulation, but has no inherent restriction to prevent folding, i.e. a non-bijective interpolating function. This chapter is concerned with the properties of the set of parameterisations that form bijective thin-plate splines, such as convexity and boundness. Methods for finding sufficient as well as necessary conditions for bijectivity are also presented.

In Chapter 6, by Magnus Fontes, a collection of statistical and mathematical tools, which are useful for the exploration of multivariate data, have been selected and are presented in a form that is meant to be particularly accessible to classically trained mathematicians. The chapter contains self-contained and streamlined introductions to principal component analysis, multidimensional scaling and statistical hypothesis testing. Within the presented mathematical framework, a general exploratory methodology for the investigation of real-world high dimensional datasets that builds on statistical and knowledge-supported visualizations is proposed. This methodology is then exemplified by applying it to several different genome-wide DNA-microarray datasets. This exploratory methodology can be expanded and developed in many directions. Recent promising advances in the theory for random matrices are presented as an example that, if further developed, could potentially provide practically useful and theoretically well-founded estimations of information content in dimension-reducing visualizations. This chapter can serve as an introduction to, and help to stimulate more research within, the interesting and rapidly expanding field of data exploration.

In Chapter 7, by Stefan Diehl, the shockwave behaviour of sedimentation in wastewater treatment is considered. Continuous sedimentation is a common industrial process for separating particles from a liquid. It is used in the chemical, mining, pulp-and-paper and food industries, and can also be found in most wastewater treatment plants, where it is a crucial sub-process of a complex biological system. The process continues to present scientific problems that lead to fundamental research in different disciplines such as mathematics and wastewater, chemical, mineral, control and automation engineering. In this chapter a selective survey of previous results within the field of pure and applied mathematics is presented, with a focus on a nonlinear convection-diffusion partial differential equation with discontinuous coefficients. In a model of a wastewater treatment plant, such an equation is coupled to a set of ordinary differential equations. New results on the steady-state solutions of such a coupled system are also presented.

Chapter 8, by Palle E. T. Jorgensen and Myung-Sin Song, is concerned with wavelets, image compression, and encoding. A family of multi-scale algorithms is developed using filter functions in higher dimensions. While the primary application is to images, i.e., processes in two dimensions, the main theorems are proved in

a more general context, allowing dimension 3 and higher. The key tool for the algorithms presented is the use of tensor products of representations of certain algebras, the Cuntz algebras O_N, from the theory of algebras of operators in Hilbert spaces. The main result offers a matrix algorithm for computing coefficients for images or signals in specific resolution subspaces. A special feature with the matrix operations used is that they involve products and the iteration of slanted matrices. Slanted matrices, while large, have many zeros, i.e., they are sparse. It is proved that as the operations increase the degree of sparseness of the matrices also increases, and as a result, only a few terms in the expansions are needed in order to achieve a good approximation of the processed image. The expansions presented are local in a strong sense. An additional advantage of using representations of the algebras O_N and tensor products is that one gets easy formulas for generating all the choices of matrices going into algorithms.

In Chapter 9, by Dorin Dutkay and Sergei Silvestrov, this fruitful line of investigation of applications of operator theory methods and operator representations of algebras to wavelet analysis and its applications is continued. Wavelet representations make it possible to apply the multi-resolution techniques of wavelet theory to a larger class of problems where self-similarity or refinement is the central phenomenon. They are used to construct wavelet bases and multi-resolutions on fractal measures and Cantor sets or on solenoids. In this chapter an open question about the irreducibility of the wavelet representation associated to the Cantor set is answered by proving that if the quadrature mirror filter does not have a constant absolute value, then the wavelet representation is reducible. The proof uses the dynamics of ergodic shifts on solenoids, Birkoff's ergodic theorem, and also concavity and Jensen's inequality for the logarithm function. Such inequalities for functions and integrals are a broad and important topic in mathematics and its applications.

The next Chapter 10, by Maria Johansson and Lars-Erik Persson, provides a review of and deeper insights into multidimensional inequalities of Hardy and Polya-Knopp types. Multidimensional Hardy-type inequalities are very important for several areas in mathematics, for example partial differential equations and homogenization theory and for their applications to e.g. tribology and material sciences. Sawyer's well-known two-dimensional Hardy-type inequality from 1985 is complemented and extended in various ways, unifying many ideas and results, and also new results are proved and interesting open questions are raised. The newest information about weight characterizations for Hardy-type operators acting between weighted Lebesque spaces are presented and discussed in this general frame. From the Sawyer paper we know that for the two-dimensional case we need three independent conditions in the case $1 < p \leq q < \infty$. One main piece of information in this chapter is that if one of the weights is of product type, then only one condition is necessary for such characterization and in fact these results also hold in the general n-dimensional case. And in this case also the similar results can be stated in the more complicated case $1 < q < p < \infty$. Finally, also the corresponding limit cases with Polya-Knopp inequalities are presented and discussed and here the results in fact hold without any restriction on the weights, which is a remarkable fact.

In the next Chapter 11, by Lars-Erik Persson, Lyazzat Sarybekova and Nazerke Tleukhanova, a new Fourier series multiplier theorem of Lizorkin type is proved. The result is given for a general strong regular system and, in particular, for the trigonometric system it implies an analogy of the original Lizorkin theorem. The important role is played by the tools, methods and results from the interpolation theory and embedding theorems for Banach spaces, such as the Peetre K-functional for pairs of Banach spaces, Marcinkiewicz interpolation theorem and interpolation and embedding properties of the Lp and related function spaces.

In Chapter 12, by Hiroyuki Osaka and Jun Tomiyama, the theme of interpolation and inequalities is continued and expanded in the direction of interplay between classes of matrix and operator monotone and convex functions and interpolation classes. A natural extension of monotonicity and convexity of functions to non-commutative spaces of matrices and operators is via standard functional calculus and the same inequality based on requirements of monotonicity and convexity but applied for the standard positivity-induced partial order on matrices and operators instead of the special case of the order on real numbers. These classes of functions have important connections to interpolation, approximation, and moment problems, as well as interesting applications for example in quantum mechanics, quantum information, automatic control and telecommunication. A deeper understanding of the relations between these classes of functions for matrices of different sizes is important for analysis and applications, and many fundamental problems remain open. Other interpolation classes of functions, which are important for applications and closely related, are the sets of all positive real-valued continuous functions, which can be interpolated on a given subset by using a positive Pick function. In this chapter, the interpolation classes are characterized in a useful way by an operator inequality, n-monotone functions are characterized from the point of Jensen's type inequality for operators, and the understanding of the mutual inclusions between these function classes is advanced with several new results and solutions of open problems. Several of Gunnar Sparr's results and open problems play an important part in this chapter.

The book closes with Chapter 13, by Sergei Silvestrov, and is concerned with the two pioneering, far-reaching and in many ways yet to be fully explored papers by Gunnar Sparr and Jaak Peetre on the interpolation of normed abelian groups and on non-commutative integration. These papers introduced important methods and tools unifying many previously known interpolation results and methods within the same framework at the same time, making it possible to expand and apply interpolation methods and results to non-commutative spaces in the ways essential to building non-commutative integration and non-commutative extensions of the function spaces. In particular, these directions are of importance in e.g. non-commutative geometry and in applications to quantum physics. While some notions and methods from these papers have been applied in various contexts, many other excellent methods and ideas presented in them have remained undiscovered and as such not been developed further. This final chapter presents a concise state of the art review of these and some other related important works of Gunnar Sparr, as well

as some related investigations into non-commutative spaces and non-commutative integration done in operator algebras and operator theory.

The main aim of this book is to stimulate new advances in the areas of mathematics represented in the volume and related directions, as well as in the addressed applications in engineering, physics, life sciences. The book consists of thirteen carefully selected and refereed contributed chapters with a shared emphasis on important methods, research directions and applications of analysis within and beyond mathematics. As mentioned at the outset, the works in this book have been collected in celebration of Professor Gunnar Sparr's 65th anniversary and over four decades of fundamental work in mathematics and its applications in various fields of mathematics and engineering education, as well as to interdisciplinary, industrial and international cooperation. This is well reflected in the topics considered in the individual chapters of the volume. This book will serve as a source of inspiration for a broad spectrum of researchers and research students, as the contributions lie at the intersection of the research directions and interests of several large research communities and research groups in modern mathematics and its applications in other branches of science, engineering and technology.

Support from the Swedish Research Council, Swedish Foundation for International Cooperation in Research and Higher Education (STINT), Swedish Royal Academy of Sciences and Crafoord Foundation during preparation of this book is gratefully acknowledged.

Lund
August 2010

Kalle Åström
Lars-Erik Persson
Sergei Silvestrov

Contents

Contributors

Kalle Åström Centre for Mathematical Sciences, Lund University, Lund, Sweden, kalle@maths.lth.se

Stefan Diehl Centre for Mathematical Sciences, Lund University, P.O. Box 118, 22100 Lund, Sweden, diehl@maths.lth.se

Dorin Ervin Dutkay Department of Mathematics, University of Central Florida, 4000 Central Florida Blvd., P.O. Box 161364, Orlando, FL 32816-1364, USA, ddutkay@mail.ucf.edu

Anders Ericsson Centre for Mathematical Sciences, Lund University, Lund, Sweden, anderse@maths.lth.se

Anders P Erikson Centre for Mathematical Sciences, Lund University, Lund, Sweden, anderspe@maths.lth.se

Magnus Fontes Centre for Mathematical Sciences, Lund University, Box 118, 22100, Lund, Sweden, fontes@maths.lth.se

Maria Johansson Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden, maria.l.johansson@ltu.se

Palle E. T. Jorgensen Department of Mathematics, The University of Iowa, Iowa City, IA 52242, USA, jorgen@math.uiowa.edu

Johan Karlsson Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden, johan.karlsson@fcc.chalmers.se

Jan J. Koenderink Man–Machine Interaction Group, EEMCS, Delft University of Technology, Delft, The Netherlands, jan.koenderink@telfort.nl

Georg Lindgren Centre for Mathematical Sciences, LTH, Lund University, Box 118, Lund, SE-22100, Sweden, georg@maths.lth.se

Hiroyuki Osaka Department of Mathematical Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan, osaka@se.ritsumei.ac.jp

Lars-Erik Persson Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden

Narvik University College, P.O. Box 385, N 8505, Narvik, Norway, larserik@sm.luth.se

Lyazzat Sarybekova L. N. Gumilyov Eurasian National University, 5 Munaitpasov St, Astana 010008, Kazakhstan, lsarybekova@yandex.ru

Sergei Silvestrov Centre for Mathematical Sciences, Lund University, Box 118, 22100 Lund, Sweden, sergei.silvestrov@math.lth.se

Division of Applied Mathematics, The School of Education, Culture and Communication, Mälardalen University, Box 883, 72123 Västerås, Sweden, sergei.silvestrov@mdh.se

Myung-Sin Song Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA, msong@siue.edu

Gunnar Sparr Centre for Mathematical Sciences, LTH, Lund University, Box 118, Lund, 22100, Sweden, gunnar@maths.lth.se

Nazerke Tleukhanova L. N. Gumilyov Eurasian National University, 5 Munaitpasov St, Astana 010008, Kazakhstan, tleukhanova@rambler.ru

Jun Tomiyama Prof. Emeritus of Tokyo Metropolitan University, 201 11-10 Nakane 1-chome, Meguro-ku, Tokyo, Japan, juntomi@med.email.ne.jp

Chapter 1
Gunnar Sparr: The Remarkable Mathematician, Entrepreneur and Ambassador for Mathematics

Lars-Erik Persson

Abstract The article describes the wide range of activities of Professor Gunnar Sparr, Lund University, in mathematics, engineering mathematics, and in the education of mathematics. Scientifically he has made lasting contributions particularly in interpolation theory and mathematical computer vision. In the latter field, his contributions are further strengthened by his being the founder of a large and internationally recognized research group. The article also describes the impact of Gunnar Sparr on industrial mathematics in the region and Sweden, manifested by numerous collaboration projects, and even the founding of a successful start-up company. It also describes his role in the Swedish educational system, by being the driving force in the forming of a successful MSc-program in Engineering Mathematics, with the mathematical sciences providing the main engineering tools.

1.1 Introduction

Mathematics is the most fantastic subject which has ever beens created by human beings. A subject which has survived all trends and developments of new areas of science. In many cases new results from mathematics have either been the direct reason or contributed in an essential way to such developments. This means in particular that mathematics is closely related to our culture, and several technological tools we nowadays use in a natural way are developed by using mathematical ideas and theories (e.g. Google search, modern fibre cables, credit cards, satellite signals, mobile phones, predictions of stock markets, predictions in the nature, pattern recognition, effective properties of composite materials, etc.).

L.-E. Persson (✉)
Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden

Narvik University College, P.O. Box 385, N 8505, Narvik, Norway
e-mail: larserik@sm.luth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_1,

It is very important for this development that there exist "entrepreneurs" with roots in pure mathematics, who also can develop and implement such new mathematics to the world around us. I claim that Gunnar Sparr is the most splendid example of such an entrepreneur we ever have had in Sweden, and he has served as a source of inspiration also for many other Swedish mathematicians with interest to work in this direction (including myself and many of my students). Gunnar Sparr can indeed be characterized as a remarkable mathematician, engineering mathematician and mathematical engineer. He is an outstanding ambassador for our wonderful subject mathematics.

It is impossible to give a complete picture of such a remarkable man and his work in a limited article like this. However, I will at least try to give a flavour of his life and work so far. Let me just conclude my introduction with a few sentences, which can be seen as an abstract of my further discussion below:

- Gunnar has indeed an unusual and varied carrier. Despite of having worked at the same place all the time, he has changed focus on his mathematical work several times.
- Gunnar is a genuine entrepreneur for mathematics both outside and inside the university.
- Gunnar has planted a number of seeds that have continued to live on their own in a remarkably successful way.
- Gunnar has always not only claimed that mathematics is useful, but also practically proved its usefulness by own research, collaborations, and even the founding of a start-up company.
- Gunnar has also had a great impact on the Swedish educational system by having been the driving force in the creation of the very successful study programs in Engineering Mathematics, which today exist both in Lund and Gothenburg.
- Gunnar's work has stretched borders and in a remarkable way helped to increase the esteem of mathematics outside the mathematical world.

1.2 Background, Family, Biographical Data

Gunnar Sparr was born 1942 in Karlskoga, but with his full genealogical table from a small place Våmhus in the province of Dalarna a bit north in Sweden. He was the only child in a family, where his father was Lars Sparr (1910–1985), an electrical engineer, and his mother was Mait Sparr (1914–2007).

Gunnar went to high school in Växjö, in the southern part of Sweden. In 1962, he started his studies at Lund Institute of Technology (LTH, a faculty of Lund University) for a MSc in Engineering (Swedish: civilingenjör). LTH then was very new, having been founded the year before. He was burning for mathematics, and in parallel to the engineering studies at LTH he studied mathematics at the Science Faculty. In 1966 he completed a MSc in Engineering Physics at LTH. After that he started PhD-studies in mathematics, within interpolation theory, with Professor Jaak Peetre as advisor. This first research period is described in the next section.

Fig. 1.1 This photo of Gunnar and his wife Annika is taken around 2005 by Professor Michael Cwikel, who is a close friend of the Sparr family

In particular, during this period he met his wife Annika (born Haaker). They married in 1971 and they got two children, Emma (born 1972) and Anna (born 1978). They have today four grandchildren. Also Annika is a well-known mathematician working at Lund University (Fig. 1.1).

Gunnar defended his thesis in 1972. In the same year he got a position as associate professor at the department of mathematics at Lund university. After a few years of continued research, he qualified for the Swedish honorary degree of "docent" in 1974. He was promoted professor of Mathematics at Lund University/LTH in 1999.

The environment in which he has worked all of his professional life, the Department of Mathematics at Lund University, is famous for having educated and hosted several very prominent mathematicians. It has a long tradition with excellence in particular in partial differential equations and functional analysis. From the start of LTH, a separate division was created to serve LTH with undergraduate teaching and research. The first professor in mathematics at LTH was Jaak Peetre, assigned in 1963. During the first decades, the two divisions of the department, belonging to LTH and the Science Faculty, respectively, developed in parallel and were very similar in structure and research topics. Successively, however, the LTH division broadened its scope, both in education and research, to include more engineering oriented topics, besides the traditional ones. It got a clear identity as belonging to a technical university. Thus, today it constitutes the fundament of an assigned engineering program, Engineering Mathematics, and rather unique research profiles in engineering mathematics, with entrepreneurship manifested e.g. by several spin-off companies. The transition, to a large extent driven by Gunnar, will be described below.

1.3 First Research Period: Interpolation Spaces, 1966–circa 1975

When having finished his MSc in Engineering Physics in 1966, Gunnar started to work for a PhD in mathematics. His advisor was Jaak Peetre and the subject was interpolation theory for Banach spaces, a subtopic of functional analysis. Jaak then was a rather new professor, in fact the youngest in Sweden when appointed, but was already advisor of several PhD students. He had succeeded to form a group with very positive climate and much cooperation. Among those who like Gunnar worked in interpolation theory were Tord Holmstedt, Jörgen Löfström, Jöran Bergh, and later Per Nilsson and Björn Jawerth. In the vicinity worked Annika Haaker, who later became Gunnar's wife. Many researchers came to visit Jaak and Lund.

The thesis project of Gunnar dealt with the generalization of interpolation theory from couples of spaces to n-tuples. He defended his now famous PhD thesis *Interpolation of several Banach spaces* in 1972. Famous because this was one starting point of a new development concerning interpolation between three or more Banach spaces. This thesis has had a great impact of the further development of the interpolation theory.

By that time, 1972, Gunnar had together with Jaak already prepared another paper, *Interpolation of normed Abelian groups*, much cited. Here Gunnar came up with the original idea of adjoining to the scale of quasi-Banach spaces L_p, $0 < p \leq \infty$, a space L_0, with a metric (not norm) defined by the measure of the support of f. He made the observation that all ordinary L_p spaces, as well as Lorentz spaces L_{pq}, could be obtained from L_0 and L_∞ by an interpolation construction. There was an analogous situation for operators, with L_0 defined by having "rank" as metric. Jaak and Gunnar realized that these cases were instances of a more general framework. Together they built such in the paper, where also several other concrete cases were studied.

This second work was then succeeded by another joint paper with Jaak, *Interpolation and non-commutative integration*, published in 1975. In this, the two cases of L_0 mentioned above were further united within a framework of gauge spaces. Also the ideas and results in these works are nowadays fundamental for the further development both of interpolation theory itself and also of some related questions concerning function spaces and mathematical physics.

After this, Gunnar devoted a couple of years to the problem of characterizing interpolation spaces for couples of weighted L_p-spaces. At that time, this was a hot problem in the field, solved only in a few special cases. Gunnar succeeded to make a general characterization in an exhaustive and deep paper, *Interpolation of weighted L_p-spaces*. For instance, it was proved that A is an interpolation space with respect to $\{L_{p_0,w_0}, L_{p_1,w_1}\}$ (weighted L_p-spaces) if and only if

$$f \in A,\ K(t,g) \leq K(t,f) \text{ for } t > 0 \Longrightarrow g \in A,\ \|g\|_A \leq C\,\|f\|_A \text{ for some } C,$$

where

$$K(t, f) = \inf_{f=f_0+f_1} \|f_0\|_{p_0,w_0} + t\|f_1\|_{p_1,w_1}.$$

Crucial ingredients in this theory were some matrix lemmas, generalizing a majorizing theorem (inequality) of Hardy, Littlewood and Polya on doubly stochastic matrices and rearrangements. Here it is worth to mention that independently, and almost simultaneously, Professor Michael Cwikel solved the same problem by a different method. Michael visited Lund at this time, and a lasting relation with both Jaak and Gunnar started. As a curiosity can be mentioned that at the same visit, Michael got aware of another overlap of his own work, namely a paper of Annika Sparr on conjugate spaces of Lorentz spaces (written 1970, unpublished for a long time, but finally published in the volume of Contemporary Mathematics 2007 devoted to Michael on his 61th birthday). Also these first results and ideas concerning K-monotonicity has had a great impact on the further development in interpolation theory and many wonderful results have been proved and disproved in more general situations.

After the L_p-paper, Gunnar started to work on an open question in interpolation theory, raised by Foias and Lions in a classical paper: Characterize all positive functions h on the positive real axis such that, for linear operators T,

$$\text{if } \|Tf\|_p \le \|f\|_p,\ \|Tf\|_{p,w} \le \|f\|_{p,w} \text{ all } f, \text{ then } \|Tf\|_{p,h(w)} \le \|f\|_{p,h(w)} \text{ all } f.$$

Since long, it was known that the two cases $p = 1$ and $p = 2$ are essentially equivalent to the theorem of Hardy-Littlewood-Polya and that of Löwner on monotone matrix functions, respectively. A theorem in the general case thus has these two non-trivial theorems as special cases. To attack this, Gunnar found a new and original proof for Löwner's theorem, with potential to generalize. Even if he didn't succeed fully in this, his paper *A new proof of Löwner's theorem on monotone matrix functions*, published in 1980, contains ideas that have initiated further studies by several authors, e.g. Sergei Silvestrov, Hiroyuki Osaka and Jun Tomiyama.

1.4 First Teaching Period, 1972–circa 1990

During the 1970s, as described above, after his PhD thesis defence Gunnar went into an intense research period. However, gradually teaching duties took over, both in volume and devotion. As so many others, he found it impossible to combine full time work in education with own serious research. He decided to go in for teaching full-heartedly, and involved himself more and more in teaching and teaching related tasks. During this period he also wrote an undergraduate text-book in linear algebra, still in use.

Around 1980, an initiative to modernize the system-oriented courses at LTH was raised by the educational board. The mathematician Sven Spanne was given a central role in this, and was the main responsible for adjusting the mathematics courses to the curricula of the engineering programs. Often courses in mathematics

at LTH, and other technical universities, were looked upon as poor-man's versions of courses at the Science Faculty, but here they got an identity of their own. The students learned to use rather advanced mathematical topics, fitting to their surrounding courses. Variants of courses in *Linear systems*, dealing with the mathematical foundations of system theory, played prominent roles in different study programs, providing both skills and motivation. Lund came to play a pioneer role in Sweden in this respect. Gunnar took a very active role in this, working close to Sven Spanne.

In particular, in 1983 Gunnar got the responsibility for a course in partial differential equations, called *Continuous Systems*. This was, and still is, the final course in the mathematics chain for the mathematically oriented programs. The ambition was to cover and integrate all the three links in the chain of engineering/physical modelling – mathematical analysis of the model equations – physical interpretation of their solutions, while maintaining a firm mathematical basis. To this end, Gunnar wrote a text-book (later revised together with Annika), covering the scope without flinching for abstraction and strict mathematical arguments. For the first time, elements of computing were introduced in the mathematics courses. The course has survived over the years, and is well respected and well appreciated, despite a well-deserved reputation of being difficult. The book is also used at other universities.

Gunnar was also heavily engaged in other upper level undergraduate courses. Over the years, he has taken initiative to and developed new and engineering oriented advanced courses in e.g. Matrix theory, Optimization, Analytic functions. Later, after having started research in image analysis, he took the initiative to and developed a basic course in image analysis. This turned out to be particularly fruitful in the building up of a research group, as described in Sect. 1.7 below.

Summing up, I claim that Gunnar and Sven here initiated and developed ideas and material for teaching of mathematics for engineers, which had and still has a great impact of such teaching at many technical universities in Sweden (including my main working place Luleå University of Technology).

1.5 Search for New Directions: Platform Building, 1985–circa 1995

In several respects, the course in Continuous Systems became a watershed for Gunnar (and indirectly for the division). In the way it was designed, applications were very visible, as well as the usefulness of mathematical theory. The course brought him into contact with very clever students, eager to learn advanced mathematics and to apply the knowledge to real world problems.

By the mid-1980s, these students were at a stage where they looked for ideas for master projects. At this time master projects in mathematics were rare. In an effort to keep these talented and well motivated students within mathematics, and

to broaden the scope of the activities of the LTH division, Gunnar contacted former fellow students, working in the industry, to get ideas for master projects. Projects on real world problems were formulated, with mathematics in the centre, in diverse fields of applications: Electronics, control theory, medicine, etc. Some examples of early industrial collaboration partners were RIFA in Stockholm, ASEA in Västerås, and Ericsson company in Stockholm. One name to be particularly mentioned in this context is Gunnar Björklund, then at the electronics company RIFA, later Ericsson Components, Stockholm.

Gunnar also had much interaction with other departments at LTH and Lund University, e.g. Automatic Control, Industrial Automation, Electrical Measurements, Applied Electronics, various Medical departments, ... Particularly important to him over the years has been Professor Karl Johan Åström at the Department of Automatic Control.

From the late 1980s, these contacts resulted in a large number of master projects. One of the first was Tord Wingren (then surname Nilsson). In a joint project with RIFA, dealing with laser trimming of thick film resistors, a very efficient algorithm was developed, possible to use in the production to speed up the process and make it more reliable. Crucial in this was an ingenious idea of Tord that made it possible to use rapidly converging series from elliptic functions for computations of conformal mappings. This was in fact the start of an extraordinary industrial carrier for Tord, first at RIFA, and later as managing director at Ericsson Mobile Platform (EMP), and still later a similar position at Samsung Europe. Tord remained a big friend of mathematics, and has e.g. played a crucial role on the funding side in the realization of the Sonja Kovalevsky days for Swedish high school students, see Sect. 1.12 below.

Another important master project was one stemming from Ericsson, where Johan Helgesson developed semi-analytic methods for the solution of heat conduction problems for electronic components. In particular, it enabled an efficient treatment of the "hot-spots" that are critical in the design of such devices.

Over the years, Gunnar has successfully supervised more than 50 master projects, to a large extent jointly with industry and other sciences. Gradually, these have got more and more directed on image applications. Many of these master's theses have kept a very high standard. Already from the beginning, when ASEA (later ABB) by the end of the 1980s during 5 years founded a prize for the best master's thesis at LTH, work supervised by Gunnar received the prize three times (Rantzer, Helgesson and Diehl). In this category of top class master's theses can also be mentioned Gunnar's student Kalle Åström, (son of Karl Johan) who somewhat later won the Swedish competition "Innovation Cup", with a work on laser guided vehicles.

The experiences from these projects came to play an important role for Gunnar. Through them, he came into contact with industrial partners and began to build an understanding of their conditions and constraints. He developed a conviction that mathematics could and should play a more active role in engineering practice and education. The role of a mathematics department at a technical university should not only be to transfer tools and knowledge from one generation to the next, and to serve as consultants to other fields. More generally, a vision of "mathematics as

technology" began to grow, which later developed into a new study programme in Engineering Mathematics, see Sect. 1.11 below.

The network of industrial and academic partners, that Gunnar started to build during this period, then grew through him and his successors. Today it has grown into a for mathematics departments unusual size. This is depicted e.g. by a large number of master projects each year, ten to twenty, jointly with industry and organizations.

In 1987, Gunnar attended the SIAM conference ICIAM'87 in Paris, with a contribution *Applications of elliptic functions in microelectronics*, describing joint works with Tord Wingren and Johan Helgesson. This became a new turning point to Gunnar, and led him back to research, but in other directions than earlier. (SIAM = Society for Industrial and Applied Mathematics.)

1.6 Engineering Mathematics as a Research Direction, 1985–Present

By the end of the 1980s, the successful experiences from the master projects showed that there was a large group of engineering students who would welcome a possibility to continue with PhD-studies in mathematics, but with an engineering profile. The then existing PhD study program in mathematics was not perfectly suited, partly because of the direction, but also because students from the engineering faculty didn't fit well to the prerequisite claims. In particular, it was difficult for them to compete with science students on vacant positions.

In this situation, rather than to try to adapt to the study plan at hand, Gunnar took the initiative to create and implement a new application oriented study plan that could live in parallel to the existing one. This was done in steps, utilizing the fact that in the Swedish system there is an intermediate Licentiate degree, about halfway to the PhD. A new study plan for this degree was developed, and was accepted in 1988. It was more flexible in the prerequisites but had about the same end claims on courses, and more freedom in the thesis work.

The first student to finalize such a Licentiate degree was Anders Rantzer (today he is professor of Automatic Control in Lund). His research project dealt with complex analysis and algebra in theoretical control and was done in collaboration with Karl Johan Åström at the control department. For his thesis, Anders was awarded the SIAM prize for "Best Student Paper" 1989.

Another early Licentiate thesis was done by Peter Juhlin. In a collaboration with the Department of Clinical Physiology he studied potential problems in electrocardiography (fitting well to his parallel background as medical doctor). Peter then went to USA, where he got a PhD in biomedical engineering.

A sequence of successfully completed Licentiate degrees followed, and the need for a continuation to a full PhD came up. To meet this, a study plan for a PhD in mathematics, directed on applied mathematics, was created. It was accepted in 1994.

In this, also mathematical modelling was counted for, with claims also on courses from neighbouring sciences. Moreover, publishing in a wide range of journals was accepted.

The first to finalize a PhD according to this concept was Stefan Diehl, with a continuation of his Licentiate thesis. This dealt with the modelling and analysis of sedimentation in waste water treatment by means of conservation laws, in a collaboration with Professor Gustaf Olsson, Department of Industrial Automation. Stefan has then continued in this problem area, and today he leads an active and successful research direction at the department.

Research work also continued along other lines, aiming at mathematical modelling of a variety of technical processes. Often this landed in applications of partial differential equations. One important problem area, already mentioned, was the thermal properties of electronic components and circuits. Work continued with funding from NMP, the Swedish National Program for Microelectronics, in collaboration with Ericsson and Sven Mattisson at the Department of Applied Electronics, LTH, and later Ericsson EMP. A kernel for a simulation program for the temperature distribution in electronic components for use at Ericsson was developed. Another project, jointly with ASEA, Västerås, dealt with models for periodic pulse control of asynchronous engines. This work was supported by ITM, Institute of Applied Mathematics.

Another important research area has been tomography, motivated by medical applications. Collaboration partners here have been Professors Kjell Lindström and Hans W. Persson at the Department of Electrical Measurements. This department has played a leading role in the development of ultrasound techniques in medicine. This time the goal was to use Doppler techniques in mammography. Mathematically this led to an attractive and rich problem setting, given the name of vector tomography. Here Peter Juhlin made an early important theoretical contribution.

Vector tomography later became the topic of the PhD theses of Kent Stråhlén and Fredrik Andersson (partly). The latter developed a sharp tool, the moment transform, casting much light on the core analytical problems. This work is related both to partial differential equations and image analysis, as is also the case with a lot of other topics, which will be described in the next section.

Inspired by the projects and students he supervised, Gunnar resumed own research, often in collaboration with them. As can be seen from the topics, they were led to make bold excursions into a wide range of mathematics. Still the projects turned out to land happily, and they attracted very clever students. In this work, Gunnar was of course helped by the solid and broad training in pure mathematics PhD-studies in Lund had given him, but also by his engineering degree.

At that time (and maybe still), this kind of work didn't give much academic credits in the mathematical community. Despite this, Gunnar felt so attracted by working in this way that he decided to continue, even to the cost of a possible academic carrier. It is noteworthy that much later this kind of bridging research became academically fashionable, and even encouraged. At that stage, in 1999, Gunnar was among the first to be promoted to professor at LTH.

1.7 Mathematical Imaging Group, 1990–Present

In 1985, Gunnar was asked by Karl Johan Åström to be a member of the committee of the PhD thesis defence of Lars Nielsen in Automatic Control. The title of the thesis was *Simplifications in Visual Servoing*. Lars demonstrated the possibility to design a control system for robots, involving the use of visual information to identify and move to objects in the scene. In particular, Lars had developed a system for marking of objects in terms of concentric triangles, making it possible to discriminate between different objects.

The heart of the matter was the discovery of a relation between certain areas in the markings, which is invariant under perspective mappings. Lars had found this by clever, but cumbersome, computing by a computer algebra system, at that time a new facility.

Gunnar succeeded to make a simple proof, possible to generalize to other configurations. This was the starting point of a fruitful collaboration on projective area invariants between Gunnar and Lars. (Lars is today a professor of Vehicular Systems at Linköping University.)

Another important contact in imaging at about the same time was Helmut Hertz, professor at the Department of Electrical Measurements. He is perhaps most famous for being a pioneer in the development of ultra-sound techniques in cardiology. Another major achievement of his is the ink-jet printer. In the 1980s, the possibility to use these for high quality reproduction of digital images had been opened. Helmut anticipated the future need for "visual" interpolation methods for digital images, that preserve edges and borders, instead of smoothing them. By Helmut, Gunnar was early led into this problem, which belongs to a problem area that still is very active (including subfields like anisotropic diffusion and level set methods).

These two contacts led Gunnar into the field of computer vision, and he started to study the subject seriously. This is a relatively new field of research, in need of seemingly unlimited amounts of mathematics and mathematical modelling. In computer vision one is concerned with the extraction and interpretation of high level information from images, like pattern recognition, motion, navigation, reconstruction, and shape. Often a guiding goal is to mimic the corresponding capacities in biological vision. Computer vision has a very rich field of applications, e.g. in robotics, medicine, telecommunication, etc. The practicians mostly come from computer science or electrical engineering, sometimes also physics, but more seldom mathematics.

A system for computer vision often contains low-level image processing components for feature extraction and representation, realized by different kinds of filtering. Very often this is the crucial step, for instance in many medical applications.

Scientifically, computer vision may be a strange bird at a mathematics department. Aware of this, already from the beginning, Gunnar was very active (and successful) in finding external funding, in order not to obstruct the core activities of the department. Luckily enough, the research agency STU, Swedish National

Board for Technological Development, at this time had started a framework program in computerized image analysis. Together with Lars Nielsen he wrote a proposal around the two applications mentioned above. Despite the fact that mathematics was far away from STU's ordinary action area, the proposal was approved. The courage of the committee, with John Graffman as handling officer, to back up this new actor should be appreciated. Much later it turned out to be well in line with the commissions of STU, since this came to be the triggering factor for the Mathematical Imaging Group, which later has played an important role for the development of image-based industry in the region.

The STU-project lasted for several years, and Gunnar got successively larger funding. This made it possible to hire a PhD-student and to start the building up of a small image laboratory at the mathematics department. The STU program also played an important role in establishing a network of Swedish researchers in computer vision and image analysis.

At this time, a hot research topic in computer vision was invariancy, notably in geometry and algebra, but also in contexts related to cognition. The work of Gunnar and Lars on projective area-invariants was well in line with this, and was recognized by the leaders in the field. In addition, Gunnar had invented the concept of affine shape, as a sharp tool to construct and compute with invariants, see Sect. 1.8 below. The concept was successively developed and presented at the European conferences in computer vision, ECCV, and was met with a lot of interest and appreciation. Thus at the conference ECCV in 1992, Gunnar's paper was selected among the few best in a separate volume. Further, Gunnar and Lars were invited to the trendsetting Reykavik meeting in 1991, organized by the research organisations ESPRIT of EC and DARPA of US. Besides established researchers like Faugeras, Eklundh, Mohr, Mundy, there were a number of younger scientists who later became eminent researchers in the field, Zisserman, van Gool, Ponce, Forsyth, ...

By this, Lund was placed on the computer vision map. Together with the leading European groups, Gunnar prepared an application to the third framework program of EC. The project was called VIVA, Viewpoint Invariant Visual Acquisition, and got accepted. The key objective was to construct and implement invariants under specific transformations, and to use them for recognition. Some of the partners in VIVA were INRIA (France), the universities in Oxford and Leuven, KTH, and the company General Electric (GEC). The project was running for the years 1993–1996, and is generally considered to have played an important role for the development of the field.

The STU-project and VIVA enabled the hiring of PhD-students and further acquiring of equipments. It enabled big steps to be taken to form what later became the Mathematical Imaging Group, MIG. An important factor in this was also that Gunnar at that time started to give an undergraduate course in image analysis, then a missing theme at LTH. To begin with it was met with some suspicion by representatives of the study programs, who thought mathematics was a wrong place for such an engineering oriented topic. However, Gunnar persisted in giving the course, which attracted many students, and became popular. After a few years the

course got a place in the study programs. It has since then provided important professional tools to a large number of LTH-engineers today working with images.

Much of the development of the group was driven by the PhD-education. Among the first to complete a PhD were Anders Heyden and Kalle Åström. Their theses attracted much attention, dealing with projective and algebraic geometry and invariant theory. Anders got an award for best student paper 1995 at the leading international conference ICCV, International Conference on Computer Vision, and Kalle's thesis was appointed the best thesis in image analysis in the Nordic countries in 1997.

Of vital importance was the fact that when Anders and Kalle had finished their PhDs, it was possible to keep them within the group. This was enabled by funding from two new projects, that were on their way to be accepted, Dynamic Vision from the Swedish research council TFR, and Cumuli from EC. An important role for the consolidation of the group was at the same time played by the project VISIT from SSF, the Swedish Foundation for Strategic Research.

Of these projects, Cumuli, Computational Understanding of Multiple Images and Applications, was running 1996–2000. It built on the advances made e.g. in VIVA on projective multi-image geometry, and was directed on applications in accurate 3D industrial measurements, high speed motion analysis, and augmented reality. Among the partners were INRIA from France, the Fraunhofer institute from Germany, and the Swedish company I.V. Image Systems. The project Dynamic Vision from TFR, 1995–1999, dealt with vision in feedback loops, trying to bridge computer vision and control. The goal was to use the new theoretical results to make possible for a robot to perform tasks like motion planning, grasping, obstacle detection and avoidance. The project VISIT was a broad one, driven together with other Swedish groups in the field, and played an important role for the cooperation among these.

At around 2000, the Mathematical Imaging Group was well established. Anders Heyden and Kalle Åström quickly had built up good reputations and soon were established independent researchers. MIG was the largest research group at the Centre for Mathematical Sciences. All the time it has lived on numerous external projects with funding from EC, SSF, TFR, Vinnova, VR, and commissioned research projects. A large number of clever students have got their PhDs in the group. Among them Fredrik Kahl stands out, who under the last few years has caught huge prestigious funding from the European Research Council ERC, in their "Starting Grant" program, and the Swedish SSF, in their "Research Leader of the Future" program. Fredrik also was awarded the prestigious Marr Prize at the international conference ICCV in 2005.

As has been described above, from the beginning the two main scientific themes for MIG were geometry and invariancy. Gradually, the profile then broadened, and today large parts of computer vision and image processing are covered. Roughly speaking, the research can be divided into geometric computer vision, cognitive vision and medical image analysis. A speciality has become applications of optimization in computer vision, through Fredrik Kahl. Today (spring 2010)

the group has more than 20 members, among them four professors, five associate professors, and around 15 post-docs and PhD-students.

A central business idea of MIG has all the time been to take advantage of being located at a mathematics department. As has been remarked earlier, in this respect it differs from most other research groups in the field, who in general come from other sciences. Computer vision and image analysis are in need of a lot of mathematical modelling and tools from a wide range of mathematics, e.g. projective, algebraic and differential geometry, linear algebra, partial and ordinary differential equations, optimisation, harmonic analysis, integral transforms, inverse problems, probability theory and numerical analysis. To have a broad such background, and the access to experts, provide an enormous advantage when working in the field.

Today the Mathematical Imaging Group constitutes the largest competence resource in image processing and image analysis at Lund University. It is active in numerous collaborations with companies and academic departments in the region. Members of the group have played crucial roles in a number of spin-off companies, e.g. Decuma, Cognimatics, Polar Rose, WeAidU, Ludesi, Ange Optimization, and are involved in several others, e.g. by sharing patents in Cellavision, Precise Biometrics, Danaher Motion .... A particular relation has been established with Axis Communications, world-leading manufacturer of network cameras, through the founder Mikael Karlsson and the R&D manager Daniel Elvin (by the way also he a former master student of Gunnar).

1.8 Second Research Period, 1990–Present

While the previous sections mainly have dealt with the building of an important group, initiated and to a large extent developed by Gunnar, this section will be more directed on Gunnar's own research in this area. As described above, when entering the field of computer vision, Gunnar's main research topics were geometry and invariance. Rather soon, he established contacts and collaborations with several of the leading researchers of the field. Very inspiring to him were the contacts with Professor Jaan Koenderink, Utrecht, with his ability to work across all science borders and see core structures. Very influential to Gunnar were also the Rosenön workshops, organized by Professor Jan-Olof Eklundh, KTH, which he attended several times.

Much of Gunnar's work the first years centred on the concept of affine shape, which he had introduced. In this, the guiding star was to find quantitative representations of geometric configurations that are independent of the underlying coordinate representations (to be compared to the analogous claim in physics). Stated in another way, this means to find representations that are invariant under affine (or other classes of) transformations. For point-configurations $\mathscr{X}$ consisting of m points $X^1, \ldots, X^m$ in n dimensions, it can be proven that the linear space

$$s(\mathscr{X}) = \left\{ \xi \mid \sum_1^m \xi_k X^k = 0 \text{ with } \sum_1^m \xi_k = 0 \right\},$$

where X^k stands for coordinates in an arbitrary affine system, has the invariancy properties wanted. The linear space $s(\mathscr{X})$ is called the *affine shape* of $\mathscr{X}$. Thus there is a one-to-one correspondence between point configurations and linear subspaces of $\mathbf{R}^m$, or equivalently, point configurations are represented by points in a Grassman manifold.

The concept of affine shape was developed successively in a series of papers, at several times presented at ECCV-conferences. It was shown to be an efficient and elegant tool for some of the basic problems of geometric computer vision. One of these, recognition, was discussed above in the context of VIVA. Another one, the structure and motion problem, deals with the recovery of the scene structure and the camera locations from a sequence of images. In this context, Gunnar was the first to introduce iterative methods, which later became very popular.

Most approaches to these problems are based on point features. However, within the framework of affine shape also more general configurations could be treated, e.g. curves. These were to play a fundamental role in the recognition problems of the next section, about Decuma.

Over the years, with an open mind and big contact network, Gunnar has been involved in a large number of collaborations. Here I only include a few examples of such contacts:

- In medicine, Gunnar and Professor Håkan Arheden from the Department of Clinical Physiology started a collaboration between the departments, aiming at developing diagnostic tools in cardiology from MR-images. Here segmentation is often a critical step. This has so far resulted in one PhD in mathematics, a few ongoing PhD-projects, and several master projects, and also a medical course for engineering students. In a wider context, Gunnar has been active in the planning group for a centre for bio-imaging in Lund, led by Professor Deniz Kirik.
- In another bio-related project, supported by SSF, Professor Dan Nilsson from the Biological Vision group and Gunnar have started a collaboration, aiming at using features of early visual processing of animals for applications in machine vision.
- Still another project, together with the Department of Ophtalmology at Malmö University Hospital, supported by VR, aimed at tools for detection of glaucoma from deflections in the nerve fibre layer of the retina, measured by optical coherence tomography. This led to a model in terms of a differential equation of eikonal type.
- Several projects and publications have dealt with image processing. In a collaboration with Professor Clemens Kaminsky, Department of Combustion Physics, non-linear diffusion filtering methods were developed for images obtained by planar laser induced flouescence spectroscopy. An EC-project Floceye, together with a company LR-Miljö, gave contributions to a system for image based control of drum thickeners in waste water treatment. A joint project with Stora-Enso AB, Karlstad, dealt with modelling and analysis of complex microstructures in paper. Tools for automated biometric systems for person identification and verification were developed in a TFR-project. An early paper in image processing used potential theory for edge preserving interpolation.

- Something completely different was a cooperation with the Swedish artist and art historian Oscar Reutersvärd, famous as inventor of "impossible pictures". The concept of affine shape is well suited to analyze such pictures, which also have some relevance in machine vision, as degenerate cases in reconstruction. (In fact, this was the subject of Anders Heyden's Licentiate thesis.) Unfortunately, because of the decease of Reutersvärd, the cooperation ended before the plans of a joint publications were realized, and was only manifested by a joint seminar at Lund Mathematical Society in 2000.

Finally, I want to mention that Gunnar's own research has several times been interlaced with the work of some PhD-student. Over the years he has been advisor or co-advisor for 14 students to a PhD, and for additional five students to a Licentiate's degree. The former group consists of Stefan Diehl, Anders Heyden, Kalle Åström, Rikard Berthilsson, Kent Stråhlen, Fredrik Kahl, Magnus Oscarsson, Björn Johansson, Henrik Malm, Henrik Stewenius, Anders Eriksson, Fredrik Andersson, Erik Bergvall, and Johan Öinert. The group of Licentiates consists of Anders Rantzer, Peter Juhlin, Mattias Nilsson, Oscar Wigelius and Charlotte Svensson. Of these 19 degrees, by today four have become professors (Rantzer, Heyden, Åström, and Kahl) and 11 work in industry (which also is a telling figure).

1.9 Entrepreneurship: Decuma, 1998–2004

An experience that significantly influenced Gunnar's later work was being a co-founder of the spin-off company Decuma. Through this he got hands-on experiences of engineering mathematics in real life.

By the end of the 1990s, Gunnar was contacted by Ericsson, through his former student Tord Wingren, about a possible collaboration project. The company saw a future need for input alternatives to keyboards for mobile devices. In particular, they were interested in transferring handwritten text on a pressure sensitive screen to ascii-code. Such techniques had appeared on the market, e.g. for pocket computers, so called "Personal Digital Assistants", PDAs. However, the technology was not very well developed.

An obvious complication in handwriting recognition is the fact that the appearance of a written letter varies from person to person, and even over time for a fixed person. By modelling such variations by some class of transformations, e.g. affine or similarities, Gunnar realized that the invariants he and others had studied could be used as identifiers of the different characters. It fitted well to the work of the PhD student Rikard Berthilsson, who studied invariancy properties of 3D curves in terms of affine shape. He already had results and algorithms, which when adjusted to the planar problem at hand, gave promising results.

Encouraged by Ericsson, they prepared a patent application around the mathematical concepts of "shape" and "proximity measures". The patent was approved in 1999. Around this, the joint-stock company Decuma AB was then founded by Gun-

nar, Rikard, and Kalle Åström. By a grant from Idéon Science Park, it got a small office. A first employee was hired in the beginning of 2000. A big and principally important step was taken late spring 2000, when a "professional" managing director was hired. The founders could now concentrate more on research and development.

The company needed financing. In a first round in the beginning of 2000 some "business angels" came in. A big step was taken later in 2000, when Mikael Karlsson, founder and director of the telecom company Axis Communications, came in through his "private" venture capital company Visionalis. He also became a member of the board, where he added a lot of professional knowledge and valuable contacts. The company grew, from the beginning mainly on the R&D side. After a while, with eight PhD in mathematical sciences, Decuma used to present itself as one of the biggest industrial employers in Sweden of doctors in mathematics.

Having started with Latin letters, the product portfolio was expanded to Asiatic ones, Japanese and Chinese. This step was of course very challenging, already because of the number of characters and their complexity. On the other hand, this same would make the value of a technique avoiding keyboard still larger, and big markets would be opened. To succeed in this, it was not sufficient to have only engineering competences. To help this, the linguist Magnus Nordenhake and a few persons with Asian native languages were hired. Assisted by them, the Decuma mathematicians Anders Holtsberg and Martin Lindberg developed recognizers for Chinese and Japanese that turned out to be very competitive, both what concerned hit-rate and speed.

The company grew also on the marketing side, and at some stage there were 27 employees in total. The products of Decuma were met with big interest in the IT-world. A first contract was signed for digital pens with the neighbour company in Lund, Anoto AB. It was followed by contracts for handheld computers with Casio (Latin) and HP (Japanese). In particular, the latter fact, that a small Swedish company was chosen by the multi-national company HP to interpret Japanese signs, was given much attention in professional circles.

The company and its founders also received several awards. Already in 1999, before the company was founded, the founders were given an award in the Swedish competition Innovation Cup. In 2000, Decuma was appointed "spin-off of the year" in the southern region by the Royal Academy of Sciences, KVA. In 2002, Rikard Berthilsson got the Chester Carlson award from the Academy of Engineering Sciences, IVA. (Chester Carlson was the founder of Xerox.) Decuma also attracted a lot of attention in Swedish press, with half- or full pages in the big newspapers and trade journals. Often they were fascinated by the fact that the founders came from mathematics (with headlines like "Swedish math geniuses teach the Japaneses to write"). In Time Magazine, December 2001, Decuma was appointed "Start-up of the week".

For a period, Decuma was kind of a figure head for Idéon, who often directed their prominent visitors to the company. The fact that the company made use of such a broad range of academic competences, in particular mathematics, was also very much liked by the university and LTH. Thus Gunnar was on several occasions

Fig. 1.2 From the ceremony in Copenhagen, where Decuma received the IST Grand Prize 2003. In the front line Thierre Breton, Director of France Telecom, and chairman of the jury, Roger Larsson, Anders Berglund and Gunnar Sparr from Decuma. Behind them Erkki Liikanen, EU Commissioner for Enterprise & Information Society, and Helge Sander, Danish Minister for Science, Technology and Innovation

called in by the vice-chancellor and deans to present the success story of Decuma to their official visitors from industry, politics, and administration.

The peak of the Decuma story was reached when the company received the EC Information Society Technologies (IST) Grand Prize 2003 for "groundbreaking products and services that represent the best of European innovation in information society technologies". This event is organised by Euro-Case, a collective of European Academies of Engineering Sciences (and is sometimes called the "European championship for IT enterprises"). Besides the prestige, it also gave a significant amount of money, 200,000 Euro, well needed (Fig. 1.2).

In fact, Decuma had constantly been accompanied by a problem: The company didn't get enough income. These were the days of the collapsing IT-bubble, and a difficult time to survive for a small company like Decuma, with a niche product for expanding markets. Already in 2001, new investors had been needed, and found in one Danish and one Swedish venture capital company. Unfortunately, these were very business oriented and came in with very short time perspectives. For instance, the R&D was substantially diminished, which of course was a big disaster for a research oriented company like Decuma.

The situation was helped a little by the IST prize, which gave a lot of favourable attention. Already before that, however, contacts had been established with Sony Inc., and now they developed rapidly. Contracts were signed making Decuma sole supplier for Sony's new PDA-device Clié, for recognition of all kinds of characters, Latin, Chinese and Japanese. Sony had big plans for Clié, in that from an already large market share for PDAs in USA, they aimed at making Clié the market leader.

Apparently Sony had still wider interests in Decuma. Thus in 2003 they went in with a large amount of money and became the biggest owner. In the same round, the Swedish governmental investment agency Industrifonden came in with an equal amount. Sony even talked about building around Decuma a "centre of excellency" within the corporation, dealing with a wide span of R&D, not only handwriting recognition. Here certainly the academic and mathematical flavour of Decuma played an important role.

However, in the global IT-decline, even Sony got into trouble. Strategic, global decisions were taken, resulting in big reorganisations. Despite what seemed to be in good progress, Sony decided to shut down their PDA division, including Clié. Decuma's services were no longer needed to them, and they lost their engagement.

By this, Decuma lost its pace as an independent company. It was difficult to make a new start, and catch new investors. Decuma was for sale, and in 2004 it was bought by the Canadian Zi-Corporation. Decuma's products became part of a bigger portfolio of mobile services. This made it easier to interact with the big players on the market, and new contracts were signed. In 2009 another step was taken, in that Zi-Corporation was bought by Nuance, who presents itself as the "leading provider of speech and imaging solutions for businesses and consumers around the world". Also Nuance has kept an office around Decuma in Lund. The development of Decuma thus has followed an industrial logic, and it must be seen as a success that the company has survived these difficult times, and even attracted big companies to establish themselves in the region.

This is thus the context in which the products of Decuma live today, being further developed and saled on a global market by a major player. Even if the founders no longer are on the train, they may feel a satisfaction to know that their efforts today runs on or even are engraved in the hardware of mobile devices on the global mass-market. At the same time, the development of Decuma should be seen as a proof of utility for mathematics and mathematically based industrial products. The experiences from Decuma were strongly guiding for Gunnar in the implementation of the Engineering Mathematics study programme, see Sect. 1.11 below.

An instructive experience from Decuma to Gunnar was that in industry you don't develop things just because it is possible. (This in contrast to academia, where to expand what is possible often is the driving force.) You may have the best product in the world, or feel able to develop it, but if no one asks for it, it deserves not necessarily to be done. These different criteria must be respected when working in engineering mathematics.

1.10 Special Adviser in a Murder Trial

A very special example of applied mathematics deserves a section of its own. It was about a forensic pattern investigation for a murder trial, getting a lot of attention in Swedish media.

The background was that an old lady was murdered in Stockholm in 1993. In a trial, one of her home helpers, Joy Rahman, was judged guilty and later imprisoned. A central role in the judgement was played by the murder tool, a textile tape, and its association to a particular Christmas tapestry. This tapestry was found in the home of another old lady in the neighbourhood, who also had Rahman as a helper.

The association between the textile tape and the tapestry was questioned, along with other components of the first judgement. A Swedish star lawyer, Peter Althin, managed to procure a new trial, which took place in 2002. (By then, Rahman thus had been imprisoned in 8 years.) In advance to the new process, Gunnar was contacted by the Swedish Prosecutor-General (Swedish: Riksåklagaren), through Agneta Blidberg (later Palme-prosecutor) about the possibility to investigate patterns on the tape and the tapestry.

Together with Fredrik Kahl and Kalle Åström, Gunnar accepted the challenge. They made use of marks in the glue layer of the tape caused by yarn patterns on the tapestry. A designed method was developed, measuring quantitatively the agreement between pattern pieces on the two objects. A very good agreement was found, where the probability that this agreement had arisen by chance could be estimated to less than 10^{-16}. (This should be compared to what is claimed for legal security in DNA tests, 10^{-6}.)

What became controversial in the new trial was the fact that the modelling also admitted the possibility of shrinking in different scales of the textile tape and the tapestry. Mathematically, this was modelled by affine transformations of the patterns. The possibility and magnitude of such shrinking was much questioned by the defence counsel Peter Althin, while the investigators claimed that the extremely low probability estimate in fact also gave evidence for that such a shrinking had occurred.

The Court of Appeal (Swedish: Svea Hovrätt) accepted the investigation, having in their final judgement a sharp formulation (in free translation): “The court finds that the investigation of Gunnar Sparr gives an unequivocal result, that leaves no room for interpretations or questionings” (Swedish: “Gunnar Sparrs undersökning ger enligt hovrättens mening ett entydigt resultat som inte lämnar utrymme för tolkningar eller ifrågasättanden”). We thus have a case, of which there are not so many, where the validity of mathematical arguments is legally established by a court. (In the new trial, despite this, Joy Rahman was acquitted, since the total body of evidence was considered too weak.) This excursion into public life was of course an unusual, but valuable, experience for a mathematician, where all sympathies of journalists and the general public seemed to be given to the accused.

1.11 Mathematical Education, 1998–Present

From the end of the 1990s, Gunnar entered a new direction, in parallel to work in research, teaching and Decuma. Again he went into education, but this time on the planning level, performing a sort of entrepreneurship within academia. He and his collaborators directed themselves at two extremes on the educational spectrum, *Advanced for Few*, and *Basic for Many*.

1.11.1 Forming of a New Study Program

During the 1990s, as described in another article in this volume, mathematics as technology had developed at LTH, both at the Centre for Mathematical Sciences (MC) and at other departments. At MC such a development could be seen not only at the division for mathematics, LTH, but also e.g. at Mathematical Statistics, especially through Professors Jan Holst and Georg Lindgren. MC was experienced in industrial and academic applications, and had good relations to neighbouring sciences. The soil for a designed education program in engineering mathematics was prepared.

In 1998, Gunnar took the initiative to a letter to the executives of LTH, motivating and suggesting a new study program. They gave a positive response, forwarding a task on further investigations to the educational board under Rune Kullberg. A working group was appointed, with Gunnar as chairman. This group also contained representatives from industry, in particular Bo Bernhardsson at Ericsson (also professor at the Department of Automatic Control), who over the years has been a very important speaking partner to Gunnar.

With Gunnar as the driving force, a plan for a new study program was made. Here he was guided by his experiences from industrial and other collaborations, described above. A firm claim was to form an engineering education. The curriculum should have breadth, reflecting the fact that mathematics is a universal instrument, and the practicing mathematical engineer should be able to overview and work in broad areas. Education should prepare to goal-oriented work, where mathematics is intertwined with engineering or other subjects (biology, economy, medicine, ...). By inside experiences through Decuma, Gunnar had seen the immense demand for mathematicians skilled in programming (or programmers trained in mathematics). The education should also foster attitudes, not to make way for engineering argument to get further and find solutions when theory lacks.

The program was planned also to give serious training in mathematical modelling and mathematical communication. Here, when later coming to the implementation phase, substantial contributions were made by Magnus Fontes, Kalle Åström, and Anders Rantzer.

Once the planning was made, a lot of lobby work followed, directed on education committees and boards. Some people were positive, and even enthusiastic, but many

were sceptical. These expressed doubts that the mathematical sciences could be the core of an engineering program. Here good arguments against could be found in the good records that MC and other mathematically oriented departments had built up.

After a thorough preparation, the idea was supported by the executives of LTH, Professors Gunilla Jönsson and Klas Malmqvist, and later also by the Vice-chancellor of Lund University. The decision to start the Engineering Mathematics program was taken in 2001.

The first students were admitted in 2002. In the first batch there were 30 students, after that around 40 per year. Gunnar was the program director 2001–2008, thus leading the planning and implementation of the whole curriculum He laid a lot of efforts and all his devotion on this, and was called "Papa Π" by the students. (Π is the program symbol of Engineering Mathematics.)

The program opens for a broad range of specializations: *Computations and simulations*, *Biological and medical modelling*, *Financial modelling*, *Environment, risk and climate*, *Signals, images and systems*. These got well-anchored through a process where a large number of colleagues from all over LTH were involved in different planning groups.

Up to 2009, around 50 students have finished their diploma, and have been very attractive on the job-market. Today they work over the whole spectrum of branches, depicted by the titles of specializations above. About a third of them have continued with PhD studies, also this in a broad range of sciences. In 2008, the idea of an Engineering Mathematics program was taken up by Chalmers in Gothenburg. This must be seen as a success for the concept, making it still more established.

1.11.2 Strengthening of Mathematics for All Programs at LTH

In Sweden, as well as in other countries, there has been an intense discussion about the decline in mathematical skills from high school. This is especially evident at technical universities, where the basis for later studies in engineering subjects risks to be weakened. At LTH, with Gunnar and the Director of Studies Lars-Christer Böiers as driving forces, it has been possible to convince the executives and educational boards to make a strengthening of the mathematics education for all programs, both what concern quantity and quality. In the realization, a number of new pedagogical initiatives were taken.

1.12 Administration, General Issues

On the department level, Gunnar has been chairman of the division for Mathematics at LTH almost all the time since the mid-1980s until his age retirement in 2007. He has taken very active part in the development and strengthening of the department.

On a national level, among others, Gunnar has been the chairman of the Swedish National Committee for Mathematics, a sub-committee of the Royal Swedish Academy of Sciences, for the period 2005–2008. Before that, he was a member of the committee since 1998. An important task for this committee is to disseminate an understanding of mathematics and its role in society, where Gunnar has been very engaged.

In particular, the arrangement "the Sonja Kovalevsky days" for high-school students should be mentioned. This is a yearly event with a large number of participants. It started in 2000, from an initiative of Gerd Brandell within the committee, strongly supported by the chairman Professor Christer Kiselman. Gunnar has played a heavy role in the realization, which was possible thanks to generous support for several years from Ericsson EMP in Lund. The planning and implementation of the days were for many years a dear concern of Gerd and Gunnar.

My final words: Thereby I hope that I have been able to give at least a flavour of the remarkable man Gunnar Sparr, and his contributions as mathematician, entrepreneur and ambassador for mathematics. Thank you Gunnar for everything you have done for our dear and fantastic subject mathematics and thank you for everything you have meant for me pers(s)onally and for my students. It is a great honour for me to be your friend and supporter.

Chapter 2
The Engineering Mathematics Study Programme in Lund: Background and Implementation

Georg Lindgren and Gunnar Sparr

Abstract Since 2002, an education programme Engineering Mathematics has existed in Lund, with the mathematical sciences providing the main engineering tools. As verified by the employments of those who have finished, the education has been successful in preparing for mathematical work in a wide range of branches, as well as for doctoral studies in a wide range of subjects. The article describes the background and the implementation of this programme.

2.1 Setting the Scene

Lund Institute of Technology (LTH), today's Faculty of Engineering within Lund University, was founded in 1961. Starting with Engineering Physics, during the subsequent years all the classical engineering programmes were established: Electrical Engineering, Mechanical Engineering, Civil Engineering, Chemical Engineering, and Architecture. During the last decade, a few programmes of new type have been introduced. One of them is Engineering Mathematics. In this article we give our personal view on its background and implementation.

2.1.1 Background

Already from the start of LTH, the mathematical sciences, Mathematics, Mathematical Statistics and Numerical Analysis, constituted a substantial part of the curriculum. Mathematics and Mathematical Statistics were represented by professors Jaak Peetre and Gunnar Blom, respectively, while Numerical

G. Lindgren (✉) · G. Sparr
Centre for Mathematical Sciences, LTH, Lund University, Box 118, Lund, SE-22100, Sweden
e-mail: georg@maths.lth.se; gunnar@maths.lth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_2,

Analysis was represented by the lecturer Torgil Ekman. When the programmes at LTH expanded with electrical, mechanical, civil, and chemical engineering, and architecture, strong positions of the mathematical courses were implemented almost unchanged, even if architects and chemists got some special treatment with separate courses.

One of the fundamental principles behind the creation of LTH was that it should have equal status as KTH and CTH, but that it should complement these in new areas of national interest, taking advantage of the nearness to Lund University. The engineering environment in which the mathematical sciences would work came to include several mathematically oriented subjects, like Automatic Control, Telecommunication Systems, Teletransmission Theory, Solid Mechanics, and Electromagnetic Fields. Representatives for these subjects were later to become instrumental for the creation of the Engineering Mathematics study programme.

The strong position on the educational side opened for early recruitment of doctoral students, most of them from the science faculty, working also as teaching assistants. During the 10 years period 1966–1976, there were 13 PhD-defenses in mathematics and mathematical statistics with advisors at LTH. Gunnar Blom, the professor in mathematical statistics, had worked in industry as a statistician and encouraged his students to work with problems from many different fields of probability and statistics. Also in that respect, LTH became "different" from KTH and CTH.

However different, the first decade of LTH was a period of defining the course packages and writing new course material and text books. In fact, at this time and before, Lund had taken the lead in Sweden in writing elementary text-books in mathematical sciences. If they look traditional today, one should have in mind that they to a large extent have formed the tradition.

2.1.2 First Steps Towards a New Role in Education

The scientific environment at LTH, in which the mathematical sciences worked, was very encouraging for the gradual change in the roles of these subjects that took place during the 1970s. Of course, we never realized, at that time, the future consequences of these changes, but in retrospect, it is easy to identify a few critical items.

The students soon realized that the mathematics courses didn't fit perfectly to the applied courses to follow. Common jokes among the first generations were about the fact that 'all' courses started with a mini-course in Laplace transforms, while these were too application oriented to fit in the mathematical courses. At the same time, several of the mathematically oriented engineering departments were very open and supportive in letting mathematics take a stronger responsibility in the curricula. To a large extent this depended on their trust in the abilities and overview of Sven Spanne, a lecturer in mathematics. Through own studies all over LTH and the university, he had an academic track record, difficult to beat. In order to modernize the Electrical Engineering and the (new) Computer Engineering programmes, Sven was asked

by the Educational Board to look over the courses in mathematics. To this end, he developed a course entitled "Linear Systems", establishing a common mathematical basis in transforms, distributions, systems of differential equations, for all system oriented sciences at LTH. At the same time, their own courses were appropriately modified. A new feature for the electrical engineering programme was that for a period the course in linear systems was synchronized with laborative work in a parallel course in electronics. A general opinion among students and teachers in the system area was that the new course in linear systems meant a tremendous gain for LTH.

This course became part of a revision where also the succeeding courses in mathematics and many related subjects were modified. In particular this was the case with the topic of partial differential equations, where courses could be found at different departments. Doubt was demonstrated at that time about the ability of mathematicians to make such a course, useful for the applications. At last a decision was made, and the task was given to Gunnar Sparr. High priority was given to make a course, called "Continuous systems", where not only the mathematical subject had priority, but also the physical backgrounds of the equations as well as the interpretation of solutions. The course came to generate a large number of master's theses, and collaboration projects with industry and other departments.

Many students wanted to study more mathematics. To meet this, a flexible system of more advanced courses developed. These courses were also very much attended by doctoral students from all over LTH, often strongly encouraged by their advisors. One could even talk about a "course on demand" system, where the subjects varied from year to year. Besides the traditional topics for engineering education, related to analysis and differential equations, also for the time more esoteric ones were represented, like algebra, graph theory, projective geometry, number theory. Some courses from this group developed to play prominent roles in different study programmes, e.g. courses in matrix theory and optimization. During these years, Jaak Peetre demonstrated much foresight, and introduced students who later became professors in different subjects with to them important professional tools.

In statistics, a new course in "Stationary Stochastic Processes" created a similar link as for linear systems to the systems oriented subjects. One of the elective courses in mathematical statistics was built around Gunnar Blom's experiences in operations research. These resulted in one compendium in "Operations Research" and one in "Stochastic Processes", containing mostly Markov processes theory and applications. But there was also one chapter devoted to stationary processes. By encouragement from Lars Holst in Uppsala, and Rolf Johannesson and Per Eriksson in Lund, Georg Lindgren expanded this, 1973, into a course in stationary processes aimed to give the statistical basis for stochastic control and signal processing. The course was introduced as compulsory for electrical engineers. Holger Rootzén gave the course a more engineering touch. The course was termed "Stokastiska processer, allmän kurs", abbreviated SPAK, a name that still remains with its descendent course. A follow up course, which dealt with time series analysis, was also introduced in the mid 1970s, with computer experiments on the department's pride, Das Gupta, a nickname for the computer HP 9830. Sven Spanne liked to

entertain himself on Gupta, and after some years computer labs were introduced also in linear and continuous systems.

The SPAK course included a few experimental lectures with real signal processing tools, noise generators, and linear filters, experiments which were performed at the signal processing department, even if the course was organized by mathematical statistics. This arrangement seems to have been quite unique for an engineering school, and it was often commented from international colleagues how strange it was to have an "electrical engineering" course taught by a mathematics department. For mathematical statistics it meant that generations of PhD students had to face some real engineering problems, and it laid the basis for much of the future development. When Computer Engineering was introduced among the LTH programmes in the mid 1980s, the SPAK course was made compulsory also there, quite naturally, in view of the growing Ericsson activities in Lund.

Numerical Analysis was not represented at professor's level in the original LTH plan. Research was considered covered by the presence of Carl-Erik Fröberg, professor in Information Sciences at the Science Faculty, and one of the creators of the electronic computing machine "Siffermaskinen i Lund", SMIL. The subject lived for almost two decades together with Computer Science, until it 1999 joined the then new Centre for Mathematical Sciences, with research funding also from LTH. After the retirement of Carl-Erik Fröberg in 1984 it took some years before the chair got a new permanent holder in Gustaf Söderlind. With him and his collaborators, research in numerical analysis was revitalized, as well as the courses offered. This was then further developed after the move to the Centre for Mathematical Sciences.

2.1.3 Setting the Platform

Organizationally, the mathematical sciences at LTH had their roots in the Engineering Physics programme. In the early 1980s it was time for a restructuring of this programme, and a working group (lead by director of studies for Engineering physics Anders Lundström, later head of the testing department at Scania) outlined a plan containing three "applied" branches in the Engineering Physics programme: Applied Mathematics, Applied Physics, Applied Mechanics. This was the first time when the mathematical sciences were assigned an independent role in the engineering education, and they were no longer only basic science – the baby *Mathematics as Technology* was born. Even if the concrete study plan was very modest in mathematical subjects, physics being very strong, modern, and expanding, it gave them a platform from which they could operate. The courses in linear systems and stationary stochastic processes linked mathematics strongly with the engineering (technology) courses in automatic control and telecommunication theory. A similar chain started with Continuous Systems, and continued with courses in optics and waves, quantum mechanics and atomic physics. There were

also strong links between Continuous Systems and other parts of the curriculum for Engineering Physics, like electromagnetic fields and solid mechanics.

2.2 Research Paves the Way to a New Engineering Programme

The development on the educational side, described above, gradually got consequences on the research side, described next.

2.2.1 Overview

During the 1980s, the mathematical sciences started to cultivate contacts with the world outside the university. Co-operation with industry grew up, often around some Master's thesis project. Several dealt with modeling and analysis of industrial processes, for instance in microelectronics together with Ericsson and RIFA. Others were done together with a variety of departments at LTH, e.g. in waste water treatments, together with the Department of Industrial Automation at LTH. Very close cooperation was established with the Department of Automatic Control, a department, which led by Karl Johan Åström over the years has been very supportive to mathematics.

A rewarding cooperation, which lasted for almost two decades, started 1984 with the Ophthalmology department in Malmö. It resulted in SITA, the "Swedish Interactive Threshold Algorithm", representing a new generation of sophisticated statistical software for glaucoma diagnosis, now present in instruments at eye clinics all over the world. Bertram Broberg, professor in Solid Mechanics at LTH, and Anders Lundström also stimulated research in Stochastic Mechanics and the application of stochastic processes in mechanical engineering. The thesis by Igor Rychlik in 1986 introduced a new way to describe stochastic fatigue loads that made a theoretical statistical analysis possible, and has now become the standard technique.

During the 1980s, several students had made master projects in mathematics around some application from industry or other sciences. Some of these wanted to continue with PhD-studies along the same lines. Possibilities were opened through a new study plan directed on applied mathematics. The first to complete a degree according to this was Anders Rantzer (who later became a professor of Automatic Control in Lund), who in 1988 completed a licentiate (about halfway to a PhD in the Swedish system). He was then followed by a sequence of PhDs and licentiates, Stefan Diehl, Anders Heyden, Kalle Åström. . . .

By support in a grant from STU, the Swedish National Board for Technical Development, for "Engineering mathematical statistics", it was possible to recruit Jan Holst back to Lund 1986 from the Danish Technical University. With him, practical engineering entered also in the statistics PhD theses. His first PhD student

in Lund defended his thesis 1990 on statistics, heat exchangers and geothermal energy. This was followed by many statistical Master's and PhD projects together with energy industry and other industry, like Sydkraft, Volvo, Scania.

Grants for many application oriented research projects were during the 1990s obtained also in mathematics from TFR, SSF, EC. By the beginning of the 1990s, a dominating topic became image analysis and computer vision, where a big and active research group was built up, mainly by external funding, with Gunnar Sparr as driving force. The interest from industry in image related problems seemed to be unlimited, contacts were taken continuously, and a lot of Master's thesis and other projects were started. Many of the performers of these were hired by the orderers, often at the Idéon Science Part in Lund, and a considerable transfer of knowledge took place. Without doubt, the image group at the Centre for Mathematical Sciences has played an important role in giving the Lund-Malmö region its strong position in image based industry. Several companies have their origin in mathematics (Decuma, Cognimatics, Polar Rose, ...).

Here Decuma has a particular standing, with a product transforming handwritten text to digital computer code. The company received a lot of attention as a spin-off company of research in mathematics. This culminated in 2002, when Decuma won the prestigious EC Information Society Technology Grand Prize, and Sony decided to make Decuma their main provider for handwriting recognition.

The industry oriented research set a mark in the Master's thesis examination. At the beginning of the present century, Mathematics was the largest Master's thesis subject in the Engineering physics program, and Mathematical statistics the fifth largest, together accounting for almost one third of the total, and 40% of the industry related ones.

2.2.2 The Trondheim Experience, SAM, ECMI, and Industrial Economy

Before we plunge into the Engineering Mathematics programme we need to mention a few more steps towards its realization.

- The Trondheim experience: In Trondheim, the Norwegian Technical University had managed a special line of study in Industrial Mathematics, as a profile within some of the traditional engineering programmes. The celebration of its 10th year of existence inspired the delegation from Lund to the conclusion that LTH had all ingredients for a successful industry oriented educational initiative. Karl Johan Åström supported the idea of more visible applied mathematics profiles within the existing programmes in Lund, and a first official proposal was sent to the LTH board.
- SAM, "Systems and Applied Mathematics": This was an initiative to present all the advanced courses at LTH on applied mathematics, automatic control, information technology, and other systems oriented subjects. Björn Wittenmark edited a nice catalogue that showed the strength of LTH in 1992.

- ECMI, "European Consortium for Mathematics in Industry": This already existing European educational initiative had nodes in Copenhagen and at Chalmers in Göteborg. Especially Jan Holst worked hard to create also a Lund node, and in 1997 Lund was officially admitted to the ECMI circle. Related to this is a double degree agreement between LTH and the Technical University of Kaiserslautern that was signed by the LTH rector in 1999 (in fact the first one for LTH).
- Industrial Economy: This new engineering programme at LTH started in 1998. Here Jan Holst, chairman of the educational board 2002–2006, worked very energetically to strengthen a structure having "Mathematical Modelling" as one of the specializations.

2.2.3 Summing Up

By the turn of the century, there thus existed successful groups in applied and industrial mathematics and statistics at LTH, also recognized economically by the faculty. These groups not only helped others as consultants, but also by successful examples demonstrated the potential of mathematics for engineering use, in new and old areas. This certainly helped to settle the doubts that may have existed about the needs and prospects for an engineering MSc program, relying on mathematics as the major tool. The culture that had been built up at the Centre for Mathematical Sciences certainly was an important component in the creation of the Engineering Mathematics programme at LTH.

2.3 The Engineering Mathematics Study Programme

During the 1990s, the idea to create a new study programme for a MSc in engineering ("civilingenjör") successively developed. This should be a specially designed programme where the mathematical sciences (mathematics, mathematical statistics and numerical analysis), and mathematical methodology in general, provide the central engineering tools. In some of the existing programmes, in particular Engineering Physics, there was already a possibility by elective courses to form a mathematical profile, but the new programme should have this character from start. In this way, it should still more prepare for mathematical work in a broad range of areas.

2.3.1 The Process

The idea to create a programme along these lines was communicated to the executives of LTH, who gave a positive feedback, asking for more decision material.

A working group was formed, with Gunnar Sparr as chairman. After much planning and lobby work, a decision to start a new programme in Engineering Mathematics was finally taken by LTH and Lund University. Starting in 2002, it was the first of its kind in Sweden (in 2008 Chalmers in Gothenburg followed the idea).

In the argumentation, a number of expected characteristics for the programme were presented. First of all it should deserve to be called an engineering education, not only to be an education in mathematics with some additional applied courses. The combination of mathematical theory and engineering subjects was expected to attract well motivated students, who not otherwise had chosen an engineering or mathematical education. A hypothesis (which later turned out to be correct) was that a programme of this kind also would attract female students. The education should have a holistic approach to mathematics as integrated in the learning process, and the programme should work as a test-bed for pedagogical ideas. The modelling and communication aspects should be kept high.

2.3.2 Vision for the Programme

Much of the background and thinking about the programme is summarized in the following programme vision, presented to the governing board of LTH a few years after start.

> The motto of the programme is *Mathematics as Technology*.
>
> By tradition, heavy use of mathematics in applications often has taken place together with physics. This pattern has changed during the last few decades. Today, in order to become a good applied mathematician, or engineering mathematician, it is not necessary to be also a physicist. The by tradition strong links between mathematics and physics have been complemented by a number of other sciences, engineering as well as others, like economy, biology and medicine. New areas have emerged, often related to information technology. It is in such borders and frontiers, new and old ones, the programme is acting by providing the students bridging competences.
>
> The most important factor for the change of scenery is the development within computer technology. This causes the borderline to what is possible to compute and simulate to advance continuously. This causes new needs for methods and algorithms, which in turn increase the claims on mathematical understanding of their behaviour and of the underlying problems. Another effect of the computer development, no less important, is that completely new technological/mathematical problem areas and opportunities have emerged, for which no completed scientific tradition to lean against is available. This has lead to a direct use of mathematics as technology, without intermediaries from other sciences.
>
> It is from this background the programme has been created, with the mathematical sciences and computer science as cornerstones. The ability to handle problems that from the beginning are not formulated in mathematical terms is trained by building and analysis of mathematical models. But this must be done in combination with other kinds of proficiency. To this end, within the programme is built a broad basis of such knowledge, with course providers from four faculties, besides LTH also the Science, Social, and Medical faculties. System thinking permeates the program. All together this forms an education that creates possibilities to use mathematical/engineering strategies and tools to attack problems in a broad spectrum of processes and systems from industry and society. These include not

only technological but also e.g. biological, economical and medical systems, and various information systems.

As a study program, Engineering Mathematics tries to enrich the pedagogical manifoldness at LTH. The programme aims at being an arena for the development of new issues in mathematics courses. Two examples of this are the new courses in Mathematical Modeling and Mathematical Communication. More than any other programme the mathematical contexts form structural elements in the process of learning, also in other subjects than the mathematical ones. The education aims to create an open attitude among the students to get engaged in complex, concrete and perhaps vaguely formulated problems, and to search to develop creative mathematical/engineering methods for their solution.

At LTH there exists a strong tradition within systems and applied mathematics, with strong links to industry in many departments. The latter also holds true for the Centre for Mathematical Sciences, a fact that is crucial for a programme like Engineering Mathematics, where the mathematical sciences have double roles, being both core subjects and engineering tools. The Centre has been the growing place for a number of innovations that have lead to companies and patents. The big contact area against the world around is illustrated by the fact that the Centre for Mathematical Sciences around the year 2000 was the biggest department what concerns Master's theses in the Engineering Physics program, with over 30% of all and almost 40% of all external ones, done with industry or organisations.

2.3.3 Implementation

The first 30 students in Engineering Mathematics were admitted in 2002. After that around 40 students have started each year. During the short time since its start, the programme (like all others at LTH) has undergone a revision, due to the Bologna process.

At LTH, the adaption to the Bologna framework has been done in terms of a 'soft' 3 + 2 model. By this is meant that the education is organized as a uniting 5 years education, with a possibility to conclude with a Bachelor Degree after 3 years, if some additional constraints are fulfilled (among them a Bachelor thesis). This structure has been crucial in the planning of the Engineering Mathematics programme, since it leaves freedom for more theory from the beginning than should be possible with a heavy claim on becoming engineers after 3 years.

The first 3 years for Engineering Mathematics are almost completely filled with mandatory courses. Roughly speaking, the 180 ECTS course credits are distributed in the following way:

- About 50% to the mathematical sciences (Mathematics, Mathematical Statistics, Numerical Analysis)
- About 10% to each one of the groups
  - Computer science – Programming
  - Mathematical modelling – Mathematical communication
  - Systems – Signals
  - Mechanics – Physics – Field theory
  - Economy – Biology – Sustainable development

This means that every student, besides a lot of mathematics, gets a very broad basic background in other sciences to use together with mathematics. Moreover, he/she in this way meets many different scientific cultures, and gets training to communicate within these.

During years four and five, the student chooses one of five master specializations:

- Biological and medical modelling
- Computations and simulations
- Environment, risk and climate
- Financial modelling
- Signals, images and systems

As can be seen, these cover a very wide spectrum, reflecting the fact that mathematics is a common language for large parts of science and technology. A claim in the design of the specializations has been that they shall have a significant exchange with the world outside the academy. Profile courses are taught by professionals in the respective fields. The programme opens up to modern areas of applications, at the same time as it makes possible to delve into mathematical theory. The education ends with a Master's thesis of 30 ECTS.

For economical reasons a programme with only around 40 students can't have too many specially designed own courses, but has to rely on courses that can be shared with other programmes. Still, it has been possible to introduce a number of profile courses for Engineering Mathematics. Within the mandatory curriculum, two of them are found already in the first year: Mathematical Modelling and Mathematical Communication. A theme for the second year is mathematical system theory, reflected by several courses in mathematics and mathematical statistics (among them descendents of the ones mentioned above). In the third year there is a special course package in numerical analysis, designed for this programme. A course in algorithm implementation trains in efficient implementation of algorithms on different platforms. The mandatory course block ends with an advanced course in mathematical modelling, where all the mathematical tools acquired so far are at disposal. In the master specializations there are specially designed courses in biological systems and quantitative human physiology. Also in economy there are courses designed for students with a strong mathematical background. Besides these, there are of course advanced mandatory and elective courses in the mathematical sciences and in other related subjects.

2.3.4 The Students

In all the information material for the programme has been emphasized that "the programme directs itself to those with a deep interest in mathematics and the use of mathematics". Inquiries among the newly arrived students show that this objective to a large extent has been accomplished. In fact, when asking for the main reason for choice of program, almost all (and markedly more than other programs), state

"interest" as the main reason. Looking at the enrollment demands, Engineering Mathematics is among the most attractive programmes at LTH. It also has had a comparatively high enrollment of women, typically 30–40%, one year even 45%. Also geographically, the enrollment from non-local regions is larger than the LTH average.

2.3.5 Employments Afterwards

Up to mid-2009, there are about 50 students who have finished for a MSc in Engineering Mathematics. Among them are 15 women. Apparently the engineering mathematicians are very attractive on the job market, they have been very successful in getting good and adequate employments. Looking at their first job, of the 50 first passed, 31 have went to industry and other occupations, while 19 have went to PhD studies.

In the first category, the following (partially overlapping) branches are represented (with number of individuals within parentheses):

- Finance and banking (6)
- Energy trading (1)
- Insurance (1)
- Industrial automation and systems consulting (3)
- Telecom (4)
- Image technology (4)
- IT software (5)
- Medical engineering (2)
- Radio (1)
- Construction engineering (1)
- Nuclear power engineering (1)
- Own enterprise, consulting (2)

The list shows that the education has prepared well for non-academic careers in a wide range of branches. Of the 31 who went to industry, 10 got their first job outside Sweden.

Also those 19 individuals who continued with PhD studies are widely spread over subjects:

> Applied Mechanics, Automatic Control, Bioengineering, Electrical Engineering, Finance, Geology, Information Theory, Material Science, Medical Imaging, Signal Processing, Solid Mechanics, and, of course, Mathematics, Mathematical Statistics, Numeric Analysis.

Also within the group of mathematical sciences, both pure theory and a wide range of applications are represented. In the latter group can be found e.g. finance, computer vision, environment, remote sensing. Of the 19 PhD students, five make their degree abroad, ten in Lund, and four somewhere else in Sweden.

2.4 Conclusion

The programme in Engineering Mathematics obviously has filled a gap in the educational system. It was built on existing use of mathematics in the engineering subjects, and it has successfully exploited the research in applicable mathematics that has grown up within the mathematical sciences. Its existence has definitely already strengthened the position of these sciences at LTH, and affected the culture within them. Hopefully it can also in the future help to open new areas and roles for mathematical education and knowledge in many sciences, industry and society. For the future, it is also up to the university to recognize and feel a long term commitment to *Mathematics as Technology* – with technology not only as engineering but as a means for improvement.

Chapter 3
Theory of "Edge-Detection"

Jan J. Koenderink

Abstract "Edge detection" is one of the most fundamental operations in image processing. Yet it remains unclear what precisely is meant by "edge" – apart from being what edge detectors detect, or by an "edge detector" – except for being that which detects edges. Many edge detection algorithms are in common use, they "find" slightly different entities and it remains unclear how one may compare their effectiveness in detecting edges, although this is commonly done. A principled theory of edge detection is offered that is based on the structure of the 2-jet of the image at a certain scale. In this theory there is no such a thing as an "edge detector", edges are defined in terms of the 2-jet as a single object.

3.1 Introduction

"Edge detection" has been one of the cornerstones of image processing from the early days on, for instance, David Marr's seminal book [1] starts with it. What *is* an "edge"? (Or any "feature" for that sake!) In the final analysis an edge is what an edge detector detects. So what is an "edge detector"? It is a "detector of edges". Clearly this state of affairs is less than satisfactory. In this paper I attempt to throw some new light on the matter.

3.1.1 Images and Differentiation

An ideal (see below) "image" is a cross-section of a trivial fiber bundle $\mathbb{E}^n \times \mathbb{R}^+$, where the Eucidean space $\mathbb{E}^n$ models the conventional "image plane" and the

J.J. Koenderink (✉)
Man–Machine Interaction Group, EEMCS, Delft University of Technology, Delft, The Netherlands
e-mail: jan.koenderink@telfort.nl

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*, Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_3,

non-negative reals $\mathbb{R}^+$ the "(image) intensity". One usually has a model for the physics that "explains" the intensity. In many cases the intensities are limited from above, thus one has the segment $\mathbb{I} = (0, 1)$ instead of $\mathbb{R}^+$. What is important here is that the image plane and the image intensity dimensions are *mutually incommensurable*. In many cases the base space is really $\mathbb{Z}^n$, i.e., the image plane is "pixellated". Here I assume that any pixellation occurs at a scale much finer than the scale at which "features" (such as edges) are sought. This is the case, for instance, in the common case of the viewing of high quality glossy prints. Even then the notion of pixellation is important though, for it suggests that one cannot assume anything like an "infinitesimal domain". In image processing "ideal images" don't occur and the issue of "differentiability" doesn't even come up.

The notion of "scale" is essential in any definition of "edge". For instance, a meteorological "front" is an edge (e.g., discontinuity in temperature) on the global scale, but takes a day to pass your town. The "inner scale" of an image (the "outer scale" being the extent of the current "region of interest") can be changed by "blurring", that is to say, by convolving the image with some rotationally symmetric, non-negative kernel $K(\mathbf{r}, \sigma)$, where the scale parameter σ is the "width" of the kernel. One requires the kernel to be isotropic and space-invariant. A more interesting requirement is that blurring should not generate detail (called "spurious resolution" in optics). Blurring should raise the level of local minima and lower the level of local maxima. It is easily shown that this limits the kernels to Gaussians [2]. Blurring then becomes equivalent to diffusion of intensity.

Taking $n = 1$ for simplicity, one obtains the kernels

$$G_0(x, \sigma) = \frac{\mathrm{e}^{-\frac{x^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma}. \tag{3.1}$$

of width σ and unit weight. The immediate temptation is to pose ("$\otimes$" denotes convolution)

$$I(x, \sigma) = I(x) \otimes G_0(x, \sigma), \tag{3.2}$$

where $I(x)$ denotes "the" image, apparently the image "at infinite resolution", clearly a non-entity. I will consider the expression as a merely symbolic one, indicating the physical operation of *sampling the image at scale* σ. The kernel represents the "point operator". When applied to the image it "takes a bite" out of it and presents one with the sample (a numerical quantity) $I(x, \sigma)$. The point operator is parameterized by location and width and yields a localized sample of the image intensity. I assume one knows the image only through such samples. The image at "infinite resolution" is a mere virtual entity that is devoid of physical meaning. The point operator is to be thought of as physical mechanism, a little machine that quietly sits there and on call queries the image for its local intensity. The title "point" is quite apt since – like in Euclid's definition[1] – it "is that which has no parts".

[1] Euclid: Elements. (ca. 300 BC).

Next consider differentiation. One is tempted to consider the following three expressions for $\partial I(x,\sigma)$ as *equivalent* because of linearity:

$$\partial I(x,\sigma) = \tag{3.3}$$
$$= \partial\{I(x)\} \otimes G_0(x,\sigma) \tag{3.4}$$
$$= \partial\{I(x) \otimes G_0(x,\sigma)\} \tag{3.5}$$
$$= I(x) \otimes \partial\{G_0(x,\sigma)\}, \tag{3.6}$$

but wait:

- The first interpretation is *senseless* because the very notion of "differentiability" of "the" image is undefined.
- The second expression is only slightly better since $I(x) \otimes G_0(x,\sigma) = I(x,\sigma)$ is at least *something*, it is again an *image*. Since we made it ourselves we have some knowledge of it (Vico's VERUM FACTUM EST [3]), but since we have only a finite datastructure of samples "differentiation" in the true sense is not applicable. As a consequence, this interpretation is *useless*.
- In contradistinction, the third interpretation makes perfect sense since one differentiates an ideal entity, the analytic function G_0. The kernel $\partial G_0(x,\sigma) = G_1(x,\sigma)$ is an operator that yields samples of the first derivative. I consider it to be the *physical implementation of a tangent vector*.

Think of the kernel $G_1(x,\sigma)$ as of a little machine that lurks at a place and that when queried takes a bite of the image and spews out a sample of the first derivative of the intensity in the x-direction. In a similar way one constructs higher derivative operators $G_n(x,\sigma)$.

The k-jet $J_k(I,\sigma)$ of the image is the set $\{I_0(x,\sigma), I_1(x,\sigma), \ldots, I_k(x,\sigma)\}$. In this paper I consider the 2-jet of images on a 1-dimensional base space ("image plane" a line, e.g., the images obtained via a linear CCD-array as in a flatbed scanner).

3.1.2 Models of "Edges"

The primordial image of an "edge" is the unit step function $U(x)$ defined by the rule $U(x > 0) = 1$, $U(x < 0) = 0$ (where $U(0)$ remains undefined). By blurring one obtains the family of blurry step functions at the scale σ

$$U(x,\sigma) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sigma\sqrt{2}}\right)\right), \tag{3.7}$$

where "erf" denotes the error function.

The first derivative at a certain scale σ of a very steep edge (of width $\mu \ll \sigma$) yields the point operator at scale $\sqrt{\sigma^2 + \mu^2} \approx \sigma$. Thus the first derivative operator acts as a "perfect edge detector" for the blurry edges. Indeed, it is no coincidence that all "edge detectors" proposed in the literature are very similar to $G_1(x, \sigma)$.

"Real" edges (whatever these might be!), that one believes to "see" in natural images like photographs, are different from $U(x, \sigma)$ in many ways. No doubt the derivative operator applied to a random image will yield a non-zero sample at virtually *any* location. Then where are the "edges"? *There exists no principled answer to this.* This is where the numerous ad hoc approaches one encounters in the literature diverge.

3.2 A Theory of Edge Detection

I consider images as cross-sections of $\mathbb{E} \times \mathbb{I}$, thus on a 1-dimensional domain and limited in intensity, and the 2-jet $J_2(\sigma)$, that is $\{I_0(x, \sigma), I_1(x, \sigma), I_2(x, \sigma)\}$. This is an apt model of the actual image processing setting. Generalization to higher dimensional substrates is immediate.

I study the neighborhood of a point, for the sake of notation I consider the origin. I assume that "images" are known *only through their samples*, i.e., through $J_2(\sigma)$ at the origin. Thus the image structure at a point is summed up through three numbers. All (typically infinitely many) images that yield the same sample are mutually indistinguishable, I call them "metameric". Metameric images agree in their Taylor expansions at the origin up to (and including) the second order, but differ in the tails of their Taylor series. Images whose initial three Taylor coefficients are identically zero yield the same samples as the uniformly zero ("black") image. Thus $J_2(\sigma)$ at the origin is blind to such black images. The space of black images has infinite cardinality, whereas the space of images "seen" by $J_2(\sigma)$ at the origin is 3-dimensional.

I define the linear operator Π as the truncation of a Taylor series at the second order. It is idempotent, thus a projection. Its kernel is the space of black images.

3.2.1 The Samples Cone and the Samples Solid

The sample corresponding to a unit impulse function of weight a at location x is $a\{G_0(x, \sigma), G_1(x, \sigma), G_2(x, \sigma)\}$ (see Fig. 3.1). Its convex hull is a *solid cone* with the unit impulse samples as generators. Since arbitrary images are convex combinations of impulses, all possible samples that one may ever encounter are apparently to be found in the *interior of the cone* (see Fig. 3.2). Uniform images of level a yield samples $\{a, 0, 0\}$, this is a half-line at the origin that lies in the interior of the cone. It may be denoted "the featureless axis". Any sample can be decomposed uniquely as the superposition of a uniform image and a suitably chosen

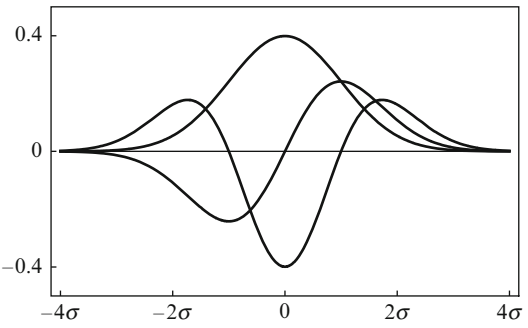


Fig. 3.1 The kernels of the 2-jet. The zeroth order kernel is the Gaussian bell-shape, the first order looks like a classical "edge detector", whereas the second order kernel looks like the classical "Laplacean operator"

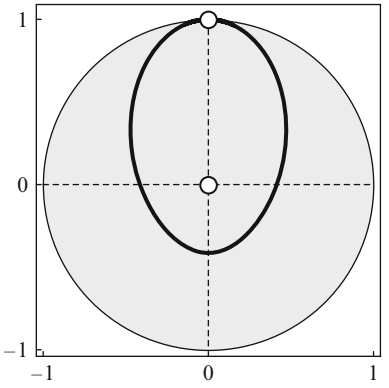


Fig. 3.2 A stereographic projection of the cone of impulses from the direction of the featureless axis. Thus the point at the center denotes the featureless axis. The circular disk denotes the hemisphere centered on the featureless axis, the point indicated on its boundary correspond to impulses at $\pm\infty$

impulse. Apparently *any* image is indistinguishable from some "point on a uniform background".

There are many alternative ways to construct images that "would explain" a given sample. For instance, there is always an image composed of three (well chosen) impulses that explains the sample. One may prescribe the location of the impulses (say at $\{-\sigma, 0, +\sigma\}$) and find unique amplitudes for any given image. There is no guarantee that these amplitudes would turn out to be all non-negative though.

However, such characterizations are less than useful because of the upper limit on the intensity levels which excludes narrow impulses. Apparently one needs to consider more appropriate descriptions.

The uniform image of the highest intensity yields the sample $\{1, 0, 0\}$ with may be called the "blank point" since it represents the overall white image. The images $I(x)$ and $1 - I(x)$ yield the mutually related samples $\{+u, +v, +w\}$ (say) and $\{1 - u, -v, -w\}$, revealing a central symmetry about the point $\{\frac{1}{2}, 0, 0\}$. One concludes that all samples must lie in the interior of the intersection of the cone at the origin and the (by the central symmetry) inverted cone at the blank point. This is a *finite volume*.

Another way to see this is to remark that all images are limited by $0 < I(x) < 1$, thus lie in the interior of the *unit hypercube* in the infinite dimensional space with

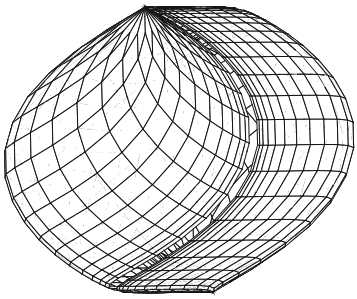


Fig. 3.3 The region of possible samples from the 2-jet. It is a convex solid with central symmetry. The boundary is composed of two mutually congruent smooth patches. The patches meet in curves where the surface has two distinct tangent planes. At two diametrically located points (corresponding to the blank and the empty image) the surface has conical singularities

continuous index x. The samples thus lie in the interior of the projection Π of this hypercube in the 3-dimensional space of samples (see Fig. 3.3).

The projection Π will conserve the central symmetry from the hypercube. This projected hypercube must lie inside the intersection of the cones considered above. Some thought reveals that the two cones must be tangent to the boundary of the volume at the origin and the blank point. By invoking Lyapunov's convexity theorem [4, 5] and Pontryagin's bang–bang principle [6] one finds that the images that correspond to boundary points are "bars", that are images with intensity either 0 or 1 and no more than 2 transitions on $(-\infty, +\infty)$ (see also Schrödinger [7]).

Although the sets of metameric images are generically of infinite cardinality, samples on the boundary of the 2-jet solid specify unique images. They are light or dark bars (2-parameter sets), left or right edges (1-parameter sets) or uniform images (2 points, the blank image and the "empty image" (the origin)). Apparently "edges" are indeed special features in the context of the 2-jet in the sense that they automatically arise as boundary elements.

3.2.2 Canonical Basis: The "Three Pixel Representation"

The 2-jet basis is not orthonormal, its Grammian is

$$\frac{1}{2\sigma\sqrt{\pi}}\begin{pmatrix} 1 & 0 & -\frac{1}{2\sigma^2} \\ 0 & \frac{1}{2\sigma^2} & 0 \\ -\frac{1}{2\sigma^2} & 0 & \frac{3}{4\sigma^4} \end{pmatrix}. \tag{3.8}$$

Thus there is some reason to look for alternative, more convenient bases.

Since the samples solid is the projection of a hypercube and inherits its symmetry, it is of interest to find its inscribed crate of maximum volume. In order to find it

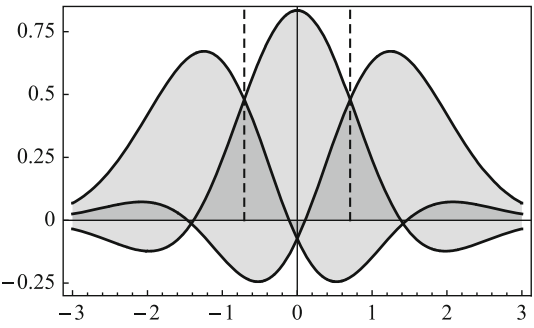


Fig. 3.4 The canonical basis functions of the 2-jet. Apart from obvious under and overshoots, these functions claim three adjacent areas, *left*, *center* and *right*. The 2-jet can be interpreted as a "three pixel sub-image"

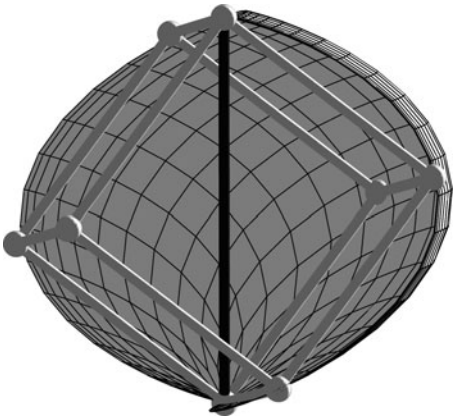


Fig. 3.5 The maximum volume inscribed crate, brought into cubical shape via a suitable affinity. Only one half of the boundary of the samples solid has been drawn in order to visualize the excellent fit

one divides the substrate into three pieces $(-\infty, -a)$, $(-a, +a)$ and $(+a, +\infty)$ for $a > 0$. The crate has volume $ae(-a^2)/2$, which assumes a unique maximum at $a = \sigma/\sqrt{2}$. The crate is spanned by the vectors ("erfc" the complementary error function)

$$\left\{ \begin{pmatrix} \frac{1}{2}\mathrm{erfc}(\frac{1}{2}) \\ (\mathrm{e}^{\frac{1}{4}}\sqrt{2\pi}\sigma)^{-1} \\ (2\mathrm{e}^{\frac{1}{4}}\sqrt{\pi}\sigma^2)^{-1} \end{pmatrix}, \begin{pmatrix} \mathrm{erfc}(\frac{1}{2}) \\ 0 \\ -(\mathrm{e}^{\frac{1}{4}}\sqrt{\pi}\sigma^2)^{-1} \end{pmatrix}, \begin{pmatrix} \frac{1}{2}\mathrm{erfc}(\frac{1}{2}) \\ (\mathrm{e}^{\frac{1}{4}}\sqrt{2\pi}\sigma)^{-1} \\ (2\mathrm{e}^{\frac{1}{4}}\sqrt{\pi}\sigma^2)^{-1} \end{pmatrix} \right\}. \tag{3.9}$$

Taking these as the basis (see Fig. 3.4), the crate turns into the unit cube and one may expect the samples solid to assume a particularly attractive shape. Indeed, the samples solid snugly fits the unit cube as its volume is only 73.5% of the volume of the circumscribed sphere (see Fig. 3.5). It looks like a "slightly bloated" or "inflated" copy of the unit cube.

In this description the image is represented as a piecewise constant function that assumes different values on the three adjacent regions $\mathcal{L}^+ = (-\infty, \sigma/\sqrt{2})$, $\mathcal{C}^+ = (-\sigma/\sqrt{2}, +\sigma/\sqrt{2})$ and $\mathcal{R}^+ = (\sigma/\sqrt{2}, +\infty)$, a "three pixel image" (hence "left", "center" and "right"). The values will be *almost* in the range $(0, 1)$, "almost"

because the sample solid is actually a "slightly inflated" version of the unit cube. The excursions outside the $(0, 1)$ range are less than $0.146\ldots$. Since such excursions only occur for "true" edges or bars, they hardly ever occur in realistic image processing settings. The combined regions $\mathcal{R}^- = \mathcal{L}^+ \cap \mathcal{C}^+$ (complement of $\mathcal{R}^+$), $\mathcal{L}^- = \mathcal{C}^+ \cap \mathcal{R}^+$ (complement of $\mathcal{L}^+$), $\mathcal{C}^- = \mathcal{R}^+ \cap \mathcal{L}^+$ (complement of $\mathcal{C}^+$) and, finally, $\mathcal{I}^- = \mathcal{L}^+ \cap \mathcal{C}^+ \cap \mathcal{R}^+$ (the blank image), with the empty set $\mathcal{E}$ (the empty image) are the remaining vertices of the crate.

The "sizes" of the pixels can be defined via the arc length of the edge loci (see below) in the canonical representation. One finds that the 5 and 95% cut-offs of the edge loci are $\pm 2.2646\ldots\sigma$. The two outer pixels have a width of approximately $1.557\ldots\sigma$, whereas the center pixel has width $1.4142\ldots\sigma$. Thus – despite the infinite domain – for practical purposes the pixels are of roughly similar size.

In the illustrations in this chapter this canonical basis will be used in absence of a remark to the contrary. The Euclidian metric of this representation can be used to define a *semimetric on the space of images*, one simply defines the inner product of two images as the inner product of their images in the canonical representation. This allows us to define the complement of the projection and consider the space of images as the direct sum of the space of "three pixel images" and the space of black images. Any image can be split uniquely in a three pixel image and a black image. The projection Π simply discards the black image and retains the three pixel image. This allows one to construct a unique canonical representative for any set of metameric images. These canonical (three pixel) representatives are typically true images (intensities in the range $(0, 1)$) except for occasional, minor under and overshoots. This observation is the basis for a simple edge detection algorithm (see below).

3.2.3 The Edge Loci

The "left edge" L_a (with $L_a(x) = 1$ iff $x < a$, zero otherwise) has the locus (see Figs. 3.6 and 3.7)

$$\left(\frac{1}{2}(1 + \mathrm{erf}(\frac{a}{\sigma\sqrt{2}}), -\frac{\mathrm{e}^{-\frac{a^2}{2\sigma^2}}}{\sigma\sqrt{2\pi}}, -\frac{a\,\mathrm{e}^{-\frac{a^2}{2\sigma^2}}}{\sigma^3\sqrt{2\pi}}\right), \tag{3.10}$$

in the 2-jet space. Of course it is more interesting to consider it in the canonical basis, though the expressions then are more cumbersome. Especially its arc length function is complicated (though numerically no problem of course). The total arc length is $3.276\ldots$, only slightly more than that of the length of the edge progression $\mathcal{L}^+\mathcal{R}^-\mathcal{I}$ (which amounts to 3).

The "right edges" $R(a)$ are simply $R(a) = I - L(a)$. The locus is the image by central symmetry of the left edge locus. The shape of the left (and by symmetry

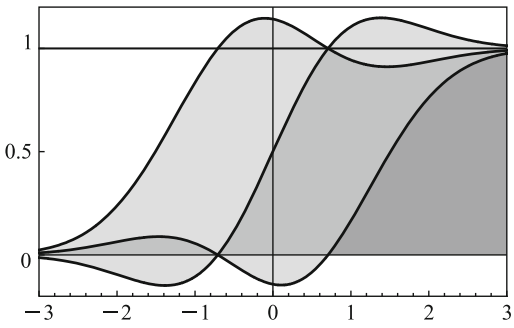


Fig. 3.6 The coordinates of the edge locus. Notice the great similarity to the coordinate of the polygonal arc $\mathcal{L}^+\mathcal{R}^-\mathcal{I}$. Differences are minor under and overshoots

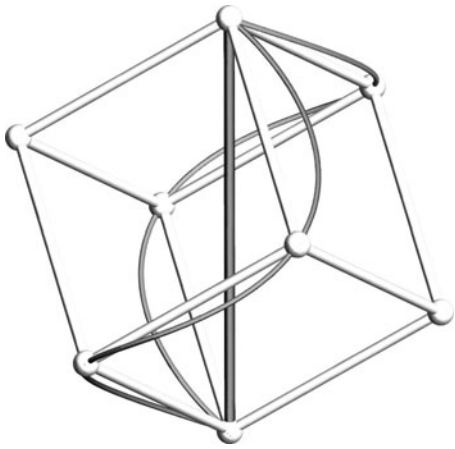


Fig. 3.7 The edge loci in relation to the maximum volume crate. Notice how the edge loci closely follow edge progressions, e.g., the *left edge* locus the polygonal arc $\mathcal{L}^+\mathcal{R}^-\mathcal{I}$

the right) edge locus closely follows that of the $\mathcal{L}^+\mathcal{R}^-\mathcal{I}$ edge progression, thus it is characterized through two pronounced curvature and torsion extrema. Looked at from the direction of the featureless axis the two edge loci smoothly connect into a characteristic, closed "figure eight" curve. The apparent "intersection" is actually due to two sharp bends (right angles), one at the empty and one at the blank point.

3.2.4 A Notion of "Complementary Locations"

Two images may be called "complementary" (here "supplementary" would perhaps be a more apt expression) if they add to the blank image. One might think of photographic negative and positive (in the old fashioned silver-based process I mean). Such images are related by central symmetry, they add to a featureless image (the blank image). In this vein, one might conceive of two impulse images adding to the equivalent of a featureless image. For this to be possible the corresponding cone generators have to be mutually coplanar with the featureless axis. It is immediate to prove that $\{a, -1/a\}$ (for $a \in \mathbb{R}$) characterizes such "complementary locations".

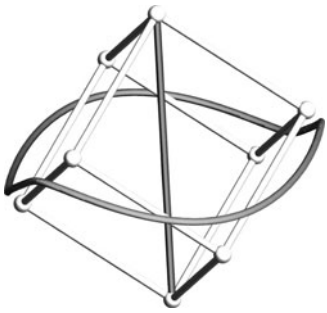


Fig. 3.8 The full bar locus. It is a smooth, closed curve that straddles the hexagonal arc $\mathcal{L}^+\mathcal{R}^-\mathcal{C}^+\mathcal{L}^-\mathcal{R}^+\mathcal{C}^-$. It is the "equator" of the samples solid, the locus at which the circumscribed cylinder with generators parallel to the featureless axis touches the samples solid

The central bar $(-1, +1)$ has complementary transitions and so has the left edge $(-\infty, 0)$ and the right edge $0, +\infty)$. These are evidently special configurations. All bars of the type $(-1/a, a)$ are special in the sense that small variations of their transition locations lie in the plane spanned by the impulse images at a and $-1/a$, thus containing a line parallel to the featureless axis. Thus the locus $(-1/a, a)$ lies on the cylinder with generators parallel to the featureless axis, circumscribed to the samples solid: These bars are the strongest features imaginable [7]! I'll refer to them as "full bars".

The locus of full bars closely straddles the edge progression $\mathcal{L}^+\mathcal{R}^-\mathcal{C}^+\mathcal{L}^-\mathcal{R}^+\mathcal{C}^-$ of length 6. It is only slightly shorter, its full length being $5.71596\ldots$. This is due to its "cutting corners", for the locus of full bars is smooth and doesn't pass through the vertices of the cube. (See Fig. 3.8.)

3.2.5 The Principled Definition of "Edge"

So what is an edge? At this point of the exposition a principled definition is possible.

Suppose one has acquired a sample from the 2-jet. Generically it will fall inside the interior of the samples solid. Together with the featureless axis the sample defines a unique halfplane. This half plane may conceivably:

1. Fail to intersect one of the edge loci
2. Intersect the left edges locus
3. Intersect the right edges locus

In case of an intersection that intersection is unique. In case 1 the sample cannot be "explained" by way of an edge, that is to say, there is no step function in the set (of infinite cardinality) of metametic images. In cases 2 or 3 there exists a unique explanation in terms of a left or right edge. The sample can be explained as the convex combination of a unique step function and a unique featureless image. Thus you obtain a step that is parameterized by two levels an the location of the transition.

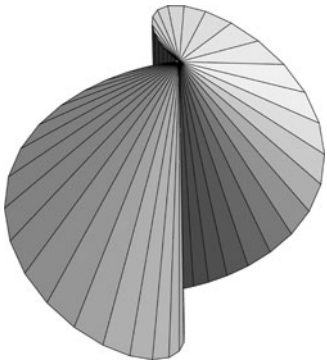


Fig. 3.9 The *left* and *right* edge regions. Either region is a convex body bounded by two smooth, ruled surfaces. The two regions meet in a line (the featureless axis) and are mutually congruent by central symmetry. Their union exhausts about half of the samples solid

Thus this is a principled definition of "edge detection":

Definition 1. *One has detected an edge if the sample of the 2-jet allows for an edge explanation. The edge is unique and characterized by two levels in the range* $\mathbb{I}^1$ *and a location in the range* $\mathbb{R}$.

Notice that the *process of edge detection* is defined, not any "edge detector". Although the first derivative kernel $G_1(x, \sigma)$ certainly *looks* like a classical edge detector as familiar from the literature, *it is nothing of the sort.* The "edge" is only defined in the 2-jet as a whole.

The samples that lead to a "left edge"-interpretation lie within the convex hull of the left edge locus (see Fig. 3.9). Both the edge locus and the featureless axis lie on its exterior. The volume is almost exactly[2] one quarter of that of the samples solid.

The complement of the union of the left and right edge volumes exists of two non-convex, mutually congruent, disjunct regions. These regions correspond to dark and light bars. There are generically infinitely "bar"-explanations possible. In the majority of cases one may point out a unique full-bar explanation. Although the issue of bars is – just like that of the edges – an interesting one, it will not be pursued here.

Thus the samples solid divides into four parts of almost the same volume (see Fig. 3.10). The parts correspond to left or right edges and light or dark bars. Featureless images are non-generic (the featureless axis, a zero-volume set) and the blank and empty images even more so (they correspond to points). The applications oriented image processor might well worry whether this type of "edge detection" is not overly generous. After all, one doesn't want to find one quarter of the pixels classified as edges, edges should be sparse![3]

[2] Since we found these volumes only numerically (the expressions becoming extremely cumbersome) we cannot exclude the possibility that the ratio might be *exactly* $1/4$.

[3] One expects edges to be curve-like, whereas the image is a surface. Thus the fraction of pixels classified as "edge" should be of the order $N^{-\frac{1}{2}}$, where N denotes the number of pixels of the image (typically 10^4–10^6).

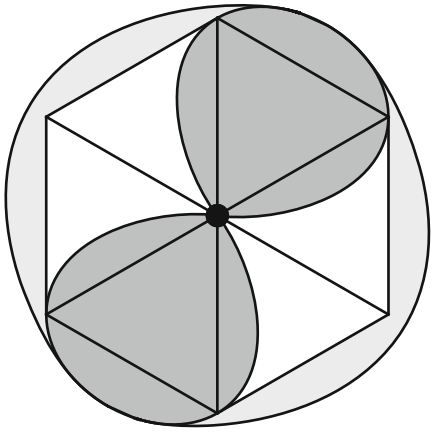


Fig. 3.10 The various regions that parcellate the samples solid as seen from the direction of the featureless axis. The *white hexagon* is the maximum volume inscribed crate, it is seen inside the (*light gray*) silhouette of the samples solid. The regions of edges are shown in *darker gray*

Fig. 3.11 An eye image (part from a larger image). The scan-line defines a 1-dimensional image used for the example

3.2.5.1 "Edges" in Applications

In applications there are many reasons to sharpen the definition of "edge":

- Samples obtained at nearby locations that lead to the detection of an edge may actually indicate the presence of "the same" edge. Since edge detections come with a location, a sophisticated algorithm might combine multilocal detections. A simpler method is to discard edge detections with locations that are too far from the location of the sample.
- Samples are always somewhat uncertain, thus a sample actually represents a finite volume in the samples solid. One should retain only the most conservative classification (i.e., "no edge", or "featureless") over this volume.

Such methods greatly cut down on the number of edge detections. The implementation of such algorithms is not necessarily simple and may require care and non-trivial decisions. This is the standard method used in practice. For most "edge detectors" one needs to set some kind of (essentially arbitrary) threshold value in order for the detector to select "edges". At zero threshold an edge detector finds and edge at every pixel, which is clearly an undesirable behavior.

In our type of edge definition even simple, straightforward implementations work quite well, even without any ad hoc measures such as arbitrary thresholding. Just for the sake of illustration, here is a simple example. Figure 3.11 defines a 1-dimensional image as a scan-line through an edge image.

A very simple algorithm is:

Algorithm I

A. Decide on a scale and compute the 2-jet for every pixel of the image;
B. Convert to the "three–pixel representation";
C. For every pixel do

C.1. Let the sample be $\{l, c, r\}$. Find the minimum of $\{l, c, r\}$, call it "lower level" m". Set $\{l, c, r\} \leftarrow \{l - m, c - m, r - m\}$;
C.2. Find the maximum of $\{l, c, r\}$, call it the "articulation a". Set $\{l, c, r\} \leftarrow \{l/a, c/a, r/a\}$. Now $\{l, c, r\}$ consists of the numbers 0, 1 and α, where $0 < \alpha < 1$;
C.2. According to the permutation, classify the pixel as:

Left edge for $\{1, \alpha, 0\}$;
Right edge for $\{0, \alpha, 1\}$;
Light bar for $\{\alpha, 1, 0\}$ or $\{0, 1, \alpha\}$;
Dark bar for $\{\alpha, 0, 1\}$ or $\{1, 0, \alpha\}$;

C.3. in case of a dark bar, set the "location" to $3(\alpha - 0.5)$, in case of a light bar, set the "location" to $3(0.5-\alpha)$. (Thus the locations are in the range $(-1.5, +1.5)$, the idea being that the pixels subtend the segments $(-1.5, -0.5)$, $(-0.5, +0.5)$ and $(+0.5, +1.5)$);

D. Find the ranges of adjacent pixels labeled "left edge", likewise find the ranges of adjacent pixels labeled "right edge". Delete ranges with location of a single sign (this includes ranges of length one);
E. Find the zero crossings of the locations as function of the pixel index;
F. Define the "edges" as these zero crossings, their nature (left or right edge) is determined, the edges jump from the lower level to the lower level plus three times the articulation (the "1" level extends over only one pixel, hence the factor 3).

This algorithm was applied to the scan-line through the eye image (Fig. 3.12). The length of the scan-line is 106 pixels. The scale parameter was taken $\sigma = 4.0$. The algorithm finds two right edges and one left edge. The result is evidently very reasonable, no further "cleaning" of any kind was required.

Notice that the edge detection process "sees" edges already from a distance as is evident from the graphs of locations. Thus edges are actually detected in a region of several times the scale at either side of "the edge".

Of course the edges indicated by the algorithm depend on the scale. In actual applications one will run the edge detection algorithm for a range of relevant scales. In the case of the example, halving the value of σ introduces an additional pair of (minor) left edges, doubling the value of σ retains only the prominent right edge, whereas quadrupling the value of σ discards all edges.

3.3 Conclusions

"Edges" can be defined in the context of the 2-jet of the image. A single sample is generically "explained" by infinitely many mutually metameric images. If the metameric set includes a step function then the sample is compatible with the

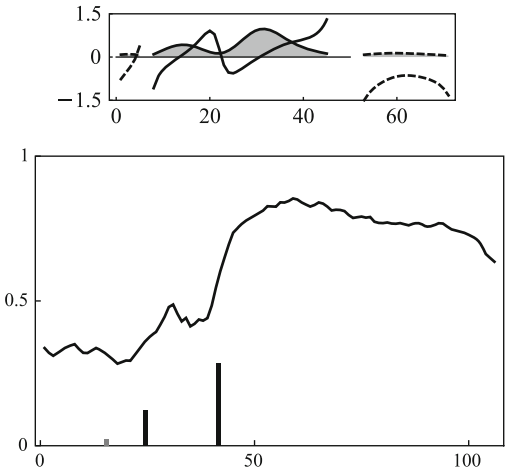


Fig. 3.12 Algorithm I applied to the scan-line of the eye image show in Fig. 3.11. In the top figure the articulations (the graphs with filling to the axis) and locations (*drawn lines* for *right edges*, *dashed lines* for *left edges*) have been plotted. In the lower figure the scan-line intensity and the edges found by the algorithm (*black* for *right*, *gray* for *left edges*) are indicated. The heights indicate the height of the jump

existence of an edge, otherwise it is not. If it is, then it makes sense to declare that one has "detected an edge". In that case one may show that the edge is uniquely determined through two distinct intensity levels and the location of the transition between these.

Notice the curious fact that this is an instance of

a theory of "edge detection" in which there does not feature any "edge detector".

The existence of an edge is signalled by the 2-jet as an integral entity. The first order differentiating kernel indeed looks much like a classical edge detector, but it is *not*. A certain activity of this operator may or may not correlate with the detection of an edge, that depends upon the activity of the other members of the 2-jet. It is not possible to define "edges" and "edge detectors" as such, despite the conventional praxis.

This theory has immediate consequences for *brain science*, where one glibly talks of "feature detectors" (including "edge detectors") as if this made sense [8]. It does *not* though, the implication being that the existence of a "feature" cannot be signalled by the activity of any single neural element. A cleaning of terminology is in order with important consequences for the basic understanding of brain functions.

References

1. Marr, D.: Vision. Freeman, San Francisco (1982)
2. Florack, L.M.J.: The Structure of Scalar Images. Kluwer, Dordrecht (1997)
3. Vico, G.: Scienza Nuova (The first new science, edited and translated by Leon Pompa). Cambridge University Press, Cambridge (1725; 2002)
4. Lyapunov, A.: Sur les fonctions-veteurs complètement additives. Bull. Acad. Sci. URSS Sér. Math. **4**, 465–478 (1940)
5. Halmos, P.: The range of vector measures. Bull. Am. Math. Soc. **54**, 416–421 (1948)
6. McShane, E.J.: The calculus of variations from the beginning of optimal control theory. SIAM J. Contr. Optim. **27**, 916–939 (1989)
7. Schrödinger, E.: Theorie der pigmente von größter leuchtkraft. Ann. Phys. **62**, 603–622 (1920)
8. Hubel, D.H.: Eye, Brain and Vision. Freeman, San Francisco (1988)

Chapter 4
Shape Modeling by Optimising Description Length Using Gradients and Parameterisation Invariance

Johan Karlsson, Anders Ericsson, and Kalle Åström

Abstract In Statistical Shape Modeling, a dense correspondence between the shapes in the training set must be established. Traditionally this has been done by hand, a process that commonly requires a lot of work and is difficult, especially in 3D. In recent years there has been a lot of work on automatic construction of Shape Models. In recent papers (Davies et al., Medical Image Computing and Computer-Assisted Intervention MICCAI'2001, pp. 57–65, 2001; Davies et al., IEEE Trans. Med. Imaging. 21(5):525–537 2002; Kotcheff and Taylor, Med. Image Anal. 2:303–314 1998) Minimum Description Length, (MDL), is used to locate a dense correspondence between shapes.

In this paper the gradient of the description length is derived. Using the gradient, MDL is optimised using steepest descent. The optimisation is therefore faster and experiments show that the resulting models are better.

To characterise shape properties that are invariant to similarity transformations, it is first necessary to normalise with respect to the similarity transformations. This is normally done using Procrustes analysis.

In this paper we propose to align shapes using the MDL criterion. The MDL based algorithm is compared to Procrustes on a number of data sets. It is concluded that there is improvement in generalisation when using MDL to align the shapes.

In this paper novel theory to prevent the commonly occurring problem of clustering under correspondence optimisation is also presented. The problem is solved by calculating the covariance matrix of the shapes using a scalar product that is invariant to mutual reparameterisations. An algorithm for implementing the ideas is proposed and compared to Thodberg's state of the art algorithm for automatic

J. Karlsson (✉)
Fraunhofer-Chalmers Research Centre for Industrial Mathematics, Göteborg, Sweden
e-mail: johan.karlsson@fcc.chalmers.se

A. Ericsson · K. Åström
Center for Mathematical Sciences, Lund University, Lund, Sweden
e-mail: anderse@maths.lth.se; kalle@maths.lth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_4,

shape modeling. The suggested algorithm is more stable and the resulting models are of higher quality according to the generalisation measure and according to visual inspection of the specificity.

4.1 Introduction

In recent years there has been a lot of work on automatic construction of Shape Models. Several different algorithms have been proposed for this purpose. The algorithms normally locate parameterisations of the shapes in the training set to get correspondences between the shapes.

There have been many suggestions on how to automate the process of building shape models or more precisely finding a dense correspondence for a set of shapes [2, 4, 23, 24, 27, 33, 37, 40]. Many approaches that correspond to human intuition attempt to locate landmarks on curves using shape features [4, 24, 37], such as high curvature. Local geometric properties, such as geodesics, have been tested for surfaces [40]. Different ways of parameterising the training shape boundaries have been proposed [2, 27]. The above cited are not clearly optimal in any sense. Many have stated the correspondence problem as an optimisation problem [3,6,13,15,18,19,22,23,29,34]. Specifically, just to mention a few, in [34] a measure is proposed and dynamic programming is applied to find the reparameterisation functions and in [3] shapes are matched using shape contexts.

In this paper we focus on the correspondence problem for shape models where the shape is the boundary of some object, i.e. a curve or a surface. Thus the correspondences should match, as one to one and onto functions, from one shape to each of all the others in the training set and the correspondences can be described by monotonous parameterisation functions.

There have been some recent interesting papers including code on the problem of matching one point set to another, [7, 8, 41]. However, these algorithms only match one shape to another, instead of working with the training set as a whole. Also, they do not enforce continuity and bijectivity. This means that the parameterisation functions do not have to be monotone and not all points need to be matched. Therefore these algorithms are not directly useful for shape modeling of curves and surfaces.

Minimum Description Length, (MDL) [13], is a paradigm that has been used in many different applications, often in connection with model optimisation. In recent papers [11, 13, 29] this paradigm is used to locate a dense correspondence between shapes by optimising MDL using for example Nelder Mead Simplex.

Being able to calculate the gradient, a variety of more efficient optimisation techniques can be considered. In this paper the gradient of the MDL with respect to the parameterisation functions is derived and a new algorithm to optimise the MDL-criteria using steepest descent is proposed and it is shown that it is more efficient than the previously proposed Nelder–Mead Simplex optimisation.

The common way to align shapes before building shape models is to do a Procrustes analysis [14, 21]. It locates the unknown similarity transformations by

minimising the sum of squared distances from the mean shape. Other methods exist. In Statistical Shape Analysis [5, 14, 28, 36] Bookstein and Kendall coordinates are commonly used to filter out the effect of similarity transformations. In this paper we propose to align shapes using MDL.

A problem when locating correspondences on shapes by optimizing over parameterisations is that correspondences of landmarks do not necessarily imply correspondence of the entire shapes. If many landmarks are placed in one point or a small region (called landmark clustering), the cost function measuring the correspondence may get a low value, but this value is based on a small part of the shape.

For the MDL-approach, a problem can be that if many landmarks are placed in one point or a small region (in this paper called landmark clustering) it is possible to get a low MDL-value without actually having good correspondences. It has been suggested that this can be prevented by using a "master example" [9], i.e. an example that is not reparameterised during optimisation. The idea is that each curve will be reparameterised to fit the "master example". This strategy breaks down if there are too many curves in the training set when optimising over all curves. In [38] a penalty is suggested if the mean of a parameterisation change over the examples in the training set is different from zero. A cost is introduced that penalises that the parameterisation functions change in the same direction on all curves. This means that the mean movement of corresponding landmarks should be zero during optimisation. Using a "master example" or a parameterisation node cost are ad hoc strategies that try to prevent clustering. The problem lies in that constraints on the parameterisation functions are needed in the optimisation.

In this paper novel theory to prevent clustering during correspondence optimisation is presented. This is done in a framework of developing a theory that is intrinsically defined for curves, as opposed to a finite sample of points along the curves. Instead of representing continuous curves using landmarks, the problem is treated analytically and numerical approximations are introduced at the last stage. The major problem here is to define shape variation in a way that is invariant to curve parameterisations. The problem is solved by calculating the covariance matrix of the shapes using a new scalar product that is invariant to global reparameterisations. This scalar product approximates equal weighting of all parts of the curves. Contrary to when using the standard scalar product it is no longer possible to neglect or give more focus to any part of the curve, i.e. clustering is prevented.

4.2 Preliminaries

4.2.1 Shape Modeling

The standard shape model describes the variation of a set of point-sets as a mean shape plus linear combinations of a set of basis vectors, represented as columns of ϕ:

$$\mathbf{x} = \mathscr{M}(\mathbf{b}) = \bar{\mathbf{x}} + \phi\mathbf{b}. \tag{4.1}$$

To get ϕ, let the i-th column vector of $\mathbf{X}$ be the configuration of landmarks (point-coordinates) for shape $\mathbf{x}_i$ when Procrustes analysis has been performed and the mean shape,

$$\bar{\mathbf{x}} = \sum_{i=1}^{n_s} \frac{1}{n_s} \mathbf{x}_i,$$

has been subtracted. If principal component analysis is applied to $\mathbf{X}$ and the principal components are put as columns in ϕ, the shapes can be described with the linear model in (4.1). A singular value decomposition of $\mathbf{X}$ gives us

$$\mathbf{X} = \mathbf{USV}^T. \tag{4.2}$$

Here $\mathbf{V}$ corresponds to ϕ in (4.1) and the diagonal of $\mathbf{S}^T\mathbf{S}$ gives the eigenvalues $\{\lambda_k | k = 1 \ldots n_s - 1\}$ corresponding to the variance in the training set along the axis described by the corresponding principal component.

4.2.2 Parameterisation Optimisation

For an illustration of how the parameterisation optimisation works, we look at shapes which are curves. The principles are the same for surfaces, but of course become more complex. Assume that a family of n_s geometric objects, represented by continuous curves

$$\mathbf{c}_i(s), i = 1, \ldots, n_s, s \in [0, 1]$$

is given. Each curve is represented using some parameterisation,

$$\mathbf{c}_i : [0, 1] \ni s \rightarrow \mathbf{c}_i(s) \in \mathbf{R}^2.$$

For simplicity, assume that the shapes are parameterised by arc length. We want to model the shape of these curves. A linear representation of the model for the continuous curves, as for the point distribution model in (4.1), would be desirable. To model the shape it is necessary to solve the correspondence problem, i.e. to find a set of reparameterisation functions $\{\gamma_i\}_{i=1}^{n_s}$, where $\gamma_i : [0, 1] \rightarrow [0, 1]$ are strictly increasing bijective functions such that $\mathbf{c}_i(\gamma_i(s))$ corresponds to $\mathbf{c}_j(\gamma_j(s))$ for all pairs (i, j) and all parameter values $s \in [0, 1]$, i.e. if $\mathbf{c}_i(\gamma_i(s))$ is at a certain anatomical point on shape i then $\mathbf{c}_j(\gamma_j(s))$ should be at the same anatomical point on shape j. Correspondence between the curves $\mathbf{c}_i(\gamma_i(s))$ and $\mathbf{c}_j(\gamma_j(s))$ is denoted

$$\mathbf{c}_i(\gamma_i(s)) :=: \mathbf{c}_j(\gamma_j(s)).$$

The same formulation can be used for closed curves by changing the interval $[0, 1]$ to the circle $\mathbf{S}^1$.

The traditional method to determine correspondence is to select a number of landmarks on each curve. From these landmarks a model $\mathscr{M}$ of shape is constructed. A cost function, such as the minimum description length (MDL) [32], is calculated from the shape variation model. Then the parameters of the basis functions controlling the parameterisation functions are optimised using a suitable optimiser to minimise this cost function. At each iteration in the optimisation procedure this results in the landmarks moving along the shapes.

4.2.3 Treating the Curves Analytically

In earlier work on locating shape correspondences [10, 12, 38] the shape curves have typically been sampled with 70–200 landmarks. During the optimisation of the correspondences, new sampling points have often been located by linear interpolation of the annotated landmarks.

Here we aim at treating the curves as continuous objects and attempting an analytical approach to the problem. At each iteration the shapes are resampled with a fixed number of sampling points defined by the current set of parameterisation functions. The resampling of the shapes is done in order to perform the numerical computations necessary to evaluate the cost function. This is done by introducing the numerical approximations at the last stage possible.

First assume that the curves $\{\mathbf{c}_i\}_{i=1}^{n_s}$ in the training set are continuous. Note that they may have an analytic representation as for example the synthetic training set with the parametric representation $\{\mathbf{c}_i(t) = (a_i cos(t), b_i sin(t))\}_{i=1}^{n_s}$, but when evaluating the curve at $t \in \mathbf{R}$ it will still be approximated up to the numerical precision of the computer. The numerical precision in MATLAB, for example, is 10^{-15}. This is not a big issue. The problem here is that a continous curve (function) is infinite dimensional.

The parameterisation functions used in this paper are built up by basis functions $p_{kl}(t)$ and parameterisation nodes n_{kl} in the following way:

$$\gamma_i(t) = t + \sum_{k=1,l=1}^{k=n_l,l=2^{(k-1)}} n_{kl}\, p_{kl}(t).$$

The basis functions $p_{kl}(t)$ in this work are tent-functions. Let $g(t)$ be the piecewise linear function,

$$g(t) = \begin{cases} t, & 0 \le t < 1/2 \\ 1-t, & 1/2 \le t \le 1 \end{cases}.$$

The width of the basis function at level k is $w = 1/2^{(k-1)}$ and

$$p_{kl}(t) = \begin{cases} 0, & 0 \le t \le w(l-1) \\ g((t - w(l-1))/w), & w(l-1) \le t \le wl \\ 0, & wl \le t \le 1 \end{cases}$$

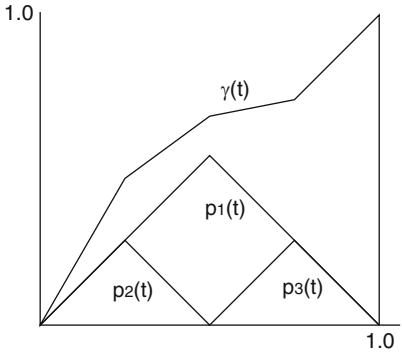


Fig. 4.1 Construction of parameterisation function γ using linear combination of tent functions

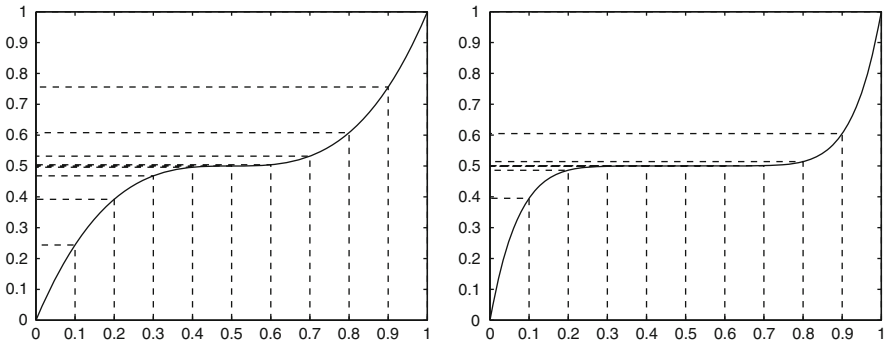


Fig. 4.2 In the two plots it can be seen how the sampling points, defined by the set $\mathscr{S}$, are mapped by two parameterisation functions. The plots describe how landmarks can cluster during optimisation of the parameterisation functions

In the rest of the paper a more simple expression for γ_i is used,

$$\gamma_i(t) = t + \sum_{l=1}^{n_n} n_{il}\, p_{il}(t).$$

During optimisation $\{n_{il}\}$ are constrained to ensure that $\{\gamma_i\}$ are strictly monotone. Note that, analogously to the curves, the parameterisation functions can also be evaluated to the numerical precision of the computer.

Figure 4.1 illustrates the construction of the parameterisation functions and Fig. 4.2 gives further illustration of how the parameterisation function works.

One way of determining ϕ from experimental data, where the curves $\{\mathbf{c}_i\}_{i=1}^{n_s}$ have been aligned for example according to the Procrustes condition (similarity transformations), is to make a singular value decomposition of the covariance matrix

$$\mathbf{C}_0 = \frac{1}{n_s - 1}\mathbf{X}\mathbf{X}^T, \tag{4.3}$$

where

$$\mathbf{X} = [\mathbf{c}_1 \circ \gamma_1 - \bar{\mathbf{c}}, \ldots, \mathbf{c}_{n_s} \circ \gamma_{n_s} - \bar{\mathbf{c}}],$$

and

$$\bar{\mathbf{c}}(s) = \frac{1}{n_s} \sum_{i=1}^{n_s} \mathbf{c}_i \circ \gamma_i(s).$$

This is straightforward in the case of finite point configurations, but for continuous curves it is more difficult since the 'column vectors' of $\mathbf{X}$ are infinite dimensional. However (with slight abuse of notation), the matrix

$$\mathbf{C} = \frac{1}{n_s - 1} \mathbf{X}^T \mathbf{X}, \tag{4.4}$$

where with infinite columns each element is really an integral, is finite dimensional and has the same non zero singular values as $\mathbf{C}_0$. Let $\mathbf{X}$ be finite dimensional for a moment. Using singular value decomposition $\mathbf{X}$ can be written $\mathbf{X} = \mathbf{USV}^T$ which means that $\mathbf{C} = \frac{1}{n_s - 1} \mathbf{VS}^2\mathbf{V}^T$ and $\mathbf{C}_0 = \frac{1}{n_s - 1} \mathbf{US}^2\mathbf{U}^T$. We thus get for $\mathbf{C}$ the principal components $\phi' = \mathbf{V}$ and for $\mathbf{C}_0$ the principal components $\phi = \mathbf{U}$. The basis ϕ can thus easily be calculated from the singular value decomposition of $\mathbf{C}$ in the following way,

$$\phi = \mathbf{U} = \mathbf{USV}^T\mathbf{VS}^{-1} = \mathbf{X}\phi'\mathbf{S}^{-1}. \tag{4.5}$$

This can now be generalised to an infinite dimensional $\mathbf{X}$ since $\mathbf{C}$ will still be finite. In order to do this (4.5) is interpreted in the following way. $\mathbf{M} = \phi'\mathbf{S}^{-1}$ is a matrix determining what combinations of the functions $\mathbf{X}_j$ (that with slight abuse of notation are columns in $\mathbf{X}$) make up the modes in ϕ. $\phi_i = \sum_j \mathbf{M}_{ji}\mathbf{X}_j$.

When optimising correspondences, the cost function must be evaluated in each iteration of the optimisation procedure. For example for MDL, $\mathbf{C}$ must be evaluated, where each element c_{ij} is a scalar product of the curve $(\mathbf{c}_i \circ \gamma_i - \bar{\mathbf{c}})$ with the curve $(\mathbf{c}_j \circ \gamma_j - \bar{\mathbf{c}})$,

$$c_{ij} = \frac{1}{n_s - 1} \int_0^1 (\mathbf{c}_i(\gamma_i(s)) - \bar{\mathbf{c}}(s)) \cdot (\mathbf{c}_j(\gamma_j(s)) - \bar{\mathbf{c}}(s)) ds. \tag{4.6}$$

To evaluate (4.6) the shapes must be sampled in a finite number of sample points, n_p. Define the set $\mathscr{S} = \{s_n : s_n = n/(n_p - 1), n = 0, \ldots, n_p - 1\}$. The sampling of the curves $\mathbf{c}_i$ can now be introduced at the last possible stage as,

$$\mathbf{c}_i(\gamma_i(\mathscr{S})).$$

To summarise, the curves $\mathbf{c}_i$ and parameterisation functions γ_i should have continous representation. The numerical approximations of (4.6) are done by sampling

$$\mathbf{c}_i(\gamma_i(\mathscr{S})).$$

at the points in $\mathscr{S}$ at the last stage possible.

In practise, the curves $\mathbf{c}_i$ are most commonly segmented from images. The precision of such curves is typically around 1,000–2,000 landmarks, typically with linear interpolation between them.

4.3 Minimum Description Length

4.3.1 Background

There are a number of different methods for selecting which model is to be considered optimal. Well established are for example AIC (An Information Criterion) and BIC (Bayes Information Criterion (BIC)) [20]. Another alternative is Ljungs Final Prediction Error (FPE) [30]. In shape modeling, Minimum Description Length (MDL), closely related to BIC, has proved to be a successful criterion, and this paper focuses on MDL. Minimum Description Length (MDL) was first introduced by Rissanen in [32]. Since then, this theory has been applied to many modeling problems in many different fields. In [9–11,13] Davies et al. applies MDL to determine the correspondences for a set of contours. The description length is a concept derived from information theory and is, in simple words, the effort that is needed to send the training set using the model bit by bit. The basic idea is to minimise the length of the message (the description length) required to transmit a full description of the training set.

The MDL – principle searches iteratively for the set of parameterisation functions γ_i that gives the cheapest message to transmit. The cost function makes a trade-off between a model that is general (can represent any instance of the object), specific (it can only represent valid instances of the object) and compact (it can represent the variation with as few parameters as possible). Davies and Cootes relate these ideas to the principle of Occam's razor: the simplest explanation generalises the best. The derivation of the description length presented here is along the lines of that in [9].

4.3.2 Deriving the Description Length

4.3.2.1 Setup

Given a training set of n_s shapes, each represented by n_p landmarks in n_d dimensions, the shape column vectors of $\mathbf{X}$ are $\mathbf{x}_i (i = 1, \ldots, n_s)$. Using the linear model from (4.1) we get

$$\mathbf{x}_i \approx \mathscr{M}(\mathbf{b}_i) = \bar{\mathbf{x}} + \phi \mathbf{b}_i = \bar{\mathbf{x}} + \sum_{j=1}^{n_m} \phi_j b_i^j, \tag{4.7}$$

where ϕ_j are the eigenvectors of the covariance matrix for the shapes and n_m is the number of shape modes.

We want to derive an expression for a cost function, deserving the name "description length". To this end, assume that the description length of the mean shape $\bar{\mathbf{x}}$ and of the eigenvectors ϕ_j are independent of the parameterisation, for a given training set. Thus only the description length of the training shapes, as represented by the coordinates b_i^j of the training shapes in shape space, need to be considered.

Dimensionality

Let the coordinates of the original shapes $\mathbf{x}_i$ in shape space, which is spanned by $\phi_1, \ldots, \phi_{n_m}$, be

$$b_i^j = (\mathbf{x}_i - \bar{\mathbf{x}})^T \phi_j, (i = 1, \ldots, n_s,\ j = 1, \ldots, n_m).$$

Since the eigenvectors are orthonormal, the total description length $\mathscr{L}$ of the set $\{b_i^j\}$, can be calculated as a sum

$$\mathscr{L} = \sum_{j=1}^{n_m} \mathscr{L}_j, \tag{4.8}$$

where $\mathscr{L}_j$ is the expression of the description length of the 1D-dataset $\mathscr{Y} = \{b_1^j, \ldots, b_{n_s}^j\}$.

Calculating the Description Length

Assume that the data can be described using a set of one-parameter Gaussian models. If the shapes in the training set are aligned using Procrustes, this is a reasonable assumption. Using Shannon's codeword length [35], given a probability density $\mathscr{P}$, the description length of a given value, $\hat{\alpha}$, encoded using a probabilistic model, is

$$-\log(\mathscr{P}(\hat{\alpha})). \tag{4.9}$$

In order to calculate the description length in bits the logarithms should be calculated in base 2. We are using the natural logarithm, which gives the description length in nats.

Quantisation

It requires an infinite amount of information to describe a real number to arbitrary accuracy. To measure the description length of a real number, the accuracy must

be limited and quantified. To describe the real number y to an accuracy δ, it is represented as an integer n, that solves

$$\min_{n\in\mathbb{N}} |y - \delta n|.$$

The approximation is then $\hat{y} = \delta n$.

The constant value δ should be related to the expected noise of the training shapes (typically one pixel/voxel in the original images from which they were annotated.)

Data Range

The range of the samples

$$\mathbf{x}_i = \left[x_i^1, \ldots, x_i^{n_d}, x_i^{n_d+1}, \ldots, x_i^{n_d n_p}\right]^T,$$

where the mean shape has already been subtracted, is defined so that all points, (in $\mathbf{R}^{n_d n_p}$), are bounded according to

$$-r \leq x_i^j \leq r, \qquad i = 1, \ldots, n_s,\ j = 1, \ldots, n_p n_d.$$

If these bounds are transformed into shape space, the range of the data, R, in shape space is obtained in the Euclidean metric as: $R^2 = r^2 + r^2 + \ldots + r^2 = n_p n_d r^2$. This implies that $|\mathbf{x}_i| \leq R = r\sqrt{n_p n_d}$, which in turn gives

$$|b_i^j| \leq |\mathbf{x}_i||\boldsymbol{\phi}_j| \leq R \quad \text{for all } i = 1, \ldots, n_s,\ j = 1, \ldots, n_m$$

since $|\boldsymbol{\phi}_j| = 1$.

4.3.2.2 The Description Length of a 1D-Dataset

To decode the encoded message the receiver must know the encoding model. It is therefore necessary to measure the description length of both the encoding model and the data. Using a two-part coding scheme, see [9], the description length is decomposed to

$$\mathscr{L} = \mathscr{L}_{param} + \mathscr{L}_{data}. \tag{4.10}$$

In the following, an expression for the description length of a data-set $\mathscr{Y} = \{\hat{y}_1, \ldots, \hat{y}_{n_s}\}$, where $\hat{y}_i$ is Gaussianly distributed, is derived. Knowing the description length of a Gaussianly distributed 1-D set $\mathscr{Y}$, the description length (4.8) of a shape model can easily be resolved.

Coding the Parameters

The description length of all parameters needed to describe the encoding model must be derived. Here the one-parameter Gaussian model is used. The frequency function of the Gaussian model is

$$f(x;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e(-\frac{x^2}{2\sigma^2}). \tag{4.11}$$

Thus it is only the σ-parameter that is needed to encode and decode the message and the only parameter that needs to be transmitted. An estimation s of σ from the quantified values of $\mathscr{Y} = \{\hat{y}_1, \ldots, \hat{y}_{n_s}\}$ is,

$$s = \sqrt{\frac{1}{n_s}\sum_{i=1}^{n_s} \hat{y}_i{}^2}.$$

Quantifying s to an accuracy of δ gives

$$\hat{s} = n\delta, n \in \mathbb{N}.$$

The range of $\hat{s}$ is $[0, s_{max}]$. Without prior knowledge, $\hat{s}$, is assumed to have the same range as the data, $s_{max} = R$. If s is assumed to be uniformly distributed in this range, the description length becomes

$$\mathscr{L}_{\hat{s}} = -\log \mathscr{P}(\hat{s}) = \log(\frac{R}{\delta}).$$

How δ can be chosen is discussed later. To be able to decode the message, the accuracy δ must be known to the receiver. First consider working with bits and the 2-logarithm. If δ is of the form $\delta = 2^k$, the description length can be calculated from

$$\mathscr{L}_\delta = 1 + |\log_2(\delta)|.$$

In order to work with the natural logarithm now instead let δ be of the form $\delta = e^k$ so that the description length can be calculated from

$$\mathscr{L}_\delta = 1 + |\log(\delta)|.$$

This yields the following total description length of the s-parameter

$$\mathscr{L}_{param} = \mathscr{L}_{\hat{s}} + \mathscr{L}_\delta = \log(\frac{R}{\delta}) + 1 + |\log(\delta)|.$$

Coding the Data

In (4.11) the frequency function of the Gaussian distribution is given. The probability, $\mathscr{P}(\hat{y})$, of x to lie in the range $\hat{y} \pm \Delta/2$ is

$$\mathscr{P}(\hat{y};\hat{s}) = \int_{\hat{y}-\frac{\Delta}{2}}^{\hat{y}+\frac{\Delta}{2}} f(x;\hat{s})dx. \tag{4.12}$$

This integral is approximated to the first order by

$$\mathscr{P}(\hat{y};\hat{s}) \approx \frac{\Delta}{\hat{s}\sqrt{2\pi}} e(-\frac{\hat{y}^2}{2\hat{s}^2}). \tag{4.13}$$

If, $\hat{s} \geq 2\Delta$, the approximation has a mean fractional error of less than $1 \pm 0.8\%$ according to Davies et al. [9]. Set therefore $s_{cut} = 2\Delta$. If $\hat{s}$ falls below s_{cut}, the approximation in (4.13) does not hold. However, since we always overestimates the description length if we set $\hat{s} = s_{cut}$, this can be used in this case. The case where all data, $\mathscr{Y}$, has the same quantified value ($\mathscr{Y} \leq \Delta$) must also be addressed. This gives us the following three different coding schemes:

1. If $\hat{s} > s_{cut}$, explicitly code the data.
2. If $\hat{s} \leq s_{cut}$, estimate $\hat{s}$ with s_{cut}.
3. If $\hat{s} \leq s_0$, which means that the range of $\mathscr{Y} \leq \Delta$, the description length is zero.

Case 1

Using (4.9) and (4.13) the code length of the data for one shape mode over all shapes is

$$\begin{aligned} \mathscr{L}_{data} &= -\sum_{i=1}^{n_s} \log(\mathscr{P}(\hat{y}_i)) = \\ & -n_s \log \Delta + \frac{n_s}{2} \log(2\pi\hat{s}^2) + \frac{1}{2\hat{s}^2} \sum_{i=1}^{n_s} \hat{y}_i{}^2 = \\ & -n_s \log \Delta + \frac{n_s}{2} \log(2\pi\hat{s}^2) + \frac{n_s s^2}{2\hat{s}^2}. \end{aligned} \tag{4.14}$$

Case 2

Estimating $\hat{s}$ by s_{cut} and putting it into (4.14) gives

$$\mathscr{L}_{data} = -\sum_{i=1}^{n_s} \log(\mathscr{P}(\hat{y}_i)) =$$

$$-n_s \log \Delta + \frac{n_s}{2} \log(2\pi s_{cut}^2) + \frac{n_s s^2}{2{s_{cut}}^2}. \tag{4.15}$$

Case 3

When the range of $\mathscr{Y} \leq \Delta$ ($\hat{s} \leq s_0$), the mean needs to be sent. But since the mean is zero it costs nothing to transmit

$$\mathscr{L}_{data} = 0.$$

An expression for the description of a 1D-dataset has now been derived. Since the total description length of a shape model can be written as a sum of description lengths of 1D-datasets, we are almost done.

Simulations

Running simulations with synthetic data, it is possible to estimate $\mathscr{L}_{data}$ for a given s at arbitrary precision. This is done by numerically solving (4.12) for a huge simulated dataset ($y_i \in N(0, s)$). In the top of Fig. 4.3 the simulated description length is compared to the DL calculated according to the analytical formulas (case 1–case 3) derived above for the description length of a dataset, where the elements are Gaussianly distributed. For this experiment $\Delta = 0.01$. It can be seen in the figure that setting $s_{cut} = 2\Delta = 0.02$ and letting s_0 be the limit where $\mathscr{Y} \leq \Delta$ (around 0.001) as proposed by Davies in [9] does not correspond well to the simulation when s is small (for case 2 and case 3). If s_0 and s_{cut} are changed to $s_{cut} = 2/5\Delta = 0.004$ and $s_0 = 1/5\Delta = 0.002$ the correspondence is much better. See bottom of Fig. 4.3 where the absolute difference between the numerically calculated description length and the two estimates are plotted.

Calculating the Optimal δ

The description length $\mathscr{L}$ for sending Gaussianly distributed data with quantified standard deviation $\hat{s} > s_{cut}$ is

$$\mathscr{L} = \mathscr{L}_{param} + \mathscr{L}_{data} =$$

$$\log(\frac{R}{\delta}) + 1 + |\log(\delta)| - n_s \log \Delta + \frac{n_s}{2} \log(2\pi\hat{s}^2) + \frac{n_s s^2}{2\hat{s}^2}$$

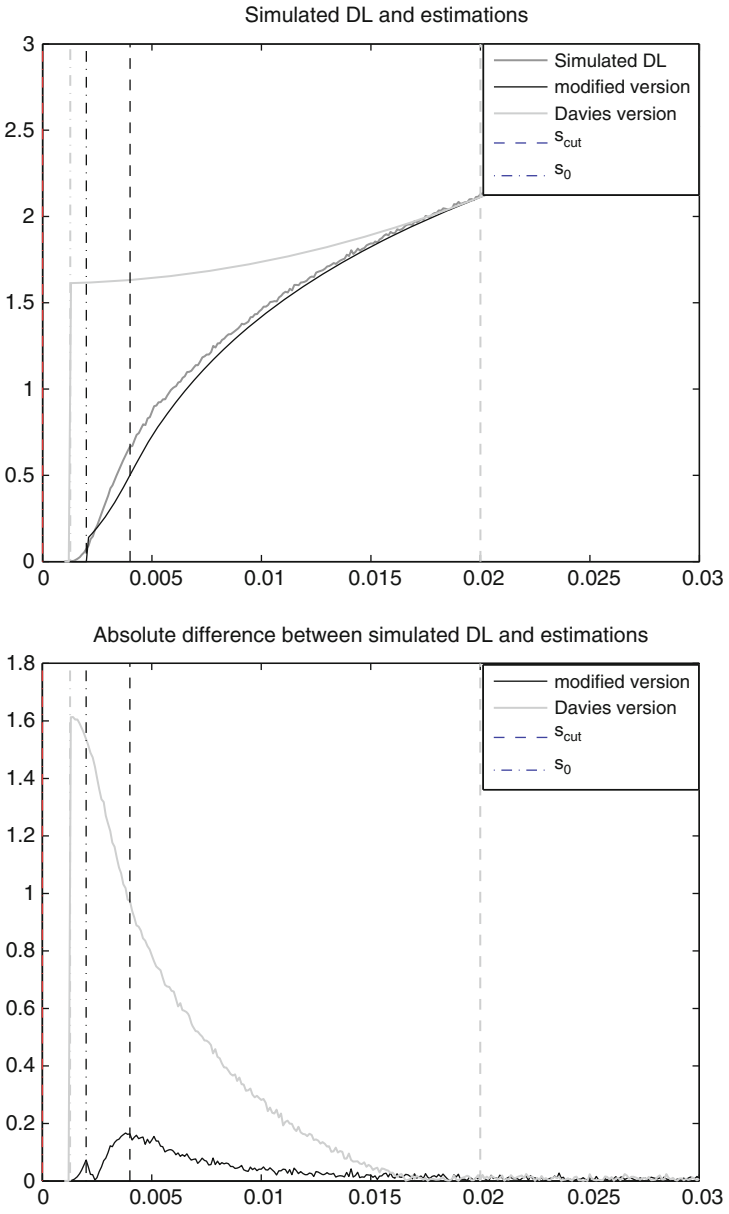


Fig. 4.3 The *top* plot shows simulated description length of Gaussianly distributed data together with Davies and a modified estimate. The *bottom* plot shows the error between the simulated and the two estimates of description length

where $\hat{s}$ is a function of s and δ. The optimal value of the accuracy δ can be found by differentiating $\mathscr{L}$ with respect to δ and setting it equal to zero. This gives

$$\delta_{opt} = \frac{s}{n}\sqrt{(\frac{n_s}{n_s - 2})}$$

if $\delta < 1$, or

$$\delta_{opt} = \frac{s}{n}$$

if $\delta \geq 1$.

The optimal δ can now be found by picking n to get the best local optimum.

In [9] $\mathscr{L}$ is approximated by

$$\mathscr{L} = K - \log(\delta) + |\log(\delta)| - \frac{n_s}{2}\log(2\pi s^2) + \frac{n_s}{2} + \frac{n_s\delta^2}{12s^2}.$$

Using this approximation gives the following optimal delta

$$\delta_o = \min(1, s\sqrt{\frac{12}{n_s}}).$$

Figure 4.4 shows how $\mathscr{L}$ depends on δ. It can be seen that δ_{opt} and δ_o differs significantly. The discontinuous behavior of the DL is caused by the quantification.

One effect of plugging in the optimal δ is that even though $\mathscr{L}_{total}$ is still continuous as a function of s, the derivative of $\mathscr{L}_{total}$ will not be continuous

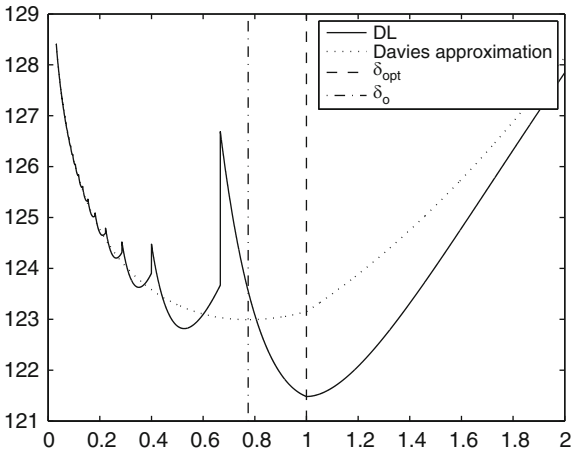


Fig. 4.4 The description length of Gaussianly distributed dataset and Davies approximation as a function of δ

anymore. Since this is a very important property and since the assumptions about the distribution of s is rough, we suggest that δ is chosen as a constant.

4.3.2.3 The Total Description Length

The total description length for the training set is given by (4.8), where $\mathscr{L}_j$ is the description length (for a 1D-dataset and one s-parameter) for the j-th component in shape space.

If Δ is sufficiently small, the quantified values approach the original values

$$\hat{y} \to y \text{ and } \hat{s}^2 \to s^2.$$

Using this fact gives the following expression for the total description length in the i-th direction. Here the description length of the parameter is added to the description length of the data as in (4.10). The index (1) and (2) refers to case 1 and case 2 respectively. We get

$$\begin{aligned}
\mathscr{L}_{i(1)} &= \log(\frac{R}{\delta}) + 1 + |\log(\delta)| - n_s \log(\Delta) + \\
&\quad \frac{n_s}{2}\log(2\pi) + n_s \log(s_i) + \frac{n_s}{2}, \\
\mathscr{L}_{i(2)} &= \log(\frac{R}{\delta}) + 1 + |\log(\delta)| - n_s \log(\Delta) + \\
&\quad \frac{n_s}{2}\log(2\pi) + n_s \log(s_{cut}) + \frac{n_s}{2}(\frac{s_i}{s_{cut}})^2.
\end{aligned} \tag{4.16}$$

Notice here, that the first five terms in the description lengths for both cases 1 and 2 are identical and considered constant. They only depend on parameters that are specific for a particular training set, but that do not depend on the parameterisation functions.

Substituting (4.16) into (4.8) gives

$$\begin{aligned}
\mathscr{L}_{total} &= F(n_s, R, \delta, \Delta) + \sum_{s_i > s_{cut}} \left(n_s \log(s_i) + \frac{n_s}{2} \right) + \\
&\quad \sum_{s_j \le s_{cut},\ s_j \ge s_0} \left(n_s \log(s_{cut}) + \frac{n_s}{2}(\frac{s_j}{s_{cut}})^2 \right),
\end{aligned} \tag{4.17}$$

where $s_i > s_{cut}$ indicates case 1 and $s_j \le s_{cut},\ \ s_j \ge s_0$ case 2. $F(n_s, R, \delta, \Delta)$ is a function that only depends on n_s, R, δ and Δ. It is constant for a given training set.

4.3.3 Simplifications of the Cost Function

Separately taking care of case 3 ($s_j \leq s_0$) does not change much of the value of the cost function in practice. The cost function can therefore be simplified to

$$\mathscr{L}_{total} = F(n_s, R, \delta, \Delta) + \sum_{s_i > s_{cut}} \left(n_s \log(s_i) + \frac{n_s}{2} \right) + \sum_{s_j \leq s_{cut}} \left(n_s \log(s_{cut}) + \frac{n_s}{2} (\frac{s_j}{s_{cut}})^2 \right).$$

Subtracting $n_s(n_s - 1)\log(s_{cut})$ from the cost function and factoring out $n_s/2$ we get,

$$\mathscr{L}_{total} = \hat{F}(n_s, R, \delta, \Delta) + \frac{n_s}{2} \left(\sum_{s_i > s_{cut}} \left(\log(\frac{s_i}{s_{cut}})^2 + 1 \right) + \sum_{s_j \leq s_{cut}} (\frac{s_j}{s_{cut}})^2 \right). \tag{4.18}$$

For a given training set, $\hat{F}(n_s, R, \Delta)$ is constant. The eigenvalue λ_i of the covariance matrix (4.3), which decides the shape modes of the linear model in (4.1), is equal to s_i^2. Substituting $s_i^2 = \lambda_i$, $s_{cut}^2 = \lambda_c$, focusing on the terms that can change under optimisation and omitting the amplitude $n_s/2$ we get the following expression for optimisation:

$$\mathscr{L} = \sum_{\lambda_i > \lambda_c} \left(\log(\frac{\lambda_i}{\lambda_c}) + 1 \right) + \sum_{\lambda_j \leq \lambda_c} \frac{\lambda_j}{\lambda_c}. \tag{4.19}$$

If $s_j \geq s_0$, it is only the amplitude $n_s/2$ and the constant $\hat{F}$ that differ (4.19) from (4.17). During optimisation the constant function $\hat{F}$ and the amplitude $n_s/2$ can be ignored. Equation (4.19) was first stated in [38] but without motivating derivation. The advantages of this cost-function are that it is simpler, more intuitive than those formerly presented and that the derivatives are continuous.

4.4 Using the Gradient

4.4.1 Background

Lately there has been much attention to MDL and its effectiveness for locating correspondences in automatic shape modeling. A successful algorithm, but a problem with this method is that the objective function is not stated explicitly. Another problem is the slow convergence of the optimisation step. Besides that, the optimisation process can easily get stuck at local minima. Due to these reasons

it can be hard to optimise. The cost function only depends on the eigenvalues λ_i of the covariance matrix (4.3), except for constants. The eigenvalues of a matrix are implicitly defined given the matrix. In previous work the Nelder-Mead Simplex method has been used to solve the optimisation problem. In general this optimisation technique is slow. In this section the theory presented in [31] is applied and the gradient of the description length (4.19) with respect to the parameterisation functions is derived along the lines in our previous work [19]. Being able to calculate the gradient, a variety of optimisation techniques can be considered. In this section a new algorithm to optimise the MDL-criteria using steepest descent is proposed and it is shown that it is more efficient than the previously proposed Nelder–Mead Simplex optimisation.

4.4.2 Computing the Derivatives of the Singular Values

In [31], theory for computing the Jacobian of the singular values decomposition is presented. A recapitulation is given here. For a more thorough description we recommend Andrew's work, in particular [1].

Equation (4.2) shows how a matrix $\mathbf{A}$ always can be factorised using Singular Value Decomposition (SVD). Here we are interested in computing the derivatives of the singular values, $\frac{\partial s_k}{\partial a_{ij}}$, for every element a_{ij} of the $M \times N$ matrix $\mathbf{A}$. Taking the derivative of (4.2) with respect to a_{ij} gives

$$\frac{\partial \mathbf{A}}{\partial a_{ij}} = \frac{\partial \mathbf{U}}{\partial a_{ij}}\mathbf{S}\mathbf{V}^T + \mathbf{U}\frac{\partial \mathbf{S}}{\partial a_{ij}}\mathbf{V}^T + \mathbf{U}\mathbf{S}\frac{\partial \mathbf{V}^T}{\partial a_{ij}}. \tag{4.20}$$

Since $\mathbf{U}$ is an orthogonal matrix, we have

$$\mathbf{U}^T\mathbf{U} = \mathbf{I} \Rightarrow \frac{\partial \mathbf{U}^T}{\partial a_{ij}}\mathbf{U} + \mathbf{U}^T\frac{\partial \mathbf{U}}{\partial a_{ij}} = \omega_{\mathbf{U}}^{ij\,T} + \omega_{\mathbf{U}}^{ij} = \mathbf{0}, \tag{4.21}$$

where $\omega_{\mathbf{U}}^{ij}$ is given by

$$\omega_{\mathbf{U}}^{ij} = \mathbf{U}^T\frac{\partial \mathbf{U}}{\partial a_{ij}}. \tag{4.22}$$

From (4.21) it is clear that $\omega_{\mathbf{U}}^{ij}$ is an anti-symmetric matrix. Similarly, an anti-symmetric matrix $\omega_{\mathbf{V}}^{ij}$ can be defined for $\mathbf{V}$ as

$$\omega_{\mathbf{V}}^{ij} = \frac{\partial \mathbf{V}^T}{\partial a_{ij}}\mathbf{V}. \tag{4.23}$$

By multiplying (4.20) by $\mathbf{U}^T$ and $\mathbf{V}$ from left and right, respectively, and using (4.22) and (4.23), one obtains

$$\mathbf{U}^T \frac{\partial \mathbf{A}}{\partial a_{ij}} \mathbf{V} = \omega_{\mathbf{U}} ij \mathbf{S} + \frac{\partial S}{\partial a_{ij}} + \mathbf{S} \omega_{\mathbf{V}} ij \,. \tag{4.24}$$

Since $\omega_{\mathbf{U}}^{ij}$ and $\omega_{\mathbf{V}}^{ij}$ are anti-symmetric matrices, all their diagonal elements are zero. Recalling that $\mathbf{S}$ is a diagonal matrix, it follows that the diagonal elements of $\omega_{\mathbf{U}}^{ij}\mathbf{S}$ and $\mathbf{S}\omega_{\mathbf{V}}^{ij}$ are zero too. Thus, (4.24) yields the derivatives of the singular values as

$$\frac{\partial s_k}{\partial a_{ij}} = u_{ik} v_{jk}, \tag{4.25}$$

since $\frac{\partial a_{kl}}{\partial a_{ij}} = 0$ for all $(k,l) \neq (i,j)$, while $\frac{\partial a_{ij}}{\partial a_{ij}} = 1$.

4.4.3 The Gradient of the Description Length

In the proposed implementation each parameterisation function γ_j $(j = 1, \ldots, n_s)$ has n_n parameterisation nodes. Parameterisation node l on curve j is denoted n_{jl}. How the parameterisation nodes affect the sampling of the curves is described in Sect. 1. We want to minimise $\mathscr{L}$ by using more sophisticated optimisation methods that use the gradient, i.e. the derivatives $\frac{\partial \mathscr{L}}{\partial n_{jl}}$. The parameters λ_k in (4.19) depend on the nodes n_{jl}. Differentiation gives

$$\frac{\partial \mathscr{L}}{\partial n_{jl}} = \sum_{\lambda_k \geq \lambda_c} \frac{1}{\lambda_k} \frac{\partial \lambda_k}{\partial n_{jl}} + \sum_{\lambda_k < \lambda_c} \frac{1}{\lambda_c} \frac{\partial \lambda_k}{\partial n_{jl}}. \tag{4.26}$$

Now, if x_{ij} is the i-th landmark on shape j and $\frac{\partial x_{ij}}{\partial n_{jl}}$ is the derivative of the i-th landmark on shape j with respect to control node n_{jl} then

$$\frac{\partial \lambda_k}{\partial n_{jl}} = \frac{\partial s_k^2}{\partial n_{jl}} = 2 s_k \frac{\partial s_k}{\partial n_{jl}} = 2 s_k \sum_i \frac{\partial s_k}{\partial x_{ij}} \frac{\partial x_{ij}}{\partial n_{jl}}.$$

Using the results from (4.25) that $\frac{\partial s_k}{\partial x_{ij}} = u_{ik} v_{jk}$, one obtains

$$\frac{\partial \lambda_k}{\partial n_{jl}} = 2 s_k \sum_i u_{ik} v_{jk} \frac{\partial x_{ij}}{\partial n_{jl}} = 2 s_k v_{jk} \mathbf{U}_k^T \frac{\partial \mathbf{X}_j}{\partial n_{jl}}. \tag{4.27}$$

Before sampling, a continuous version of $\mathbf{X}_j$ is

$$\mathbf{X}_j(t) = \mathbf{x}_j(\gamma_j(t)) - \bar{\mathbf{x}}(\bar{\gamma}(t)),$$

where

$$\bar{\mathbf{x}} = \frac{1}{n_s} \sum \mathbf{x}_j(\gamma_j(t))$$

and

$$\bar{\gamma} = \frac{1}{n_s} \sum \gamma_j(t)$$

which means that differentiation of $\mathbf{X}_j(t)$ with respect to n_{jl} becomes

$$\frac{\partial \mathbf{X}_j(t)}{\partial n_{jl}} = \frac{\partial \mathbf{x}_j(\gamma_j(t))}{\partial n_{jl}} - \frac{\partial \bar{\mathbf{x}}(\bar{\gamma}(t))}{\partial n_{jl}} =$$

$$\mathbf{x}_j' \frac{\partial \gamma_j(t)}{\partial n_{jl}} - \bar{\mathbf{x}}' \frac{\partial \bar{\gamma}(t)}{\partial n_{jl}} = \mathbf{x}_j' p_{jl} - \bar{\mathbf{x}}' \bar{p}_l, \tag{4.28}$$

$$\bar{p}_l = \frac{1}{n_s} \sum_j p_{jl}(t), \tag{4.29}$$

In the experimental validation the elements $\frac{\partial x_{ij}}{\partial n_{jl}}$ of $\mathbf{x}_j'$ is estimated using difference approximation.

4.4.4 Algorithm for Minimising the Description Length

If the gradient of the cost function is known for a specific optimisation problem, it generally pays off to use more sophisticated optimisation techniques than the Nelder–Mead Simplex method. Here steepest descent is proposed. An overview of the proposed algorithm is presented in the boxed Algorithm 1 and each step is explained in more detail below.

Algorithm 1 Algorithm for minimising the description length

Input: The training set, which correspondences are to be optimised $\mathbf{c}_i$, i $= 1, \ldots, n_s$
Output: The optimised correspondences $\mathbf{x}_i = \mathbf{c}_i(\gamma_i)$, i $= 1, \ldots, n_s$

1. INITIALISATION
   Initially the reparameterisation functions are set to arc-length parameterisation.
2. **Repeat** lines 3–6 until convergence according to some stop criterium
3. RESAMPLE SHAPES
   The curves are resampled according to the current parameterisation nodes
4. RESCALE AND ALIGN SHAPES
   The curves are aligned.
5. CALCULATE DL AND THE GRADIENT OF DL
   Calculate the gradient of the DL with respect to the parameterisation nodes n_{jl}.
6. UPDATE PARAMETERISATIONS
   Search for a local minima in the gradient direction.
7. **end** of repeat loop

1. Initialisation

Initially the number of nodes, the number of landmarks and a few optimisation parameters are set to appropriate values. The reparameterisation functions are set to arc-length and each curve is initialised to arc-length parameterisation.

2. Start loop

3. Resample shapes

The curves are resampled as explained in Sect. 1.

4. Rescale and align Shapes

After all the curves have been resampled they must be aligned and rescaled according to for example the Procrustes alignment. When Procrustes is applied all landmarks are equally weighted. Therefore the Procrustes method perform best if the landmarks are approximately equally distributed around the shapes. This is important to bear in mind.

5. Calculate DL and the gradient of DL

In this step the gradient according to (4.27) is calculated for all parameterisation nodes.

6. Update parameterisations

A search for a local minimum is performed in the gradient direction. When the position of the nearest local minima in the gradient direction has been estimated, all parameterisation nodes are updated at the same time.

7. Check convergence

Once the parameterisation nodes have been updated the algorithm checks for convergence and if not cenverged starts over at 3.

4.5 Alignment Using MDL

4.5.1 Introduction

To characterise shape properties that are invariant to similarity transformations, it is first necessary to normalise with respect to the similarity transformations

from the annotated configurations. This is normally done using Procrustes analysis, that is minimising the sum of squared distances between the corresponding landmarks under similarity transformations. One problem with this approach is that the algorithm then tries to optimise two different goal functions at the same time. In the alignment step the sum of squared distances is minimised and then in the parameterisation evaluation step the description length is optimised. This may cause problems for the optimiser. It makes more sense to do the alignment using the same goal function that is being used for the parameterisation evaluation step, so that the goal function is optimised with respect to parameterisation and alignment.

In this section we propose to align shapes using the Minimum Description Length (MDL) criterion (4.19). This is built on our previous paper [16]. We have seen in the previous section how MDL can be used to locate correspondences. The theory presented in [31] was recapitulated and the gradient of the description length [19] with respect to the parameterisation nodes was derived. In this section the theory is applied in order to align shapes using the Minimum Description Length (MDL) criterion. The experiments are done for 2D-shapes, but in principle it would work for any dimension. Here it is shown among other things that the Procrustes alignment with respect to rotation is not optimal with respect to description length.

4.5.2 Optimisation

When the gradient of an objective function is known for a specific optimisation problem, it generally pays off to use more sophisticated optimisation techniques that use the gradient. The Description Length is here minimised using Gauss–Newton. In order to do this the derivative of the description length with respect to rotation is derived. This derivation is analogous to the derivation of the gradient of the description length with respect to the parameterisation nodes. Initially the configurations are translated to the origin. Then the shapes are normalised so that the Euclidian norm is one for all shapes. The translation of all the shapes to the origin turns out to be optimal also for the DL alignment. It could be interesting to also optimise scale but the global scale must then be preserved. This means that it would be necessary to optimise under constraints. In this work only rotation is optimised using Gauss–Newton.

4.5.3 The Gradient of the DL

Assume that the shapes are in 2D.

Let n_s shapes $\mathbf{x}_1, \ldots, \mathbf{x}_{n_s}$ in complex coordinates be centred at the origin and with the scale normalised so that the Euclidian norm is one. In complex coordinates a shape $\mathbf{x}_j$ is rotated as $\mathbf{x}_j e^{i\theta_j}$. Differentiating (4.19) with respect to θ_j, we get

$$\frac{\partial DL}{\partial \theta_j} = \sum_{\lambda_k \geq \lambda_c} \frac{1}{\lambda_k} \frac{\partial \lambda_k}{\partial \theta_j} + \sum_{\lambda_j < \lambda_c} \frac{1}{\lambda_c} \frac{\partial \lambda_k}{\partial \theta_j}. \tag{4.30}$$

Let's derive $\frac{\partial \lambda_k}{\partial \theta_j}$ and thereby the gradient. Let the j-th column of $\mathbf{X}$ be the configuration of landmarks for shape j after rotation according to θ_j,

$$\mathbf{X} = \left[\mathbf{x}_1 e^{i\theta_1}, \ldots, \mathbf{x}_{n_s} e^{i\theta_{n_s}} \right].$$

Let $\mathbf{Y}$ be the matrix holding the deviations from the mean shape,

$$\mathbf{Y} = \mathbf{X} - \overline{\mathbf{X}},$$

where each column in $\overline{\mathbf{X}}$ is the mean shape $\bar{\mathbf{x}}$,

$$\bar{\mathbf{x}} = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathbf{x}_j e^{i\theta_j}.$$

If we now apply a principal component analysis to, we can describe our shapes with the linear model in (4.1). A singular value decomposition of $\mathbf{Y}$ gives us $\mathbf{Y} = \mathbf{USV}^T$. Here $\mathbf{V}$ corresponds to ϕ in (4.1) and the diagonal matrix $\mathbf{S}^T\mathbf{S}$ holds the eigenvalues λ_k.

Now, if y_{ij} is the i-th landmark on shape j and $\frac{\partial y_{ij}}{\partial \theta_j}$ is the derivative of the i-th landmark on shape j with respect to the rotation θ_j then

$$\begin{aligned} \frac{\partial \lambda_k}{\partial \theta_j} &= \frac{\partial s_k^2}{\partial \theta_j} = 2 s_k \frac{\partial s_k}{\partial \theta_j} = \\ 2 s_k \sum_{pq} \frac{\partial s_k}{\partial y_{pq}} \frac{\partial y_{pq}}{\partial \theta_j} &= 2 s_k \sum_{pq} u_{pk} v_{qk} \frac{\partial y_{pq}}{\partial \theta_j}. \end{aligned} \tag{4.31}$$

Here it is used that $\frac{\partial s_k}{\partial y_{pq}} = u_{pk} v_{qk}$, where u_{pk} and v_{qk} are elements in $\mathbf{U}$ and $\mathbf{V}$, see Sect. 4.4.2. Moreover,

$$\begin{aligned} \frac{\partial y_{pq}}{\partial \theta_j} &= \frac{\partial \mathbf{x}_q(p) e^{i\theta_q}}{\partial \theta_j} - \frac{1}{n_s} \sum_l \frac{\partial \mathbf{x}_l(p) e^{i\theta_l}}{\partial \theta_j} = \\ & \begin{cases} i(1 - \frac{1}{n_s}) \mathbf{x}_j(p) e^{i\theta_j} & q = j \\ -i \frac{1}{n_s} \mathbf{x}_j(p) e^{i\theta_j} & q \neq j \end{cases}, \end{aligned} \tag{4.32}$$

$$\text{since} \quad \frac{\partial \mathbf{x}_q(p) e^{i\theta_q}}{\partial \theta_j} = \begin{cases} i \mathbf{x}_j(p) e^{i\theta_j} & q = j \\ 0 & q \neq j \end{cases}.$$

If n_s is reasonably large (this is assumed in our implementation), the second term can be ignored so that we get

$$\frac{\partial y_{pq}}{\partial \theta_j} = \begin{cases} i\mathbf{x}_j(p)e^{i\theta_j} & q = j \\ 0 & q \neq j \end{cases}.$$

Then $\frac{\partial \lambda_k}{\partial \theta_j}$ can be written as

$$\frac{\partial \lambda_k}{\partial \theta_j} = 2s_k i v_{jk} \mathbf{U}_k^T \mathbf{X}_j, \tag{4.33}$$

where $\mathbf{U}_k$ is the k-th column in $\mathbf{U}$ and $\mathbf{X}_j$ is the j-th column in $\mathbf{X}$. Note that the result is very similar to the final equation (4.27) in the previous section.

4.6 Parameterisation Invariance

A problem when locating correspondences on shapes by optimizing over parameterisations is that correspondences of landmarks do not necessarily imply correspondence of the entire shapes. If many landmarks are placed in one point or a small region (called landmark clustering), the cost function measuring the correspondence may get a low value, but this value is based on a small part of the shape. This problem arises in many situations where shapes need to be reparameterised.

One problem with the MDL approach is that if many of the sampled points are concentrated to a particular region on all curves, a very low description length is achieved. This can be partially prevented by using a "master example", i.e. an example that is not reparameterised during optimisation. The idea is that each curve will reparameterise to fit to the "master example". This strategy breaks down if there are too many curves in the training set. A node cost can also be applied as suggested in [38]. The node cost penalises that corresponding nodes move in the same direction on all curves during the optimisation. The mean movement of corresponding nodes should be zero during optimisation. In this section a scalar product that is invariant to global reparameterisation is introduced. This is based on our previous paper [26].

In this section it is shown that the infimum of the description length (DL) in the standard formulation is always zero. The global minimum for the standard formulation is determined by reparameterising all the shapes so that all landmarks are positioned in approximately the same point. This means that the global minimum corresponds to the worst possible solution for landmark correspondence. Using the new scalar product when calculating the covariance matrix makes the optimisation well defined in the sense that the infimum of the DL does not correspond to moving all the landmarks to approximately the same point. With this formulation

the tendency for clustering of landmarks (placing all landmarks at one point or in a small region) is significantly reduced.

The following deals with 2D-curves. Similar principles could be applied to 3D-surfaces.

4.6.1 Difficulties with Parameterisation Dependent Methods

There are several potential problems with optimising the correspondences of continous curves as explained in Sect. 4.2.2. One problem is that it is not certain that the minimisation problem is well defined. It might very well be the case that the infimum is not attained at any point, or that the global minimum does not correspond to a meaningful solution. In such cases, one might hope that there are local minima that are meaningful, but it is not even certain that there are any such local minima.

The standard shape model describes the variation of a set of points as linear combinations of a set of basis vectors

$$\mathbf{x} = \bar{\mathbf{x}} + \phi \mathbf{b}.$$

The goal here is to derive a shape theory that is intrinsically defined for curves and is independent of parameterisations. It is reasonable to aim at a linear model, i.e.

$$\mathbf{c}(s) = \bar{\mathbf{c}}(s) + \sum_{k=1}^{n_m} \mathbf{b}_k \phi_k(s) = \bar{\mathbf{c}}(s) + \phi(s)\mathbf{b}, s \in [0, 1].$$

ϕ consists of the principal components of the formally infinite dimensional covariance matrix of the shapes in the training set.

As seen above, a possible criterion for solving the correspondence problem $(\mathbf{c}_i(\gamma_i(s))) := \mathbf{c}_j(\gamma_j(s))$ for all i,j) is the description length of the shape model. The reparameterisation functions $\{\gamma_i\}_{i=1}^{n_s}$ are located by minimizing the description length.

The description length of a shape model $\mathscr{M}$ is

$$DL(\mathscr{M}) = \sum_{\lambda_i \geq \lambda_c} (1 + \log \frac{\lambda_i}{\lambda_c}) + \sum_{\lambda_i < \lambda_c} \frac{\lambda_i}{\lambda_c} + K,$$

where the scalars λ_i are the eigenvalues of the covariance matrix $\mathbf{C}$ of the shapes in the training set, the scalar λ_c is a cutoff constant and K is a scalar, which is independent of the parameterisations. The constant K can be ignored during optimisation and the following cost function is used,

$$F(\mathbf{C}) = \sum_{\lambda_i \geq \lambda_c} (1 + \log \frac{\lambda_i}{\lambda_c}) + \sum_{\lambda_i < \lambda_c} \frac{\lambda_i}{\lambda_c}, \tag{4.34}$$

where $\mathbf{C}$ is the covariance matrix defined above and the elements of $\mathbf{C}$ are scalar products between the shapes in the training set, as in (4.6).

The scalar product depends on the reparameterisation functions γ_i and γ_j. How a parameterisation function can distribute the weights on different parts of a curve is illustrated in Fig. 4.2. In this figure it can be seen how two parameterisation functions put more weight on the middle part of the curve. Using the standard scalar product in this case, it is possible that $|\gamma_i(t_n) - 0.5| \leq \epsilon$ for all sample points t_n. This means that $\mathbf{c}_i(s), s \in (0, 0.5 - \epsilon) \cup (0.5 + \epsilon, 1)$ is not taken into account in the calculation of the scalar product.

Even if the entire shape is considered to some extent, the result is not independent of curve parameterisations. By changing all parameterisations γ_i with a mutual parameterisation γ, so that $\tilde{\gamma}_i = \gamma_i \circ \gamma$, one effectively puts different weights at different parts of the curves without changing the correspondences. The same problem occurs when a discrete set of landmarks is used as in [10, 12]. The most weight is put on that part of the curve which has the most landmarks. The problem with this is that the optimisation scheme to locate correspondences becomes ill-posed as illustrated by the following theorem.

Theorem 4.1. *For $F(\mathbf{C})$ defined by (4.34),*

$$\inf_{\gamma_1,\ldots,\gamma_{n_s}} F(\mathbf{C}) = 0.$$

The infimum is attained when the reparameterisation functions approach a measure with all mass at one point.

Proof. Assume that a training set of continuous curves $\mathbf{c}_i(t), i = 1, \ldots, n_s, t \in [0, 1]$, is given and that the reparameterisation functions $\gamma_i, i = 1, \ldots, n_s$ put the shapes in correspondence,

$$\mathbf{c}_i(\gamma_i(s)) :=: \mathbf{c}_j(\gamma_j(s)).$$

Now, construct a reparameterisation function γ such that

$$\gamma(t) = \begin{cases} \frac{\epsilon}{1-\epsilon} t, & 0 \leq t \leq 1 - \epsilon \\ \epsilon + \frac{1-\epsilon}{\epsilon}(t - (1 - \epsilon)), & 1 - \epsilon \leq t \leq 1 \end{cases}.$$

This function is illustrated in Fig. 4.5. The correspondence of the shapes is retained if all curves are reparameterised with the parameterisation function γ. If now $\epsilon \to 0$ all weight will be situated at the start of the curves so that effectively all curves approach points. Performing Procrustes on these shapes will result in that in the limit $\epsilon \to 0$ all shapes coincide with the mean shape. This means that the covariance matrix $\mathbf{C}$ will be the zero matrix. Hence all eigenvalues of $\mathbf{C}$ are zero and $F(\mathbf{C}) = 0$.

□

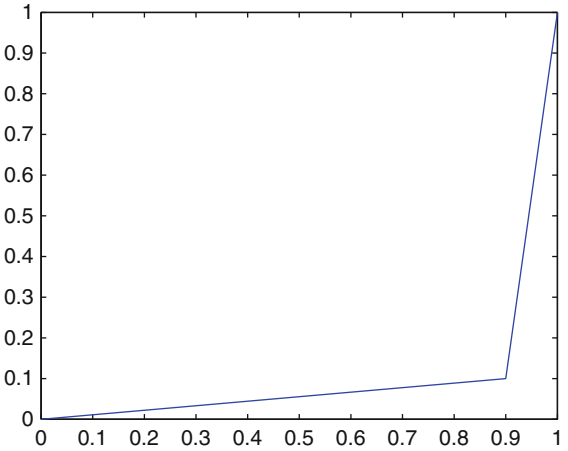


Fig. 4.5 Mutual reparameterisation function. Here $\epsilon = 0.1$

This minimum can easily be found. Assume that the curves are in correspondence. The optimiser can now attempt to concentrate points on parts of the curve, while keeping correspondence, by performing mutual reparameterisations, i.e reparameterising all curves by the same parameterisation function. This shows that the minimisation criterion is not well defined. A good global optimiser would find the global minimum. Even if the global minimum is not found it is still possible to reduce $F(\mathbf{C})$ by concentrating the landmarks on parts of the curves with low variance. This means that it is possible to reduce the DL in two ways, both by finding correspondences and by performing a mutual reparameterisation. By mutual reparameterisation is meant a reparameterisation function that is applied to all curves in the training set. Such a mutual reparameterisation function can put most of the weight at one point or a small section of the curve. Therefore previous algorithms for automatic landmark placement on curves tend to gather points together.

In order to avoid this several authors have presented preliminary solutions such as keeping the parameterisation of the first curve unchanged or penalising "bad" reparameterisations. Even such methods have difficulties. For example, if one keeps the parameterisation of the first curve unchanged it is still possible to reparameterise the other $n - 1$ curves so as to put more weight to a particular part of the curves. If many curves ignore the mismatch with the unchanged shape, their clustering can lead to a lower description length.

4.6.2 A Parameterisation Invariant Method

To improve the algorithm discussed above, a new scalar product, which is invariant under mutual reparameterisations is proposed. This removes the undesired way to reduce the DL.

4.6.2.1 Defining the New Scalar Product

Definition 4.1. Let $\{\mathbf{c}_i\}_{i=1}^{n_s}$ be curves centered at the origin and let them be parameterised with $\{\gamma_i\}_{i=1}^{n_s}$ and let them have their mean curve $\bar{\mathbf{c}}(s) = 1/n_s \sum_{i=1}^{n_s} \mathbf{c}_i(\gamma_i(s))$ subtracted. A new scalar product between $\mathbf{c}_i$ and $\mathbf{c}_j$ is defined by

$$\mathbf{c}_i \cdot \mathbf{c}_j = \frac{1}{n_s} \sum_{k=1}^{n_s} \int_0^1 \mathbf{c}_i \left(\gamma_i \circ \gamma_k^{-1}(s)\right) \mathbf{c}_j \left(\gamma_j \circ \gamma_k^{-1}(s)\right) ds. \tag{4.35}$$

Intuitively, what happens is that if the parameterisation functions γ_i put more weight into one part of the shapes, the composition of γ_i with the inverse parameterisation functions γ_k^{-1} neutralises this so that the curves are sampled more equally. This is illustrated in Fig. 4.6.

This also means that it is not possible to neglect any part of the curves, since it can be seen that there is always one term that gives arc-length parameterisation for $\mathbf{c}_i$ and one term that gives arc-length parameterisation for $\mathbf{c}_j$.

4.6.2.2 Proving Invariance

Theorem 4.2. *The scalar product (4.35) in Definition 4.1 is invariant under mutual reparameterisations.*

Proof. Let $\{\gamma_i\}_{i=1}^{n_s}$ be a set of parameterisation functions such that $\mathbf{c}_i \circ \gamma_i :=: \mathbf{c}_j \circ \gamma_j$, $(i = 1 \ldots n_s, j = 1 \ldots n_s)$. Let γ be an arbitrary reparameterisation function and let $\tilde{\gamma}_i(s) = \gamma_i \circ \gamma$ and $\tilde{\gamma}_j(s) = \gamma_j \circ \gamma$. Then $\mathbf{c}_i(\tilde{\gamma}_i(s)) :=: \mathbf{c}_j(\tilde{\gamma}_j(s))$ still holds. From the definition of the proposed scalar product we get,

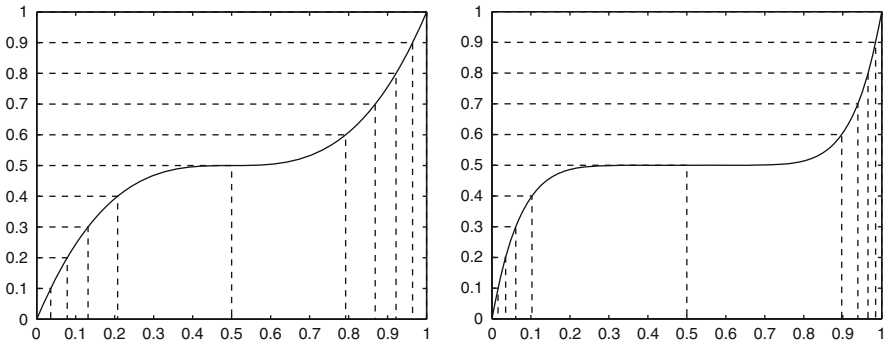


Fig. 4.6 In the two plots it can be seen how the sampling points, defined by the set $\mathscr{S}$, are mapped from the y-axis to the x-axis by two inverse parameterisation functions

$$\mathbf{c}_i(\tilde{\gamma}_i(s)) \cdot \mathbf{c}_j(\tilde{\gamma}_j(s)) =$$

$$\frac{1}{n_s} \sum_{k=1}^{n_s} \int_0^1 \mathbf{c}_i \left(\tilde{\gamma}_i \circ \tilde{\gamma}_k{}^{-1}(s)\right) \mathbf{c}_j \left(\tilde{\gamma}_j \circ \tilde{\gamma}_k{}^{-1}(s)\right) ds =$$

[since $\tilde{\gamma}_i(s) = \gamma_i \circ \gamma$ and $\tilde{\gamma}_j(s) = \gamma_j \circ \gamma$]

$$= \frac{1}{n_s} \sum_{k=1}^{n_s} \int_0^1 \mathbf{c}_i \left(\gamma_i \circ \gamma \circ \gamma^{-1} \circ \gamma_k^{-1}(s)\right) \cdot$$

$$\mathbf{c}_j \left(\gamma_j \circ \gamma \circ \gamma^{-1} \circ \gamma_k^{-1}(s)\right) ds =$$

[using that $\gamma \circ \gamma^{-1}$ is the identity function]

$$= \frac{1}{n_s} \sum_{k=1}^{n_s} \int_0^1 \mathbf{c}_i \left(\gamma_i \circ \gamma_k^{-1}(s)\right) \mathbf{c}_j \left(\gamma_j \circ \gamma_k^{-1}(s)\right) ds =$$

$$\mathbf{c}_i(\gamma_i(s)) \cdot \mathbf{c}_j(\gamma_j(s)).$$

□

4.6.2.3 Discussion

The scalar product would actually be invariant to mutual parameterisations using only one term in the sum in (4.35) in Definition 4.1.

Two terms may also seem like a natural choice and using two terms it is natural to choose the terms that parameterise $\mathbf{c}_i$ and $\mathbf{c}_j$ to arc-length respectively.

$$\mathbf{c}_i \cdot \mathbf{c}_j = \frac{1}{2} \int_0^1 \mathbf{c}_i \left(\gamma_i \circ \gamma_j^{-1}(s)\right) \mathbf{c}_j(s) + \mathbf{c}_i(s) \mathbf{c}_j \left(\gamma_j \circ \gamma_i^{-1}(s)\right) ds. \tag{4.36}$$

However, even though totally mutual reparameterisations are neutralised in (4.36), or even by using only one term, it is still possible to find parameterisations that decrease the DL by gathering landmarks together in different ways on different curves.

For example, this can be done if the curves are grouped so that landmarks are clustered in different ways in the different groups. The scalar product (4.36) cancels out the effect for the case where $\mathbf{c}_i$ and $\mathbf{c}_j$ belong to the same group. For $\mathbf{c}_i$ and $\mathbf{c}_j$ belonging to different groups the effect is not cancelled out.

Theoretically clustering at some level is still possible even when all the terms are used as in (4.35), but using all the terms makes it hard for the optimisation algorithm to locate such parameterisations and also, since the scalar product is a mean over all parameterisation functions, the effect would be small. Therefore it is better to use (4.35).

However, if time is an issue, using only one or two terms may be desirable.

To calculate the scalar product numerically it is of course necessary to sample the curves. These sample points are just a step in the evaluation and should not be confused with landmarks in the traditional sense. How this is done is explained in Sect. 4.2.3. From the set $\mathscr{S} = \{s_n : s_n = n/(n_p - 1), n = 0, \ldots, n_p - 1\}$, which is the set that define the sample points, the set $\mathscr{T}_i = \gamma_i^{-1}(\mathscr{S})$ is retrieved. This is shown schematically in Fig. 4.2. $\mathscr{T}_i$ and $\mathscr{T}_j$ decide what points that are used in the numerical calculation of the scalar product in (4.36). Since all the sets $\mathscr{T}_i (i = 1, \ldots, n_s)$ are used in (4.35), corresponding points across the entire training set are in this case used to calculate the different c_{ij}. This is another advantage of (4.35).

4.7 Experimental Validation

4.7.1 Experimental Validation of Using the Gradient

In this section, the proposed algorithm for correspondence optimisation using the gradient, described in Sect. 4.4, is validated on five data sets, illustrated in Fig. 4.7.

Thodberg's implementation of MDL [38] has been used for the comparison. MATLAB source code and test data are available from `hht@imm.dtu.dk`.

Seven parameterisation nodes have been used for the reparameterisations in all experiments. To evaluate the description length for the current parameterisation functions each curve is sampled with 64 landmarks. The parameter initialisations for the Nelder-Mead and the steepest descent algorithm were set to identical values.

The convergence speed of the proposed algorithm and the technique proposed in [38] using Nelder-Mead optimisation are compared on the five data sets, see Fig. 4.8 where the convergence rate (in seconds) of the description length using the two methods is plotted. It can be seen that the use of the steepest descent algorithm increases the convergence rate considerably. A result of applying steepest descent is that the cost function decreases in each iteration. The algorithm converges in roughly 50–100 iterations to a local minimum.

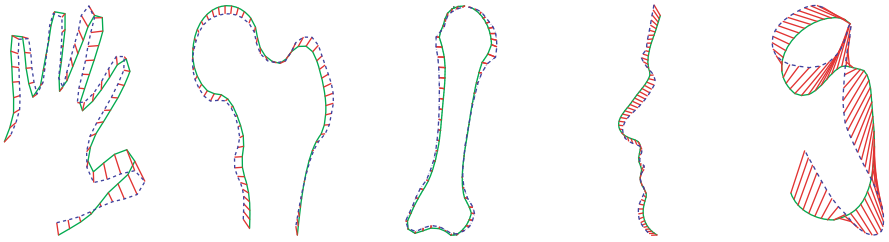


Fig. 4.7 The mean (*solid line*) and the first mode of variation (*dashed line*) of the optimised models (by the steepest descent algorithm) is plotted for the five datasets

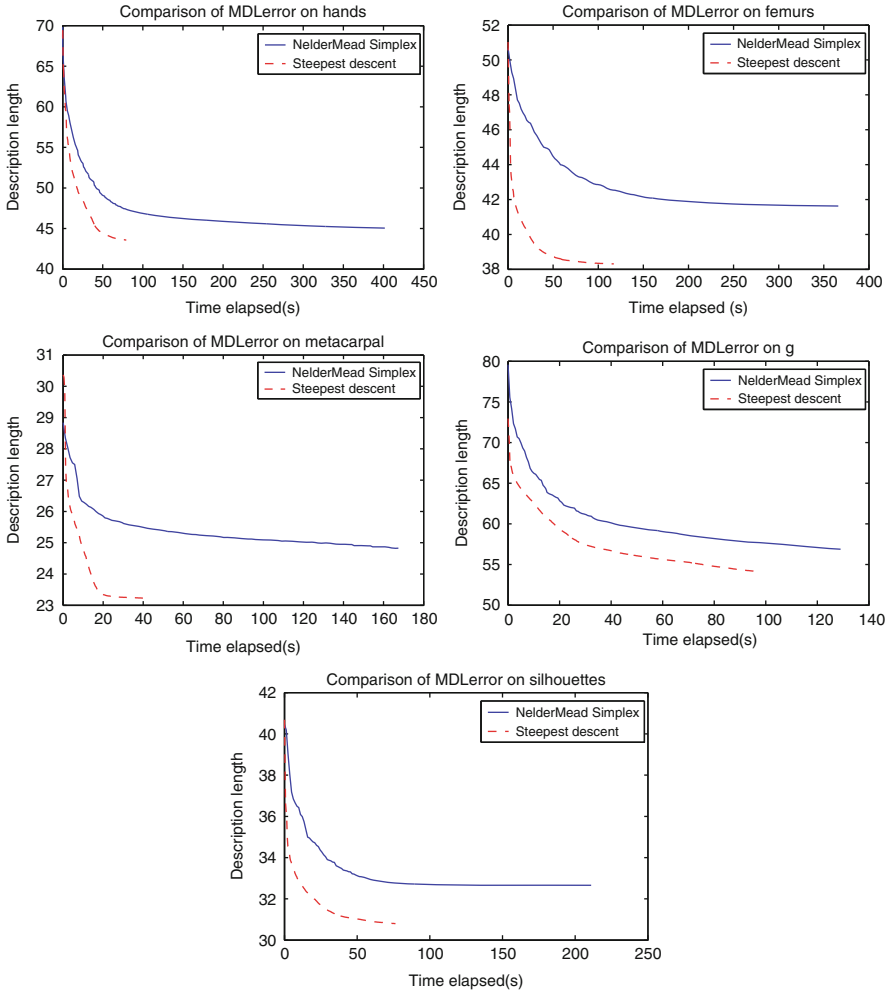


Fig. 4.8 The convergence rate (in seconds) of the description length for the five models using steepest descent (*dashed*) and Nelder–Mead Simplex (*solid*)

The appearance of several local minima may cause problem during optimisation. The algorithm locates a local minimum, which is close to the initialisation parameterisation, but there is no guarantee for this being a global minimum. Due to these facts it is necessary to compare the quality of the model achieved using the proposed steepest descent optimisation and the model achieved using Nelder–Mead Simplex optimisation. The quality of the models is measured as the mean square error in leave-one-out reconstructions. By leave-one-out means that a model is built with all but one sample. The model is then fitted to this unseen sample and the mean square

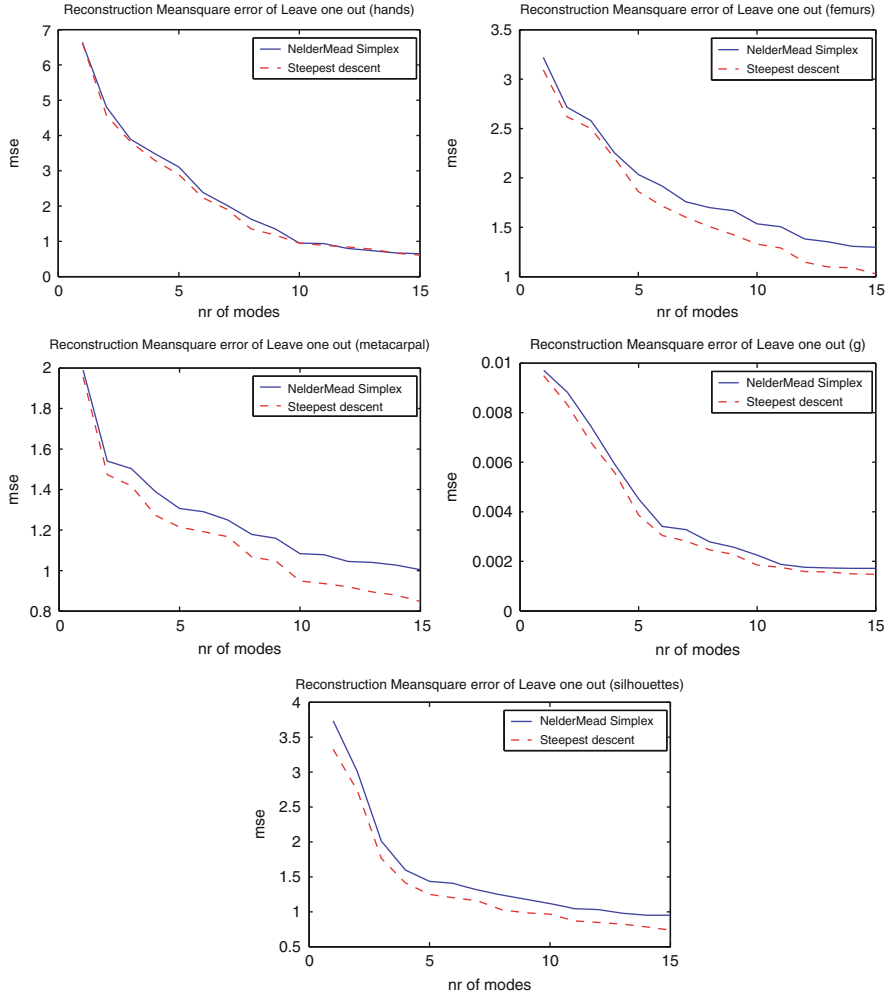


Fig. 4.9 The mean square approximation error of the five models after steepest descent (*dashed*) and Nelder–Mead (*solid*) optimisation is plotted against number of modes used. This measures the models ability to generalise

error is measured. The result is shown in Fig. 4.9. The plot shows the mean square approximation error against the number of modes used. This measures the ability of the model to represent unseen shape samples or, in other words, the models ability to *generalise*.

In Fig. 4.9 we can see that when using steepest descent, models that generalise better are achieved for all examples compared to when using Nelder–Mead Simplex.

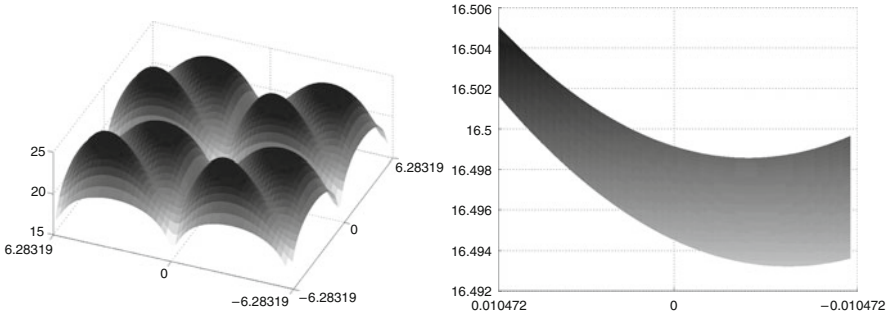


Fig. 4.10 The description length cost function. In the upper figure the range on each axis is -2π to 2π. The lower figure zooms in on the origin showing that the minimum of the 3D surface is not at the origin

4.7.2 Experimental Validation of Alignment Using MDL

The MDL based algorithm is compared to Procrustes on seven data sets.

The experiments were conducted in the following way. Given a dataset, the centre of gravity was moved to the origin for all shapes and scale was normalised to one according to the Euclidian norm. The rotation of the shapes was initialised according to the Procrustes condition. The rotation was then optimised using Gauss-Newton, minimising the Description Length.

In Fig. 4.10 the typical behavior of the cost function can be seen. Here, for visualisation, two rotations (x and y axis) have been optimised to align three shapes. In the top figure it can be seen that the minimum is a well defined global minimum (it seems to be several minima, but this is due to that the range goes from -2π to 2π in x and y). The bottom figure zooms in on the origin (the origin corresponds to the Procrustes solution). It can be seen that the minimum is not at the origin and thus the Procrustes solution does not coincide with the description length solution. When more shapes are aligned, projections of the cost function to lower dimensions look similar to these plots.

We validate the algorithm on seven real data sets, see Fig. 4.14. The contours were sampled with 64–128 landmarks using arc-length parameterisation.

The quality of the models was measured as the mean square error in leave-one-out reconstructions. This is defined as the models generalisation ability. The model is built with all but one example and then fitted to the unseen example. The results are shown in Fig. 4.11. The plots show the mean squared approximation error against the number of modes used. This measures the ability of the model to represent unseen shape instances of the object class.

For all examples we get models that give the same or lower error when using the description length criterion compared to Procrustes alignment. This means that the models generalise better. The improvements are small but consistent. The computational cost increases of course using this alignment compared to Procrustes,

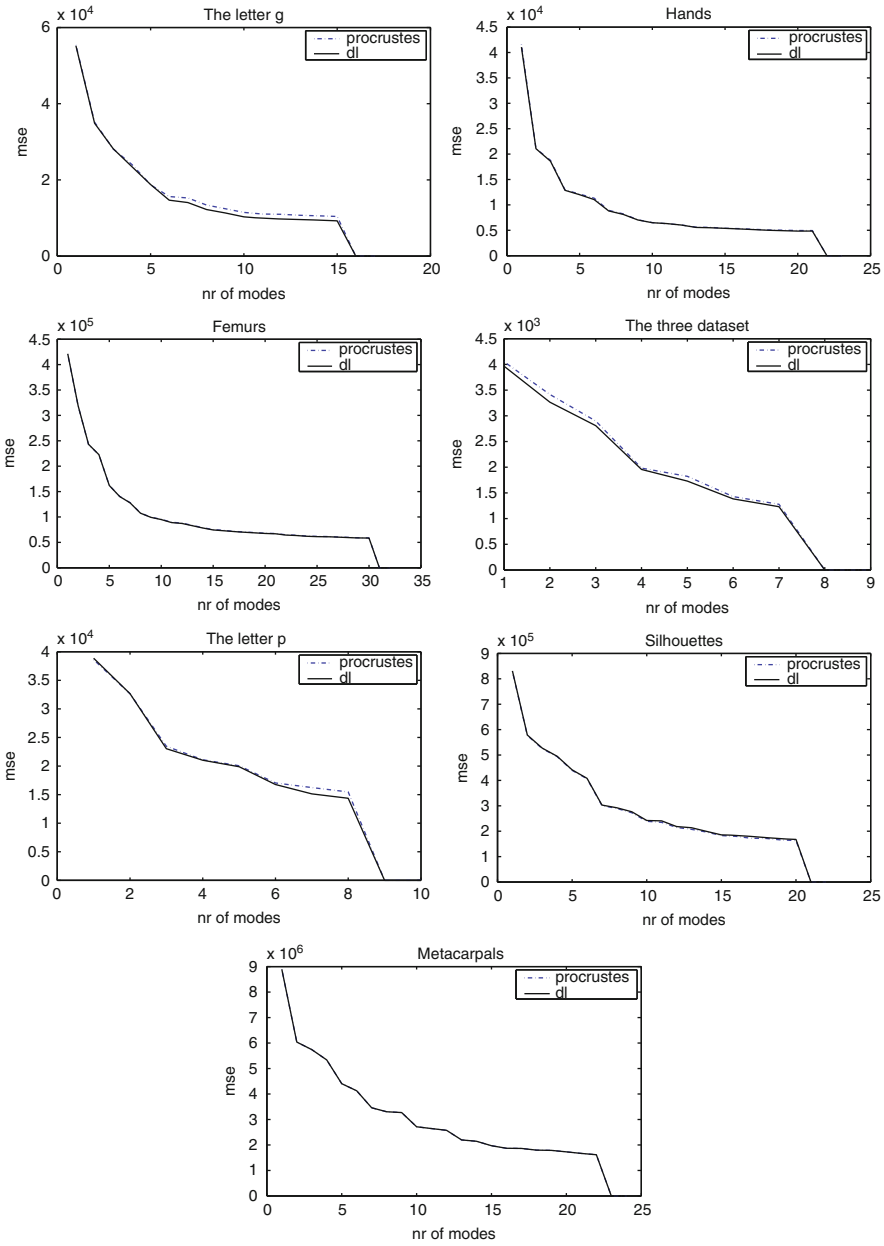


Fig. 4.11 The mean squared error of leave one out reconstructions of the g-dataset, the hand dataset, the femur dataset, the three dataset, the 90 p shapes, the 22 silhouettes and the 24 metacarpals

but it is often worth while spending time on building a model of the highest possible quality. Once the model has been built it can be used over and over again.

One reason why the description length alignment does not get even better results can be that when there are many shapes the path to the minimum is difficult for the optimiser to follow. The derivatives also get numerically unstable close to the minimum. Another conclusion could be that Procrustes alignment, in its simplicity, is a very good method for aligning shapes.

Sometimes the difference can be big. In Fig. 4.12 is an example of when the Procrustes alignment goes visibly wrong. It is a synthetic example with 24 shapes, each built up by 128 landmarks. The shapes are rectangles with a bump at different locations on the top side. For a human it would be natural to align the boxes and let the bump be misaligned. These shapes are built up with a majority of landmarks around the bumps and therefore both algorithms will give large emphasis on the alignment of the bump. Note that this data only has one shape mode (the position of the bump) and this mode is linear. Therefore perfect alignment should give zero mean squared error on the leave one out reconstruction using just one mode. In this example the model built from description length aligned box bumps gets almost zero error on the first shape mode, but the model built from Procrustes aligned shapes needs two modes to get a small error, see Fig. 4.13. The description length aligned boxes also correspond better to how humans would align the shapes, as can be seen in Fig. 4.12.

4.7.3 Experimental Validation of Parameterisation Invariant Scalar Product

In this section the algorithm using the parameterisation invariant scalar product from Sect. 4.6 is validated on the hand-, femur- and g- data sets.

Models built using the proposed scalar product are compared to models built using the standard scalar product. For the standard scalar product the algorithm uses parameterisation node cost [38] to prevent clustering of sample points.

Thodberg's implementation of MDL [38] has again been used for the comparison.

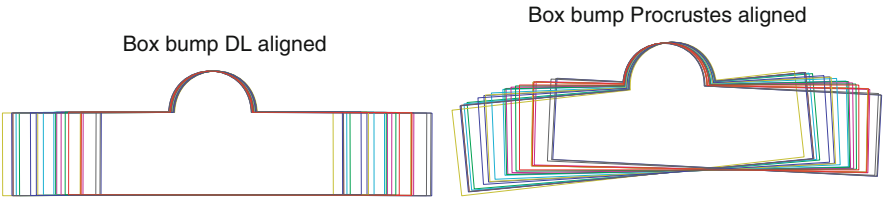


Fig. 4.12 A synthetic example shows that Procrustes alignment can fail (*right*). Note that the description length approach succeeds (*left*)

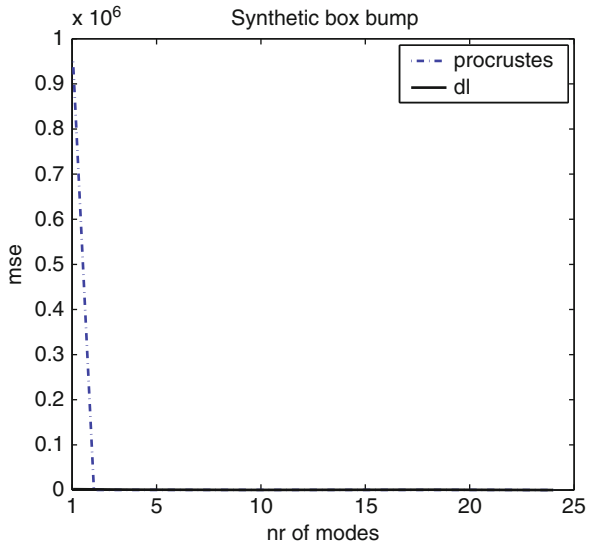


Fig. 4.13 The generalisation ability of the synthetic model. It shows that the description length aligned succeeds in locating the only shape mode and the Procrustes alignment needs two

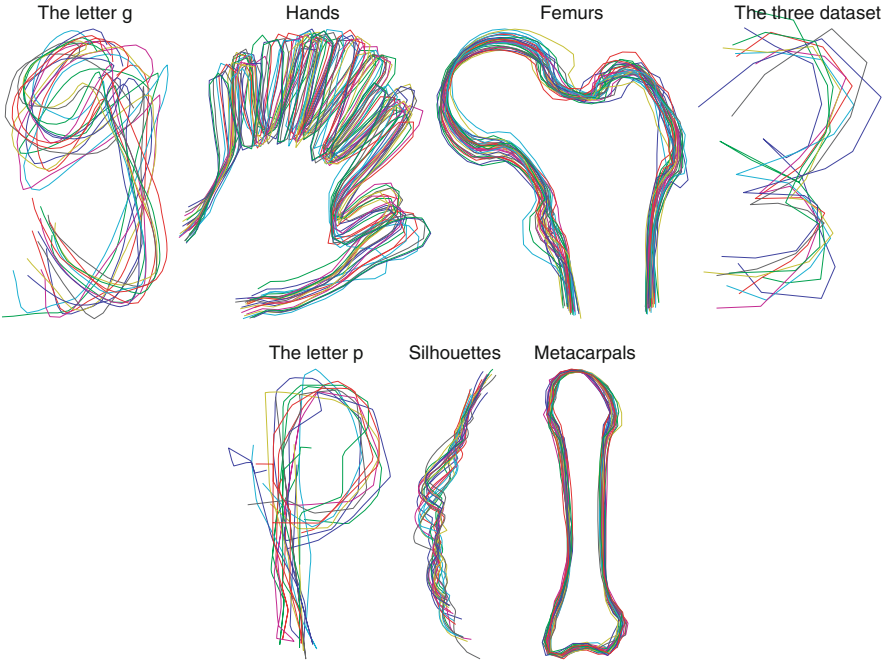


Fig. 4.14 The description length aligned datasets

Seven control nodes have been used for the reparameterisations in all experiments. To evaluate the description length for a given set of parameterisations the curves must be sampled. For the proposed scalar product the sampling of each parameterisation function $\gamma_i (i = 1, \ldots, n_s)$ results in n_p sample points. This means that for the proposed scalar product the number of sample points are $n_p n_s$. Since the number of landmarks n_p were set to 65, this results in between 1,105 to 2,080 sample points for the three different datasets. To be able to compare the algorithms the number of landmarks were set to 2,048 for the standard scalar product. Other initialisations and parameters for the optimisation algorithms were identical in all cases.

The generality of the models is measured as the mean square error in leave-one-out reconstructions. The error between the original curve and the model approximation curve is calculated by integrating the squared distance between the curves by arc-length. The result is shown in Fig. 4.15. The plot shows the mean square approximation error against the number of modes used. This measures the ability of the model to represent unseen shape samples or, in other words, the models ability to *generalise*. It can be seen that the new scalar product gives models that generalise better than models built using the standard scalar product, even if node cost penalties are used.

The models built using the different scalar products are also evaluated using specificity. By specificity of the models is meant that shapes generated by reasonable

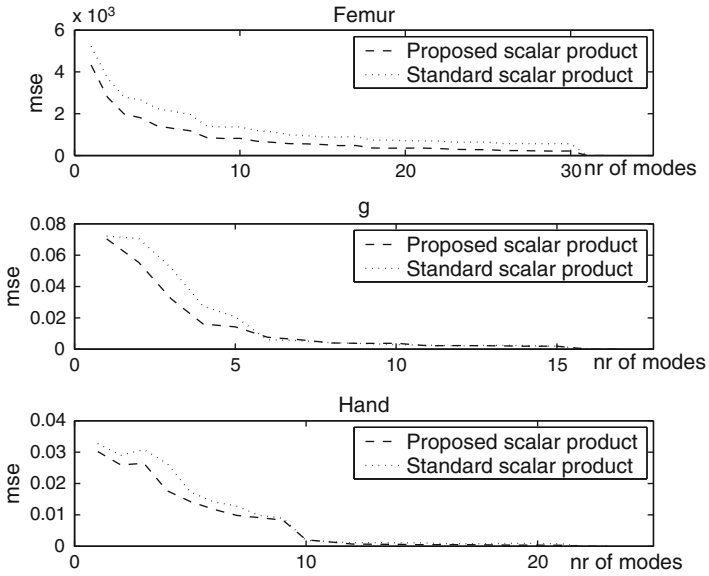


Fig. 4.15 The mean square approximation error of the six models is plotted against the number of modes used. The *top* figure shows the models of femurs, the *middle* figure shows the models of g:s and the *bottom* figure shows the models of hands

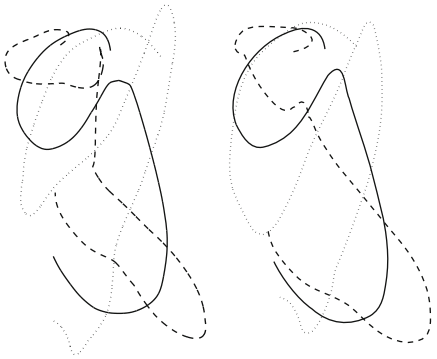


Fig. 4.16 The mean (*solid line*) and the first mode of variation (*dashed* and *dotted lines*) of the optimised models. The model on the left is optimised using the standard scalar product and the model on the right is optimised using the proposed scalar product

parameter values (not more than ± 3 std.) should be representative for the training set. In [17,25] it is shown that the quantitative measure of specificity is problematic. A better way to evaluate the specificity is by plotting shapes modelled at equal parameter values by the different models that are to be compared. By increasing the parameter value up to 3 std. it can often be seen which model is more specific. In Fig. 4.16, the mean shape and the mean shape plus and minus three standard deviations of the first shape mode is plotted for each model. The specificity evaluation is presented for the most difficult dataset, the g-shapes. In the figure, it can be seen that the shapes generated by the proposed algorithm are more representative. Look at the left plot (standard scalar product used) and note, for example, the sharp bend in middle of the dashed curve and the not sharp enough corner at the end of the first arc of the mean shape (the solid curve). That this corner is not sharp is a sign that the correspondences at this point are matching badly. Also when comparing the two dotted g:s it can be seen that the one to the right (proposed scalar product used) looks more natural. For the other datasets the differences in specificity of the two models are somewhat harder to distinguish.

4.8 Summary and Conclusions

In this paper a way to minimise the description length that is more efficient than the previously used Nelder–Mead Simplex technique has been presented. The gradient of the description length with respect to parameterisation has been derived and a steepest descent method is proposed to minimise the MDL-criteria. The proposed algorithm has been compared to the state of the art algorithm proposed in [39]. For all test cases the models using the proposed method have better generalisation ability and the convergence rate has increased considerably.

In this paper we present a new way to align shapes. The rotation is located by minimising the description length. We derive the gradient of the description length with respect to the rotation and propose to use Gauss–Newton to minimise the

MDL-criterion. We have compared the proposed algorithm to Procrustes alignment and shown that better models can be achieved using the proposed method.

In this paper landmark clustering is avoided by introducing a new scalar product that is invariant to mutual reparameterisations. The optimisation of the shape model becomes invariant to mutual reparameterisations and is therefore better suited for finding correspondences. It is shown that the global minimum of the optimisation problem using the standard scalar product is located by reparameterising all the shapes to a single point. Using the proposed scalar product this problem is avoided and a more robust and stable algorithm is achieved. The algorithm is compared to a state of the art algorithm, which uses ad hoc solutions to prevent clustering of landmarks. The comparison shows that the achieved models using the proposed scalar product are more general for all the three tested datasets. The model of the most difficult g-dataset turns out to be more specific than the model built using the standard scalar product. It is concluded that the proposed formulation not only treats the optimisation problem in a more mathematically rigorous way, but it also results in better models than the former ad-hoc reparameterisation node cost.

Acknowledgements We thank H. Thodberg (IMM, DTU) for the silhouettes and B. Kimia et.al. for the images of the sharks, birds, flight birds, rats and forks.

This work has been financed by the SSF sponsored project 'Vision in Cognitive Systems' (VISCOS) and by UMAS and the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Centre for Medical Innovations (CMI) at the Karolinska Institute.

References

1. Andrew, A., Chu, E., Lancaster, P.: Derivatives of eigenvalues and eigenvectors of matrix functions. SIAM93 J. Matrix Anal. Appl. **14**, 903–926 (1993)
2. Baumberg, A., Hogg, D.: Learning flexible models from image sequences. In: Proceedings of the European Conference on Computer Vision, ECCV'94, pp. 299–308. (1994)
3. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(24), 509–522 (2002)
4. Benayoun, A., Ayache, N., Cohen, I.: Adaptive meshes and nonrigid motion computation. In: Proceedings of the International Conference on Pattern Recognition, ICPR'94, pp. 730–732. (1994)
5. Bookstein, F.L.: Size and shape spaces for landmark data in two dimensions. Statist. Sci. **1**(2), 181–242 (1986)
6. Bookstein, F.L.: Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. Med. Image Anal. **3**, 225–243 (1999)
7. Chui, H., Rangarajan, A.: A feature registration framework using mixture models. In: IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), pp. 190–197. (2000)
8. Chui, H., Rangarajan, A.: A new algorithm for non-rigid point matching. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. II, pp. 44–51. (2000)
9. Davies, R.H.: Learning shape: Optimal models for analysing natural variability. Ph.D. thesis, Division of Imaging Science and Biological Engineering, University of Manchester, Manchester, England (2002)

10. Davies, R.H., Cootes, T.F., Taylor, C.J.: A minimum description length approach to statistical shape modeling. In: Information Processing in Medical Imaging, pp. 50–63. (2001)
11. Davies, R.H., Cootes, T.F., Waterton, J.C., Taylor, C.J.: An efficient method for constructing optimal statistical shape models. In: Medical Image Computing and Computer-Assisted Intervention MICCAI'2001, pp. 57–65. (2001)
12. Davies, R.H., Twining, C.J., Cootes, T.F., Allen, P.D., Taylor, C.J.: Shape discrimination in the hippocampus using an mdl model. In: Information Processing in Medical Imaging, pp. 684–695. (2003)
13. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.: A minimum description length approach to statistical shape modeling. IEEE Trans. Med. Imaging **21**(5), 525–537 (2002)
14. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, New York (1999)
15. Ericsson, A.: Automatic shape modelling and applications in medical imaging. Licentiate thesis, Department of Mathematics, Lund Institute of Technology, Centre for Mathematical Sciences, Lund, Sweden (2003)
16. Ericsson, A., Karlsson, J.: Aligning shapes by minimising the description length. In: Proceedings of the Scandinavian Conference on Image Analysis, SCIA'05, vol. 3540/2005, pp. 709–718. Joensuu, Finland (2005)
17. Ericsson, A., Karlsson, J.: Benchmarking of algorithms for automatic correspondence localisation. In: Proceedings of the British Machine Vision Conference, BMVC'06. Edinburgh, Great Britain, vol. 2, pp. 759–768 (2006)
18. Ericsson, A., Åström, K.: An affine invariant deformable shape representation for general curves. In: Proceedings of the International Conference on Computer Vision, ICCV'03, Nice, France (2003)
19. Ericsson, A., Åström, K.: Minimizing the description length using steepest descent. In: Proceedings of the British Machine Vision Conference, Norwich, United Kingdom (2003)
20. Forsyth, D.A., Ponce, J.: Computer Vision. Prentice Hall, New Jersey (2003)
21. Gower, J.C.: Generalized procrustes analysis. Psychometrica **40**, 33–50 (1975)
22. Hill, A., Taylor, C.J.: Automatic landmark generation for point distribution models. In: Proceedings of the British Machine Vision Conference, pp. 429–438 (1994)
23. Hill, A., Taylor, C.J.: A framework for automatic landmark indentification using a new method of nonrigid correspondence. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 241–251 (2000)
24. Kambhamettu, C., Goldgof, D.B.: Points correspondences recovery in non-rigid motion. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR'92, pp. 222–237 (1992)
25. Karlsson, J., Ericsson, A.: A geodesic ground truth correspondence measure for benchmarking. In: Proceedings of the International Conference on Pattern Recognition, ICPR'06. Hong Kong, China, vol. 3 (2006)
26. Karlsson, J., Ericsson, A., Åström, K.: Parameterisation invariant statistical shape models. In: Proceedings of the International Conference on Pattern Recognition, Cambridge, UK (2004)
27. Kelemen, A., Szekely, G., Gerig, G.: Elastic model-based segmentation of 3d neuroradiological data sets. IEEE Trans. Med. Imaging **18**(10), 828–839 (1999)
28. Kendall, D.G., Barden, D., Carne, T.K., Le, H.: Shape and Shape Theory. Wiley, Chichester (1999)
29. Kotcheff, A.C.W., Taylor, C.J.: Automatic construction of eigenshape models by direct optimization. Med. Image Anal. **2**, 303–314 (1998)
30. Ljung, L.: System Identification. Prentice Hall, New Jersey (1998)
31. Papadopoulo, T., Lourakis, M.: Estimating the jacobian of the singular value decomposition. In: Proceedings of the European Conference on Computer Vision, ECCV'00, pp. 555–559. (2000)
32. Rissanen, J.: Modeling by shortest data description. Automatica **14**, 465–471 (1978)
33. Rueckert, D., Frangi, F., Schnabel, J.A.: Automatic construction of 3d-statistical deformation models using nonrigid registration. In: Medical Image Computing and Computer-Assisted Intervention MICCAI'2001, pp. 77–84. (2001)

34. Sebastian, T., Klein, P., Kimia, B.: Constructing 2d curve atlases. In: IEEE Workshop on Mathematical Methods in Biomedical Image Analysis, pp. 70–77. (2000)
35. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Techn. J. **27**, 379–423; 623–656 (1948)
36. Small, C.G.: The Statistical Theory of Shape. Springer, New York, (1996)
37. Tagare, H.D.: Shape-based nonrigid correspondence with application to heart motion analysis. IEEE Trans. Med. Imaging **18**, 570–579 (1999)
38. Thodberg, H.H.: Minimum description length shape and appearance models. In: Image Processing Medical Imaging, IPMI'03, pp. 51–62. (2003)
39. Thodberg, H.H.: Minimum description length shape and appearance models. Technical report: IMM Technical Report 2003-01, Technical University of Denmark, Lyngby (2003)
40. Wang, Y., Peterson, B.S., Staib, L.H.: Shape-based 3d surface correspondence using geodesics and local geometry. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, CVPR'00, pp. 644–651. (2000)
41. Zheng, Y., Doermann, D.: Robust point matching for non-rigid shapes: A relaxation labeling based approach. Technical report: Lamp-tr-117, University of Maryland, College Park (2004)

Chapter 5
On the Bijectivity of Thin-Plate Splines

Anders P. Erikson and Kalle Åström

Abstract The thin-plate spline (TPS) has been widely used in a number of areas such as image warping, shape analysis and scattered data interpolation. Introduced by Bookstein (IEEE Trans. Pattern Anal. Mach. Intell. 11(6):567–585 1989), it is a natural interpolating function in two dimensions, parameterized by a finite number of landmarks. However, even though the thin-plate spline has a very intuitive interpretation as well as an elegant mathematical formulation, it has no inherent restriction to prevent folding, i.e. a non-bijective interpolating function. In this chapter we discuss some of the properties of the set of parameterizations that form bijective thin-plate splines, such as convexity and boundness. Methods for finding sufficient as well as necessary conditions for bijectivity are also presented. The methods are used in two settings (a) to register two images using thin-plate spline deformations, while ensuring bijectivity and (b) group-wise registration of a set of images, while enforcing bijectivity constraints.

5.1 Introduction

Thin-plate splines are a class of widely used non-rigid spline mapping functions. It is a natural choice of interpolating function in two dimensions and has been a commonly used tool for over a decade. Introduced and developed by Duchon [6] and Meinguet [11] and popularized by Bookstein [1], its attractions include an elegant mathematical formulation along with a very natural and intuitive physical interpretation.

The thin-plate spline framework can also be employed in a deformation setting, that is mappings from $\mathbb{R}^m$ to $\mathbb{R}^m$. This is accomplished by the combination of

A.P. Erikson (✉) · K. Åström
Centre for Mathematical Sciences, Lund University, Lund, Sweden
e-mail: anderspe@maths.lth.se; kalle@maths.lth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_5,

several thin-plate spline interpolants. Here we restrict ourselves to $m = 2$. If instead of understanding the displacement of the thin metal plate as occurring orthogonally to the (x_1, x_2)-plane view them as displacements of the x_1- or x_2-position of the point constraints. With this interpretation, a new function $\phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ can be constructed from two thin-plate splines, each describing the x_1- and x_2-displacements respectively.

In spite of its appealing algebraic formulation, thin-plate spline mappings do have drawbacks and, disregarding computational and numerical issues, one in particular. Namely, bijectivity is never assured. In computer vision, non-linear mappings in $\mathbb{R}^2$ of this sort are frequently used to model deformations in images. The basic assumption is that all the images contain similar structures and therefore there should exist mappings between pairs of images that are both one-to-one and onto. Hence bijective mappings are of interest.

This work is an attempt at characterizing the set of bijective thin-plate spline mappings. It contains a formulation of how to describe this set, as well as proofs of many of its properties. It also includes a discussion of some experimentally derived indications of other attributes of this set, such as boundedness and convexity, as well as methods for finding sufficient conditions for bijectivity.

The paper is organized as follows. In Sect. 5.2, the definition of the thin-plate spline is given and properties of thin-plate spline interpolation is given. In Sect. 5.3, the bijectivity constrain on thin-plate splines are studied and the properties of the set of deformed control points that produce bijective thin-plate spline deformations are presented in Sect. 5.4. In Sect. 5.5, it is shown how these constraint can be used for pairwise image registration using thin-plate splines. In Sect. 5.6 is presented a method for automatic generation of active appearance models. In this method a whole stack of images is simultaneously co-registered to a common active appearance model using bijectivity constrained thin-plate spline deformations.

5.2 Thin-Plate Splines

Consider a thin metal plate extending to infinity in all directions. At a finite number of discrete positions $\mathbf{t}_i \in \mathbb{R}^2$, $i = 1 \ldots n$, the plate is at fixed heights z_i, see Fig. 5.1. The metal plate will take the form that minimizes its *bending energy*. In two dimensions the bending energy of a plate described by a function $g(x, y)$ is proportional to

$$J(g) = \int\int_{\mathbb{R}^2} \left(\left(\frac{\partial^2 g}{\partial x^2} \right)^2 + 2 \left(\frac{\partial^2 g}{\partial x \partial y} \right)^2 + \left(\frac{\partial^2 g}{\partial y^2} \right)^2 \right) dx dy. \tag{5.1}$$

Consequently, the metal plate will be described by the function that minimizes (5.1) under the point constraints $g(\mathbf{t}_i) = z_i$. It was proven by Duchon [6] that if such a function exists it is unique.

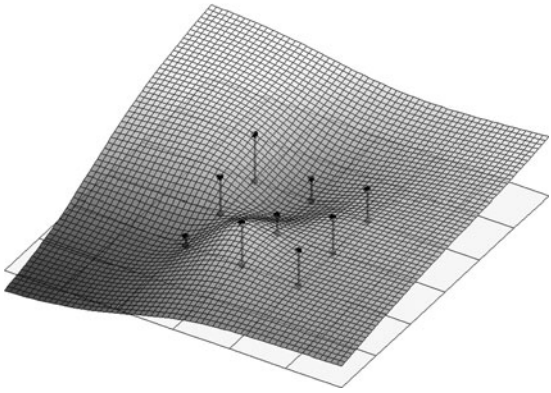


Fig. 5.1 The shape of a thin metal plate constrained to lie at some distances above a ground plane at nine different locations

Given n point constraints $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2 \dots \mathbf{t}_n)$, along with the corresponding displacements $\mathbf{z} = (z_1, z_2, \dots, z_n)$, $z_i \in \mathbb{R}$. Define

$$\sigma(h) = \begin{cases} ||h||^2 \log(||h||), & ||h|| > 0, \\ 0, & ||h|| > 0, \end{cases} \tag{5.2}$$

where $||\cdot||$ is the Euclidian vector norm.

Definition 5.1. A thin-plate spline function $g_z : \mathbb{R}^2 \Rightarrow \mathbb{R}$ is a minimizer of (5.1) iff it can be written on the following form

$$\begin{aligned} g_{\mathbf{T},\mathbf{z}}(\mathbf{x}) &= \sum_{i=1}^{n} \delta_i \sigma(\mathbf{x} - \mathbf{t}_i) + a_1 + a_2 x_1 + a_3 x_2 = \\ &= \underbrace{\begin{bmatrix} \delta_1 & \delta_2 & \dots & \delta_n \end{bmatrix}}_{\delta^T} \underbrace{\begin{bmatrix} \sigma(\mathbf{x} - \mathbf{t}_1) \\ \sigma(\mathbf{x} - \mathbf{t}_2) \\ \vdots \\ \sigma(\mathbf{x} - \mathbf{t}_n) \end{bmatrix}}_{\mathbf{s}(\mathbf{x})} + \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} = \\ &= \begin{bmatrix} \delta^T & a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix}, \end{aligned} \tag{5.3}$$

where $g_{\mathbf{T},\mathbf{z}}(\mathbf{x})$ δ_i, a_i satisfy

$$g_{\mathbf{T},\mathbf{z}}(\mathbf{t}_i) = z_i, \tag{5.4}$$

$$\sum_{i=1}^{n} \delta_i = \sum_{i=1}^{n} \delta_i t_{ix} = \sum_{i=1}^{n} \delta_i t_{iy} = 0. \tag{5.5}$$

Combining (5.3)–(5.5) the thin-plate spline can be found by solving the equations

$$\begin{bmatrix} \mathbf{s}(\mathbf{t}_1) & 1 & t_{11} & t_{12} \end{bmatrix} \begin{bmatrix} \delta^T \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = z_1,$$

$$\vdots$$

$$\begin{bmatrix} \mathbf{s}(\mathbf{t}_n) & 1 & t_{n1} & t_{n2} \end{bmatrix} \begin{bmatrix} \delta^T \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = z_n, \tag{5.6}$$

$$\begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \delta \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = 0, \tag{5.7}$$

$$\begin{bmatrix} \mathbf{T}^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \delta \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = 0. \tag{5.8}$$

With the symmetric $n \times n$ matrix S defined by $S_{ij} = \sigma(\mathbf{t}_i - \mathbf{t}_j)$ we can write (5.6)–(5.8)

$$\underbrace{\begin{bmatrix} S & \mathbf{1}_n & \mathbf{T} \\ \mathbf{1}_n^T & 0 & 0 \\ \mathbf{T}^T & 0 & 0 \end{bmatrix}}_{\Gamma} \begin{bmatrix} \delta^T \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} \mathbf{z} \\ 0 \\ 0 \end{bmatrix}. \tag{5.9}$$

If $\mathbf{t}_1, \dots, \mathbf{t}_n$ are not collinear the symmetric matrix Γ is of full rank (see [7]) and (5.9) has the unique solution

$$\begin{bmatrix} \delta^T \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \Gamma^{-1} \begin{bmatrix} \mathbf{z} \\ 0 \\ 0 \end{bmatrix}. \tag{5.10}$$

Consequently, with the following partition of Γ^{-1}

$$\Gamma^{-1} = \begin{bmatrix} \Gamma^{11} & \Gamma^{12} \\ \Gamma^{21} & \Gamma^{22} \end{bmatrix},$$
$$\Gamma^{11},\ n \times n$$
$$\Gamma^{12} = (\Gamma^{21})^T,\ n \times 3$$
$$\Gamma^{22},\ 3 \times 3$$

the thin-plate spline can be defined.

Definition 5.2. A thin-plate spline under point constraints $\mathbf{T}$ and $\mathbf{z}$ can be written

$$\begin{aligned} g_{\mathbf{T},\mathbf{z}}(\mathbf{x}) &= \begin{bmatrix} \delta^T & a_1 & a_2 & a_3 \end{bmatrix} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} = \\ &= \left(\Gamma^{-1} \begin{bmatrix} \mathbf{z} \\ 0 \\ 0 \end{bmatrix} \right)^T \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{z}^T & 0 & 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} = \\ &= \begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & \mathbf{x} \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \mathbf{z}. \end{aligned} \tag{5.11}$$

Thin-plate splines of this form has a number of desirable properties. They are both continuous and smooth interpolants. Equivariance under similarity transformations also holds.

Lemma 5.1. *The thin-plate spline are equivariant under similarity transformations* $\Pi : \mathbb{R}^2 \to \mathbb{R}^2$ *of* $\mathbf{T}$.

$$g_{\Pi(\mathbf{T}),\mathbf{z}}(\mathbf{x}) = g_{\mathbf{T},\mathbf{z}}(\Pi^{-1}(\mathbf{x})), \tag{5.12}$$

where

$$\Pi(\mathbf{T}) = \alpha(\mathbf{T} + [\mathbf{1}_n \psi_1 \mathbf{1}_n \psi_2])R,$$
$$R \in O(2),$$

$$\alpha \in \mathbb{R},$$
$$\psi \in \mathbb{R}^2.$$

Lemma 5.1 fits nicely in with the metal plate analogy. Rotation, scaling and translation of the location of the point constraints should not affect the bending of the plate but solely result in a corresponding alteration of the plate. From our intuitive understanding of this approach it is expected that the interpolation by such a transformed spline should be equal to a transformation of the original interpolation spline, which is exactly what this lemma confirms.

Finally, for the matrix Γ^{12}, the following also holds

Lemma 5.2. *If Γ is the matrix associated with a thin-plate spline mapping with point-constraints* $\mathbf{T}$ *then with*

$$\Gamma^{12} = \begin{bmatrix} \Gamma_0^{12} & \Gamma_1^{12} & \Gamma_2^{12} \end{bmatrix}.$$

It holds that

$$\begin{aligned} &(\Gamma_0^{12})^T \mathbf{1}_n = 1,\ (\Gamma_1^{12})^T \mathbf{1}_n = 0,\ (\Gamma_2^{12})^T \mathbf{1}_n = 0, \\ &(\Gamma_1^{12})^T \mathbf{T}_1 = 1,\ (\Gamma_0^{12})^T \mathbf{T}_1 = 0,\ (\Gamma_2^{12})^T \mathbf{T}_1 = 0, \\ &(\Gamma_2^{12})^T \mathbf{T}_2 = 1,\ (\Gamma_0^{12})^T \mathbf{T}_1 = 0,\ (\Gamma_1^{12})^T \mathbf{T}_1 = 0. \end{aligned} \tag{5.13}$$

Proof.

$$I = \Gamma^{-1}\Gamma = \begin{bmatrix} \Gamma^{11} & \Gamma^{12} \\ \Gamma^{21} & \Gamma^{22} \end{bmatrix} \begin{bmatrix} S & \mathbf{1}_n & \mathbf{T} \\ \mathbf{1}_n^T & 0 & 0 \\ \mathbf{T}^T & 0 & 0 \end{bmatrix} =$$

$$= \begin{bmatrix} \Gamma^{11} S + \Gamma^{12} \begin{bmatrix} \mathbf{1}_n^T \\ \mathbf{T}^T \end{bmatrix} & \Gamma^{11} \begin{bmatrix} \mathbf{1}_n & \mathbf{T} \end{bmatrix} \\ \\ \Gamma^{21} S + \Gamma^{22} \begin{bmatrix} \mathbf{1}_n^T \\ \mathbf{T}^T \end{bmatrix} & \Gamma^{21} \begin{bmatrix} \mathbf{1}_n & \mathbf{T} \end{bmatrix} \end{bmatrix}$$

$\Rightarrow$

$$\Gamma^{21} \begin{bmatrix} \mathbf{1}_n & \mathbf{T} \end{bmatrix} = \begin{bmatrix} \Gamma_0^{12^T} \\ \Gamma_1^{12^T} \\ \Gamma_2^{12^T} \end{bmatrix} \begin{bmatrix} \mathbf{1}_n & \mathbf{T}_1 & \mathbf{T}_2 \end{bmatrix} =$$

$$= \begin{bmatrix} \Gamma_0^{12^T} \mathbf{1}_n & \Gamma_0^{12^T} \mathbf{T}_1 & \Gamma_0^{12^T} \mathbf{T}_2 \\ \Gamma_1^{12^T} \mathbf{1}_n & \Gamma_1^{12^T} \mathbf{T}_1 & \Gamma_1^{12^T} \mathbf{T}_2 \\ \Gamma_2^{12^T} \mathbf{1}_n & \Gamma_2^{12^T} \mathbf{T}_1 & \Gamma_2^{12^T} \mathbf{T}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

□

The thin-plate spline formulation can easily be generalized into higher dimension interpolants. With a different bending energy function, and its associated fundamental solution (5.2), the above lemmas can be extended under this generalisation. For more details see [15].

5.2.1 Pair of Thin-Plate Spline Mappings

The thin-plate spline framework can also be employed in a deformation setting, that is mappings from $\mathbb{R}^m$ to $\mathbb{R}^m$. This is accomplished by the combination of several thin-plate spline interpolants. In this section we do however restrict ourselves to $m = 2$.

If instead of understanding the displacement of the thin metal plate as occurring orthogonally to the (x_1, x_2)-plane view them as displacements of the x_1- or x_2-position of the point constraints. With this interpretation, a new function $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ can be constructed from two thin-plate splines, each describing the x_1- and x_2-displacements respectively.

Definition 5.3. Given a set of target points $\mathbf{T} = [\mathbf{T}_1\ \mathbf{T}_2] = \begin{bmatrix} \mathbf{t}_1 \\ \vdots \\ \mathbf{t}_n \end{bmatrix}$, $\mathbf{t}_i \in \mathbb{R}^2$ and a set of destination points $\mathbf{Y} = [\mathbf{Y}_1\ \mathbf{Y}_2] = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix}$, $\mathbf{y}_i \in \mathbb{R}^2$ A pair of thin-plate splines mapping $\phi_{\mathbf{T},\mathbf{Y}} : \mathbb{R}^2 \to \mathbb{R}^2$ is the bivariate function $\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}) = (g_1(\mathbf{x}), g_2(\mathbf{x}))$, where $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ are two thin-plate spline interpolants ensuring the point constraints $\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{T}) = \mathbf{Y}$.

The two thin-plate splines satisfying these constraints are

$$g_1(\mathbf{x}) = g_{\mathbf{T},\mathbf{Y}_1}(\mathbf{x}) = \begin{bmatrix} \mathbf{Y}_1^T\ 0\ 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} \tag{5.14}$$

and

$$g_2(\mathbf{x}) = g_{\mathbf{T},\mathbf{Y}_2}(\mathbf{x}) = \begin{bmatrix} \mathbf{Y}_2^T\ 0\ 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix}. \tag{5.15}$$

Since we know that $g_1(\mathbf{T}) = \mathbf{Y}_1$ and $\phi_2(\mathbf{T}) = \mathbf{Y}_2$ it follows that $\phi_{\mathbf{T},\mathbf{Y}}(T) = (g_1(\mathbf{T}), g_2(\mathbf{T})) = (\mathbf{Y}_1, \mathbf{Y}_2) = Y$.

Using (5.11), such a pair of thin-plate splines mapping under point constraints $\mathbf{T}$ and $\mathbf{Y}$ is given by

$$\begin{aligned}\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}) &= (g_1(\mathbf{x}), g_2(\mathbf{x})) = [g_1(\mathbf{x}) \;\; g_2(\mathbf{x})]) = \\ &= \left[\begin{bmatrix} \mathbf{Y}_1^T & 0 & 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} \;\; \begin{bmatrix} \mathbf{Y}_2^T & 0 & 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} \right] = \\ &= \left(\begin{bmatrix} \mathbf{Y}_1^T & 0 & 0 \\ \mathbf{Y}_2^T & 0 & 0 \end{bmatrix} \Gamma^{-1} \begin{bmatrix} \mathbf{s}(\mathbf{x}) \\ 1 \\ x_1 \\ x_2 \end{bmatrix} \right)^T = \begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \mathbf{Y}. \quad (5.16)\end{aligned}$$

Deformations of this type inherits many of the properties of the underlying thin-plate spline interpolants. Firstly, pair of thin-plate spline mappings are continuous, smooth and surjective interpolants. The domain of these mappings is all of $\mathbb{R}^2$ and at infinity $\phi_{\mathbf{T},\mathbf{Y}}$ is purely affine. Equivariance holds, not only on $\mathbf{T}$ of Lemma 5.1 but also on $\mathbf{Y}$.

Lemma 5.3. *Thin-plate spline mappings are equivariant under affine transformations $\Xi : \mathbb{R}^2 \to \mathbb{R}^2$ of $\mathbf{Y}$, i.e.*

$$\phi_{\mathbf{T},\Xi(\mathbf{Y})}(\mathbf{x}) = \Xi(\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x})), \quad (5.17)$$

where

$$\begin{aligned}\Xi(\mathbf{Y}) = \mathbf{Y}\Psi + \begin{bmatrix} \mathbf{1}_n\psi_1 & \mathbf{1}_n\psi_2 \end{bmatrix}, \\ \Psi \in \mathbb{R}^{2\times 2}, \\ \psi \in \mathbb{R}^2.\end{aligned}$$

Proof.

$$\begin{aligned}\phi_{\mathbf{T},\Xi(\mathbf{Y})}(\mathbf{x}) = \begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \left(\mathbf{Y}\Psi + \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix} \psi \right) = \\ \left(\begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \mathbf{Y} \right) \Psi + \\ + \begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix} \psi = [\text{using Lemma 5.2}] = \\ = \left(\begin{bmatrix} \mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2 \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \mathbf{Y} \right) \Psi + \psi = \Xi(\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}))\end{aligned}$$

□

5.3 Bijectivity Constraints on Thin-plate Spline Mappings

In spite of its appealing algebraic formulation presented in the previous section, thin-plate spline mappings do have drawbacks and, disregarding computational and numerical issues, one in particular. Namely, bijectivity is never assured. In computer vision, non-linear mappings in $\mathbb{R}^2$ of this sort are frequently used to model deformations in images. The basic assumption is that all the images contain similar structures and therefore there should exist mappings between pairs of images that are both one-to-one and onto. Hence bijective mappings are required.

From Sect. 5.2.1 we have a deformation $\phi_{\mathbf{T},\mathbf{Y}}$ that, for a given set of n control points $\mathbf{T}$, is parameterized (linearly) by the destination points $\mathbf{Y}$. It is of interest knowing which $\mathbf{Y}$ gives a bijective deformation, i.e the set

$$\Omega_{\mathbf{T}} = \{\mathbf{Y} \in \mathbb{R}^{2n} | \phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}) \text{ is bijective}\}.$$

Such a mapping $\phi : \mathbb{R}^2 \to \mathbb{R}^2$ is locally bijective at a point $\mathbf{x} \in \mathbb{R}^2$ iff its functional determinant $|J(\phi)|$ is non-zero. Here

$$|J(\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}))| = \begin{vmatrix} \frac{\partial \phi_1}{\partial x_1} & \frac{\partial \phi_1}{\partial x_2} \\ \frac{\partial \phi_2}{\partial x_1} & \frac{\partial \phi_2}{\partial x_2} \end{vmatrix} \tag{5.18}$$

using (5.16)

$$\begin{aligned} \frac{\partial \phi_1}{\partial x_1} = \frac{\partial}{\partial x_1}\left(\left[\begin{bmatrix}\mathbf{s}(\mathbf{x})^T & 1 & x_1 & x_2\end{bmatrix}\begin{bmatrix}\Gamma^{11} \\ \Gamma^{21}\end{bmatrix}\mathbf{Y}_1\right]\right) = \\ \left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + 0 \cdot \Gamma_0^{12^T} + 1 \cdot \Gamma_1^{12^T} + 0 \cdot \Gamma_2^{12^T}\right)\mathbf{Y}_1 = \\ \left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)\mathbf{Y}_1, \end{aligned} \tag{5.19}$$

and similarly

$$\frac{\partial \phi_2}{\partial x_1} = \left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)\mathbf{Y}_2, \tag{5.20}$$

$$\frac{\partial \phi_1}{\partial x_2} = \left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)\mathbf{Y}_1, \tag{5.21}$$

$$\frac{\partial \phi_2}{\partial x_2} = \left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)\mathbf{Y}_2. \tag{5.22}$$

where

$$\mathbf{s}'_{\mathbf{x}_i}(\mathbf{x}) = \frac{\partial}{\partial x_i}\begin{bmatrix}\sigma(\mathbf{x}-\mathbf{t}_1)\\ \sigma(\mathbf{x}-\mathbf{t}_2)\\ \vdots \\ \sigma(\mathbf{x}-\mathbf{t}_n)\end{bmatrix} = \begin{bmatrix}(x_i - t_{1i})\,(1+\log(||\mathbf{x}-\mathbf{t}_1||))\\ (x_i - t_{2i})\,(1+\log(||\mathbf{x}-\mathbf{t}_2||))\\ \vdots \\ (x_i - t_{ni})\,(1+\log(||\mathbf{x}-\mathbf{t}_n||))\end{bmatrix} =$$

$$= x_i\mathbf{1}_n + \mathbf{T}_i + \begin{bmatrix}(x_i - t_{1i})\,(\log(||\mathbf{x}-\mathbf{t}_1||))\\ (x_i - t_{2i})\,(\log(||\mathbf{x}-\mathbf{t}_2||))\\ \vdots \\ (x_i - t_{ni})\,(\log(||\mathbf{x}-\mathbf{t}_n||))\end{bmatrix}. \tag{5.23}$$

Inserting into (5.18) yields

$$|J(\phi_{\mathbf{T},\mathbf{Y}}(\mathbf{x}))| =$$

$$\left|\begin{matrix}\left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)\mathbf{Y}_1 & \left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)\mathbf{Y}_1\\ \left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)\mathbf{Y}_2 & \left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)\mathbf{Y}_2\end{matrix}\right| =$$

$$= \underbrace{\left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)}_{=\mathbf{b}_1(\mathbf{x})^T}\mathbf{Y}_1\underbrace{\left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)}_{=\mathbf{b}_2(\mathbf{x})^T}\mathbf{Y}_2 -$$

$$\underbrace{\left(\mathbf{s}'_{\mathbf{x}_2}(\mathbf{x})^T\Gamma^{11} + \Gamma_2^{12^T}\right)}_{=\mathbf{b}_2(\mathbf{x})^T}\mathbf{Y}_1\underbrace{\left(\mathbf{s}'_{\mathbf{x}_1}(\mathbf{x})^T\Gamma^{11} + \Gamma_1^{12^T}\right)}_{=\mathbf{b}_1(\mathbf{x})^T}\mathbf{Y}_2 =$$

$$= \mathbf{Y}_1^T\underbrace{\left(\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T\right)}_{D_{\mathbf{T}}(\mathbf{x})}\mathbf{Y}_2 =$$

$$= \frac{1}{2}\begin{bmatrix}\mathbf{Y}_1^T & \mathbf{Y}_2^T\end{bmatrix}\underbrace{\begin{bmatrix}0 & D_{\mathbf{T}}(\mathbf{x})\\ D_{\mathbf{T}}(\mathbf{x})^T & 0\end{bmatrix}}_{B(\mathbf{x})}\underbrace{\begin{bmatrix}\mathbf{Y}_1\\ \mathbf{Y}_2\end{bmatrix}}_{\hat{\mathbf{Y}}} =$$

$$= \frac{1}{2}\hat{\mathbf{Y}}^T B(\mathbf{x})\hat{\mathbf{Y}} = \frac{1}{2}h_{\mathbf{T}}(\mathbf{Y},\mathbf{x}). \tag{5.24}$$

Using Lemma 5.2, $\mathbf{b}_i(\mathbf{x})$ can be simplified

$$\mathbf{b}_i(\mathbf{x})^T = \left(\Gamma^{11}\mathbf{s}'_{\mathbf{x}_i}(\mathbf{x}) + \Gamma_1^{12}\right) =$$

$$= \Gamma^{11}\left(x_i\mathbf{1}_n + \mathbf{T}_i + \underbrace{\begin{bmatrix}(x_i - t_{1i})\,(\log(||\mathbf{x}-\mathbf{t}_1||))\\ (x_i - t_{2i})\,(\log(||\mathbf{x}-\mathbf{t}_2||))\\ \vdots \\ (x_i - t_{ni})\,(\log(||\mathbf{x}-\mathbf{t}_n||))\end{bmatrix}}_{\gamma_i(\mathbf{x})}\right) + \Gamma_i^{12} =$$

$$= \begin{bmatrix} \Gamma^{11}\mathbf{1}_n = 0 \\ \Gamma^{11}\mathbf{T}_i = 0 \end{bmatrix} = \Gamma^{11}\gamma_i(\mathbf{x}) + \Gamma_i^{12}. \tag{5.25}$$

Each point $\mathbf{x} \in \mathbb{R}^2$ gives a quadratic constraint on $\mathbf{Y}$, ($\hat{\mathbf{Y}}^T B_{\mathbf{T}}(\mathbf{x})\hat{\mathbf{Y}} \neq 0$) for local bijectivity. In order to simplify notation, $\mathbf{Y}$ will be used to denote its vectorized version $\hat{\mathbf{Y}}$ as well. The intended form of $\mathbf{Y}$ should be clear from the context. Since $\phi_{\mathbf{T},\mathbf{Y}}$ is a continuous mapping, for it to be globally bijective this constraint must either be $> 0,\ \forall \mathbf{x} \in \mathbb{R}^2$ or $< 0,\ \forall \mathbf{x} \in \mathbb{R}^2$.
The set $\Omega_{\mathbf{T}}$ can thus be written

$$\Omega_{\mathbf{T}} = \{\mathbf{Y} \in \mathbb{R}^{2n} | \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} > 0,\ \forall \mathbf{x} \in \mathbb{R}^2 \text{ or } \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} < 0,\ \forall \mathbf{x} \in \mathbb{R}^2\}.$$

Defining

$$\Omega_{\mathbf{T}}^+ = \{\mathbf{Y} \in \mathbb{R}^{2n} | \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} > 0,\ \forall \mathbf{x} \in \mathbb{R}^2\}$$

and with $\Omega_{\mathbf{T}}^-$ defined similarly, one can write $\Omega_{\mathbf{T}} = \Omega_{\mathbf{T}}^+ \cup \Omega_{\mathbf{T}}^-$. Seeing that if $\mathbf{Y} \in \Omega_{\mathbf{T}}^+$ then $\begin{bmatrix} -I & 0 \\ 0 & I \end{bmatrix}\mathbf{Y} \in \Omega_{\mathbf{T}}^-$, it does, without loss of generality, suffice to examine $\Omega_{\mathbf{T}}^+$. Hence, references to bijective thin-plate spline mappings will from here on be with respect to the set $\Omega_{\mathbf{T}}^+$.

The sought after set is evidently the intersection of an infinite number of high-dimensional quadratic forms each on the form (5.24). In an attempt at visualisation one can take 2-dimensional intersections of these constraints and plot the resulting quadratic curve for any number of points in $\mathbb{R}^2$, see Fig. 5.2.

This is clearly a somewhat impractical representation of $\Omega_{\mathbf{T}}^+$, an implicit representation would be preferable. That is a function $e(\mathbf{Y})$ such that

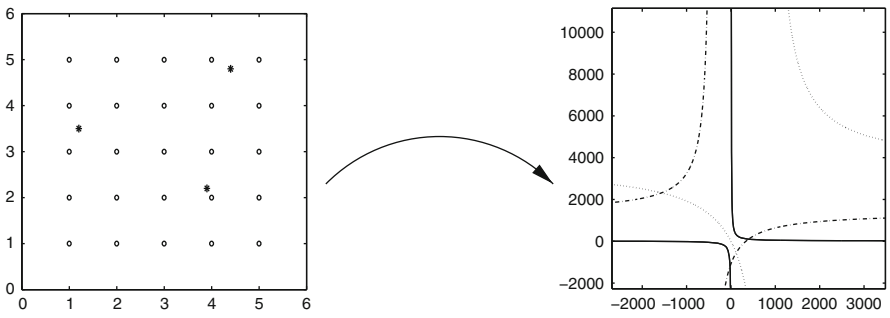


Fig. 5.2 The constraints imposed by three points in $\mathbb{R}^2$ on a 2-D affine subset of $\mathbb{R}^{2n}$. *Left*: The source configuration with three arbitrarily chosen points in $\mathbb{R}^2$ marked. *Right*: The three resulting quadratic constraints

$$\begin{cases} e(\mathbf{Y}) > 0 \Leftrightarrow Y \in \Omega_{\mathbf{T}}^{+}, \\ e(\mathbf{Y}) \leq 0 \Leftrightarrow Y \notin \Omega_{\mathbf{T}}^{+}. \end{cases}$$

Such an implicit representation of $\Omega_{\mathbf{T}}^{+}$ is contained in the affine variety defined by the envelope equations

$$\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} = 0, \tag{5.26}$$

$$\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})'_{x_1}\mathbf{Y} = 0, \tag{5.27}$$

$$\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})'_{x_2}\mathbf{Y} = 0. \tag{5.28}$$

This comes from the fact that (5.24) form a family of quadrics in $\mathbb{R}^{2n}$, parametrised by points in $\mathbb{R}^2$. Then the implicit representation must be a subset of the envelope of these functions.

An alternative way of viewing these equations is that points on the boundary of $\Omega_{\mathbf{T}}^{+}$ must be global minimizers of $h_{\mathbf{T}}(\mathbf{x}, \mathbf{Y}^*)$ over $\mathbb{R}^2$ with global minima 0. With this interpretation (5.26)–(2) are the first-order-conditions for such a minima.

However, the task of solving this system of log-linear equations is a formidable one that has yet not been accomplished.

5.3.1 Properties of $\Omega_{\mathbf{T}}^{+}$

Despite the high degree of complexity that $\Omega_{\mathbf{T}}^{+}$ possesses there are still a number of properties that can be identified. Firstly, the set in question actually is of a very familiar shape

Lemma 5.4. *The closure of* $\Omega_{\mathbf{T}}^{+}$, $cl(\Omega_{\mathbf{T}}^{+})$ *is*
(i) a generalised double cone
(ii) star-convex around 0
(iii) connected

Proof. It is only necessary to show that $cl(\Omega_{\mathbf{T}}^{+})$ is a cone since this implies star-convex around 0 and star-convex implies connected.

The closure of $\Omega_{\mathbf{T}}^{+}$ can be written $cl(\Omega_{\mathbf{T}}^{+}) = \{\mathbf{Y} \in \mathbb{R}^{2n} | \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} \geq 0, \ \forall \mathbf{x} \in \mathbb{R}^2\}$ then for any $y \in cl(\Omega_{\mathbf{T}}^{+})$ obviously λy, $\lambda \in \mathbb{R}$, is also in $\in cl(\Omega_{\mathbf{T}}^{+})$, hence $cl(\Omega_{\mathbf{T}}^{+})$ is a double cone. With a similar reasoning it can easily be shown that $\Omega_{\mathbf{T}}^{+}$ is a double cone with the origin removed. □

The defining matrix of the quadratic constraints $B_{\mathbf{T}}(\mathbf{x})$ and its subordinate $D_{\mathbf{T}}(\mathbf{x})$ are also surprisingly simple in their form. Some of their characteristics can be summed up in the following two lemmas

Lemma 5.5. *The $n \times n$ matrix $D_{\mathbf{T}}(\mathbf{x}) = \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T$ defined in Sect. 5.3*

1. *Is non-zero for all $\mathbf{x} \in \mathbb{R}^2 \Rightarrow \mathbf{b}_1(\mathbf{x})$ and $\mathbf{b}_2(\mathbf{x})$ are never parallel,*
2. *Is zero-diagonal,*
3. *Is skew-symmetric,*
4. *Column rank 2,*
5. *Has eigenvalues $\lambda = \pm i\lambda_D$, $\lambda_D = \sqrt{\mathbf{b}_1(\mathbf{x})\mathrm{d}_{\mathbf{T}}(\mathbf{x})\mathbf{b}_2(\mathbf{x})}$.*

Proof.

1. The matrix $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ defines the bijectivity constraints on $\mathbf{Y}$ for a given point $\mathbf{x} \in \mathbb{R}^2$. If there exist an $\mathbf{x}$ such that $\mathrm{d}_{\mathbf{T}}(\mathbf{x}) = 0$ then $\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} = 0$, for any destination configuration $\mathbf{Y}$. The thin-plate spline mapping is never locally bijective around that point regardless the choice of $\mathbf{Y}$. However, since we know that setting $\mathbf{Y} = \mathbf{T}$ gives the identity mapping, which is bijective. From this contradiction it is concluded that $D_{\mathbf{T}}(\mathbf{x})$ must be non-zero for all $\mathbf{x} \in \mathbb{R}^2$.

   If $\mathbf{b}_1(\mathbf{x})$ or $\mathbf{b}_2(\mathbf{x})$ are parallel then $\mathrm{d}(\mathbf{x}) = 0$ and the implication follows.
2. The matrix $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ is zero-diagonal since

$$(D_{\mathbf{T}}(\mathbf{x}))_{ii} = (\mathbf{b}_1(\mathbf{x}))_i(\mathbf{b}_2(\mathbf{x}))_i - (\mathbf{b}_2(\mathbf{x}))_i(\mathbf{b}_1(\mathbf{x}))_i = 0.$$

3. It is skew-symmetric as

$$\begin{aligned} D_{\mathbf{T}}(\mathbf{x})^T &= \left(\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T\right)^T = \\ &\mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T - \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T = \\ &= -\left(\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T\right) = -D_{\mathbf{T}}(\mathbf{x}). \end{aligned}$$

4. The column rank follows from that each column in $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ are the linear combination of non-zero vector non-parallel vectors $\mathbf{b}_1(\mathbf{x})$ and $\mathbf{b}_2(\mathbf{x})$.
5. Assuming that the eigenvectors of $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ can be written $v = \mathbf{b}_1(\mathbf{x}) + \alpha\mathbf{b}_2(\mathbf{x})$. The eigenvalue problem then becomes

$$\begin{aligned} \mathrm{d}_{\mathbf{T}}(\mathbf{x})v &= \lambda v, \\ \left(\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T\right)\left(\mathbf{b}_1(\mathbf{x}) + \alpha\mathbf{b}_2(\mathbf{x})\right) &= \\ &= \lambda\left(\mathbf{b}_1(\mathbf{x}) + \alpha\mathbf{b}_2(\mathbf{x})\right), \\ \left(\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}) + \alpha\mathbf{b}_2(\mathbf{x})^T\mathbf{b}_2(\mathbf{x})\right)\mathbf{b}_1(\mathbf{x}) + & \\ \left(-\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_1(\mathbf{x}) - \alpha\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x})\right)\mathbf{b}_2(\mathbf{x}) &= \lambda\mathbf{b}_1(\mathbf{x}) + \lambda\alpha\mathbf{b}_2(\mathbf{x}). \end{aligned}$$

Then for equality the following must hold

$$\begin{cases} \mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}) + \alpha\mathbf{b}_2(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}) = \lambda, \\ -\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_1(\mathbf{x}) - \alpha\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}) = \lambda\alpha. \end{cases}$$

Eliminating α gives

$$\lambda^2 + \Big((\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_1(\mathbf{x}))(\mathbf{b}_2(\mathbf{x})^T\mathbf{b}_2(\mathbf{x})) - (\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}))^2\Big) = 0$$

$$\lambda = \pm\sqrt{\underbrace{-\Big((\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_1(\mathbf{x}))(\mathbf{b}_2(\mathbf{x})^T\mathbf{b}_2(\mathbf{x})) - (\mathbf{b}_1(\mathbf{x})^T\mathbf{b}_2(\mathbf{x}))^2\Big)}_{\geq 0 \text{ (Cauchy-Schwarz)}}} =$$

$$= \pm i\sqrt{\mathbf{b}_1(\mathbf{x})^T \mathrm{d}_{\mathbf{T}}(\mathbf{x})\mathbf{b}_2(\mathbf{x}))}.$$

Since $\mathbf{b}_1(\mathbf{x})$ and $\mathbf{b}_2(\mathbf{x})$ are never parallell λ_D is always non-zero. Hence the two non-zero eigenvalues of $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ must be $\pm i\lambda_D$. It can be noted that as these eigenvalues are both imaginary and sums to zero ($i\lambda_D + (-i\lambda_D) = 0 = Tr(\mathrm{d}_{\mathbf{T}}(\mathbf{x}))$) in accordance with (ii) and (iii). □

Lemma 5.6. $B_{\mathbf{T}}(\mathbf{x}) = \begin{bmatrix} 0 & D_{\mathbf{T}}(\mathbf{x}) \\ -D_{\mathbf{T}}(\mathbf{x}) & 0 \end{bmatrix}$ *is a zero-diagonal, symmetric* $2n \times 2n$ *matrix with column rank 4 and eigenvalues* $\pm\lambda_D$

Proof. If v and u are the eigenvectors to $\mathrm{d}_{\mathbf{T}}(\mathbf{x})$ with eigenvalues $i\lambda_D$ and $-i\lambda_D$ it is readily shown that $\begin{bmatrix} v \\ iv \end{bmatrix}$, $\begin{bmatrix} -v \\ iv \end{bmatrix}$, $\begin{bmatrix} u \\ iu \end{bmatrix}$ and $\begin{bmatrix} -u \\ iu \end{bmatrix}$ are eigenvectors to $B_{\mathbf{T}}(\mathbf{x})$ with eigenvalues $-\lambda_D$, λ_D, λ_D and $-\lambda_D$ respectively.

$$\begin{bmatrix} 0 & D_{\mathbf{T}}(\mathbf{x}) \\ -D_{\mathbf{T}}(\mathbf{x}) & 0 \end{bmatrix}\begin{bmatrix} v \\ iv \end{bmatrix} = \begin{bmatrix} ii\lambda_D v \\ -i\lambda_D v \end{bmatrix} = -\lambda_D \begin{bmatrix} v \\ iv \end{bmatrix}$$

Similarly it can be shown for the remaining three. Zero-diagonality, symmetry and column rank follows trivially from the preceding lemma. □

The matrix $B_{\mathbf{T}}(\mathbf{x})$ is evidently of high dimension and low rank. Its vector- and null-space both vary with $\mathbf{x}$. A linear subspace of $\mathbb{R}^{2n}$ that is a subset of the null space of $B_{\mathbf{T}}(\mathbf{x})$ for all $\mathbf{x} \in \mathbb{R}^2$, as well as parts of the affine variety of the quadratic equation $B_{\mathbf{T}}(\mathbf{x})$ defines, can nevertheless be found.

Lemma 5.7. *The function* $h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x})$ *of (5.24) is the zero-function,*

$$h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x}) = \mathbf{Y}^T B(\mathbf{x})\mathbf{Y} = 0, \ \forall \mathbf{x} \in \mathbb{R}^2,$$

if

$$\mathbf{Y} \in \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix} \bigcup N,$$

where

$$N = \left\{ v = \begin{bmatrix} \mu w \\ \nu w \end{bmatrix} \mid w \in \mathbb{R}^n,\ \mu, \nu \in \mathbb{R}^2 \right\}.$$

Proof. If $\mathbf{Y} \in \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix}$ then $\mathbf{Y} = \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix} \tilde{\mathbf{Y}}$.
Using Lemma 5.2 it follows that

$$h_{\mathbf{T}}(\tilde{\mathbf{Y}}, \mathbf{x}) = \tilde{\mathbf{Y}}^T \begin{bmatrix} \mathbf{1}_n^T & 0 \\ 0 & \mathbf{1}_n^T \end{bmatrix} B_{\mathbf{T}}(\mathbf{x}) \begin{bmatrix} \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n \end{bmatrix} \tilde{\mathbf{Y}} =$$

$$= \tilde{\mathbf{Y}}^T \begin{bmatrix} -\mathbf{1}_n^T D_{\mathbf{T}}(\mathbf{x}) \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n^T D_{\mathbf{T}}(\mathbf{x}) \mathbf{1}_n \end{bmatrix} \tilde{\mathbf{Y}}.$$

Expanding gives

$$\mathbf{1}_n^T D_{\mathbf{T}}(\mathbf{x}) \mathbf{1}_n = \mathbf{1}_n^T (b_1(\mathbf{x}) b_2(\mathbf{x})^T - b_2(\mathbf{x}) b_1(\mathbf{x})^T) \mathbf{1}_n =$$

$$= (\mathbf{1}_n^T b_1(\mathbf{x}))(b_2(\mathbf{x})^T \mathbf{1}_n) - (\mathbf{1}_n^T b_2(\mathbf{x}))(b_1(\mathbf{x})^T \mathbf{1}_n)$$

$$= \begin{bmatrix} \mathbf{1}_n^T b_1(\mathbf{x}) = \mathbf{1}_n^T (\Gamma^{11} \gamma_1(\mathbf{x}) + \Gamma_1^{12}) = \mathbf{1}_n^T \Gamma^{11} \gamma_1(\mathbf{x}) + \mathbf{1}_n^T \Gamma_1^{12} = 0 \\ \mathbf{1}_n^T b_2(\mathbf{x}) = \mathbf{1}_n^T (\Gamma^{11} \gamma_2(\mathbf{x}) + \Gamma_2^{12}) = \mathbf{1}_n^T \Gamma^{11} \gamma_2(\mathbf{x}) + \mathbf{1}_n^T \Gamma_2^{12} = 0 \end{bmatrix} = 0$$

$$\Rightarrow h_{\mathbf{T}}(\tilde{\mathbf{Y}}, \mathbf{x}) = \tilde{\mathbf{Y}}^T \begin{bmatrix} -\mathbf{1}_n^T D_{\mathbf{T}}(\mathbf{x}) \mathbf{1}_n & 0 \\ 0 & \mathbf{1}_n^T D_{\mathbf{T}}(\mathbf{x}) \mathbf{1}_n \end{bmatrix} \tilde{\mathbf{Y}} = 0$$

If $\mathbf{Y} \in N$ then $\mathbf{Y} = \begin{bmatrix} \mu w \\ \nu w \end{bmatrix}$ and

$$h_{\mathbf{T}}\left(\begin{bmatrix} \mu w \\ \nu w \end{bmatrix}, \mathbf{x} \right) = \begin{bmatrix} \mu w^T & \nu w^T \end{bmatrix} B_{\mathbf{T}}(\mathbf{x}) \begin{bmatrix} \mu w \\ \nu w \end{bmatrix} =$$

$$= \begin{bmatrix} \mu w^T & \nu w^T \end{bmatrix} \begin{bmatrix} 0 & D_{\mathbf{T}}(\mathbf{x}) \\ -D_{\mathbf{T}}(\mathbf{x}) & 0 \end{bmatrix} \begin{bmatrix} \mu w \\ \nu w \end{bmatrix} =$$

$$= \mu \nu w^T D_{\mathbf{T}}(\mathbf{x}) w - \nu \mu w^T D_{\mathbf{T}}(\mathbf{x}) w = 0$$

□

Next we address the issues of boundedness and convexity.

5.3.2 Boundedness

Obviously, from the equivariance property of Lemma 5.3, the set in question is indeed unbounded. Since the composition of bijective deformations is also bijective, any bijective target configuration can be deformed by any mapping from the unbounded set of bijective affine transformations $\mathbb{R}^2 \to \mathbb{R}^2$ and still belong to $\Omega^+_{\mathbf{T}}$.

However, if the affine transformations are disregarded is $\Omega^+_{\mathbf{T}}$ still unbounded? By studying one-dimensional intersections of the set, it can be shown (for specific configurations $\mathbf{T}$) that the set is indeed bounded if this restriction is introduced. Consider the subset E of configurations in which the first three points have the same positions as corresponding points in $\mathbf{T}$, i.e. E is formed by perturbing all but the first three points of $\mathbf{T}$. Define the subset $\Omega^+_{\mathbf{T},E}$ as

$$\Omega^+_{\mathbf{T},E} = \{\mathbf{Y} \in E | h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x}) > 0,\ \forall \mathbf{x} \in \mathbb{R}^2\}.$$

These are the configurations in E which gives bijective thin-plate-spline transformations. Now study one-dimensional affine subspaces of E containing $\mathbf{T}$, i.e.

$$E_{\mathbf{d}} = \{\mathbf{T} + s\mathbf{d} | s \in \mathbb{R}\},$$

where $\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$ represents changes in configurations. Here $\mathbf{d}$ must be zero at the elements corresponding to the fixed points so that $E_{\mathbf{d}} \subset E$. This intersection of $E_{\mathbf{d}}$ with $\Omega^+_{\mathbf{T},E}$ is

$$\Omega^+_{\mathbf{T},E_d} = \{\mathbf{Y} \in E_{\mathbf{d}} | h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x}) > 0,\ \forall \mathbf{x} \in \mathbb{R}^2\}.$$

Here

$$h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x}) = h_{\mathbf{T}}(\mathbf{T} + s\mathbf{d}, \mathbf{x}) = a_{\mathbf{d}}(\mathbf{x})s^2 + b_{\mathbf{d}}(\mathbf{x})s + c_{\mathbf{d}}(\mathbf{x}).$$

Since $h_{\mathbf{T}}(\mathbf{Y}, \mathbf{x})$ is quadratic in its first argument, for each point $\mathbf{x} \in \mathbb{R}^2$, we thus get a quadratic constraint on s. Here the coefficients of the second order constraints are given by

$$\begin{aligned} a_{\mathbf{d}}(x) &= \mathbf{d}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{d}, \\ b_{\mathbf{d}}(x) &= 2\mathbf{T}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{d}, \\ c_{\mathbf{d}}(x) &= \mathbf{T}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{T}. \end{aligned} \tag{5.29}$$

Lemma 5.8. *The function $b_{\boldsymbol{d}}(\mathbf{x})$ can be simplified as*

$$b_{\boldsymbol{d}}(\mathbf{x}) = [\mathbf{b}_1(\mathbf{x})^T \ \ \mathbf{b}_2(\mathbf{x})^T]\boldsymbol{d} = \mathbf{b}(\mathbf{x})^T\boldsymbol{d}.$$

The function $c_{\boldsymbol{d}}(\mathbf{x})$ is independent of both $\boldsymbol{d}$ and $\mathbf{x}$. In fact

$$c_{\boldsymbol{d}}(\mathbf{x}) = 2.$$

Proof. Using Lemma 5.2 gives

$$\begin{aligned}
\mathbf{T}_1^T \mathbf{b}_1(\mathbf{x}) &= \mathbf{T}_1^T (\Gamma^{11}\gamma_1(\mathbf{x}) + \Gamma_1^{12^T}) = \mathbf{T}_1^T \Gamma_1^{12^T} = 1,\\
\mathbf{T}_1^T \mathbf{b}_2(\mathbf{x}) &= \mathbf{T}_1^T (\Gamma^{11}\gamma_2(\mathbf{x}) + \Gamma_2^{12^T}) = \mathbf{T}_1^T \Gamma_2^{12^T} = 0,\\
\mathbf{T}_2^T \mathbf{b}_1(\mathbf{x}) &= \mathbf{T}_2^T (\Gamma^{11}\gamma_1(\mathbf{x}) + \Gamma_1^{12^T}) = \mathbf{T}_2^T \Gamma_1^{12^T} = 0,\\
\mathbf{T}_y^T \mathbf{b}_2(\mathbf{x}) &= \mathbf{T}_2^T (\Gamma^{11}\gamma_2(\mathbf{x}) + \Gamma_2^{12^T}) = \mathbf{T}_3^T \Gamma_2^{12^T} = 1.
\end{aligned}$$

So

$$\begin{aligned}
\mathbf{T}_1^T D_{\mathbf{T}}(\mathbf{x}) &= \mathbf{T}_1^T (\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T) = \mathbf{b}_2(\mathbf{x})^T,\\
\mathbf{T}_2^T D_{\mathbf{T}}(\mathbf{x}) &= \mathbf{T}_2^T (\mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T) = -\mathbf{b}_1(\mathbf{x})^T.
\end{aligned}$$

This implies that

$$\begin{aligned}
b_{\mathbf{d}}(x) &= 2\mathbf{T}^T B_{\mathbf{T}}(x)\mathbf{d} = [\mathbf{T}_1^T \;\; \mathbf{T}_2^T] B_{\mathbf{T}}(\mathbf{x})\mathbf{d} = [-\mathbf{T}_2^T D_{\mathbf{T}}(\mathbf{x}) \;\; \mathbf{T}_1^T D_{\mathbf{T}}(\mathbf{x})]\mathbf{d} =\\
&= [\mathbf{b}_1(\mathbf{x})^T \;\; \mathbf{b}_2(\mathbf{x})^T]\mathbf{d} = \mathbf{b}(\mathbf{x})^T\mathbf{d}
\end{aligned}$$

and

$$c_{\mathbf{d}}(\mathbf{x}) = 2\mathbf{T}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{T} = [\mathbf{b}_1(\mathbf{x})^T \;\; \mathbf{b}_2(\mathbf{x})^T]\mathbf{T} = \mathbf{b}_1(\mathbf{x})^T\mathbf{T}_1 + \mathbf{b}_2(\mathbf{x})^T\mathbf{T}_2 = 2. \quad \square$$

A sufficient condition on the boundedness of $\Omega^+_{\mathbf{T},E_d}$ is that there exists a point $\mathbf{x} \in \mathbb{R}^2$ such that $a_{\mathbf{d}}(\mathbf{x}) < 0$ since this will limit the distance s for which the spline mapping is bijective. To prove that $\Omega^+_{\mathbf{T},E}$ is unbounded it is sufficient to show that $\Omega^+_{\mathbf{T},E_d}$ is bounded for every direction $\mathbf{d}$, i.e. that $a_{\mathbf{d}}(\mathbf{x})$ never can be non-negative.

Here we need to study the space of all functions $a_{\mathbf{d}}(\mathbf{x})$ as the direction d is varied.

Lemma 5.9. *Given a thin-plate-spline defined by n separate control points, assume that the first three points define an affine basis. All possible functions $a_{\boldsymbol{d}}(\mathbf{x})$ given by (5.29) lie in the $D = (n+1)^2 - n$ dimensional space A of functions spanned by functions $a_{ij}(\mathbf{x})$,*

$$a_{ij}(\mathbf{x}) = f_{1,i}(\mathbf{x})\, f_{2,j}(\mathbf{x}), \tag{5.30}$$

where

$$f_1(\mathbf{x}) = \begin{bmatrix} \gamma_1(\mathbf{x}) \\ 1 \end{bmatrix},$$

$$f_2(\mathbf{x}) = \begin{bmatrix} \gamma_2(\mathbf{x}) \\ 1 \end{bmatrix}.$$

Proof. The function $a_{\mathbf{d}}(\mathbf{x})$ can be written

$$\begin{aligned} a_{\mathbf{d}}(\mathbf{x}) &= \mathbf{d}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{d} = 2\mathbf{b}_1(\mathbf{x})^T(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T)\mathbf{b}_1(\mathbf{x})^T = \\ &= 2(\gamma_1(\mathbf{x})^T\Gamma^{11} + (\Gamma_1^{12})^T)\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)(\Gamma^{11}\gamma_2(\mathbf{x}) + \Gamma_2^{12}) = \\ &= \gamma_1(\mathbf{x})^T \underbrace{\Big(\Gamma^{11}\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)\Gamma^{11}\Big)}_{\text{zero-diagonal}} \gamma_2(\mathbf{x}) + \\ &\quad + 2(\Gamma_1^{12})^T\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)\gamma_2(\mathbf{x}) - 2(\Gamma_2^{12})^T\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)\gamma_1(\mathbf{x}) + \\ &\quad + 2(\Gamma_1^{12})^T\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)\Gamma_2^{12} = \sum_{i=1}^{n+1}\sum_{j=1}^{n+1}\alpha_{ij}\, f_{1,i}(\mathbf{x})\, f_{2,j}(\mathbf{x}) = \\ &= \sum_{i=1}^{n+1}\sum_{j=1}^{n+1}\alpha_{ij}\, a_{ij}(\mathbf{x}) \end{aligned}$$

With $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ defined as above. As the matrix $\Gamma^{11}\Big(\mathbf{d}_1\mathbf{d}_2^T - \mathbf{d}_2\mathbf{d}_1^T\Big)\Gamma^{11}$ is zero-diagonal, $\alpha_{ij} = 0$ for all $i = j$, except for $i = j = 1$ giving the dimension of A. □

Theorem 5.1. *For a number of grids T, including rectangular regular grids of $l \times m$ with $l < 10$ and $m < 10$, the set of perturbations that leave three of the corner points fixed and gives bijective thin-plate splines is bounded in all directions for which $\mathbf{d}_1$ is not parallel to $\mathbf{d}_2$.*

Proof. The proof follows from explicit study of the basis functions $a_{ij}(\mathbf{x})$ for these grids. Thus for a given point configuration $\mathbf{T}$ and assuming that three of the points in $\mathbf{T}$ constitute an affine basis it is possible to calculate a basis A of functions which contain all possible functions $a_{\mathbf{d}}(\mathbf{x})$ with $\mathbf{d}$ leaving the affine basis fixed. By studying the feasibility of the convex set

$$\{z \in \mathbb{R}^D,\ z \neq 0 |, Az \geq 0, (\mathbf{1}^T A)^T z = 1\}$$

with A containing as rows the D basis functions sampled at a discrete number of points. It can be shown that there exists no non-negative functions in A, except the zero function. The only directions $\mathbf{d}$ for which $a_{\mathbf{d}}(\mathbf{x})$ is constantly equal to zero are those with which $\mathbf{d}_1$ and $\mathbf{d}_2$ are parallel. □

5.3.3 Convexity

In certain computer vision applications it is desirable to find deformations that map two or more images onto each other optimally. In optimization theory the main issue is not that of linearity and nonlinearity, but convexity and nonconvexity, any convex properties of $\Omega_{\mathbf{T}}^{+}$ is therefore of great interest. Since the set in question is the intersection of an infinite number of non-convex sets, it would be expected that $\Omega_{\mathbf{T}}^{+}$ is non-convex. This is also the case.

Lemma 5.10. *In general $\Omega_{\mathbf{T}}^{+}$ is not a convex set.*

Proof. Proof by counter-example. For $\Omega_{\mathbf{T}}^{+}$ to be convex then for any $\mathbf{y}_1, \mathbf{y}_2 \in \Omega_{\mathbf{T}}^{+}$ the line $\lambda\mathbf{y}_1 + (1-\lambda)\mathbf{y}_2$ must also be in $\Omega_{\mathbf{T}}^{+}$ for $0 \leq \lambda \leq 1$. A simple counter-example where this convexity requirement is not met can be found by choosing $\mathbf{T}$ to be a regular 3×3 rectangular grid, and $\mathbf{y}_1$, $\mathbf{y}_2$ slightly altered versions of $\mathbf{T}$, see Fig. 5.3. □

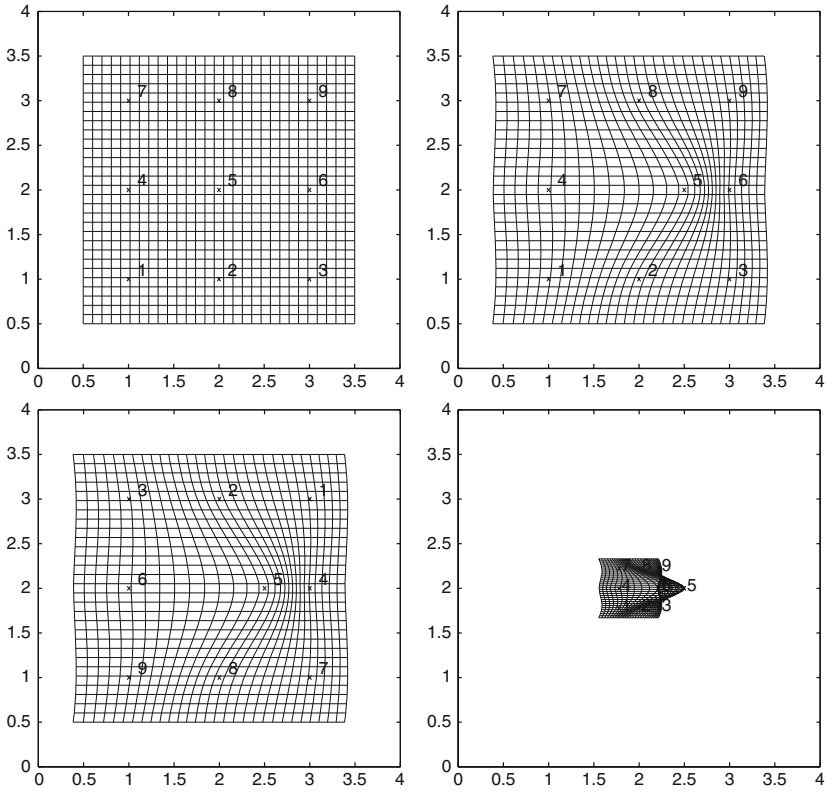


Fig. 5.3 A simple example illustrating the non-convexity of $\Omega_{\mathbf{T}}^{+}$ *Top left*: Source configuration $\mathbf{T}$, *Top right*: Target configuration $\mathbf{y}_1$, *Bottom left*: Target configuration $\mathbf{y}_2$, *Bottom right*: Target configuration $\mathbf{y}_{1+2} = \lambda\mathbf{y}_1 + (1-\lambda)\mathbf{y}_2$, $\lambda = 0.4$. Clearly $\mathbf{y}_1$ and $\mathbf{y}_2$ are bijective but $\mathbf{y}_{1+2}$ is not

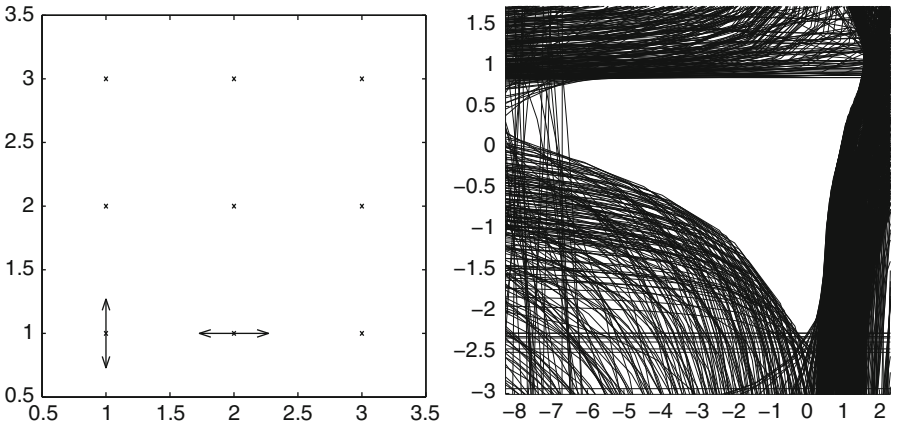


Fig. 5.4 Example of non-convex intersection with $\Omega_{\mathbf{T}}^{+}$ under affine restriction. Here only the two left-most points in the *bottom row* are permitted to move in one dimension, as indicated by the *arrows*. The resulting set is clearly non-convex

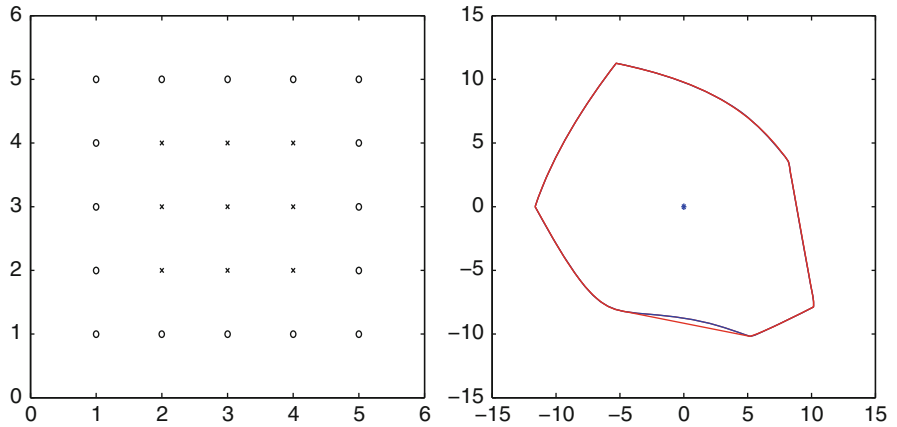


Fig. 5.5 An example of a non-convex intersection with $\Omega_{\mathbf{T}}^{+}$ with the restriction that only points in the interior of the convex hull of $\{\mathbf{t}_i\}$, $i = 1, \ldots, n$ were permitted to move

Adopting the approach of disregarding affine transformations from Sect. 5.3.2 does not make the set display any convex characteristics. As in the proof of the preceding lemma a counterexample can easily be constructed, see Fig. 5.4.

As it was observed that these non-convex intersections often involved points on the boundary of the convex hull of $\{\mathbf{t}_i\}$, $i = 1, \ldots, n$. The idea was to examine if not permitting points on this boundary to move would ensure convexity. This proved not to be the case, a example of this can be seen in Fig. 5.5.

Apparently $\Omega_{\mathbf{T}}^{+}$ is a highly non-convex set. However, there are restrictions for which convexity can be achieved.

Lemma 5.11. *The set $\Omega^+_{\mathbf{T},E}$ is convex if the affine subspace $E \subseteq N$, with N defined as in Lemma 5.7.*

Proof. With

$$\Omega^+_{T,E} = \{\mathbf{Y} \in E | \mathbf{Y}^T B(\mathbf{x})\mathbf{Y} > 0\ , \forall \mathbf{x} \in \mathbb{R}^2\},$$
$$E = \{U + Ey | E \in \mathbb{R}^{2n \times l},\ U \in \mathbb{R}^{2n},\ y \in \mathbb{R}^l\}$$

we get $Y = U + Ey$. Consequently, using Lemma 5.7

$$\mathbf{Y}^T B(\mathbf{x})\mathbf{Y} = (U + Ey)^T B(\mathbf{x})(U + Ey) =$$
$$= y^T (\underbrace{E^T B(\mathbf{x}) E}_{=0}) y + 2U^T B(\mathbf{x}) y + U^T B(\mathbf{x}) U = 2U^T B(\mathbf{x}) y + U^T B(\mathbf{x}) U.$$

The set $\Omega^+_{T,E}$ is now defined by linear constraints, a polytope and is therefore convex. □

Corollary 5.1. *The feasible bijective thin-plate spline mappings when only displacing one target point location make up a convex set.*

Proof. This follows trivially from Lemma 5.11 as the corresponding affine subset is contained in N. □

Finally, there are strong indications that $\Omega^+_{\mathbf{T}}$ is star-convex around $\mathbf{T}$. That is, that the intersection of $\Omega^+_{\mathbf{T}}$ and any affine one-dimensional subspace of $\mathbb{R}^{2n}$ containing $\mathbf{T}$ is convex. However, proving this statement still remains open.

5.4 Sufficient Conditions for Bijectivity

Given the complexity of the set of bijective thin-plate spline deformations, the enterprise of finding the defining expressions analytically is a formidable one. Instead one can use numerical methods to derive conditions on $\Omega^+_{\mathbf{T}}$. By finding subsets of $\Omega^+_{\mathbf{T}}$, through different relaxation methods, sufficient conditions for bijectivity can be obtained. In this section we discuss some of these conditions.

5.4.1 Maximum-Volume Inscribed Sphere

A sufficient condition for bijectivity could be a sphere S contained in $\Omega^+_{\mathbf{T}}$, so that if $\mathbf{Y} \in S \Rightarrow \mathbf{Y} \in \Omega^+_{\mathbf{T}}$. Obviously the larger the volume of the sphere contained in $\Omega^+_{\mathbf{T}}$ is the better the sufficient condition would be.

Let S_R be a sphere with radius R defined by an quadratic inequality

$$S_R = \left\{ d \in \mathbb{R}^{2n} \| \; -\frac{1}{R^2} d^T d + 1 > 0 \right\}. \tag{5.31}$$

Using the notation from Sect. 5.3, $\Omega_{\mathbf{T}}^+$ is the intersection of quadrics on the form

$$C(\mathbf{x}) = \left\{ d \in \mathbb{R}^{2n} \mid d^T B_{\mathbf{T}}(\mathbf{x}) d + 2\mathbf{b}(\mathbf{x})^T d + 2 > 0, \; \mathbf{x} \in \mathbb{R}^2 \right\}$$

it is clear that $S_R \subset \Omega_{\mathbf{T}}^+$ if $S \subset C(\mathbf{x}), \; \forall x \in \mathbb{R}^2$

Theorem 5.2. *A thin-plate spline mapping $\phi_{\mathbf{T},\mathbf{Y}} : \mathbb{R}^2 \to \mathbb{R}^2$ with n point constraints $\mathbf{T}$ and $\mathbf{Y}$ is bijective if*

$$||\mathbf{Y} - \mathbf{T}|| < R = \frac{1}{\sqrt{\max_{\mathbf{x} \in \mathbb{R}^2} (\lambda_M(\mathbf{x}))}}, \tag{5.32}$$

that is if $\mathbf{Y}$ is inside a sphere centered at $\mathbf{T}$ with radius R. Here $\lambda_M(\mathbf{x})$ are the eigenvalues of the matrix

$$M(\mathbf{x}) = \begin{bmatrix} \mathbf{b}_1(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T & \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T \\ \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T & \mathbf{b}_2(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T \end{bmatrix} \tag{5.33}$$

Proof. The S-procedure, a commonly used method for dealing with quadratic constraints [12] , gives that S_R is in $C(\mathbf{x})$ if there exist a $\tau \geq 0$ such that

$$\begin{bmatrix} B_{\mathbf{T}}(\mathbf{x}) & \mathbf{b}(\mathbf{x}) \\ \mathbf{b}(\mathbf{x})^T & 2 \end{bmatrix} - \tau \begin{bmatrix} -\frac{1}{R^2} I & 0 \\ 0 & 1 \end{bmatrix} \succeq 0$$

$$\begin{bmatrix} B_{\mathbf{T}}(\mathbf{x}) + \tau \frac{1}{R^2} I & \mathbf{b}(\mathbf{x}) \\ \mathbf{b}(\mathbf{x})^T & 2 - \tau \end{bmatrix} \succeq 0$$

By the Schur complement, this is equivalent to

$$\left(B_{\mathbf{T}}(\mathbf{x}) + \tau \frac{1}{R^2} I \right) - \frac{1}{2 - \tau} \mathbf{b}(\mathbf{x}) \mathbf{b}(\mathbf{x})^T \succeq 0,$$

$$0 \leq \tau \leq 2.$$

Setting $\tau = 1$ gives

$$B_{\mathbf{T}}(\mathbf{x}) + \frac{1}{R^2} I - \mathbf{b}(\mathbf{x})\mathbf{b}(\mathbf{x})^T =$$

$$= \begin{bmatrix} 0 & \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T - \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T \\ \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T - \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T & 0 \end{bmatrix} -$$

$$-\begin{bmatrix} \mathbf{b}_1(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T & \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T \\ \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T & \mathbf{b}_2(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T \end{bmatrix} + \frac{1}{R^2} I =$$

$$= \frac{1}{R^2} I - \underbrace{\begin{bmatrix} \mathbf{b}_1(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T & \mathbf{b}_1(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T \\ \mathbf{b}_2(\mathbf{x})\mathbf{b}_1(\mathbf{x})^T & \mathbf{b}_2(\mathbf{x})\mathbf{b}_2(\mathbf{x})^T \end{bmatrix}}_{M}(\mathbf{x}) \succeq 0.$$

This holds if $\frac{1}{R^2}$ is greater than the largest eigenvalue of $M(\mathbf{x})$, or

$$R \leq \frac{1}{\sqrt{\max_{\mathbf{x}\in\mathbb{R}^2}(\lambda_M(\mathbf{x}))}}. \qquad \square$$

Even though this theorem provides a simple sufficient condition for bijectivity it does require the computation of a large number eigenvalues. As eigenvalue computation involving large matrices is a notoriously arduous task, it should be avoided as much as possible. Fortunately, a closer look at the matrix $M(\mathbf{x})$ from the preceding theorem reveals a relatively simple expression of the largest eigenvalue of such matricises.

Theorem 5.3. *The largest eigenvalue of a matrix on the form*

$$M = \begin{bmatrix} uu^T & vu^T \\ uv^T & vv^T \end{bmatrix},$$

where $v,\ u \in \mathbb{R}^n$ *and not parallel, is equal to*

$$\lambda_{max} = \frac{1}{2}\left(u^T u + v^T v + \sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2}\right). \tag{5.34}$$

Proof. Assuming that the symmetric rank 4 matrix M has eigenvectors on the form $\begin{bmatrix} u + av \\ cu + dv \end{bmatrix}$. Then finding the eigenvalues of M means solving

$$\begin{bmatrix} vv^T & uv^T \\ vu^T & uu^T \end{bmatrix}\begin{bmatrix} u + av \\ cu + dv \end{bmatrix} = \lambda \begin{bmatrix} u + av \\ cu + dv \end{bmatrix}.$$

Multiplying gives

$$\begin{bmatrix} uu^T u + auu^T v + cvu^T u + dvu^T v \\ uv^t u + auv^T v + cvv^T u + dvv^T v \end{bmatrix} = \lambda \begin{bmatrix} u + av \\ cu + dv \end{bmatrix},$$

$$\begin{bmatrix} (u^T u + au^T v)u + (cu^T u + du^T v)v \\ (v^T u + av^T v)u + (cv^T u + dv^T v)v \end{bmatrix} = \begin{bmatrix} \lambda u + \lambda av \\ \lambda cu + \lambda dv \end{bmatrix}.$$

For equality

$$\begin{cases} (u^T u + a u^T v) = \lambda, \\ (cu^T u + d u^T v) = \lambda a, \\ (v^T u + a v^T v) = \lambda c, \\ (c v^T u + d v^T v) = \lambda d \end{cases}$$

must hold.

Continuing solving the equation system yields

$$\begin{cases} (u^T u + a u^T v) = \lambda, \\ (cu^T u + d u^T v) = (a u^T u + a^2 u^T v), \\ (v^T u + a v^T v) = cu^T u + acu^T v, \\ (c v^T u + d v^T v) = d u^T u + ad u^T v. \end{cases}$$

Eliminating λ from the second equation gives

$$c = \frac{v^T u + a v^T v}{u^T u + a u^T v}.$$

Inserted into the third and fourth equations

$$\begin{cases} ((\frac{v^T u + a v^T v}{u^T u + a u^T v}) u^T u + d u^T v) = a u^T u + a^2 u^T v, \\ ((\frac{v^T u + a v^T v}{u^T u + a u^T v}) v^T u + d v^T v) = d u^T u + ad u^T v. \end{cases}$$

Now solving for d

$$d = v^T u \frac{v^T u + a v^T v}{(u^T u + a u^T v)(u^T u + a u^T v - v^T v)},$$

and back-substitution gives a single polynomial equation in a

$$\begin{aligned} (u^T v)^3 a^4 + (u^T v)^2 (3(u^T u) - v^T v) a^3 + 3u^T v((u^T u)^2 - (u^T u)(v^T v)) a^2 + \\ + ((u^T u)^3 - (u^T u)^2 v^T v - u^T u (u^T v)^2 - (u^T u)^2 v^T v + u^T u (v^T v)^2 - \\ - v^T v (u^T v)^2) a - (u^T u)^2 u^T v + (u^T u)(v^T v)(u^T v) - (u^T v)^3 = 0, \end{aligned}$$

which can be factorized into

$$\begin{aligned} &\Big((u^T v) a^2 + (u^T u - v^T v) a - u^T v \Big) \\ &\quad \Big((u^T v)^2 a^2 + 2(u^T v)(u^T u) a + ((u^T u)^2 + (u^T v)^2 - (u^T u)(v^T v)) \Big) = 0. \end{aligned}$$

The first parenthesis gives

$$a_{1,2} = -\frac{u^T u - v^T v}{2u^T v} \pm \sqrt{\left(\frac{u^T u - v^T v}{2u^T v}\right)^2 + 1} =$$
$$= \frac{-(u^T u - v^T v) \pm \sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2}}{2u^T v}$$

and the second

$$a_{3,4} = -\frac{u^T u}{u^T v} \pm \sqrt{\left(\frac{u^T u}{u^T v}\right)^2 - \frac{(u^T u)^2 + (u^T v)^2 - (u^T u)(v^T v)}{(u^T v)^2}} =$$
$$= \frac{-u^T u \pm \sqrt{(u^T v)^2 - (u^T u)(v^T v)}}{u^T v}.$$

Finally with $\lambda = u^T u + a u^T v$ the eigenvalues of M can written

$$\lambda_{1,2} = u^T u + a_{1,2} u^T v = \frac{1}{2}\left(u^T u + v^T v \pm \sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2}\right)$$
$$\lambda_{3,4} = \pm u^T u + a_{3,4} u^T v = \sqrt{(u^T u)(v^T v) - (u^T v)^2}$$

Since M is a rank 4 matrix and $\lambda_1, \ldots, \lambda_4 \neq 0$ the initial assumption on the form of the eigenvectors is correct and the non-zero eigenvalues of M are the ones given above. It only remains to determine which of these eigenvalues is the largest. Obviously $\lambda_1 \geq \lambda_2$ and $\lambda_3 \geq \lambda_4$. Comparing λ_1 and λ_3

$$\lambda_1^2 - \lambda_3^2 = \frac{1}{4}\left((u^T u + v^T v) + \sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2}\right)^2 -$$
$$- \left((u^T u)(v^T v) - (u^T v)^2\right) = \frac{1}{2}(u^T u)^2 - (u^T u)(v^T v) + \frac{1}{2}(v^T v)^2 +$$
$$+ (u^T v)^2 + \frac{1}{2}(u^T u + v^T v)\sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2} =$$
$$= \frac{1}{2}\underbrace{\left((u^T u) - (v^T v)\right)^2}_{\geq 0} + \underbrace{(u^T v)^2}_{\geq 0} +$$
$$+ \underbrace{\frac{1}{2}(u^T u + v^T v)}_{\geq 0}\underbrace{\sqrt{(u^T u - v^T v)^2 + 4(u^T v)^2}}_{\geq 0} \geq 0$$
$$\Rightarrow \lambda_1^2 \geq \lambda_3^2.$$

Since $\lambda_1, \lambda_3 > 0$ this implies that $\lambda_1 > \lambda_3$. □

Applying this result to Theorem 5.2 results in the following corollary.

Corollary 5.2. *The largest eigenvalue of the matrix* $M(\mathbf{x})$ *from Theorem 5.2 is given by*

$$\lambda_M(\mathbf{x}) = \frac{1}{2}\Big(b_1(\mathbf{x})^T b_1(\mathbf{x}) + b_2(\mathbf{x})^T b_2(\mathbf{x}) + \sqrt{(b_1(\mathbf{x})^T b_1(\mathbf{x}) - b_2(\mathbf{x})^T b_2(\mathbf{x}))^2 + 4(b_1(\mathbf{x})^T b_2(\mathbf{x}))^2}\Big). \tag{5.35}$$

Remark 5.1. It can be noted that two of the smaller eigenvalues of $M(\mathbf{x})$ are identical to the eigenvalues of the matrix $B_{\mathbf{T}}(\mathbf{x})$.

$$\lambda_{3,4} = \pm\sqrt{(b_1(\mathbf{x})^T b_1(\mathbf{x}))(b_2(\mathbf{x})^T b_2(\mathbf{x})) - (b_1(\mathbf{x})^T b_2(\mathbf{x}))^2 -} = \text{[see Lemma 5.6]} = \pm\lambda_D.$$

An example of a maximum-volume sphere conditions for a generic source configuration $\mathbf{T}$ is shown in Fig. 5.6.

5.4.2 Maximum-Volume Inscribed Ellipsoid

The condition from the previous section can be improved by instead finding the maximum-volume ellipsoid $\mathscr{E} = \{d \in \mathbb{R}^{2n} \mid d^T A d + 2a^T d - 1 < 0\}$ inscribed in $\Omega_{\mathbf{T}}^+$. Finding such extremal volume ellipsoids can be formulated as optimization problems [10, 14].

$$\max \quad \text{volume of } E$$
$$\text{s.t. } E \subset C(\mathbf{x}), \forall\ \mathbf{x} \in \mathbb{R}^2.$$

However, since there are finitely many variables and an infinite number of constraints this is a semi-infinite program [8]. In order to avoid this we approximate $\Omega_{\mathbf{T}}^+$ by the intersection of a finite subset of these constraints.

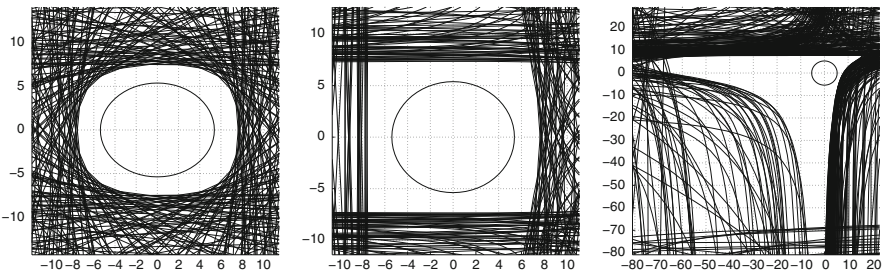


Fig. 5.6 The intersection of three arbitrarily chosen hyperplanes and $\Omega_{\mathbf{T}}^+$ along with the resulting maximum-volume sphere condition of Theorem 5.2

$$\tilde{C}_i = \left\{ d \in \mathbb{R}^{2n} \mid d^T B_i d + 2 b_i^T d + 2 > 0,\ i = 1 \dots L,\ \ B_i = B_{\mathbf{T}}(\mathbf{x}_i),\ \mathbf{x}_i \in \mathbb{R}^2 \right\}$$

Using that the volume of E is proportional to $\log(\det A)$ the maximum-volume inscribed ellipsoid optimization problem can be formulated

Lemma 5.12. *The ellipsoid* $\mathscr{E} = \left\{ d \in \mathbb{R}^{2n} \mid d^T A^* d + 2 a^{*T} d - 1 < 0 \right\}$ *where* A^* *and* a^* *are the global optimizers of*

$$\begin{aligned} &\min \ \log(\det A) \\ &\text{s.t.} \begin{bmatrix} B_i & b_i \\ b_i^T & 2 \end{bmatrix} - \tau_i \begin{bmatrix} -A & a \\ a^T & 1 \end{bmatrix} \succeq 0 \\ &\tau_i \geq 0 \\ &i = 1 \dots L \end{aligned}$$

is the maximum-volume ellipsoid inscribed in $\bigcap_{i=1}^{L} \tilde{C}_i$.

Proof. The volume of an ellipsoid on the given form is inversely proportional to $\log \det(A)$. The constraints follows directly from the S-procedure, see the proof of Theorem 5.2. □

This is a non-linear program with a convex objective function and bilinear matrix inequality constraints. It can been shown [14] that it is a convex program if $\Omega_{\mathbf{T}}^{+}$ is a convex set. Following that this formulation is less constrained than Lemma 5.2, the ellipsoid $\mathscr{E}$ should provide a superior sufficient constraint on $\mathbf{Y}$ for bijectivity. However, the disadvantage of this approach is that it involves a more computationally complex optimization problem.

5.4.3 Improving Sufficient Conditions for Bijectivity

The sufficient conditions derived in Sects. 5.4.1 and 5.4.2 are on a very compact and simple form but can in cases be overly tight. Using properties discussed in Sect. 5.3.1, such convex bounded quadratic constraints can be further improved while still keeping their appealing representation.

First the following lemma that connects the null space of $B_{\mathbf{T}}(\mathbf{x})$ to bijective target configurations is formulated.

Lemma 5.13. *If* $\mathbf{Y}$ *gives a bijective bijective mapping, that is* $\mathbf{Y} \in \Omega_{\mathbf{T}}^{+}$, *so do all points in the hyperplane* $\mathbf{Y} + \begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix}$, where $\mu, \nu \in \mathbb{R}$.

Proof. If $\mathbf{Y} \in \Omega_{\mathbf{T}}^{+}$ then $\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} > 0, \forall x \in \mathbb{R}^2$.

$$(\mathbf{Y} + \begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix})^T B_{\mathbf{T}}(\mathbf{x})(\mathbf{Y} + \begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix}) =$$

$$= \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} + 2\mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x}) \left(\begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right) +$$

$$+ \left(\begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right)^T B_{\mathbf{T}}(\mathbf{x}) \left(\begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} \right) =$$

$$= [\text{from Lemma 5.7 we know that } \begin{bmatrix} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n \end{bmatrix} \begin{bmatrix} \mu \\ \nu \end{bmatrix} \text{ is in the null space of}$$

$$B_{\mathbf{T}}(\mathbf{x}) \text{for all } x \in \mathbb{R}^2] = \mathbf{Y}^T B_{\mathbf{T}}(\mathbf{x})\mathbf{Y} > 0, \ \forall x \in \mathbb{R}^2. \tag{5.36}$$

□

To each bijective configuration there is an entire set of associated configurations, guaranteed also to be bijective. This, in conjunction with the cone properties of $\Omega_{\mathbf{T}}^{+}$, allows for the extension of any convex, bounded quadratic sufficient constraint as in the ensuing theorem.

Theorem 5.4. *If the ellipsoid* $\mathscr{E} = \{y | y^T A y + 2a^T y + c < 0, \ y \in \mathbb{R}^{2n}\}$ *is contained in* $\Omega_{\mathbf{T}}^{+}$ *then so is the set*

$$K = \{y | y^T \tilde{A} y < 0, \ y \in \mathbb{R}^{2n}\} \tag{5.37}$$

where

$$\tilde{A} = \Big((a^T G^{-T} E E^T G^{-1} a + c)(A - GEE^T G^T) -$$

$$- (I - GEE^T G^{-1}) a a^T (I - (GEE^T G^{-1})^T) \Big) \tag{5.38}$$

and

$$E = \begin{bmatrix} \frac{1}{\sqrt{n}} \mathbf{1}_n & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{n}} \mathbf{1}_n \end{bmatrix}.$$

Here G is the upper-triangular matrix from the Cholesky-factorization of $A = GG^T$. The set K is a double cone with the origin removed, it contains $\mathscr{E}$ and is also in $\Omega_{\mathbf{T}}^{+}$.

$$\mathscr{E} \subset K \subset \Omega_{\mathbf{T}}^{+}$$

Proof. From the cone property of $\Omega_{\mathbf{T}}^{+}$ from Lemma 5.4, we know that if $y \in \Omega_{\mathbf{T}}^{+}$ then the entire line $\lambda y, \lambda \in \mathbb{R}$ is also in $\Omega_{\mathbf{T}}^{+}$, except at the origin. Combined with

Lemma 5.13 this means that if $y \in \Omega_{\mathbf{T}}^{+}$ then the linear hull

$$L_y = \left\{ \begin{bmatrix} \mathbf{1}_n & \mathbf{0} & y \\ \mathbf{0} & \mathbf{1}_n & \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \\ \nu \end{bmatrix} \mid \forall\, \lambda, \mu, \nu \in \mathbb{R} \right\}$$

is a subset of $\Omega_{\mathbf{T}}^{+}$.

An open ball S centered at m with radius r can be written

$$S = \{y | (y-m)^T (y-m) < r^2,\ y \in \mathbb{R}^{2n}\}.$$

That is, a point y is in S if its distance to m is less than r.

If $S \subset \Omega_{\mathbf{T}}^{+}$ then $y \in \Omega_{\mathbf{T}}^{+}$ if L_y intersects S, i.e. the distance from m to L_y is less than r. An orthogonal basis for L_y can be written

$$F = \begin{bmatrix} \underbrace{\begin{matrix} \frac{1}{\sqrt{k}}\mathbf{1}_k & \mathbf{0} \\ \mathbf{0} & \frac{1}{\sqrt{k}}\mathbf{1}_k \end{matrix}}_{E} & \tilde{y} \end{bmatrix}.$$

Where

$$\tilde{y} = \frac{(I - EE^T)}{\sqrt{y^T (I - EE^T)^T (I - EE^T) y}} y = \frac{(I - EE^T)}{\sqrt{y^T (I - EE^T) y}} y.$$

The distance $d(m, L_y)$ between m and the hyperplane L_y is the length of the vector v

$$v = m - FF^T m = (I - FF^T)m,$$

thus we obtain

$$\begin{aligned} d(m, L_y)^2 = v^T v = m^T (I - FF^T)^T (I - FF^T) m = \\ = m^T (I - 2FF^T + FF^T FF^T) m = m^T (I - 2FF^T + FF^T) m = \\ = m^T (I - FF^T) m. \end{aligned}$$

The constraint $d^2 < r$ then becomes

$$\begin{aligned} d(m, L_y)^2 = m^T (I - FF^T) m = \\ = m^T \left(I - \begin{bmatrix} E & \frac{(I-EE^T)}{\sqrt{y^T (I-EE^T) y}} y \end{bmatrix} \begin{bmatrix} E^T \\ \left(\frac{(I-EE^T)}{\sqrt{y^T (I-EE^T) y}} y \right)^T \end{bmatrix} \right) m = \end{aligned}$$

$$= m^T \left((I - EE^T) - \frac{(I - EE^T) y y^T (I - EE^T)^T}{y^T (I - EE^T) y} \right) m < r^2.$$

Simplifying

$$\begin{aligned}
&m^T (I - EE^T) m y^T (I - EE^T) y - \\
&\quad - m^T (I - EE^T) y y^T (I - EE^T)^T m < r^2 y^T (I - EE^T) y \\
&y^T ((m^T (I - EE^T) m)(I - EE^T) - \\
&\quad (I - EE^T) m m^T (I - EE^T)^T - r^2 (I - EE^T)) y < 0 \\
&y^T ((m^T (I - EE^T) m - r^2)(I - EE^T) - \\
&\quad - (I - EE^T) m m^T (I - EE^T)^T) y < 0 \\
&y^T \Big((m^T (I - EE^T) m - r^2) I - (I - EE^T) m m^T \Big) \Big(I - EE^T \Big) y < 0 \qquad (5.39)
\end{aligned}$$

Equation (5.39) can then be generalised to handle ellipsoidal constraints on the form

$$\mathscr{E} = \{ y | y^T A y + 2 a^T y + c < 0,\ y \in \mathbb{R}^{2n} \}.$$

Where A is a symmetric and positive definite matrix so it has a Cholesky decomposition $A = GG^T$ as well as an inverse. Using this we can write

$$\begin{aligned}
&y^T A y + 2 a^T y + c = (y + A^{-1} a)^T A (y + A^{-1} a) + \underbrace{(-a^T A^{-1} a + c)}_{\tilde{c}} = \\
&= (G^T y + G^T (GG^T)^{-1} a)^T (G^T y + G^T (GG^T)^{-1} a) + \tilde{c} = \\
&= (\underbrace{G^T y}_{\tilde{y}} + \underbrace{G^{-1} a}_{\tilde{m}})^T (G^T y + G^{-1} a) + \tilde{c} = \\
&= (\tilde{y} + \tilde{m})^T (\tilde{y} + \tilde{m}) + \tilde{c}.
\end{aligned}$$

Inserting this into (5.39) with $r^2 = -\tilde{c}$ gives

$$\begin{aligned}
&\tilde{y}^T \Big((\tilde{m}^T (I - EE^T) \tilde{m} + \tilde{c}) I - (I - EE^T) \tilde{m} \tilde{m}^T \Big) \Big(I - EE^T \Big) \tilde{y} = \\
&= y^T G \Big(\big(a^T G^{-T} (I - EE^T) G^{-1} a - a^T A^{-1} a + c \big) I - \\
&\quad - (I - EE^T) G^{-1} a a^T G^{-T} \Big) \Big(I - EE^T \Big) G^T y = \\
&= y^T \Big((a^T G^{-T} EE^T G^{-1} a + c)(A - GEE^T G^T) - \\
&\quad - (I - GEE^T G^{-1}) a a^T (I - (GEE^T G^{-1})^T) \Big) y < 0. \qquad (5.40)
\end{aligned}$$

□

In the case of the maximum-volume inscribed sphere this results in

Corollary 5.3. *A thin-plate spline mapping* $\phi_{\mathbf{T},\mathbf{Y}} : \mathbb{R}^2 \to \mathbb{R}^2$ *with n point constraints* $\mathbf{T}$, $\mathbf{Y}$ *and* $c = \frac{1}{\sqrt{\max_{\mathbf{x}\in\mathbb{R}^2}(\lambda_M(\mathbf{x}))}}$, *as defined in Theorem 5.2, is bijective if*

$$Y^T\Big((\mathbf{T}^T EE^T\mathbf{T} + c) - (I - EE^T)\mathbf{T}\mathbf{T}^T\Big)(I - EE^T)\mathbf{Y} < 0. \tag{5.41}$$

Proof. Follows trivially from insertion of Theorem 5.2 into (5.37). □

5.5 Two View Registration Using TPS

This section addresses the problem of image registration. It is the process of geometrically aligning two or more images. Image registration has been the subject of extensive research over the last decade, see [16]. This field is widely applied in computer vision, remote sensing and medical imaging.

The approach presented here is based on the thin-plate spline mapping described in the previous section. Using this mapping we wish to find dense and bijective correspondences between pairs of images. The underlying hypothesis is that the image pairs contain similar structures and therefore there should exist mappings between pairs of images that are both one-to-one and onto, i.e. bijective.

The contribution of this section is in addition to highlighting of some additional interesting properties of the thin-plate spline mapping also the incorporation of sufficient quadratic conditions for bijectivity into an image registration framework. A description of how to combine this into a simple but efficient algorithm based on a least squares minimization formulation is also provided.

5.5.1 Thin-Plate Spline Based Image Registration

The registration of two images requires finding the deformation of one image that makes it as similar as possible to the other image. Here, the non-linear deformation used is the thin-plate spline mapping and the similarity function is simply the sum of squared differences in gray-level intensity.

Denote the image to be warped $I(\mathbf{x})$, the reference image $I_{ref}(\mathbf{x})$ and the thin-plate spline mapping by $\phi_{\mathbf{T}}(\mathbf{x}, \mathbf{Y})$. (Remark: We have slightly changed the notation for the thin-plate spline mapping to emphasize that we now see ϕ as a function of the destination configuration $\mathbf{Y}$ as well. These are the variables the similarity measure later will be optimized over). Introducing the finite set $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ of points where the two images are to be compared, typically all the pixel positions of the reference image, the similarity function can then be written

$$f(\mathbf{Y}) = \sum_{i=1}^{N} (r_i(\mathbf{Y}))^2 = \sum_{i=1}^{N} (I(\phi_{\mathbf{T}}(x_i, \mathbf{Y})) - I_{ref}(\mathbf{x}_i))^2. \tag{5.42}$$

Minimizing such a sum of squares is a frequently occurring problem and a number of methods exist that take advantage of its particular structure.

The Gauss–Newton method addresses this in a very simple but appealing manner. This iterative algorithm converges linearly towards the minima if the starting point is sufficiently close. With the Jacobian of $r(\mathbf{Y}) = [r_1(\mathbf{Y}) \dots r_N(\mathbf{Y})]$ defined as the $N \times 2n$ matrix $(J(\mathbf{Y}))_{ij} = (\frac{\partial r_i}{\partial \mathbf{Y}_j})$, the gradient and Hessian of $f(\mathbf{Y})$ in (5.42) can be written

$$\nabla f(\mathbf{Y}) = 2J(\mathbf{Y})^T r_i(\mathbf{Y}), \tag{5.43}$$

$$H(\mathbf{Y}) = J(\mathbf{Y})^T J(\mathbf{Y}) + 2\sum_{i=1}^{N} r_i(\mathbf{Y}) \nabla^2 r_i(\mathbf{Y}). \tag{5.44}$$

In order to avoid having to compute $\nabla^2 r_i(\mathbf{Y})$ in every iteration the second part of (5.44) is assumed small and is simply neglected.

$$H(\mathbf{Y}) \approx \tilde{H}(\mathbf{Y}) = J(\mathbf{Y})^T J(\mathbf{Y}).$$

This also ensures that $H(\mathbf{Y})$ is positive semidefinite. Now by approximating $f(\mathbf{Y})$ by its second-order Taylor expansion near $\mathbf{Y}_k$ we get

$$f(\mathbf{Y}) \approx f(\mathbf{Y}_k) + \nabla f(\mathbf{Y}_k)^T (\mathbf{Y} - \mathbf{Y}_k) + \frac{1}{2}(\mathbf{Y} - \mathbf{Y}_k)^T \tilde{H}(\mathbf{Y}_k)(\mathbf{Y} - \mathbf{Y}_k) = \tilde{f}(\mathbf{Y}). \tag{5.45}$$

The unconstrained minimization of this quadratic approximation of the objective function $\tilde{f}(\mathbf{Y})$ is then, in its original formulation, performed by the normal equation

$$\mathbf{Y}_{k+1} = \mathbf{Y}_k - (J(\mathbf{Y}_k)^T J(\mathbf{Y}_k))^{-1} J(\mathbf{Y}_k) r_i(\mathbf{Y}_k). \tag{5.46}$$

By applying this method iteratively $\mathbf{Y}_k$ should then converge to a local minima of $f(\mathbf{Y})$.

However, since we want to minimize (5.42) over bijective mappings only, a slight alteration of this method is required. From Sect. 5.4 we can obtain convex quadratic sufficient constraints on $\mathbf{Y}$ for bijectivity of the mapping $\phi_{\mathbf{T}}(\mathbf{x}, \mathbf{Y})$ on the form

$$\mathbf{Y}^T A \mathbf{Y} + b^T \mathbf{Y} + c > 0.$$

As the minimization of (5.45) is now no longer unconstrained the final step of the original Gauss–Newton method is replaced by the quadratically constrained,

quadratic program

$$\begin{aligned} \min \quad & \tilde{f}(\mathbf{Y}_k) = f(\mathbf{Y}_k) + \nabla f(\mathbf{Y}_k)^T(\mathbf{Y} - \mathbf{Y}_k) + \\ & + \tfrac{1}{2}(\mathbf{Y} - \mathbf{Y}_k)^T \tilde{H}(\mathbf{Y}_k)(\mathbf{Y} - \mathbf{Y}_k) \\ s.t. \quad & \mathbf{Y}^T A \mathbf{Y} + b^T \mathbf{Y} + c > 0. \end{aligned}$$

The solution $\mathbf{Y}^*$ of this optimization is taken as the next point in the iteration.

Each iteration of this modified Gauss–Newton method requires the computation of $r(\mathbf{Y}) = [r_1(\mathbf{Y}) \dots r_N(\mathbf{Y})]^T$ and $J(\mathbf{Y})$. This can be done very efficiently, using (5.3) the mapping of all points in X can be written

$$\begin{aligned} \begin{bmatrix} \phi_{\mathbf{T}}(\mathbf{x}_1, \mathbf{Y}) \\ \vdots \\ \phi_{\mathbf{T}}(\mathbf{x}_N, \mathbf{Y}) \end{bmatrix} = & \\ = \underbrace{\begin{bmatrix} \begin{bmatrix} \mathbf{s}(\mathbf{x}_1)^T & 1 & x_{11} & x_{12} \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \\ \vdots \\ \begin{bmatrix} \mathbf{s}(\mathbf{x}_n)^T & 1 & x_{(N)1} & x_{(N)2} \end{bmatrix} \begin{bmatrix} \Gamma^{11} \\ \Gamma^{21} \end{bmatrix} \end{bmatrix}}_{H_{\mathbf{T},X}} \mathbf{Y} = & \\ = H_{\mathbf{T},X} \mathbf{Y}. & \end{aligned} \tag{5.47}$$

Since the $N \times 2n$ matrix $H_{\mathbf{T},X}$ is not dependent of $\mathbf{Y}$ it can be precomputed, reducing the computation of the mapping of X by $\phi(\mathbf{Y}_k)$ to a single matrix multiplication. This then allows for an efficient calculation of the deformed image. The Jacobian of $r(\mathbf{Y})$ is also needed.

$$\begin{aligned} (J(\mathbf{Y}))_{ij} = \frac{\partial r_i}{\partial \mathbf{Y}_j} = \frac{\partial}{\partial \mathbf{Y}_j} \left(I(\phi(x_i, \mathbf{Y})) - I_{ref} \right) = & \\ = \frac{\partial}{\partial \mathbf{Y}_j} I(\phi(x_i, \mathbf{Y})) = & \\ = I'_x(\phi(x_i, \mathbf{Y})) \frac{\partial}{\partial \mathbf{Y}_j} \phi_1(x_i, \mathbf{Y}) + I'_y(\phi(x_i, \mathbf{Y})) \frac{\partial}{\partial \mathbf{Y}_j} \phi_2(x_i, \mathbf{Y}). & \end{aligned} \tag{5.48}$$

Where $I'_x(\mathbf{x})$ and $I'_y(\mathbf{x})$ are the horizontal and vertical components of the gradient of $I(\mathbf{x})$. Furthermore, since the mapping $\phi_{\mathbf{T}}(x, \mathbf{Y})$ is linear in $\mathbf{Y}$ its partial derivatives are all constant

$$\phi_{\mathbf{T}}(X, \mathbf{Y}) = [\phi_1(X, \mathbf{Y}) \, \phi_2(X, \mathbf{Y})] = H_{\mathbf{T},X}[\mathbf{Y}_1 \, \mathbf{Y}_2] =$$

$$= [H_{\mathbf{T},X}\mathbf{Y}_1 \; H_{\mathbf{T},X}\mathbf{Y}_2] \Rightarrow$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{Y}_j}\phi_1(x_i, \mathbf{Y}) = \begin{cases} (H_{\mathbf{T},X})_{ij} & n \geq j \geq 1 \\ 0 & j > n \end{cases} \Rightarrow$$

$$\Rightarrow \frac{\partial}{\partial \mathbf{Y}_j}\phi_1(x_i, \mathbf{Y}) = \left(\begin{bmatrix} H_{\mathbf{T},X} \\ 0 \end{bmatrix} \right)_{ij} \tag{5.49}$$

and similarly

$$\frac{\partial}{\partial \mathbf{Y}_j}\phi_2(x_i, \mathbf{Y}) = \left(\begin{bmatrix} 0 \\ H_{\mathbf{T},X} \end{bmatrix} \right)_{ij} \tag{5.50}$$

Equation (5.48) can be computed through componentwise multiplications of elements from $I'_x(\phi(x_i, \mathbf{Y}))$, $I'_y(\phi(x_i, \mathbf{Y}))$ and $H_{\mathbf{T},X}$ Combining all of the above then enables us to write the proposed algorithm as in Algorithm 1.

Algorithm 2 Algorithm for thin-plate spline based image registration

Input: Choose an starting point $\mathbf{Y}_0$ for the algorithm. Either by employing some coarse search method or by simply selecting $\mathbf{Y}_0 = \mathbf{T}$, the unit deformation.

1. For a given thin-plate spline source configuration $\mathbf{T}$ and a pair of images I and I_{ref} to be compared at a finite number of positions $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ compute the following:
2. - The gradient of image I.

$$\nabla I = (\frac{\partial}{\partial x} I, \; \frac{\partial}{\partial y} I) = [I'_x, I'_y].$$

3. - The matrix $H_{\mathbf{T},X}$ from (5.47).
4. - The quadratic bijectivity constraints on $\mathbf{Y}$ for $\mathbf{T}$, according to Sect. 5.4.
5. $k = 0$
6. Repeat steps 7–14 until convergence
7. Compute $\phi^k_{\mathbf{T}}(X, \mathbf{Y}_k) = H_{\mathbf{T},X}\mathbf{Y}_k$
8. Find $I(\phi^k_{\mathbf{T}}(X, \mathbf{Y}_k))$, $I(\phi^k_{\mathbf{T}}(X, \mathbf{Y}_k))$ and $I(\phi^k_{\mathbf{T}}(X, \mathbf{Y}_k))$
9. Calculate the residual $r_i(\mathbf{Y}_k) = I(\phi^k_{\mathbf{T}}(X, \mathbf{Y}_k)) - I_{ref}$
10. Use (5.48) to determine the Jacobian $J(\mathbf{Y}_k)$
11. Compute the gradient and the approximated Hessian of $f(\mathbf{Y})$ of (5.42).

$$\nabla f(\mathbf{Y}_k) = 2J(\mathbf{Y}_k)^T r_i(\mathbf{Y}_k)$$

$$\tilde{H}(\mathbf{Y}) = J(\mathbf{Y}_k)^T J(\mathbf{Y}_k)$$

12. **Optimization.** Find the solution $\mathbf{Y}^*$ to the quadratically constrained quadratic program

$$\min \; \tilde{f}_k(\mathbf{Y}), s.t. \; \mathbf{Y}^T A\mathbf{Y} + b^T\mathbf{Y} + c > 0.$$

(remark: if bijectivity is not desired then $\mathbf{Y}^* = \mathbf{Y}_{k+1}$ of (5.46).)

13. **Parameter update.** Set $\mathbf{Y}_{k+1} = \mathbf{Y}^*$ and $k = k + 1$.
14. $k = k + 1$

5.5.2 *Experimental Results*

We applied the suggested registration algorithm on three different types of images. First, a pair of simple, artificially constructed images. Second, two magnetic resonance images of a human brain, the types of images in medical imaging where image registration techniques are commonly applied. Finally, we attempted the registration of a pair of images of human faces. In this case the initial assumption of dense one-to-one mappings does not necessarily hold, as self-occlusion can easily occur for these types of images. However, bijective registrations of natural objects like faces is still of great interest, for instance in the automatic construction of the Active Appearance Models of [2].

For these experiments a source configuration $\mathbf{T}$ as a regular rectangular 10×10 grid was used, see Fig. 5.7. The quadratic constraint was pre-computed and used in all three instances. The images used were roughly 100×100 pixels in size. The results can be seen in Figs. 5.8, 5.9 and 5.10.

In these three experiments the algorithm converges to at least a satisfactory registration of the image pairs. The artificial images are overlayed very accurately, as would be expected. The images of the faces were also successfully registered, differences are slight but distinguishable. It is believed that this is caused by fundamental dissimilarities between the images, such as inconsistent lighting. However, in the case of the two magnetic resonance images of a human brain the registration process is not entirely successful. Some of the discernable features does not seem to have been correctly aligned. It is assumed that this is caused by a shortcoming inherent in the algorithm. Firstly, and this was briefly mentioned earlier, some of the assumptions made by the Gauss-Newton method, on which our approach is based, requires that the initial starting point of the algorithm is sufficiently close to the global optima. What constitutes sufficiently close is

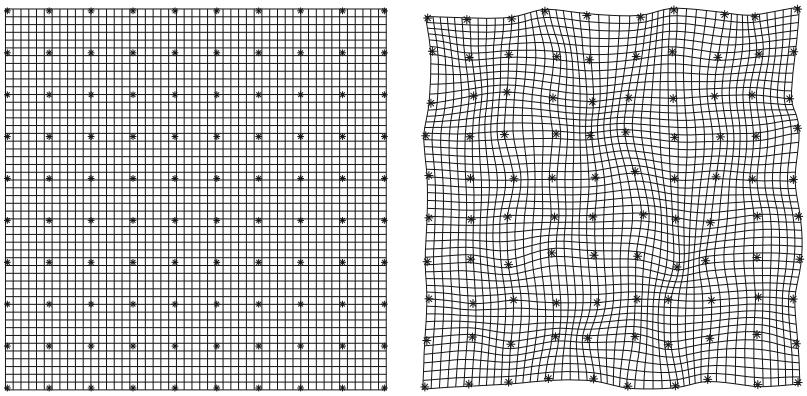


Fig. 5.7 The source configuration used in the experiments and an example bijective deformation of $\mathbb{R}^2$

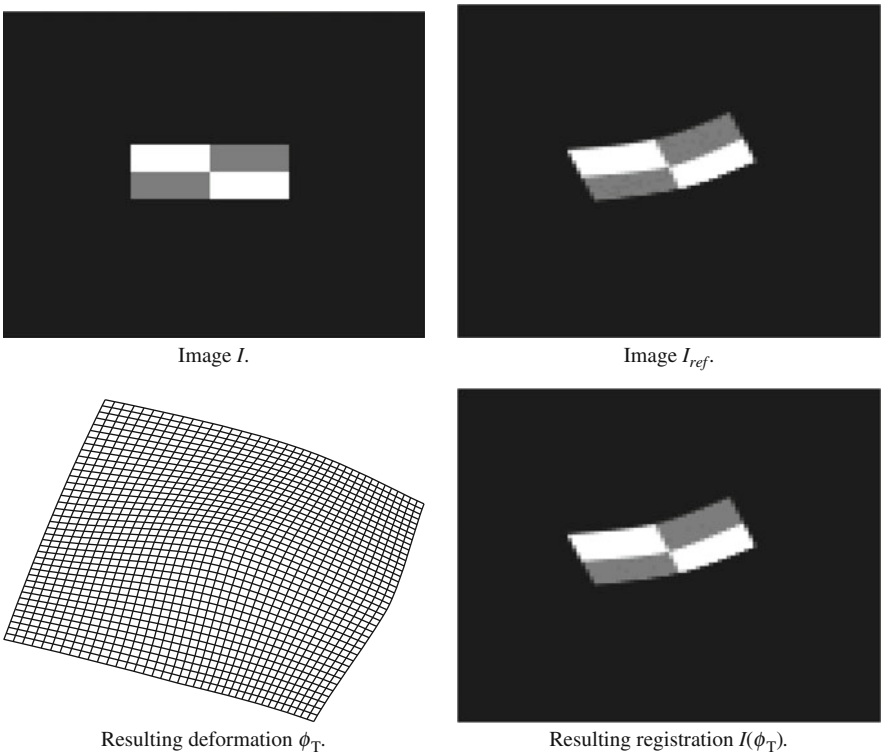


Fig. 5.8 Registration of a pair of simple artificial images

debatable but is a requirement for the method to converge to successfully. Secondly, a 10×10 grid thin-plate spline mapping can only parametrize a subset of all bijective deformations of $\mathbb{R}^2$ and in addition, since the bijectivity conditions of Sect. 5.4 are sufficient but not necessary, we can only reach a subset of this set. This means that our method is perhaps better suited for image registrations requiring smaller deformations. Nevertheless, we do believe that the results presented here the still indicates the applicability of such an algorithm.

5.6 AAAM

This section is concerned with groupwise image registration, the simultaneous alignment of a large number of images. As opposed to pairwise registration the choice of a reference image is not equally obvious, therefore an alternate approach must be taken.

Groupwise registration has received equivalent amounts of attention from the research community as pairwise registration. It has been especially addressed in

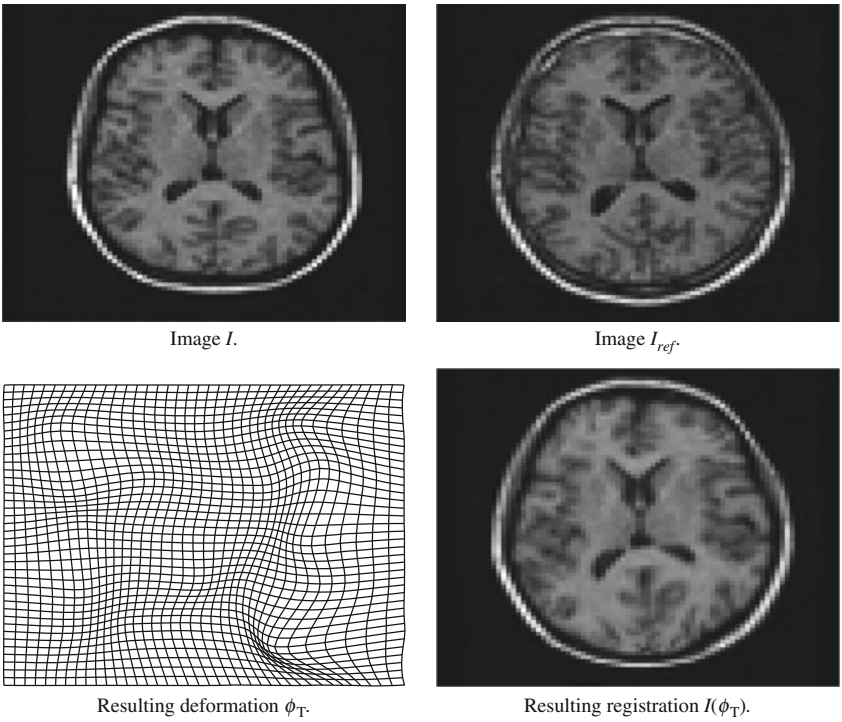


Fig. 5.9 Registration of a pair of brain MR images

shape analysis under the name Procrustes analysis, [5]. The areas of application are still remote sensing, medical imaging and computer vision, but now the aggregation of images allows for a greater understanding of their underlying distribution.

The focus in this section is towards a specific task, the use of image registration to automatically construct deformable models for image analysis.

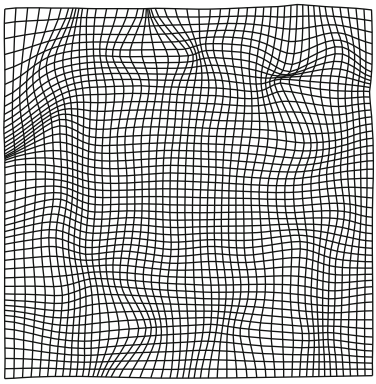
5.6.1 Automatic Active Appearance Model Generation Through Groupwise Image Registration

The outset in this section, that of automatic model construction, is approached by attempting to extend the algorithm of the previous section to handle several images. The method chosen for representing deformable models was the widely used Active Appearance Model approach.

Owing to the resemblance between registration of shapes and of images, as formulated here, many of the issues encountered in this section have been considered by the shape analysis community [3] and a number of the ideas presented here are influenced by existing shape matching techniques.

Image I.

Image I_{ref}.



Resulting deformation ϕ_{T}.

Resulting registration $I(\phi_{\mathrm{T}})$.

Fig. 5.10 Registration of a pair of images of faces

5.6.2 Active Appearance Models

Active Appearance Models (AAM) is a statistical method, introduced by Cootes et al. [2], for interpreting images. From the shape and texture of objects of interest in a number of images, compact models based on the distribution of these features are formed.

The texture, or appearance of the objects are the gray-level image intensities and their shape are usually represented by a finite number of points of correspondence through the entire set of images.

Then, using principal component analysis, eigenspace representations of these two descriptors are extracted. Depending on the application, the shape parameters are generally pre-aligned to eliminate effects from translation, scaling and rotation. By applying yet another principal component analysis, this time to the shape- and appearance parameters combined, an even more concise model describing the joint variability of the objects of interest is achieved. The resulting active appearance

model is a compact description of a deformable model based on prior knowledge of the essential characteristics of the object at hand. Through image synthesis, that is by fitting an AAM to unseen images, this approach can be used in a wide variety of image analysis applications.

There is however one disadvantage to this method. The required correspondence calls for manual annotation of landmarks across the entire set of training images. A both tedious and exhausting undertaking. Here an alternative approach is suggested, the automatic generation of Active appearance Models through groupwise image registration.

5.6.3 Groupwise Image Registration

Consider a set of N images $I_1, \ldots, I_N$, a groupwise registration of this set implies finding deformations $\theta_1, \ldots, \theta_N$, $\theta_l : \mathbb{R}^2 \to \mathbb{R}^2$ that maximizes the similarity between the corresponding deformed images $I_1(\theta_1), \ldots, I_N(\theta_N)$. Since registration is carried out with Active Appearance Models in mind, similarity is defined as to what degree an eigenspace method can represent the registered images. Using the squared distance to the eigenspace as a measurement of how well one image is represented by such a statistical model, the total dissimilarity between images $I_1(\theta_1), \ldots, I_N(\theta_N)$ can be written

$$S(\theta_1, \ldots, \theta_N) = \sum_{l=1}^{N} (\text{distance between image } I_l(\theta_l) \text{ and } E)^2 = \\ = \sum_{l=1}^{N} ||(I - EE^T)\hat{I}_l(\theta_l)||^2. \tag{5.51}$$

Here E is the M-dimensional orthogonal basis for a conventional eigenspace representation. The columns of E are the eigenvectors corresponding to the M largest eigenvalues of the covariance matrix of the statistical distribution of the image vectors. As in the preceding section, comparison is made at a finite number of locations in $\mathbb{R}^2$, the set of such locations is written as $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$. The notation $I_l(\mathbf{x})$ is used to represent images both on matrix- and vector-form, the intended form should be evident from the context.

Though this formulation of groupwise registration has the advantage of simplicity, it is actually ill-posed. A global optima of (5.51) is achieved by mapping X bijectively onto one and the same pixel in each image. As this results in zero covariance between the deformed images $S(\theta_1, \ldots, \theta_N)$ will be equal to zero. This is also an issue in shape analysis and has been identified and addressed by [4, 9, 13]. Here it is simply ignored, the assumption is that if the initial starting point of the algorithm is sufficiently good the degenerate solution will not be attained but instead the optimizer used will terminate in the desired local optima. This vagueness stems

from the underlying problem itself. What constitutes as similar objects in images is highly subjective. Hence, the formulation of a method for automatically finding and aligning areas with similar appearance in a number of images will be equally ambiguous.

With the given problem statement we can move on to the proposed method for finding local minima to (5.51). A direct optimizing of this objective function is impractical as this would involve a very large number of variables, N times the number of parameters needed to describe each deformation θ_l. Instead an iterative approach is proposed, by sequentially attending to each image individually, the number of variables in each optimization step can be greatly reduced. That is the repeated minimization of functions

$$S_l = ||(I - EE^T)I_l(\theta_l(X))||^2 = \sum_{j=1}^{p} \Big((I - EE^T)I_l(\theta_l(X))\Big)_j^2. \tag{5.52}$$

Using the thin-plate spline mapping to represent the mappings by, with the notation of the previous section, $\theta_l = \phi_{\mathbf{T}}(\mathbf{x}, \mathbf{Y}_l)$ along with the sum of squares formulation of (5.52) allows for much of the algorithm of the previous section to be adopted in groupwise image registration. The assertions made regarding bijective deformation in pairwise image registration are still valid and are hence also applied here. The residual for image l becomes

$$r_l(\mathbf{Y}_l) = (I - EE^T)I_l(\phi(\mathbf{x}, \mathbf{Y}_l)). \tag{5.53}$$

and the corresponding Jacobian

$$\begin{aligned}(J_l(\mathbf{Y}_l))_{ij} &= \frac{\partial r_{li}}{\partial \mathbf{Y}_{lj}} = \frac{\partial}{\partial \mathbf{Y}_{lj}} \Big((I - EE^T)I_l(\phi(x_i, \mathbf{Y}_l))\Big)_i = \\ &= \left((I - EE^T)\frac{\partial}{\partial \mathbf{Y}_{lj}} I_l(\phi(x_i, \mathbf{Y}_l))\right)_i \Rightarrow \\ J_l(\mathbf{Y}_l) &= (I - EE^T)\tilde{J}_l(\mathbf{Y}_l).\end{aligned} \tag{5.54}$$

Here

$$\begin{aligned}(\tilde{J}_l(\mathbf{Y}_l))_{ij} &= \frac{\partial}{\partial \mathbf{Y}_{lj}} I_l(\phi(\mathbf{x}_i, \mathbf{Y}_l)) = \\ &= I'_{lx}(\phi(\mathbf{x}_i, \mathbf{Y}_l))\frac{\partial}{\partial \mathbf{Y}_{lj}}\phi_1(\mathbf{x}_i, \mathbf{Y}_l) + I'_{ly}(\phi(\mathbf{x}_i, \mathbf{Y}_l))\frac{\partial}{\partial \mathbf{Y}_{lj}}\phi_2(\mathbf{x}_i, \mathbf{Y}_l).\end{aligned} \tag{5.55}$$

Where $I'_{lx}(\mathbf{x})$ and $I'_{lx}(\mathbf{x})$ as the gradient of $I_l(\mathbf{x})$ and $\frac{\partial \phi_1}{\partial \mathbf{Y}_{lj}}$ and $\frac{\partial \phi_2}{\partial \mathbf{Y}_{lj}}$ defined as in (5.49–5.50).

By adhering to the least square formulation used in the previous section, the algorithm for pairwise image registration can be readily extended to handle groupwise registration as defined here. Neither does this extension make the required computations significantly more demanding, resulting in an algorithm of comparable computational complexity per iteration, see Algorithm 2.

Algorithm 3 Algorithm for thin-plate spline based groupwise image registration

Input: Choose starting points $\mathbf{Y}_1^0, \ldots, \mathbf{Y}_N^0$ for the algorithm. Compute the initial eigenspace representation E^0 by finding the eigenvectors corresponding to the M largest eigenvalues of the covariance matrix of
$\left[I_1(\phi_{\mathbf{T}}(X, \mathbf{Y}_1^0)) \ \ldots \ I_N(\phi_{\mathbf{T}}(X, \mathbf{Y}_N^0)) \right]$.

1. For a given thin-plate spline source configuration $\mathbf{T}$ and N images $I_1, \ldots, I_N$ to be compared at a finite number of positions $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ compute the following:
2. - The gradient of all images.

$$\nabla I_l = \left(\frac{\partial}{\partial x} I_l, \ \frac{\partial}{\partial y} I_l\right) = [I'_{lx}, I'_{ly}].$$

3. - The matrix $H_{\mathbf{T},X}$ from (5.47) of the previous section.
4. - The quadratic bijectivity constraints on $\mathbf{Y}$ for $\mathbf{T}$, according to Sect. 5.4.
5. (Note that both $H_{\mathbf{T},X}$ and the bijectivity conditions are independent of which image they are applied to.
6. Set $k = 0$
7. Repeat steps 8–16 until convergence
8. **for** each image l from 1 to N **do** steps 9 to 14
9. Compute $\phi_{\mathbf{T}}(X, \mathbf{Y}_l^k) = H_{\mathbf{T},X} \mathbf{Y}_l^k$
10. Find $I_l(\phi_{\mathbf{T}}(X, \mathbf{Y}_l^k))$, $I'_{lx}(\phi_{\mathbf{T}}(X, \mathbf{Y}_l^k))$ and $I'_{ly}(\phi_{\mathbf{T}}(X, \mathbf{Y}_l^k))$
11. Calculate the residual $r_l(\mathbf{Y}_l^k) = (I - E^0(E^0)^T) I_l(\phi_{\mathbf{T}}(X, \mathbf{Y}_l^k))$
12. Use (5.54) to determine the Jacobian $J_l(\mathbf{Y}_k)$
13. Compute the gradient and the approximated Hessian of $S_l(\mathbf{Y}_l^k)$ of (5.52)

$$\nabla S_l(\mathbf{Y}_l^k) = 2 J_l(\mathbf{Y}_l^k)^T r_l(\mathbf{Y}_l^k)$$

$$\tilde{H}_l(\mathbf{Y}_l^k) = J_l(\mathbf{Y}_l^k)^T J_l(\mathbf{Y}_l^k)$$

14. **Optimization.** Find the solution $\mathbf{Y}^*$ to the quadratically constrained quadratic program

$$\begin{aligned} \min \quad & S_l(\mathbf{Y}_l^k) + \nabla S_l(\mathbf{Y}_l^k)^T (\mathbf{Y} - \mathbf{Y}_l^k) + \\ & + \tfrac{1}{2} (\mathbf{Y} - \mathbf{Y}_l^k)^T \tilde{H}_l(\mathbf{Y}_l^k)(\mathbf{Y} - \mathbf{Y}_l^k) \\ s.t. \quad & \mathbf{Y}^T A \mathbf{Y} + b^T \mathbf{Y} + c > 0 \end{aligned}$$

Parameter update. Set $\mathbf{Y}_l^{k+1} = \mathbf{Y}^*$

15. end of for loop
16. $k = k + 1$. **Update the eigenspace representation.** Compute E^k from the covariance matrix of $\left[I_1(\phi_{\mathbf{T}}(X, \mathbf{Y}_1^k)) \ \ldots \ I_N(\phi_{\mathbf{T}}(X, \mathbf{Y}_N^k)) \right]$
17. end of repeat loop

Groupwise image registration as used in this section is an off-line process, hence even simple termination criteria will suffice.

5.6.4 Experimental Results

The proposed algorithm was tested on a set consisting of 400 portrait-style images of male faces, see Fig. 5.11. A thin-plate spline mapping with 100 control points, evenly spaced on a regular square 10-by-10 grid, was used. As the faces were fairly centered in the images, the initial deformations $\mathbf{Y}_1^0, \ldots, \mathbf{Y}_{400}^0$ were all set to the identity mapping, centered near the middle of the images, see Fig. 5.12.

The dimension of the eigenspace representation was set to $M = 30$. A set of 1,600 points on a 40×40 grid were used as the set of locations for comparison X.

The algorithm described in the preceding section was applied to the set of images at hand, with the above parameters. A termination criterion simply limiting the number of iterations to 200 was used. The proposed method did converge and sample results can be seen in Fig. 5.13.

These results are representative of the entire resulting groupwise registration and do indicate the potential of the proposed approach. This can be further realized by examining the evolution of the mean of the registered images after each iteration, $I_{mean}^k = \sum_{l=1}^{M} I_l(\phi_{\mathbf{T}}(X^k, \mathbf{Y}_l))$, see Fig. 5.14. Here the increased degree of geometric alignment is clearly seen. To quantify the performance of this algorithm further is difficult, since, as discussed earlier, what is meant by similarity within a set of images is unclear so is the evaluation of groupwise image registration algorithms.

Fig. 5.11 A sample of the dataset used

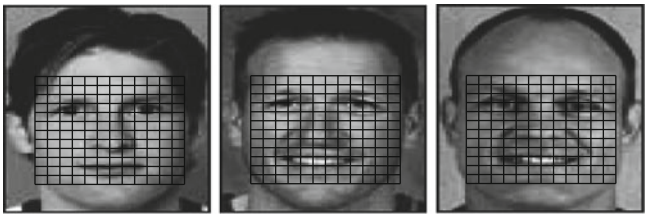


Fig. 5.12 Examples of the initial deformations $\mathbf{Y}_l^0$

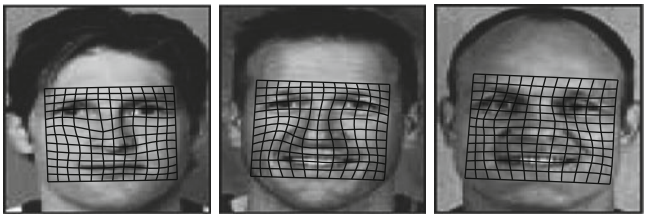


Fig. 5.13 Resulting registration for the sample images

Fig. 5.14 The evolution of the mean image I^k_{mean}, $(k = 1, 10, \ldots, 200)$

Nevertheless, as the outset was the automatic construction of active appearance models, an indication of the quality of the resulting registration could be achieved by examining the performance of the models they produce.

Constructing active appearance models using the proposed approach is extremely straightforward. The required distributions of shape and appearance are given directly by the parameters of the thin-plate spline mappings $\mathbf{Y}_l$ and the deformed images $I_l(\phi_{\mathbf{T}}(X, \mathbf{Y}_l))$.

Using the 400 aligned images an active appearance-like model was constructed. In contrast to [2], here the shape and appearance representations were kept separated in order to be able to ensure bijective deformations in the fitting process as well. With F and E as the eigenspace basis for shape and appearance respectively. The

deformation parameters $\mathbf{Y}$ for an individual mapping can be written as

$$\mathbf{Y} = Fy. \tag{5.56}$$

Since this constitutes a subset of $\Omega_{\mathbf{T}}^{+}$, new and hopefully improved bijectivity conditions ($\tilde{A}$, $\tilde{b}$ and $\tilde{c}$) can be computed. Using the notation from the definition of the registration algorithm the fitting of an active appearance model onto an image $I(\mathbf{x})$ is formulated as minimizing

$$S(y) = ||(I - EE^T)I(\phi_{\mathbf{T}}(X, Fy))||^2 = \sum_{j=1}^{p} \Big((I - EE^T)I(\phi_{\mathbf{T}}(X, Fy)) \Big)_j^2. \tag{5.57}$$

under the condition of bijective deformations. This is solved by the repeated solution of

$$\begin{aligned} \min \quad & S(y^k) + \nabla S(y^k)^T (y - y^k) + \\ & + \tfrac{1}{2}(y - y^k)^T \tilde{H}(y^k)(y - y^k) \\ s.t. \quad & y^T \tilde{A} y + \tilde{b}^T y + \tilde{c} > 0 \end{aligned}$$

An example model-fitting procedure on an image not present in the set of registered images is shown in Fig. 5.15. Further examples of model adaptations are shown in Figs. 5.16 and 5.17. These images should be read as follows. The top left images shows the original image with the boundary of the deformed points superimposed. The resulting deformation can be seen at the top right image. The middle row shows, to the left the deformed image $I(\phi_{\mathbf{T}}(X, Fy))$ and to the right its eigenspace representation $EE^T I(\phi_{\mathbf{T}}(X, Fy))$. At the bottom left is the image $I(\phi_{\mathbf{T}}^{-1}(\phi_{\mathbf{T}}(X, Fy), Fy))$, this adds the same interpolation errors introduced in the fitting procedure to the original image as well. This makes the evaluation of the quality of the resulting model fit more unprejudiced. Finally, the bottom right shows the fitted active appearance model overlayed on the original image.

5.7 Conclusion

Even though this work does not provide a complete theory on the set of bijective thin-plate spline mappings, it does contain a formulation of how to characterize this set, as well as proofs of many of its properties. It also includes a discussion of some experimentally derived indications of other attributes of this set, as well as methods for finding sufficient conditions for bijectivity. Future work includes finding such conditions analytically as well as attempting to further determine its convexity and boundness properties.

Starting image.

Fitting sequence.

Resulting fit.

Fig. 5.15 An example AAM-fitting. The current model superimposed onto the original after number of different iterations of the proposed fitting algorithm

A method for performing pairwise registration of images has also been presented. An algorithm, based on the thin-plate spline mapping, for efficiently finding the necessary deformation is proposed. Experiments on three different types of images with promising results were presented.

Improvements are still achievable. In order to overcome the drawback of the Gauss–Newton method an initial stage to the algorithm should be added. One that performs a larger-scale optimization, for instance over affine deformations only, providing a better starting point for the thin-plate spline mapping optimization. The number and distribution of the control points should also be investigated.

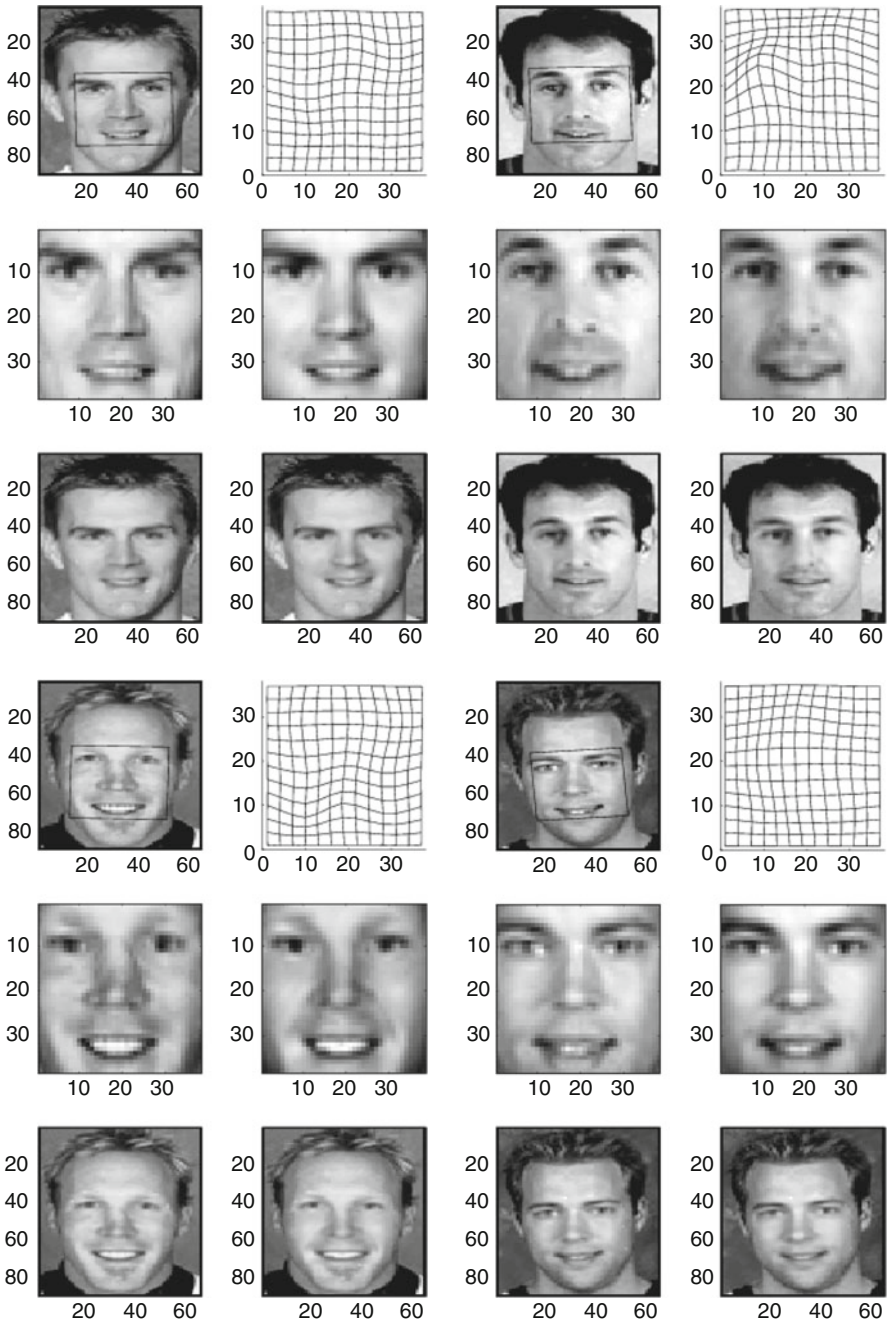


Fig. 5.16 Example AAM fittings

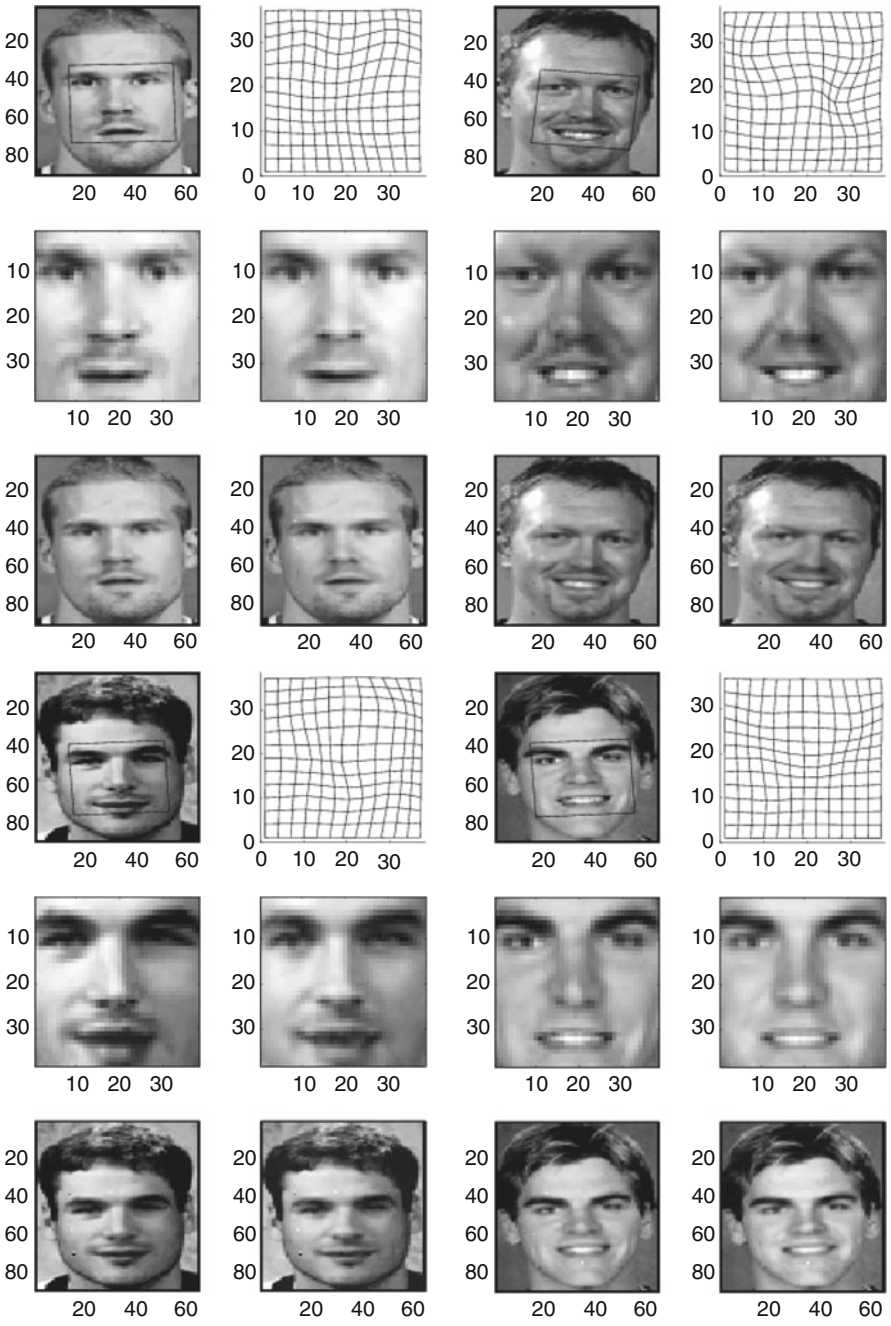


Fig. 5.17 Example AAM fittings

More points parametrizes a larger subset of the bijective deformations. Obviously improving the sufficient bijectivity constraints could also enhance the performance of the algorithm. A different objective function than (5.42) might also improve on our method. As the current similarity function, based directly on gray-level intensity, can be overly sensitive to noise, using, for instance, the minimum description length method of [4] might be preferable. Finally, a more efficient representation of the matrix $H_{\mathbf{T},X}$ should be examined, as its size grows quadratically with the size of the image even for moderately large images the matrix can become unmanageable.

Additionally, a method for carrying out non-linear geometric alignment of a large number images, especially geared towards the automatic generation of Active Appearance Models, has been proposed. By adhering to the sum of squares formulation of the previous section much of the techniques used there could effortlessly be extended to groupwise image registration. The suggested algorithm was tested on a data set of faces and the results were presented. As the nature of the problem is such that the evaluation of its performance is highly subjective, in addition to its ill-posed problem statement. These issues should be addressed by adopting ideas from shape analysis, where similar topics have been investigated. Nevertheless, as the initial results are convincing the presented approach does show promise.

References

1. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Proceedings of the 5th European Conference on Computer Vision. Freiburg, Germany (1998)
3. Cootes, T.: Statistical models of shape and appearance. Technical Report, Imaging Science and Biomedical Engineering (2004)
4. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.: A minimum description length approach to statistical shape modeling. IEEE Trans. Med. Imaging **21**(5), 525–537 (2002)
5. Dryden, I.L., Mardia, K.V.: Statistical Shape Analysis. Wiley, New York (1998)
6. Duchon, J.: Splines minimizing rotation-invariant semi-norms in sobolev spaces. In: Schempp, W., Zeller, K. (eds.) Constructive theory of functions of several variables, pp. 85–100. Springer, Berlin (1977)
7. Green, P.J., Silverman, B.W.: Nonparametric regression and generalized linear models. Number 58 in Monographs on Statistics and Applied Probability. Chapman & Hall, London (1994)
8. Hettich, R., Kortanek, K.O.: Semi-infinite programming: Theory, methods, and applications. SIAM Rev. **35**(3), 380–429 (1993)
9. Ericsson, A., Karlsson, J., Åström, K.: Parameterisation invariant statistical shape models. In: Proceedings of the International Conference on Pattern Recognition. Cambridge, UK (2004)
10. Boyd, S., Vandenberghe, L., Wu, S.W.: Determinant maximization with linear matrix inequality constraints. SIAM J. Matrix Anal. Appl. **19**(2), 499–533 (1998)
11. Meinguet, J.: Multivariate interpolation at arbitrary points made simple. J. Appl. Math. Phys. **30**, 292–304 (1979)
12. Feron, E., Boyd, S., El Ghaoui, L., Balakrishnan, V.: Linear matrix inequalities in system and control theory. Society for Industrial and Applied Mathematics, Philadelphia (1994)

13. Thodberg, H.H.: Minimum description length shape and appearance models. In: Image processing medical imaging, IPMI 2003 (2003)
14. Vandenberghe, L., Boyd, S.: Convex Optimization. Cambridge University Press, Cambridge (2004)
15. Wahba, G.: Spline models for observational data. Society for Industrial and Applied Mathematics, Philadelphia (1990)
16. Zitova, B., Flusser, J.: Image registration methods: a survey. Image Vis. Comput. **21**(11), 977–1000 (2003)

Chapter 6
Statistical and Knowledge Supported Visualization of Multivariate Data

Magnus Fontes

Abstract In the present work we have selected a collection of statistical and mathematical tools useful for the exploration of multivariate data and we present them in a form that is meant to be particularly accessible to a classically trained mathematician. We give self contained and streamlined introductions to principal component analysis, multidimensional scaling and statistical hypothesis testing. Within the presented mathematical framework we then propose a general exploratory methodology for the investigation of real world high dimensional datasets that builds on statistical and knowledge supported visualizations. We exemplify the proposed methodology by applying it to several different genomewide DNA-microarray datasets. The exploratory methodology should be seen as an embryo that can be expanded and developed in many directions. As an example we point out some recent promising advances in the theory for random matrices that, if further developed, potentially could provide practically useful and theoretically well founded estimations of information content in dimension reducing visualizations. We hope that the present work can serve as an introduction to, and help to stimulate more research within, the interesting and rapidly expanding field of data exploration.

6.1 Introduction

In the scientific exploration of some real world phenomena a lack of detailed knowledge about governing first principles makes it hard to construct well-founded mathematical models for describing and understanding observations. In order to gain some preliminary understanding of involved mechanisms and to be able to make some reasonable predictions we then often have to recur to purely statistical models. Sometimes though, a stand alone and very general statistical approach fails

M. Fontes (✉)
Centre for Mathematical Sciences, Lund University, Box 118, SE-22100, Lund, Sweden
e-mail: fontes@maths.lth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_6,

to exploit the full exploratory potential for a given dataset. In particular a general statistical model a priori often does not incorporate all the accumulated field-specific expert knowledge that might exist concerning a dataset under consideration. In the present work we argue for the use of a set of statistical and knowledge supported visualizations as the backbone of the exploration of high dimensional multivariate datasets that are otherwise hard to model and analyze. The exploratory methodology we propose is generic but we exemplify it by applying it to several different datasets coming from the field of molecular biology. Our choice of example application field is in principle anyhow only meant to be reflected in the list of references where we have consciously striven to give references that should be particularly useful and relevant for researchers interested in bioinformatics. The generic case we have in mind is that we are given a set of observations of several different variables that presumably have some interrelations that we want to uncover. There exist many rich real world sources giving rise to interesting examples of such datasets within the fields of e.g. finance, astronomy, meteorology or life science and the reader should without difficulty be able to pick a favorite example to bear in mind.

We will use separate but synchronized Principle Component Analysis (PCA) plots of both variables and samples to visualize datasets. The use of separate but synchronized PCA-biplots that we argue for is not standard and we claim that it is particularly advantageous, compared to using traditional PCA-biplots, when the datasets under investigation are high dimensionsional. A traditional PCA-biplot depicts both the variables and the samples in the same plot and if the dataset under consideration is high dimensional such a joint variable/sample plot can easily become overloaded and hard to interpret. In the present work we give a presentation of the linear algebra of PCA accentuating the natural inherent duality of the underlying singular value decomposition. In addition we point out how the basic algorithms easily can be adapted to produce nonlinear versions of PCA, so called multidimensional scaling, and we illustrate how these different versions of PCA can reveal relevant structure in high dimensional and complex real world datasets. Whether an observed structure is relevant or not will be judged by knowledge supported and statistical evaluations.

Many present day datasets, coming from the application fields mentioned above, share the statistically challenging peculiarity that the number of measured variables (p) can be very large ($10^4 \leq p \leq 10^{10}$), while at the same time the number of observations (N) sometimes can be considerably smaller ($10^1 \leq N \leq 10^3$). In fact all our example datasets will share this so called "large p small N" characteristic and our exploratory scheme, in particular the statistical evaluation, is well adapted to cover also this situation. In traditional statistics one usually is presented with the reverse situation, i.e. "large N small p", and if one tries to apply traditional statistical methods to "large p small N" datasets one sometimes runs into difficulties. To begin with, in the applications we have in mind here, the underlying probability distributions are often unknown and then, if the number of observations is relatively small, they are consequently hard to estimate. This makes robustness of employed statistical methods a key issue. Even in cases when we assume that we know the underlying probability distributions or when we use very robust statistical

methods the "large p small N" case presents difficulties. One focus of statistical research during the last few decades has in fact been driven by these "large p small N" datasets and the possibility for fast implementations of statistical and mathematical algorithms. An important example of these new trends in statistics is multiple hypothesis testing on a huge number of variables. High dimensional multiple hypothesis testing has stimulated the creation of new statistical tools such as the replacement of the standard concept of p-value in hypothesis testing with the corresponding q-value connected with the notion of false discovery rate, see [12, 13, 47, 48] for the seminal ideas. As a remark we point out that multivariate statistical analogues of classical univariate statistical tests sometimes can perform better in multiple hypothesis testing, but then a relatively small number of samples normally makes it necessary to first reduce the dimensionality of the data, for instance by using PCA, in order to be able to apply the multivariate tests, see e.g. [14, 31, 32] for ideas in this direction. In the present work we give an overview and an introduction to the above mentioned statistical notions.

The present work is in general meant to be one introduction to, and help to stimulate more research within, the field of data exploration. We also hope to convince the reader that statistical and knowledge supported visualization already is a versatile and powerful tool for the exploration of high dimensional real world datasets. Finally, "Knowledge supported" should here be interpreted as "any use of some extra information concerning a given dataset that the researcher might possess, have access to or gain during the exploration" when analyzing the visualization. We illustrate this knowledge supported approach by using knowledge based annotations coming with our example datasets. We also briefly comment on how to use information collected from available databases to evaluate or preselect groups of significant variables, see e.g. [11, 16, 30, 41] for some more far reaching suggestions in this direction.

6.2 Singular Value Decomposition and Principal Component Analysis

Singular value decomposition (SVD) was discovered independently by several mathematicians towards the end of the nineteenth century. See [46] for an account of the early history of SVD. Principal component analysis (PCA) for data analysis was then introduced by Pearson [24] in 1901 and independently later developed by Hotelling [23]. The central idea in classical PCA is to use an SVD on the column averaged sample matrix to reduce the dimensionality in the data set while retaining as much variance as possible. PCA is also closely related to the Karhunen–Loève expansion (KLE) of a stochastic process [28, 33]. The KLE of a given centered stochastic process is an orthonormal L^2-expansion of the process with coefficients that are uncorrelated random variables. PCA corresponds to the empirical or sample version of the KLE, i.e. when the expansion is inferred from samples. Noteworthy here is the Karhunen–Loève theorem stating that if the underlying process is

Gaussian, then the coefficients in the KLE will be independent and normally distributed. This is e.g. the basis for showing results concerning the optimality of KLE for filtering out Gaussian white noise.

PCA was proposed as a method to analyze genomewide expression data by Alter et al. [1] and has since then become a standard tool in the field. Supervised PCA was suggested by Bair et al. as a regression and prediction method for genomewide data [8, 9, 17]. Supervised PCA is similar to normal PCA, the only difference being that the researcher preconditions the data by using some kind of external information. This external information can come from e.g. a regression analysis with respect to some response variable or from some knowledge based considerations. We will here give an introduction to SVD and PCA that focus on visualization and the notion of using separate but synchronized biplots, i.e. plots of both samples and variables. Biplots displaying samples and variables in the same usually twodimensional diagram have been used frequently in many different types of applications, see e.g. [15, 20–22] but the use of separate but synchronized biplots that we present is not standard. We finally describe the method of multidimensional scaling which builds on standard PCA, but we start by describing SVD for linear operators between finite dimensional euclidean spaces with a special focus on duality.

6.2.1 Dual Singular Value Decomposition

Singular value decomposition is a decomposition of a linear mapping between euclidean spaces. We will discuss the finite dimensional case and we consider a given linear mapping $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$.

Let $\mathbf{e_1}, \mathbf{e_2}, \ldots, \mathbf{e_N}$ be the canonical basis in $\mathbf{R}^N$ and let $\mathbf{f_1}, \mathbf{f_2}, \ldots, \mathbf{f_p}$ be the canonical basis in $\mathbf{R}^p$. We regard $\mathbf{R}^N$ and $\mathbf{R}^p$ as euclidean spaces equipped with their respective canonical scalar products, $(\cdot, \cdot)_{\mathbf{R}^N}$ and $(\cdot, \cdot)_{\mathbf{R}^p}$, in which the canonical bases are orthonormal.

Let $L^* : \mathbf{R^p} \longrightarrow \mathbf{R}^N$ denote the adjoint operator of L defined by

$$(L(\mathbf{u}), \mathbf{v})_{\mathbf{R}^p} = (\mathbf{u}, L^*(\mathbf{v}))_{\mathbf{R}^N} \quad ; \mathbf{u} \in \mathbf{R}^N \; ; \; \mathbf{v} \in \mathbf{R}^p \, . \tag{6.1}$$

Observe that in applications $L(\mathbf{e_k})$, $k = 1, 2, \ldots, N$, normally represent the arrays of observed variable values for the different samples and that $L^*(\mathbf{f_j})$, $j = 1, 2, \ldots, p$, then represent the observed values of the variables. In our example data sets, the unique $p \times N$ matrix X representing L in the canonical bases, i.e.

$$X_{jk} = (\mathbf{f_j}, L(\mathbf{e}_k))_{\mathbf{R^p}} \quad ; j = 1, 2, \ldots, p \; ; \; k = 1, 2, \ldots, N \, ,$$

contains measurements for all variables in all samples. The transposed $N \times p$ matrix X^T contains the same information and represents the linear mapping L^* in the canonical bases.

The goal of a dual SVD is to find orthonormal bases in $\mathbf{R}^N$ and $\mathbf{R}^p$ such that the matrices representing the linear operators L and L^* have particularly simple forms.

We start by noting that directly from (6.1) we get the following direct sum decompositions into orthogonal subspaces

$$\mathbf{R}^N = Ker\, L \oplus Im\, L^*$$

(where $Ker\, L$ denotes the kernel of L and $Im\, L^*$ denotes the image of L^*) and

$$\mathbf{R}^p = Im\, L \oplus Ker\, L^* .$$

We will now make a further dual decomposition of $Im\, L$ and $Im\, L^*$.

Let r denote the rank of $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$, i.e. $r = dim\,(Im\, L) = dim\,(Im\, L^*)$. The rank of the positive and selfadjoint operator $L^* \circ L : \mathbf{R}^N \longrightarrow \mathbf{R}^N$ is then also equal to r, and by the spectral theorem there exist values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and corresponding orthonormal vectors $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}$, with $\mathbf{u^k} \in \mathbf{R}^N$, such that

$$L^* \circ L(\mathbf{u^k}) = \lambda_k^2 \mathbf{u^k} \quad ; \; k = 1, 2, \ldots, r\,. \tag{6.2}$$

If $r < N$, i.e. $dim\,(Ker\, L) > 0$, then zero is also an eigenvalue for $L^* \circ L : \mathbf{R}^N \longrightarrow \mathbf{R}^N$ with multiplicity $N - r$.

Using the orthonormal set of eigenvectors $\{\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}\}$ for $L^* \circ L$ spanning $Im\, L^*$, we define a corresponding set of dual vectors $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}$ in $\mathbf{R}^p$ by

$$L(\mathbf{u^k}) =: \lambda_k \mathbf{v^k} \quad ; \; k = 1, 2, \ldots, r\,. \tag{6.3}$$

From (6.2) it follows that

$$L^*(\mathbf{v^k}) = \lambda_k \mathbf{u^k} \quad ; \; k = 1, 2, \ldots, r \tag{6.4}$$

and that

$$L \circ L^*(\mathbf{v^k}) = \lambda_k^2 \mathbf{v^k} \quad ; \; k = 1, 2, \ldots, r\,. \tag{6.5}$$

The set of vectors $\{\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}\}$ defined by (6.3) spans $Im\, L$ and is an orthonormal set of eigenvectors for the selfadjoint operator $L \circ L^* : \mathbf{R}^p \longrightarrow \mathbf{R}^p$. We thus have a completely dual setup and canonical decompositions of both $\mathbf{R}^N$ and $\mathbf{R}^p$ into direct sums of subspaces spanned by eigenvectors corresponding to the distinct eigenvalues. We make the following definition.

Definition 6.1. A dual singular value decomposition system for an operator pair (L, L^*) is a system consisting of numbers $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$ and two sets of orthonormal vectors, $\{\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}\}$ and $\{\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}\}$ with $r = rank\,(L) = rank\,(L^*)$, satisfying (6.2)–(6.5) above.

The positive values $\lambda_1, \lambda_2, \ldots, \lambda_r$ are called *the singular values* of (L, L^*). We will call the vectors $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}$ principal components for $Im\, L^*$ and the vectors $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}$ principal components for $Im\, L$.

Given a dual SVD system we now complement the principal components for $Im\, L^*$, $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}$, to an orthonormal basis $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^N}$ in $\mathbf{R}^N$ and the principal components for $Im\, L$, $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}$, to an orthonormal basis $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^p}$ in $\mathbf{R}^p$.

In these bases we have that

$$(\mathbf{v^j}, L(\mathbf{u^k}))_{\mathbf{R}^p} = (L^*(\mathbf{v^j}), \mathbf{u^k})_{\mathbf{R}^N} = \begin{cases} \lambda_k \delta_{jk} & \text{if } j,k \leq r \\ 0 & \text{otherwise}\,. \end{cases} \tag{6.6}$$

This means that in these ON-bases $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$ is represented by the diagonal $p \times N$ matrix

$$\begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{6.7}$$

where D is the $r \times r$ diagonal matrix having the singular values of $(L,\, L^*)$ in descending order on the diagonal. The adjoint operator L^* is represented in the same bases by the transposed matrix, i.e. a diagonal $N \times p$ matrix.

We translate this to operations on the corresponding matrices as follows. Let U denote the $N \times r$ matrix having the coordinates, in the canonical basis in $\mathbf{R^N}$, for $\mathbf{u^1}, \mathbf{u^2}, \ldots \mathbf{u^r}$ as columns, and let V denote the $p \times r$ matrix having the coordinates, in the canonical basis in $\mathbf{R^p}$, for $\mathbf{v^1}, \mathbf{v^2}, \ldots \mathbf{v^r}$ as columns. Then (6.6) is equivalent to

$$X = VDU^T \quad \text{and} \quad X^T = UDV^T\,.$$

This is called a dual singular value decomposition for the pair of matrices $(X,\, X^T)$.

Notice that the singular values and the corresponding separate eigenspaces for $L^* \circ L$ as described above are canonical, but that the set $\{\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}\}$ (and thus also the connected set $\{\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}\}$) is not canonically defined by $L^* \circ L$. This set is only canonically defined up to actions of the appropriate orthogonal groups on the separate eigenspaces.

6.2.2 *Dual Principal Component Analysis*

We will now discuss how to use a dual SVD system to obtain optimal approximations of a given operator $L : \mathbf{R}^N \longrightarrow \mathbf{R^p}$ by operators of lower rank. If our goal is to visualize the data, then it is natural to measure the approximation error using a unitarily invariant norm, i.e. a norm $\|\cdot\|$ that is invariant with respect to unitary transformations on the variables or on the samples, i.e.

$$\|L\| = \|V \circ L \circ U\| \quad \text{for all } V \text{ and } U \text{ s.t. } V^*V = Id \text{ and } U^*U = Id\,. \tag{6.8}$$

Using an SVD, directly from (6.8) we conclude that such a norm is necessarily a symmetric function of the singular values of the operator. We will present

results for the L^2–norm of the singular values, but the results concerning optimal approximations are actually valid with respect to any unitarily invariant norm, see e.g. [34] and [39] for information in this direction. We omit proofs, but all the results in this section are proved using SVDs for the involved operators.

The Frobenius (or Hilbert–Schmidt) norm for an operator $L : \mathbf{R}^N \longrightarrow \mathbf{R}^{\mathbf{p}}$ of rank r is defined by

$$\|L\|_F := \left\{ \sum_{k=1}^{r} \lambda_k^2 \right\}^{1/2},$$

where λ_k, $k = 1, 2, \ldots, r$ are the singular values of (L, L^*).

Now let $\mathscr{M}_{n\times n}$ denote the set of real $n \times n$ matrices. We then define the set of orthogonal projections in $\mathbf{R}^n$ of rank $s \leq n$ as

$$\mathscr{P}_s^n := \{\Pi \in \mathscr{M}_{n\times n} \; ; \; \Pi^* = \Pi \; ; \; \Pi \circ \Pi = \Pi \; ; \; rank(\Pi) = s\}.$$

One important thing about orthogonal projections is that they never increase the Frobenius norm, i.e.

Lemma 6.1. *Let $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$ be a given linear operator. Then*

$$\|\Pi \circ L\|_F \leq \|L\|_F \quad \textit{for all } \Pi \in \mathscr{P}_s^p$$

and

$$\|L \circ \Pi\|_F \leq \|L\|_F \quad \textit{for all } \Pi \in \mathscr{P}_s^N.$$

Using this Lemma one can prove the following approximation theorems.

Theorem 6.1. *Let $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$ be a given linear operator. Then*

$$\sup_{\Pi^p \in \mathscr{P}_s^p \; ; \; \Pi^N \in \mathscr{P}_s^N} \|\Pi^p \circ L \circ \Pi^N\|_F = \sup_{\Pi \in \mathscr{P}_s^p} \|\Pi \circ L\|_F =$$

$$= \sup_{\Pi \in \mathscr{P}_s^N} \|L \circ \Pi\|_F = \left\{ \sum_{k=1}^{\min(s,r)} \lambda_k^2 \right\}^{1/2} \tag{6.9}$$

and equality is attained in (6.9) by projecting onto the $\min(s, r)$ *first principal components for $Im\, L$ and $Im\, L^*$.*

Theorem 6.2. *Let $L : \mathbf{R}^N \longrightarrow \mathbf{R}^p$ be a given linear operator. Then*

$$\inf_{\Pi^p \in \mathscr{P}_s^p \; ; \; \Pi^N \in \mathscr{P}_s^N} \|L - \Pi^p \circ L \circ \Pi^N\|_F = \inf_{\Pi \in \mathscr{P}_s^p} \|L - \Pi \circ L\|_F =$$

$$= \inf_{\Pi \in \mathscr{P}_s^N} \|L - L \circ \Pi\|_F = \left\{ \sum_{k=\min(s,r)+1}^{\max(s,r)} \lambda_k^2 \right\}^{1/2} \tag{6.10}$$

and equality is attained in (6.10) by projecting onto the $\min(s, r)$ *first principal components for* $Im\, L$ *and* $Im\, L^*$.

We loosely state these results as follows.

6.2.2.1 Projection Dictum

Projecting onto a set of first principal components maximizes average projected vector length and also minimizes average projection error.

We will briefly discuss interpretation for applications. In fact in applications the representation of our linear mapping L normally has a specific interpretation in the original canonical bases. Assume that $L(\mathbf{e_k})$, $k = 1, 2, \ldots, N$ represent samples and that $L^*(\mathbf{f_j})$, $j = 1, 2, \ldots, p$ represents variables. To begin with, if the samples are centered, i.e.

$$\sum_{k=1}^{N} L(\mathbf{e_k}) = 0\,,$$

then $\|L\|_F^2$ corresponds to the statistical *variance* of the sample set. The basic projection dictum can thus be restated for sample-centered data as follows.

6.2.2.2 Projection Dictum for Sample-Centered Data

Projecting onto a set of first principal components maximizes the variance in the set of projected data points and also minimizes average projection error.

In applications we are also interested in keeping track of the value

$$X_{jk} = (\mathbf{f_j}, L(\mathbf{e_k}))\,. \tag{6.11}$$

It represents the jth variable's value in the kth sample.

Computing in a dual SVD system for (L, L^*) in (6.11) we get

$$X_{jk} = \lambda_1(\mathbf{e_k}, \mathbf{u^1})(\mathbf{f_j}, \mathbf{v^1}) + \cdots + \lambda_r(\mathbf{e_k}, \mathbf{u^r})(\mathbf{f_j}, \mathbf{v^r})\,. \tag{6.12}$$

Now using (6.3) and (6.4) we conclude that

$$X_{jk} = \frac{1}{\lambda_1}(\mathbf{e_k}, L^*(\mathbf{v^1}))(\mathbf{f_j}, L(\mathbf{u^1})) + \cdots + \frac{1}{\lambda_r}(\mathbf{e_k}, L^*(\mathbf{v^r}))(\mathbf{f_j}, L(\mathbf{u^r}))\,.$$

Finally this implies the fundamental *biplot formula*

$$X_{jk} = \frac{1}{\lambda_1}(L(\mathbf{e_k}), \mathbf{v^1})(L^*(\mathbf{f_j}), \mathbf{u^1}) + \cdots + \frac{1}{\lambda_r}(L(\mathbf{e_k}), \mathbf{v^r})(L^*(\mathbf{f_j}), \mathbf{u^r})\,. \tag{6.13}$$

We now introduce the following scalar product in $\mathbf{R}^r$

$$(\mathbf{a}, \mathbf{b})_\lambda := \frac{1}{\lambda_1} a_1 b_1 + \cdots + \frac{1}{\lambda_r} a_r b_r \quad ; \ \mathbf{a}, \mathbf{b} \in \mathbf{R}^r \, .$$

Equation (6.13) thus means that if we express the sample vectors in the basis $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}$ for $Im\, L$ and the variable vectors in the basis $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}$ for $Im\, L^*$, then we get the value of X_{jk} simply by taking the $(\cdot, \cdot)_\lambda$-scalar product in $\mathbf{R}^r$ between the coordinate sequence for the kth sample and the coordinate sequence for the jth variable.

This means that if we work in a synchronized way in $\mathbf{R}^r$ with the coordinates for the samples (with respect to the basis $\mathbf{v^1}, \mathbf{v^2}, \ldots, \mathbf{v^r}$) and with the coordinates for the variables (with respect to the basis $\mathbf{u^1}, \mathbf{u^2}, \ldots, \mathbf{u^r}$) then the *relative positions* of the coordinate sequence for a variable and the coordinate sequence for a sample in $\mathbf{R}^r$ have a very precise meaning given by (6.13).

Now let $S \subset \{1, 2, \ldots, r\}$ be a subset of indices and let $|S|$ denote the number of elements in S. Then let $\Pi_S^p : \mathbf{R}^p \longrightarrow \mathbf{R}^p$ be the orthogonal projection onto the subspace spanned by the principal components for $Im\, L$ whose indices belong to S. In the same way let $\Pi_S^N : \mathbf{R}^N \longrightarrow \mathbf{R}^N$ be the orthogonal projection onto the subspace spanned by the principal components for $Im\, L^*$ whose indices belong to S.

If $L(\mathbf{e_k})$, $k = 1, 2, \ldots, N$, represent samples, we will say that $\Pi_S^p \circ L(\mathbf{e_k})$, $k = 1, 2, \ldots, N$, represent *S-approximative samples*, and correspondingly if $L^*(\mathbf{f^j})$, $j = 1, 2, \ldots, p$, represent variables then $\Pi_S^N \circ L^*(\mathbf{f_j})$, $j = 1, 2, \ldots, p$, represent *S-approximative variables*.

We will interpret the matrix element

$$X_{jk}^S := (\mathbf{f_j}, \Pi_S^p \circ L(\mathbf{e_k})) \tag{6.14}$$

as representing the jth S-approximative variable's value in the kth S-approximative sample.

By the biplot formula (6.13) for the operator $\Pi_S^p \circ L$ we actually have

$$X_{jk}^S = \sum_{m \in S} \frac{1}{\lambda_m} (L(\mathbf{e_k}), \mathbf{v^m})(L^*(\mathbf{f_j}), \mathbf{u^m}) \, . \tag{6.15}$$

If $|S| \leq 3$ we can visualize our approximative samples and approximative variables working in a synchronized way in $\mathbf{R}^{|S|}$ with the coordinates for the approximative samples and with the coordinates for the approximative variables. The *relative positions* of the coordinate sequence for an approximative variable and the coordinate sequence for an approximative sample in $\mathbf{R}^{|S|}$ then have the very precise meaning given by (6.15).

Naturally the information content of a biplot visualization depends in a crucial way on the approximation error we make. The following result gives the basic error estimates.

Theorem 6.3. *With notations as above we have the following projection error estimates*

$$\sum_{j=1}^{p}\sum_{k=1}^{N}|X_{jk}-X_{jk}^{S}|^2=\sum_{i\notin S}|\lambda_i|^2 \tag{6.16}$$

$$\sup_{j=1,\ldots,p\,;\,k=1,\ldots,N}|X_{jk}-X_{jk}^{S}|\leq\sup_{i\notin S}|\lambda_i|\,. \tag{6.17}$$

We will use the following statistics for measuring *projection content*:

Definition 6.2. With notations as above, the L^2*-projection content* connected with the subset S is by definition

$$\alpha_2(S):=\frac{\sum_{i\in S}|\lambda_i|^2}{\sum_{i=1}^{r}|\lambda_i|^2}\,.$$

We note that, in the case when we have sample centered data, $\alpha_2(S)$ is precisely the quotient between the amount of variance that we have "captured" in our projection and the total variance. In particular if $\alpha_2(S)=1$ then we have captured all the variance. Theorem 6.3 shows that we would like to have good control of the distributions of eigenvalues for general covariance matrices. We will address this issue for random matrices below, but we already here point out that we will estimate projection information content, or the signal to noise ratio, in a projection of real world data by comparing the observed L^2-projection content and the L^2-projection contents for corresponding randomized data.

6.2.3 Nonlinear PCA and Multidimensional Scaling

We begin our presentation of multidimensional scaling by looking at the reconstruction problem, i.e. how to reconstruct a dataset given only a proposed covariance or distance matrix. In the case of a covariance matrix, the basic idea is to try to factor a corresponding sample centered SVD or slightly rephrased by taking the square root of the covariance matrix.

Once we have established a reconstruction scheme we note that we can apply it to any proposed "covariance" or "distance" matrix, as long as they have the correct structure, even if they are artificial and a priori are not constructed using euclidean transformations on an existing data matrix. This opens up the possibility for using "any type" of similarity measures between samples or variables to construct artificial covariance or distance matrices.

We consider a $p\times N$ matrix X where the N columns $\{\mathbf{x_1},\ldots,\mathbf{x_N}\}$ consist of values of measurements for N samples of p variables. We will throughout this section assume that $p\geq N$. We introduce the $N\times 1$ vector

$$\mathbf{1} = [1, 1, \ldots, 1]^T ,$$

and we recall that the $N \times N$ covariance matrix of the data matrix $\mathbf{X} = [\mathbf{x_1}, \ldots, \mathbf{x_N}]$ is given as

$$\mathbf{C}(\mathbf{x_1}, \ldots, \mathbf{x_N}) = (\mathbf{X} - \frac{1}{N}\mathbf{X}\,\mathbf{1}\,\mathbf{1}^T)^T(\mathbf{X} - \frac{1}{N}\mathbf{X}\,\mathbf{1}\,\mathbf{1}^T) .$$

We will also need the (squared) distance matrix defined by

$$\mathbf{D}_{jk}(\mathbf{x_1}, \ldots, \mathbf{x_N}) := |\mathbf{x_j} - \mathbf{x_k}|^2 \quad ; \quad j, k = 1, 2, \ldots, N .$$

We will now consider the problem of reconstructing a data matrix X given only the corresponding covariance matrix or the corresponding distance matrix. We first note that since the covariance and the distance matrix of a data matrix X both are invariant under euclidean transformations in $\mathbf{R}^p$ of the columns of X, it will, if at all, only be possible to reconstruct the $p \times N$ matrix X modulo euclidean transformations in $\mathbf{R^p}$ of its columns.

We next note that the matrices $\mathbf{C}$ and $\mathbf{D}$ are connected. In fact we have

Proposition 6.1. *Given data points* $\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}$ *in* $\mathbf{R}^p$ *and the corresponding covariance and distance matrices,* $\mathbf{C}$ *and* $\mathbf{D}$*, we have that*

$$\mathbf{D}_{jk} = \mathbf{C}_{jj} + \mathbf{C}_{kk} - 2\mathbf{C}_{jk} . \tag{6.18}$$

Furthermore

$$\mathbf{C}_{jk} = \frac{1}{2N}\sum_{i=1}^{N}(\mathbf{D}_{ij} + \mathbf{D}_{ik}) - \frac{1}{2}\mathbf{D}_{jk} - \frac{1}{2N^2}\sum_{i,m=1}^{N}\mathbf{D}_{im} , \tag{6.19}$$

or in matrixnotation,

$$\mathbf{C} = -\frac{1}{2}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{N}\right)\mathbf{D}\left(I - \frac{\mathbf{1}\mathbf{1}^T}{N}\right) . \tag{6.20}$$

Proof. Let

$$\mathbf{y_i} := \mathbf{x_i} - \frac{1}{N}\sum_{j=1}^{N}\mathbf{x_j} \quad ; \; i = 1, 2, \ldots, N .$$

Note that

$$\mathbf{C}_{jk} = \mathbf{y_j^T}\,\mathbf{y_k} \quad \text{and} \quad \mathbf{D}_{jk} = |\mathbf{y_j} - \mathbf{y_k}|^2 ,$$

and that $\sum_{j=1}^{N}\mathbf{y_j} = 0$.

Equality (6.18) above is simply the polarity condition

$$\mathbf{D}_{jk} = |\mathbf{y_j}|^2 + |\mathbf{y_k}|^2 - 2\mathbf{C}_{jk} . \tag{6.21}$$

Moreover, since

$$\sum_{j=1}^{N} \mathbf{C}_{jk} = 0 \quad \text{and} \quad \sum_{k=1}^{N} \mathbf{C}_{jk} = 0\,,$$

by summing over both j and k in (6.21) above we get

$$\sum_{j=1}^{N} |\mathbf{y_j}|^2 = \frac{1}{2N} \sum_{j,k=1}^{N} \mathbf{D}_{jk}\,. \tag{6.22}$$

On the other hand, by summing only over j in (6.21) we get

$$\sum_{j=1}^{N} \mathbf{D}_{jk} = N|\mathbf{y_k}|^2 + \sum_{j=1}^{N} |\mathbf{y_j}|^2\,. \tag{6.23}$$

Combining (6.22) and (6.23) we get

$$|\mathbf{y_k}|^2 = \frac{1}{N} \sum_{j=1}^{N} \mathbf{D}_{jk} - \frac{1}{2N^2} \sum_{j,k=1}^{N} \mathbf{D}_{jk}\,.$$

Plugging this into (6.21) we finally conclude that

$$\mathbf{D}_{jk} = \frac{1}{N} \sum_{i=1}^{N} (\mathbf{D}_{ij} + \mathbf{D}_{ik}) - \frac{1}{N^2} \sum_{i,j=1}^{N} \mathbf{D}_{ij} - 2\mathbf{C}_{jk}\,.$$

This is (6.19). □

Now let $\mathscr{M}_{N\times N}$ denote the set of all real $N \times N$ matrices. To reconstruct a $p \times N$ data matrix $X = [\mathbf{x_1}, \ldots, \mathbf{x_N}]$ from a given $N \times N$ covariance or $N \times N$ distance matrix amounts to invert the mappings:

$$\Phi : \mathbf{R}^p \times \cdots \times \mathbf{R}^p \ni (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \mapsto \mathbf{C}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \in \mathscr{M}_{N\times N},$$

and

$$\Psi : \mathbf{R}^p \times \cdots \times \mathbf{R}^p \ni (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \mapsto \mathbf{D}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \in \mathscr{M}_{N\times N}\,.$$

In general it is of course impossible to invert these mappings since both Φ and Ψ are far from surjectivity and injectivity.

Concerning injectivity, it is clear that both Φ and Ψ are invariant under the euclidean group $E(p)$ acting on $\mathbf{R}^p \times \cdots \times \mathbf{R}^p$, i.e. under transformations

$$(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \mapsto (S\mathbf{x_1} + \mathbf{b}, S\mathbf{x_2} + \mathbf{b}, \ldots, S\mathbf{x_n} + \mathbf{b}),$$

where $\mathbf{b} \in \mathbf{R}^p$ and $S \in O(p)$.

This makes it natural to introduce the quotient manifold

$$(\mathbf{R}^p \times \cdots \times \mathbf{R}^p)/E(p)$$

and possible to define the induced mappings $\overline{\Phi}$ and $\overline{\Psi}$, well defined on the equivalence classes and factoring the mappings Φ and Ψ by the quotient mapping. We will write

$$\overline{\Phi} : ([\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]) \mapsto \mathbf{C}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N})$$

and

$$\overline{\Psi} : ([\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]) \mapsto \mathbf{D}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) .$$

We shall show below that both $\overline{\Phi}$ and $\overline{\Psi}$ are injective.

Concerning surjectivity of the maps Φ and Ψ, or $\overline{\Phi}$ and $\overline{\Psi}$, we will first describe the image of Φ. Since the images of Φ and Ψ are connected through Proposition 6.1 above this implicitly describes the image also for Ψ. It is theoretically important that both these image sets turn out to be closed and convex subsets of $\mathscr{M}_{N\times N}$. In fact we claim that the following set is the image set of $\overline{\Phi}$:

$$\mathscr{P}_{N\times N} := \left\{A \in \mathscr{M}_{N\times N} \ ; \ A^T = A \ , \ A \geq 0 \ , \ A\,\mathbf{1} = 0\right\} .$$

To begin with it is clear that the image of Φ is equal to the image of $\overline{\Phi}$ and that it is included in $\mathscr{P}_{N\times N}$, i.e.

$$(\mathbf{R}^p \times \cdots \mathbf{R}^p)/E(p) \ni ([\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]) \mapsto \mathbf{C}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \in \mathscr{P}_{N\times N} .$$

The following proposition implies that $\mathscr{P}_{N\times N}$ is the image set of $\overline{\Phi}$ and it is the main result of this subsection.

Proposition 6.2. *The mapping $\overline{\Phi}$:*

$$(\mathbf{R}^p \times \cdots \times \mathbf{R}^p)/E(p) \ni ([\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}]) \mapsto \mathbf{C}(\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_N}) \in \mathscr{P}_{N\times N}$$

is a bijection.

Proof. If $\mathbf{A} \in \mathscr{P}_{N\times N}$ we can, by the spectral theorem, find a unique symmetric and positive $N \times N$ matrix $\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N]$ (the square root of $\mathbf{A}$) with $rank(\mathbf{B}) = rank(\mathbf{A})$ such that $\mathbf{B}^2 = \mathbf{A}$ and $\mathbf{B}\,\mathbf{1} = 0$. □

We now map the points $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_N)$ in $\mathbf{R}^N$ isometrically to points $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)$ in $\mathbf{R}^p$. This is trivially possible since $p \geq N$. The corresponding covariance matrix $\mathbf{C}(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ will be equal to $\mathbf{A}$. This proves surjectivity. That the mapping $\overline{\Phi}$ is injective follows directly from the following lemma.

Lemma 6.2. *Let $\{\mathbf{y}_k\}_{k=1}^N$ and $\{\tilde{\mathbf{y}}_k\}_{k=1}^N$ be two sets of vectors in $\mathbf{R}^p$. If*

$$\mathbf{y}_k^T \mathbf{y}_j = \tilde{\mathbf{y}}_k^T \tilde{\mathbf{y}}_j \quad \textit{for } j,k = 1, 2, \ldots, N ,$$

then there exists an $\mathbf{S} \in \mathbf{O}(p)$ *such that*

$$\mathbf{S}(\mathbf{y}_k) = \tilde{\mathbf{y}}_k \quad \textit{for } k = 1, 2, \ldots, N\,.$$

Proof. Use the Gram–Schmidt orthogonalization procedure on both sets at the same time. □

We will in our explorations of high dimensional real world datasets below use "artificial" distance matrices constructed from geodesic distances on carefully created graphs connecting the samples or the variables. These distance matrices are converted to unique corresponding covariance matrices which in turn, as described above, give rise to canonical elements in $(\mathbf{R}^p \times \cdots \times \mathbf{R}^p)/E(p)$. We then pick sample centered representatives on which we perform PCA. In this way we can visualize low dimensional "approximative graph distances" in the dataset. Using graphs in the sample set constructed from a k nearest neighbors or a locally euclidean approximation procedure, this approach corresponds to the ISOMAP algorithm introduced by Tenenbaum et al. [51]. The ISOMAP algorithm can as we will see below be very useful in the exploration of DNA microarray data, see Nilsson et al. [37] for one of the first applications of ISOMAP in this field.

We finally remark that if a proposed artificial distance or covariance matrix does not have the correct structure, i.e. if for example a proposed covariance matrix does not belong to $\mathscr{P}_{N\times N}$, we begin by projecting the proposed covariance matrix onto the unique nearest point in the closed and convex set $\mathscr{P}_{N\times N}$ and then apply the scheme presented above to that point.

6.3 The Basic Statistical Framework

We will here fix some notation and for the non-statistician reader's convenience at the same time recapitulate some standard multivariate statistical theory. In particular we want to stress some basic facts concerning robustness of statistical testing.

Let $\mathscr{S}$ be the sample space consisting of all possible samples (in our example datasets equal to all trials of patients) equipped with a probability measure $P : 2^{\mathscr{S}} \longrightarrow [0, +\infty]$ and let $X = (X_1, \ldots, X_p)^T$ be a random vector from $\mathscr{S}$ into $\mathbf{R}^\mathbf{p}$. The coordinate functions $X_i : \mathscr{S} \longrightarrow \mathbf{R}$, $i = 1, 2, \ldots, p$ are random variables and in our example datasets they represent the expression levels of the different genes.

We will be interested in the law of X, i.e. the induced probability measure $P(X^{-1}(\cdot))$ defined on the measurable subsets of $\mathbf{R}^\mathbf{p}$. If it is absolutely continuous with respect to Lebesgue measure then there exists a probability density function (pdf) $f_X(\cdot) : \mathbf{R}^\mathbf{p} \longrightarrow [\mathbf{0}, \infty)$ that belongs to $\mathscr{L}^1(\mathbf{R}^p)$ and satisfies

$$P(\{s \in \mathscr{S}\ ;\ X(s) \in A\}) = \int_A f_X(x)\, dx \tag{6.24}$$

for all events (i.e. all Lebesgue measurable subsets) $A \subset \mathbf{R^p}$. This means that the pdf $f_X(\cdot)$ contains all necessary information in order to compute the probability that an event has occurred, i.e. that the values of X belong to a certain given set $A \subset \mathbf{R^p}$.

All statistical inference procedures are concerned with trying to learn as much as possible about an at least partly unknown induced probability measure $P(X^{-1}(\cdot))$ from a given set of N observations, $\{\mathbf{x^1}, \mathbf{x^2}, \ldots, \mathbf{x^N}\}$ (with $\mathbf{x}^i \in \mathbf{R}^p$ for $i = 1, 2, \ldots, N$), of the underlying random vector X.

Often we then assume that we know something about the *structure* of the corresponding pdf $f_X(\cdot)$ and we try to make statistical inferences about the *detailed form* of the function $f_X(\cdot)$.

The most important probability distribution in multivariate statistics is the multivariate normal distribution. In $\mathbf{R^p}$ it is given by the p-variate pdf $n : \mathbf{R^p} \longrightarrow (\mathbf{0}, \infty)$ where

$$n(x) := (2\pi)^{-p/2} |\Gamma|^{-1/2} e^{-\frac{1}{2}(x-\mu)^T \Gamma^{-1}(x-\mu)} \quad ; \; x \in \mathbf{R^p} \,. \tag{6.25}$$

It is characterized by the symmetric and positive definite $p \times p$ matrix Γ and the p-column vector μ, and $|\Gamma|$ stands for the absolute value of the determinant of Γ. If a random vector $X : \mathscr{S} \longrightarrow \mathbf{R^p}$ has the p-variate normal pdf (6.25) we say that X has the $\mathbf{N}(\mu, \Gamma)$ distribution. If X has the $\mathbf{N}(\mu, \Gamma)$ distribution then *the expected value* of X is equal to μ i.e.

$$\mathscr{E}(X) := \int_{\mathscr{S}} X(s)\, dP = \mu, \tag{6.26}$$

and *the covariance matrix* of X is equal to Γ, i.e.

$$\mathscr{C}(X) := \int_S (X - \mathscr{E}(X))(X - \mathscr{E}(X))^T \, dP = \Gamma. \tag{6.27}$$

Assume now that $X^1, X^2, \ldots, X^N$ are given independent and identically distributed (i.i.d.) random vectors. A *test statistic* $\mathscr{T}$ is then by definition a function $(X^1, X^2, \ldots, X^N) \mapsto \mathscr{T}(X^1, X^2, \ldots, X^N)$. Two important test statistics are *the sample mean* vector of a sample of size N

$$\overline{X}^N := \frac{1}{N} \sum_{i=1}^{N} X^i \,,$$

and *the sample covariance matrix* of a sample of size N

$$S^N := \frac{1}{N-1} \sum_{i=1}^{N} (X^i - \overline{X})(X^i - \overline{X})^T \,.$$

If $X^1, X^2, \ldots, X^N$ are independent and $\mathbf{N}(\mu, \Gamma)$ distributed, then the mean $\overline{X}^N$ has the $\mathbf{N}(\boldsymbol{\mu}, \frac{1}{N}\Gamma)$ distribution. In fact this result is asymptotically robust with respect to the underlying distribution. This is a consequence of the well known and celebrated central limit theorem:

Theorem 6.4. *If the random p vectors $X^1, X^2, X^3, \ldots$ are independent and identically distributed with means $\mu \in \mathbf{R^p}$ and covariance matrices Γ, then the limiting distribution of*

$$(N)^{1/2}\left(\overline{X}^N - \mu\right)$$

as $N \longrightarrow \infty$ is $\mathbf{N}(0, \Gamma)$.

The central limit theorem tells us that, if we know nothing and still need to assume some structure on the underlying p.d.f., then asymptotically the $N(\mu, \Gamma)$ distribution is the only reasonable assumption. The distributions of different statistics are of course more or less sensitive to the underlying distribution. In particular the standard univariate Student t-statistic, used to draw inferences about a univariate sample mean, is very robust with respect to the underlying probability distribution. In for example the study on statistical robustness [40] the authors conclude that:

...the two-sample t-test is so robust that it can be recommended in nearly all applications.

This is in contrast with many statistics connected with the sample covariance matrix. A central example in multivariate analysis is the set of eigenvalues of the sample covariance matrix. These statistics have a more complicated behavior. First of all, if $X^1, X^2, \ldots, X^N$ with values in $\mathbf{R}^p$ are independent and $\mathbf{N}(\mu, \Gamma)$ distributed then the sample covariance matrix is said to have a Wishart distribution $\mathbf{W}_p(N, \Gamma)$. If $N > p$ the Wishart distribution is absolutely continuous with respect to Lebesgue measure and the probability density function is explicitly known, see e.g. Theorem 7.2.2. in [3]. If $N \gg p$ then the eigenvalues of the sample covariance matrix are good estimators for the corresponding eigenvalues of the underlying covariance matrix Γ, see [2] and [3]. In the applications we have in mind we often have the reverse situation, i.e. $p \gg N$, and then the eigenvalues for the sample covariance matrix are far from consistent estimators for the corresponding eigenvalues of the underlying covariance matrix. In fact if the underlying covariance matrix is the identity matrix it is known (under certain growth conditions on the underlying distribution) that if we let p depend on N and if $p/N \longrightarrow \gamma \in (0, \infty)$ as $N \longrightarrow \infty$, then the largest eigenvalue for the sample covariance matrix tends to $(1 + \sqrt{\gamma})^2$, see e.g. [54], and not to 1 as one maybe could have expected. This result is interesting and can be useful, but there are many open questions concerning the asymptotic theory for the "large p, large N case", in particular if we go beyond the case of normally distributed data, see e.g. [7, 18, 25, 26] and [29] for an overview of the current state of the art. To estimate the information content or signal to noise ratio in our PCA plots we will therefore rely mainly on randomization tests and not on the (not well enough developed) asymptotic theory for the distributions of eigenvalues of random matrices.

6.4 Controlling the False Discovery Rate

When we perform for example a Student t-test to estimate whether or not two groups of samples have the same mean value for a specific variable we are performing a hypothesis test. When we do the same thing for a large number of variables at the same time we are testing one hypothesis for each and every variable. It is often the case in the applications we have in mind that tens of thousands of features are tested at the same time against some null hypothesis H_0, e.g. that the mean values in two given groups are identical. To account for this multiple hypotheses testing, several methods have been proposed, see e.g. [53] for an overview and comparison of some existing methods. We will give a brief review of some basic notions.

Following the seminal paper by Benjamini and Hochberg [12], we introduce the following notation. We consider the problem of testing m null hypotheses H_0 against the alternative hypothesis H_1. We let m_0 denote the number of true nulls. We then let R denote the total number of rejections, which we will call the total number of statistical discoveries, and let V denote the number of false rejections. In addition we introduce stochastic variables U and T according to Table 6.1.

The false discovery rate was loosely defined by Benjamini and Hochberg as the expected value $E(\frac{V}{R})$. More precisely the false discovery rate is defined as

$$FDR := E(\frac{V}{R} \mid R > 0)\, P(R > 0)\,. \tag{6.28}$$

The false discovery rate measures the proportion of Type I errors among the statistical discoveries. Analogously we define corresponding statistics according to Table 6.2. We note that the FNDR is precisely the proportion of Type II errors among the accepted null hypotheses, i.e. the non-discoveries. In the datasets that we encounter within bioinformatics we often suspect $m_1 << m$ and so if R, which we can observe, is relatively small, then the FNDR is controlled at a low level. As pointed out in [38], apart from the FDR which measures the proportion of false positive discoveries, we usually are interested in also controlling the FNR, i.e. we do not want to miss too many true statistical discoveries. We will address this by using visualization and knowledge based evaluation to support the statistical analysis.

In our exploration scheme presented in the next section we will use the step down procedure on the entire list of p-values for the statistical test under consideration suggested by Benjamini and Hochberg in [12] to control the FDR. We will also use the q-value, computable for each separate variable, introduced by Storey, see [47] and [48]. The q-value in our analyses is defined as the lowest FDR for which the particular hypothesis under consideration would be accepted under the

Table 6.1 Test statistics

	Accept H_0	Reject H_0	Total
H_0 true	U	V	m_0
H_1 true	T	S	m_1
	$m - R$	R	m

Table 6.2 Statistical discovery rates

Expected value	Name
$E(V/R)$	False discovery rate (FDR)
$E(T/(m-R))$	False negative discovery rate (FNDR)
$E(T/(T+S))$	False negative rate (FNR)
$E(V/(U+V))$	False positive rate (FPR)

Benjami–Hochberg step down procedure. A practical and reasonable threshold level for the q-value to be informative that we will use is $q < 0.2$.

6.5 The Basic Exploration Scheme

We will look for significant signals in our data set in order to use them e.g. as a basis for variable selection, sample classification and clustering.

If there are enough samples one should first of all randomly partition the set of samples into a *training set* and a *testing set*, perform the analysis with the training set and then validate findings using the testing set. This should then ideally be repeated several times with different partitions. With very few samples this is not always feasible and then, in addition to statistical measures, one is left with using knowledge based evaluation. One should then remember that the ultimate goal of the entire exploration is to add pieces of new knowledge to an already existing knowledge structure. To facilitate knowledge based evaluation, the entire exploration scheme is throughout guided by visualization using PCA biplots.

When looking for significant signals in the data, one overall rule that we follow is:

- Detect and then remove the strongest present signal.

  Detect a signal can e.g. mean to find a sample cluster and a connected list of variables that discriminate the cluster. We can then for example (re)classify the sample cluster in order to use the new classification to perform more statistical tests. After some statistical validation we then often remove a detected signal, e.g. a sample cluster, in order to avoid that a strong signal obscures a weaker but still detectable signal in the data. Sometimes it is of course convenient to add the strong signal again at a later stage in order to use it as a reference.
  We must constantly be aware of the possibility of outliers or artifacts in our data and so we must:

- Detect and remove possible artifacts or outliers.

  An artifact is by definition a detectable signal that is unrelated to the basic mechanisms that we are exploring. An artifact can e.g. be created by different experimental setups, resulting in a signal in the data that represents different experimental conditions. Normally if we detect a suspected artifact we want to, as far as possible, eliminate the influence of the suspected artifact on our data. When we do this we must be aware that we normally reduce the degrees of freedom in our data. The most common case is to eliminate a single nominal factor resulting

in a splitting of our data in subgroups. In this case we will mean-center each group discriminated by the nominal factor, and then analyze the data as usual, with an adjusted number of degrees of freedom.
The following basic exploration scheme is used

- Reduce noise by PCA and variance filtering. Assess the signal/noise ratio in various low dimensional PCA projections and estimate the projection information contents by randomization.
- Perform statistical tests. Evaluate the statistical tests using the FDR, randomization and permutation tests.
- Use graph-based multidimensional scaling (ISOMAP) to search for signals/clusters.

The above scheme is iterated until "all" significant signals are found and it is guided and coordinated by synchronized PCA-biplot visualizations.

6.6 Some Biological Background Concerning the Example Datasets

The rapid development of new biological measurement methods makes it possible to explore several types of genetic alterations in a high-throughput manner. Different types of microarrays enable researchers to simultaneously monitor the expression levels of tens of thousands of genes. The available information content concerning genes, gene products and regulatory pathways is accordingly growing steadily. Useful bioinformatics databases today include the Gene Ontology project (GO) [5] and the Kyoto Encyclopedia of Genes and Genomes (KEGG) [27] which are initiatives with the aim of standardizing the representation of genes, gene products and pathways across species and databases. A substantial collection of functionally related gene sets can also be found at the Broad Institute's Molecular Signatures Database (MSigDB) [35] together with the implemented computational method Gene Set Enrichment Analysis (GSEA) [36, 50]. The method GSEA is designed to determine whether an a priori defined set of genes shows statistically significant and concordant differences between two biological states in a given dataset.

Bioinformatic data sets are often uploaded by researchers to sites such as the National Centre for Biotechnology Information's database Gene Expression Omnibus (GEO) [19] or to the European Bioinformatics Institute's database ArrayExpress [4]. In addition data are often made available at local sites maintained by separate institutes or universities.

6.7 Analysis of Microarray Data Sets

There are many bioinformatic and statistical challenges that remain unsolved or are only partly solved concerning microarray data. As explained in [42], these include normalization, variable selection, classification and clustering. This state of

affairs is partly due to the fact that we know very little in general about underlying statistical distributions. This makes statistical robustness a key issue concerning all proposed statistical methods in this field and at the same time shows that new methodologies must always be evaluated using a knowledge based approach and supported by accompanying new biological findings. We will not comment on the important problems of normalization in what follows but refer to e.g. [6] where different normalization procedures for the Affymetrix platforms are compared. In addition, microarray data often have a non negligible amount of missing values. In our example data sets we will, when needed, impute missing values using the K-nearest neighbors method as described in [52]. All visualizations and analyses are performed using the software Qlucore Omics Explorer [44].

6.7.1 Effects of Cigarette Smoke on the Human Epithelial Cell Transcriptome

We begin by looking at a gene expression dataset coming from the study by Spira et al. [45] of effects of cigarette smoke on the human epithelial cell transcriptome. It can be downloaded from National Center for Biotechnology Informations (NCBI) Gene Expression Omnibus (GEO) (DataSet GDS534, accession no. GSE994). It contains measurements from 75 subjects consisting of 34 current smokers, 18 former smokers and 23 healthy never smokers. The platform used to collect the data was Affymetrix HG-U133A Chip using the Affymetrix Microarray suite to select, prepare and normalize the data, see [45] for details.

One of the primary goals of the investigation in [45] was to find genes that are responsible for distinguishing between current smokers and never smokers and also investigate how these genes behaved when a subject quit smoking by looking at the expression levels for these genes in the group of former smokers. We will here focus on finding genes that discriminate the groups of current smokers and never smokers.

- We begin our exploration scheme by estimating the signal/noise ratio in a sample PCA projection based on the three first principal components.

  We use an SVD on the data correlation matrix, i.e. the covariance matrix for the variance normalized variables. In Fig. 6.1 we see the first three principal components for ImL and the 75 patients plotted. The first three principal components contain 25% of the total variance in the dataset and so for this 3-D projection $\alpha_2(\{1,2,3\}, obsr) = 0.25$. Using randomization we estimate the expected value for a corresponding dataset (i.e. a dataset containing the same number of samples and variables) built on independent and normally distributed variables to be approximately $\alpha_2(\{1,2,3\}, rand) = 0.035$. We have thus captured around 7 times more variation than what we would have expected if the variables were independent and normally distributed. This indicates that we do have strong signals present in the dataset.

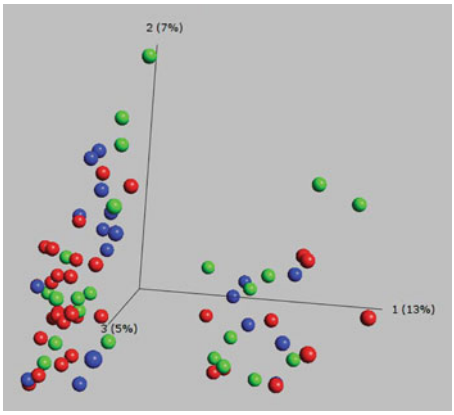


Fig. 6.1 The 34 current smokers (*red*), 18 former smokers (*blue*) and 23 never smokers (*green*) projected onto the three first principal components. The separation into two groups is not associated with any supplied clinical annotation and is thus a suspected artifact

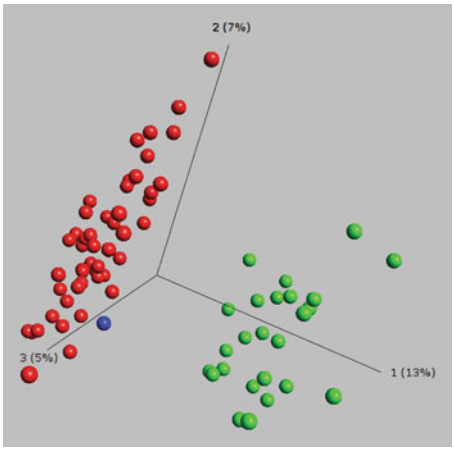


Fig. 6.2 The *red* samples have high description numbers ($\geq$58) and the *green* samples have low description numbers ($\leq$54). The *blue* sample has number 5

- Following our exploration scheme we now look for possible outliers and artifacts.

  The projected subjects are colored according to smoking history, but it is clear from Fig. 6.1 that most of the variance in the first principal component (containing 13% of the total variance in the data) comes from a signal that has a very weak association with smoking history. We nevertheless see a clear splitting into two subgroups. Looking at supplied clinical annotations one can conclude that the two groups are not associated to gender, age or race traits. Instead one finds that all the subjects in one of the groups have low subject description numbers whereas all the subjects except one in the other group have high subject description numbers. In Fig. 6.2 we have colored the subjects according to description number. This suspected artifact signal does not correspond to any in the dataset (Dataset GDS 534, NCBIs GEO) supplied clinical annotation. One can hypothesize that the description number could reflect for instance the order in which the samples were gathered and thus could be an artifact. Even if the

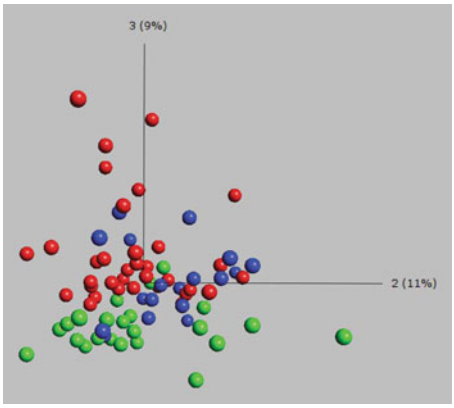


Fig. 6.3 We have filtered by variance keeping the 630 most variable genes. It is interesting to see that the third principle component containing 9% of the total variance separates the current smokers (*red*) from the never smokers (*green*) quite well

two groups actually correspond to some interesting clinical variable, like disease state, that we should investigate separately, we will consider the splitting to be an artifact in our investigation. We are interested in using all the assembled data to look for genes that discriminate between current smokers and never smokers. We thus eliminate the suspected artifact by mean-centering the two main (artifact) groups. After elimination of the strong artifact signal, the first three principal components contain "only" 17% of the total variation.

- Following our exploration scheme we filter the genes with respect to variance visually searching for a possibly informative three dimensional projection.

  When we filter down to the 630 most variable genes, the three first principal components have an L^2-projection content of $\alpha_2(\{1,2,3\}) = 0.42$, whereas by estimation using randomization we would have expected it to be 0.065. The projection in Fig. 6.3 is thus probably informative. We have again colored the samples according to smoking history as above. The third principal component, containing 9% of the total variance, can be seen to quite decently separate the current smokers from the never smokers. We note that this was impossible to achieve without removing the artifact signal since the artifact signal completely obscured this separation.

- Using the variance filtered list of 630 genes as a basis, following our exploration scheme, we now perform a series of Student t-tests between the groups of current smokers and never smokers, i.e. $34 + 23 = 57$ different subjects.

  For a specific level of significance we compute the 3-dimensional (i.e. $S = \{1,2,3\}$) L^2-projection content resulting when we keep all the rejected null hypotheses, i.e. statistical discoveries. For a sequence of t-tests parameterized by the level of significance we now try to find a small level of significance

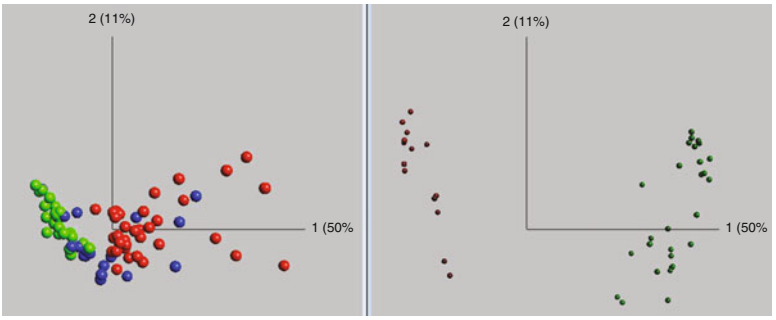


Fig. 6.4 A synchronized biplot showing samples to the *left* and variables to the *right*

and at the same time an observed L^2-projection content with a large quotient compared to the expected projection content estimated by randomization. We supervise this procedure visually using three dimensional PCA-projections looking for visually clear patterns. For a level of significance of 0.00005, leaving a total of 43 genes (rejected nulls) and an FDR of 0.0007 we have $\alpha_2(\{1,2,3\}, obsr) = 0.71$ whereas the expected projection content for randomized data $\alpha_2(\{1,2,3\}, rand) = 0.21$. We have thus captured more than 3 times of the expected projection content and at the same time approximately $0.0007 \times 43 = 0.0301$ genes are false discoveries and so with high probability we have found 43 potentially important biomarkers. We now visualize all 75 subjects using these 43 genes as variables. In Fig. 6.4 we see a synchronized biplot with samples to the left and variables to the right. The sample plot shows a perfect separation of current smokers and never smokers. In the variable plot we see genes (green) that are upregulated in the current smokers group to the far right. The top genes according to q-value for the Student t-test between current smokers and never smokers, that are upregulated in the current smokers group and downregulated in the never smokers group, are given in Table 6.3. In Table 6.4 we list the top genes that are downregulated in the current smokers group and upregulated among the never smokers.

6.7.2 Analysis of Various Muscle Diseases

In the study by Bakay et al. [10] the authors studied 125 human muscle biopsies from 13 diagnostic groups suffering from various muscle diseases. The platforms used were Affymetrix U133A and U133B chips. The dataset can be downloaded from NCBIs Gene Expression Omnibus (DataSet GDS2855, accession no. GSE3307). We will analyze the dataset looking for phenotypic classifications and also looking for biomarkers for the different phenotypes.

Table 6.3 Top genes upregulated in the current smokers group and downregulated in the never smokers group

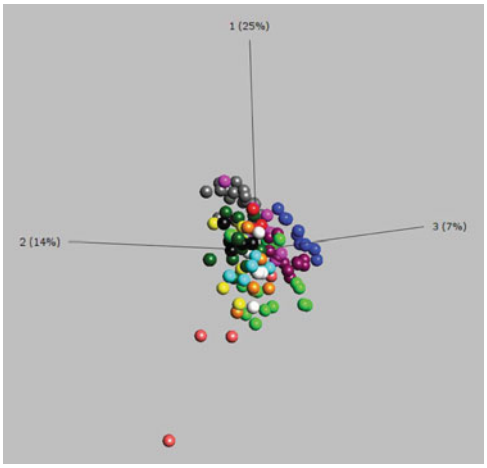
Gene symbol	q-value
NQO1	5.59104067771824e-08
GPX2	2.31142232391279e-07
ALDH3A1	2.31142232391279e-07
CLDN10	3.45691439169953e-06
FTH1	4.72936617815058e-06
TALDO1	4.72936617815058e-06
TXN	4.72936617815058e-06
MUC5AC	3.77806345774405e-05
TSPAN1	4.50425200297664e-05
PRDX1	4.58227420582093e-05
MUC5AC	4.99131989472012e-05
AKR1C2	5.72678146958168e-05
CEACAM6	0.000107637125805187
AKR1C1	0.000195523829628407
TSPAN8	0.000206106293159401
AKR1C3	0.000265342898771159

Table 6.4 Top genes downregulated in the current smokers group and upregulated in the never smokers group

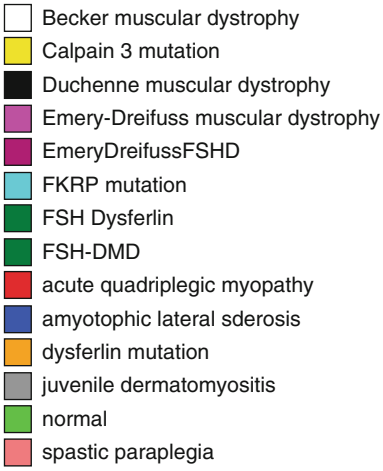
Gene symbol	q-value
MT1G	4.03809377378893e-07
MT1X	4.72936617815058e-06
MUC5B	2.38198903402317e-05
CD81	3.1605221864278e-05
MT1L	3.1605221864278e-05
MT1H	3.1605221864278e-05
SCGB1A1	4.50425200297664e-05
EPAS1	4.63861480935914e-05
FABP6	0.00017793865432854
MT2A	0.000236481909692626
MT1P2	0.000251264650053933

- We first use 3-dimensional PCA-projections of the samples of the data correlation matrix, filtering the genes with respect to variance and visually searching for clear patterns.

  When filtering out the 300 genes having most variability over the sample set we see several samples clearly distinguishing themselves and we capture 46% of the total variance compared to the, by randomization estimated, expected 6%. The plot in Fig. 6.5 thus contains strong signals. Comparing with the color legend we conclude that the patients suffering from spastic paraplegia (Spg) contribute a strong signal. More precisely, three of the subjects suffering from the variant Spg-4 clearly distinguish themselves, while the remaining patient in the Spg-group suffering from Spg-7 falls close to the rest of the samples.



(a) A three dimensional PCA plot capturing 46% of the total variance.



(b) Color legend

Fig. 6.5 PCA-projection of samples based on the variance filtered top 300 genes. Three out of four subjects in the group Spastic paraplegia (Spg) clearly distinguish themselves. These three suffer from Spg-4, while the remaining Spg-patient suffers from Spg-7

Table 6.5 Top genes upregulated in the Spastic paraplegia group (Spg-4)

Gene symbol	q-value
RAB40C	0.0000496417
SFXN5	0.000766873
CLPTM1L	0.00144164
FEM1A	0.0018485
HDGF2	0.00188435
WDR24	0.00188435
NAPSB	0.00188435
ANKRD23	0.00188435

- We perform Student t-tests between the spastic paraplegia group and the normal group.

  As before we now, in three dimensional PCA-projections, visually search for clearly distinguishable patterns in a sequence of Student t-tests parametrized by level of significance, while at the same time trying to obtain a small FDR. At a level of significance of 0.00001, leaving a total of 37 genes (rejected nulls) with an FDR of 0.006, the first three principal components capture 81% of the variance compared to the, by randomization, expected 31%. Table 6.5 lists the top genes upregulated in the group spastic paraplegia. We can add that these genes are all strongly upregulated for the three particular subjects suffering from Spg-4, while that pattern is less clear for the patient suffering from Spg-7.

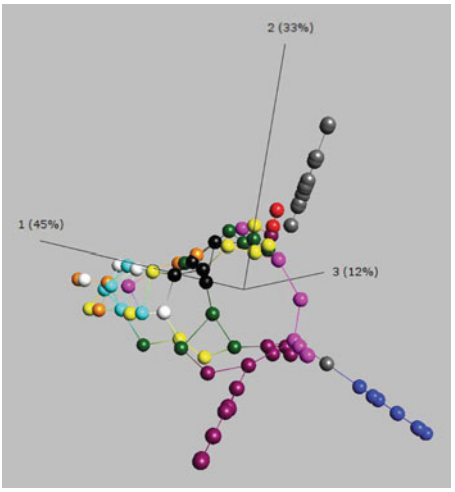


Fig. 6.6 Effect of the ISOMAP-algorithm. We can identify a couple of clusters corresponding to the groups juvenile dermatomyositis, amyotophic lateral sclerosis, acute quadriplegic myopathy and Emery Dreifuss FSHD

- In order to find possibly obscured signals, we now remove the Spastic paraplegia group from the analysis.

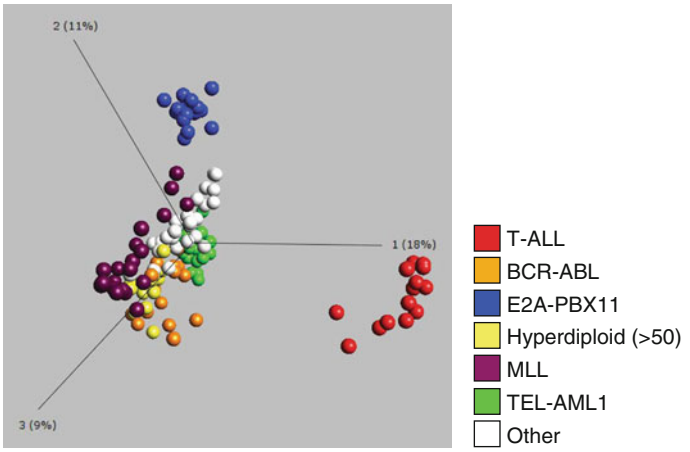
  We also remove the Normal group from the analysis since we really want to compare the different diseases. Starting anew with the entire set of genes, filtering with respect to variance, we visually obtain clear patterns for the 442 most variable genes. The first three principal components capture 46% of the total variance compared to the, by randomization estimated, expected 6%.

- Using these 442 most variable genes as a basis for the analysis, we now construct a graph connecting every sample with its two nearest (using euclidean distances in the 442-dimensional space) neighbors.

  As described in the section on multidimensional scaling above, we now compute geodesic distances in the graph between samples, and construct a resulting distance (between samples) matrix. We then convert this distance matrix to a corresponding covariance matrix and finally perform a PCA on this covariance matrix. The resulting plot (together with the used graph) of the so constructed three dimensional PCA-projection is depicted in Fig. 6.6.

  Comparing with the Color legend in Fig. 6.5, we clearly see that the groups juvenile dermatomyositis, amyotophic lateral sclerosis, acute quadriplegic myopathy and also Emery–Dreifuss FSHD distinguish themselves.

  One should now go on using Student t-tests to find biomarkers (i.e. genes) distinguishing these different groups of patients, then eliminate these distinct groups and go on searching for more structure in the dataset.



(a) Projection capturing 38% of the total variance. (b) Color legend

Fig. 6.7 Variance filtered PCA-projection of the correlation datamatrix based on 873 genes. The group T-ALL clearly distinguish itself

6.7.3 Pediatric Acute Lymphoblastic Leukemia

We will finally analyze a dataset consisting of gene expression profiles from 132 different patients, all suffering from some type of pediatric acute lymphoblastic leukemia (ALL). For each patient the expression levels of 22282 genes are analyzed. The dataset comes from the study by Ross et al. [43] and the primary data are available at the St. Jude Children's Research Hospital's website [49]. The platform used to collect this example data set was Affymetrix HG-U133 chip, using the Affymetrix Microarray suite to select, prepare and normalize the data.

As before we start by performing an SVD on the data correlation matrix visually searching for interesting patterns and assessing the signal to noise ratio by comparing the actual L^2-projection content in the real world data projection with the expected L^2-projection content in corresponding randomized data.

- We filter the genes with respect to variance, looking for strong signals.

  In Fig. 6.7 we see a plot of a three dimensional projection using the 873 most variable genes as a basis for the analysis. We clearly see that the group T-ALL is mainly responsible for the signal resulting in the first principal component occupying 18% of the total variance. In fact by looking at supplied annotations we can conclude that all of the other subjects in the dataset are suffering from B-ALL, the other main ALL type.

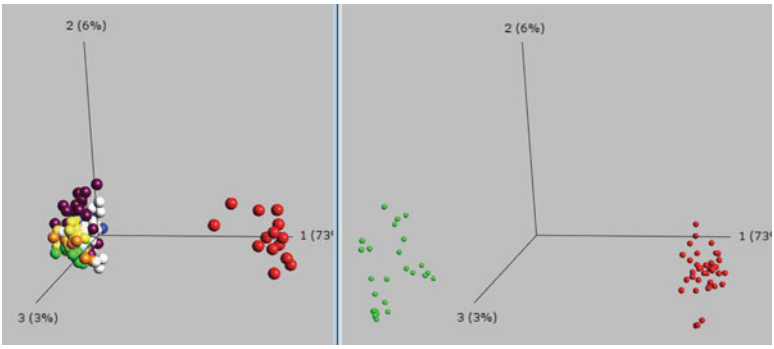


Fig. 6.8 FDR = 1.13e-24. A synchronized biplot showing samples to the *left* and genes to the *right*. The genes are colored according to their expression level in the T-ALL group. *Red*= upregulated and *green*= downregulated

- We now perform Student t-tests between the group T-ALL and the rest. We parametrize by level of significance and visually search for clear patterns.

  In Fig. 6.8 we see a biplot based on the 70 genes that best discriminate between T-ALL and the rest. The FDR is extremely low $FDR = 1.13\text{e-}24$ telling us that with a very high probability the genes found are relevant discoveries. The most significantly upregulated genes in the T-ALL group are
  CD3D, CD7, TRD@, CD3E, SH2D1A and TRA@.
  The most significantly downregulated genes in the T-ALL group are
  CD74, HLA-DRA, HLA-DRB, HLA-DQB and BLNK.
  By comparing with gene-lists from the MSig Data Base (see [35]) we can see that the genes that are upregulated in the T-ALL group (CD3D, CD7 and CD3E) are represented in lists of genes connected to lymphocyte activation and lymphocyte differentiation.

- We now remove the group T-ALL from the analysis and search for visually clear patterns among three dimensional PCA-projections filtrating the genes with respect to variance.

  Starting anew with the entire list of genes, filtering with respect to variance, a clear pattern is obtained for the 226 most variable genes. We capture 43% of the total variance as compared to the expected 6.5%. We thus have strong signals present.

- Using these 226 most variable genes as a basis for the analysis, we now construct a graph connecting every sample with its two nearest neighbors.

  We now perform the ISOMAP-algorithm with respect to this graph. The resulting plot (together with the used graph) of the so constructed three dimensional PCA-projection is depicted in Fig. 6.9. We can clearly distinguish the groups

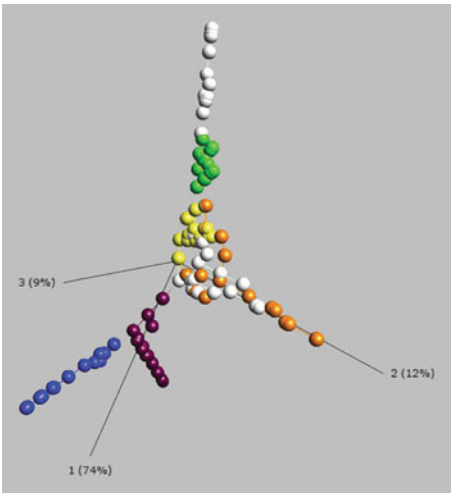


Fig. 6.9 Effect of the ISOMAP-algorithm. We can clearly see the groups E2A-PBX1, MLL and TEL-AML1. The group TEL-AML1 is connected to a subgroup of the group called Other (*white*)

E2A-PBX1, MLL and TEL-AML1. The group TEL-AML1 is connected to a subgroup of the group called Other. This subgroup actually corresponds to the Novel Group discovered in the study by Ross et al. [43]. Note that by using ISOMAP we discovered this Novel subgroup only by variance filtering the genes showing that ISOMAP is a useful tool for visually supervised clustering.

Acknowledgements I first of all would like to thank Gunnar Sparr for being a role model for me, and many other young mathematicians, of a pure mathematician that evolved into contributing serious applied work. Gunnar's help, support and general encouragement have been very important during my own development within the field of mathematical modeling. I sincerely thank Johan Råde for helping me to learn almost everything I know about data exploration. Without him the here presented work would truly not have been possible. Applied work is best done in collaboration and I am blessed with Thoas Fioretos as my long term collaborator within the field of molecular biology. I am grateful for what he has tried to teach me and I hope he is willing to continue to try. Finally I thank Charlotte Soneson for reading this work and, as always, giving very valuable feed-back.

References

1. Alter, O., Brown, P., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Natl. Acad. Sci. **97**(18), 10101–10106 (2000)
2. Anderson, T.W.: Asymptotic theory for principal component analysis. Ann. Math. Stat. **34**, 122–148 (1963)
3. Anderson, T.W.: An Introduction to Multivariate Statistical Analysis, 3rd edn. Wiley, Hoboken, NJ (2003)
4. The European Bioinformatics Institute's database ArrayExpress: http://www.ebi.ac.uk/microarray-as/ae/

5. Ashburner, M., et al.: The gene ontolgy consortium. Gene Ontology: Tool for the unification of biology. Nat. Genet. **25**, 25–29 (2000)
6. Autio, R., et al.: Comparison of Affymetrix data normalization methods using 6,926 experiments across five array generations. BMC Bioinform. **10**, suppl.1 S24 (2009)
7. Bai, Z.D.: Methodologies in spectral analysis of large dimensional random matrices, a review. Statist. Sin. **9**, 611–677 (1999)
8. Bair, E., Tibshirani, R.: Semi-supervised methods to predict patient survival from gene expression data. PLOS Biol. **2**, 511–522 (2004)
9. Bair, E., Hastie, T., Paul, D., Tibshirani, R.: Prediction by supervised principle components. J. Am. Stat. Assoc. **101**, 119–137 (2006)
10. Bakay, M., et al.: Nuclear envelope dystrophies show a transcriptional fingerprint suggesting disruption of Rb-MyoD pathways in muscle regeneration. Brain **129**(Pt 4), 996–1013 (2006)
11. Barry, W.T., Nobel, A.B., Wright, F.A.: A statistical framework for testing functional categories in microarray data. Ann. Appl. Stat. **2**(1), 286–315 (2008)
12. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B **57**, 289–300 (1995)
13. Benjamini, Y., Hochberg, Y.: On the adaptive control of the false discovery rate in multiple testing with independent statistics. J. Edu. Behav. Stat. **25**, 60–83 (2000)
14. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Stat. **29**, 1165–1188 (2001)
15. Ter Braak, C.J.F.: Interpreting canonical correlation analysis through biplots of structure correlations and weights. Psychometrika **55**(3), 519–531 (1990)
16. Chen, X., Wang, L., Smith, J.D., Zhang, B.: Supervised principle component analysis for gene set enrichment of microarray data with continuous or survival outcome. Bioinformatics **24**(21), 2474–2481 (2008)
17. Debashis, P., Bair, E., Hastie, T., Tibshirani, R.: "Preconditioning" for feature selection and regression in high-dimensional problems. Ann. Stat. **36**(4), 1595–1618 (2008)
18. Diaconis, P.: Patterns in eigenvalues: The 70th Josiah Willard Gibbs Lecture. Bull. AMS **40**(2), 155–178 (2003)
19. National Centre for Biotechnology Information's database Gene Expression Omnibus (GEO): http://www.ncbi.nlm.nih.gov/geo/
20. Gabriel, K.R.: The biplot graphic display of matrices with application to principal component analysis. Biometrika **58**, 453–467 (1971)
21. Gabriel, K.R.: Biplot. In: Kotz, S., Johnson, N.L.: (eds.) Encyclopedia of Statistical Sciences, vol. 1, pp. 263–271. Wiley, New York (1982)
22. Gower, J.C., Hand, D.J.: Biplots. Monographs on Statistics and Applied Probability 54. Chapman & Hall, London (1996)
23. Hotelling, H.: Analysis of a complex of statistical variables into principal components. J. Educ. Psychol. **24**, 417–441; 498–520 (1933)
24. Pearson, K.: On lines and planes of closest fit to systems of points in space. Phil. Mag. **2**(6), 559–572 (1901)
25. Johnstone, I.M.: On the distribution of the largest eigenvalue in principle components analysis. Ann. Stat. **29**(2), 295–327 (2001)
26. Johnston, I.M.: High dimensional statistical inference and random matrices. Proceedings of the International congress of Math. Madrid, Spain 2006, (EMS 2007).
27. Kanehisa, M., Goto, S.: KEGG:Kyoto Encyclopedia of Genes and Genomes. Nucleic Acid Res. **28**, 27–30 (2000)
28. Karhunen, K.: Über lineare Methoden in der Wahrscheinlichkeitsrechnung. Ann. Acad. Sci. Fennicae. Ser. A. I. Math.-Phys. **37**, 1–79 (1947)
29. El Karoui, N.: Spectrum estimation for large dimensional covariance matrices using random matrix theory. Ann. Stat. **36**(6), 2757–2790 (2008)
30. Khatri, P., Draghici, S.: Ontological analysis of gene expression data: Current tools, limitations, and open problems. Bioinformatics **21**(18), 3587–3595 (2005)

31. Kim, B.S., et al.: Statistical methods of translating microarray data into clinically relevant diagnostic information in colorectal cancer. Bioinformatics **21**, 517–528 (2005)
32. Kong, S.W., Pu, T.W., Park, P.J.: A multivariate approach for integrating genome-wide expression data and biological knowledge. Bioinformatics **22**(19), 2373–2380 (2006)
33. Loève, M.: Probability theory, vol. II, 4th edn. Graduate Texts in Mathematics, vol. 46. Springer, New York (1978). ISBN 0-387-90262-7.
34. Mirsky, L.: Symmetric gauge functions and unitarily invariant norms. Q. J. Math. **11**(1), 50–59 (1960)
35. The Broad Institute's Molecular Signatures Database (MSigDB): http://www.broadinstitute.org/gsea/msigdb/
36. Mootha, V.K., et al.: Pgc-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat. Genet. **34**, 267–273 (2003)
37. Nilsson, J., Fioretos, T., Höglund, M., Fontes, M.: Approximate geodesic distances reveal biologically relevant structures in microarray data. Bioinformatics **20**(6), 874–880 (2004)
38. Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A., Ploner, A.: False discovery rate, sensitivity and sample size for microarray studies Bioinformatics **21**(13), 3017–3024 (2005)
39. Rao, C.R.: Separation theorems for singular values of matrices and their applications in multivariate analysis. J. Multivar. Anal. **9**, 362–377 (1979)
40. Rasch, D., Teuscher, F., Guiard, V.: How robust are tests for two independent samples? J. Stat. Plann. Inference **137**, 2706–2720 (2007)
41. Rivals, I., Personnaz, L., Taing, L., Potier, M.-C.: Enrichment or depletion of a GO category within a class of genes: Which test? Bioinformatics **23**(4), 401–407 (2007)
42. Rocke, D.M., Ideker, T., Troyanskaya, O., Queckenbush, J., Dopazo, J.: Editorial note: Papers on normalization, variable selection, classification or clustering of microarray data. Bioinformatics **25**(6), 701–702 (2009)
43. Ross, M.E., et al.: Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. Blood **102**(8), 2951–2959 (2003).
44. Qlucore Omics Explorer, Qlucore AB, www.qlucore.com
45. Spira, A., et al.: Effects of Cigarette Smoke on the Human Airway Epithelial Cell Transcriptome. Proc. Natl. Acad. Sci. **101**(27), 10143–10148 (2004)
46. Stewart, G.W.: On the early history of the singular value decomposition. SIAM Rev. **35**(4), 551–566 (1993)
47. Storey, J.D.: A direct approach to false discovery rates. J.R. Stat. Soc. Ser. B **64**, 479–498 (2002)
48. Storey, J.D., Tibshirani, R.: Statistical significance for genomewide studies. Proc. Natl. Acad. Sci. USA **100**, 9440–9445 (2003)
49. St. Jude Children's Research Hospital: http://www.stjuderesearch.org/data/ALL3/index.html
50. Subramanian, A., et al.: Gene set enrichment analysis: A knowledgebased approach for interpreting genome wide expression profiles. Proc. Natl. Acad. Sci. USA **102**, 15545–15550 (2005)
51. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290**, 2319–2323 (2000)
52. Troyanskaya, O., et al.: Missing value estimatin methods for DNA microarrays. Bioinformatics **17**(6), 520–525 (2001)
53. Yin, Y., Soteros, C.E., Bickis, M.G.: A clarifying comparison of methods for controlling the false discovery rate. J. Stat. Plan. Inference **139**, 2126–2137 (2009)
54. Yin, Y.Q., Bai, Z.D., Krishnaiah, P.R.: On the limit of the largest eigenvalue of the large dimensional sample covariance matrix. Probab. Theory Relat. Field **78**, 509–521 (1988)

Chapter 7
Shock-Wave Behaviour of Sedimentation in Wastewater Treatment: A Rich Problem

Stefan Diehl

Abstract A common industrial process for separating particles from a liquid is continuous sedimentation, which is used in the chemical, mining, pulp-and-paper and food industries. It can also be found in most wastewater treatment plants, where it is a crucial subprocess of a complex biological system. The process has provided, and will continue to provide, scientific problems that lead to fundamental research in different disciplines such as mathematics, wastewater, chemical, mineral, control and automation engineering. A selective survey of previous results within the field of pure and applied mathematics is presented with focus on a nonlinear convection-diffusion partial differential equation with discontinuous coefficients. In a model of a wastewater treatment plant, such an equation is coupled to a set of ordinary differential equations. Some new results on the steady-state solutions of such a coupled system are also presented.

7.1 Introduction

7.1.1 Problem Origin

In this paper, we describe how a real-world problem has initiated research in different disciplines, such as the wastewater, chemical, mineral, control and automation engineering, but we focus on applied and pure mathematics. The investigations comprise modelling, well-posedness issues, numerical analysis, inverse problems (identification of constitutive relations) and control problems.

The origin is a common industrial process that has been used for a about a century: *continuous sedimentation*, a separation process in which a suspension is

S. Diehl (✉)
Centre for Mathematical Sciences, Lund University, P.O. Box 118, SE-221 00 Lund, Sweden
e-mail: diehl@maths.lth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_7,

separated into clarified liquid and thickened slurry by gravity under a continuous flow. It is used in most wastewater treatment plants and in the chemical, mining, pulp-and-paper and food industries. The process is known to behave in a nonlinear way and it is therefore difficult to model and control.

Despite several idealized assumptions, the conservation of mass leads inevitably to a nonlinear convection-diffusion partial differential equation (PDE), which has coefficients that are discontinuous functions of the one-dimensional spatial variable:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\big(F(u,x,t)\big) = \frac{\partial}{\partial x}\left(\gamma(x)\frac{\partial}{\partial x}D(u)\right) + s(t)\delta(x). \tag{7.1}$$

It should be interpreted in the weak sense. The unknown concentration $u(x,t)$ is a function of depth x and time t, s is a source function modelling the inflow of suspension and δ is the Dirac measure. The convective flux function F is composed of a constitutive relation on the settling of particles combined with two volumetric flow rates. The diffusion function D models the compression of the network of particles at high concentrations and γ is a characteristic function of the interval that corresponds to the vertical extension of the sedimentation vessel. In the case $D \equiv 0$, (7.1) becomes a *first-order hyperbolic model*; a conservation law that has solutions with shock waves. In the general case, the diffusion term is nonzero only when the solution u is greater than a critical concentration, above which the particles are in contact with each other. Then (7.1) is referred to as the *second-order parabolic model*, which is strongly degenerate. A key problem is that discontinuities appear in the solution, partly in regions where the equation is hyperbolic, partly on the interface between hyperbolic-parabolic regions, partly at locations where the coefficients have spatial discontinuities. The tools for the hyperbolic type of equation has evolved only since the 1990s and for the degenerate parabolic equation the last decade. For general overviews of hyperbolic conservation laws, see e.g. [15, 140].

7.1.2 How it Started

In 1987, I was an undergraduate student, seeking an interesting problem for the master thesis project. My teacher and, as it should turn out, supervisor and colleague for many years ahead Professor Gunnar Sparr had several suggestions. One of these was the problem of modelling continuous sedimentation in wastewater treatment, introduced to us by Professor Gustaf Olsson, an internationally driving force within the wastewater treatment research field (see e.g. his latest book contribution [151]). The master's thesis [67] got an award and resulted in the publication [86]. This was the starting point of a series of contributions within the different disciplines mentioned above. A key ingredient in my licentiate's thesis [68] in 1992 and my PhD thesis [69] in 1995 was a uniqueness condition for the first-order hyperbolic model, with which local results on existence and uniqueness were established with

support from Gunnar Sparr, see [70–72, 85]. A decade later, the well-posedness for global solutions was established and also extended to the second-order model (7.1) by Bürger, Karlsen and colleagues in [29, 30, 128]. Recently, their uniqueness result has been improved with a new uniqueness condition for equations of the type (7.1), which is in fact a generalization of the one in my PhD thesis, see Diehl [83].

7.1.3 Outline of Paper

This paper is organized as follows. The physical processes are described in Sect. 7.2 and the goals of applied mathematics research when attacking the connected problems. Described from the author's point of view, a survey of previous results can be found in Sect. 7.3. Section 7.4 is more technical, presenting some new results on the steady-state solutions of a simple model of the activated sludge process (ASP) in a wastewater treatment plant.

7.2 The Processes and Goals of Applied Mathematics

7.2.1 Batch Sedimentation

Consider a homogeneous suspension of particles in a liquid in a closed cylindrical column. *Batch sedimentation* means that the particles settle in the liquid under the force of gravity. A result of a simple experiment by the author is shown in Fig. 7.1.

A batch sedimentation test was recorded on video with an image every 4 s. Clearly visible is the uppermost declining interface, a shock wave, between the clear liquid and suspension. If the initial concentration is not too high, there is also a discontinuity rising from the bottom, which cannot usually be detected by the eye.

If the suspension contains incompressible particles, e.g. monosized glass beads, the increasing layer at the bottom consists of the maximum concentration of particles. This implies that the rising discontinuity from the bottom is a straight line until it meets the uppermost interface, and after that time point, there is a final stationary discontinuity between clear liquid and the maximum packing concentration. This behaviour is captured by a hyperbolic first-order model PDE, for which the characteristics are straight lines and shock waves appear in the solution, see Sect. 7.3.1.

The particles of most industrial suspensions are, however, compressible. Above a critical concentration, where the particles are in constant contact with each other, compressive forces appear and the model equation is parabolic. For a batch settling test, this means that the discontinuity rising from the bottom becomes concave and that the concentration of the bottom sediment increases with depth due to the weight of the material above.

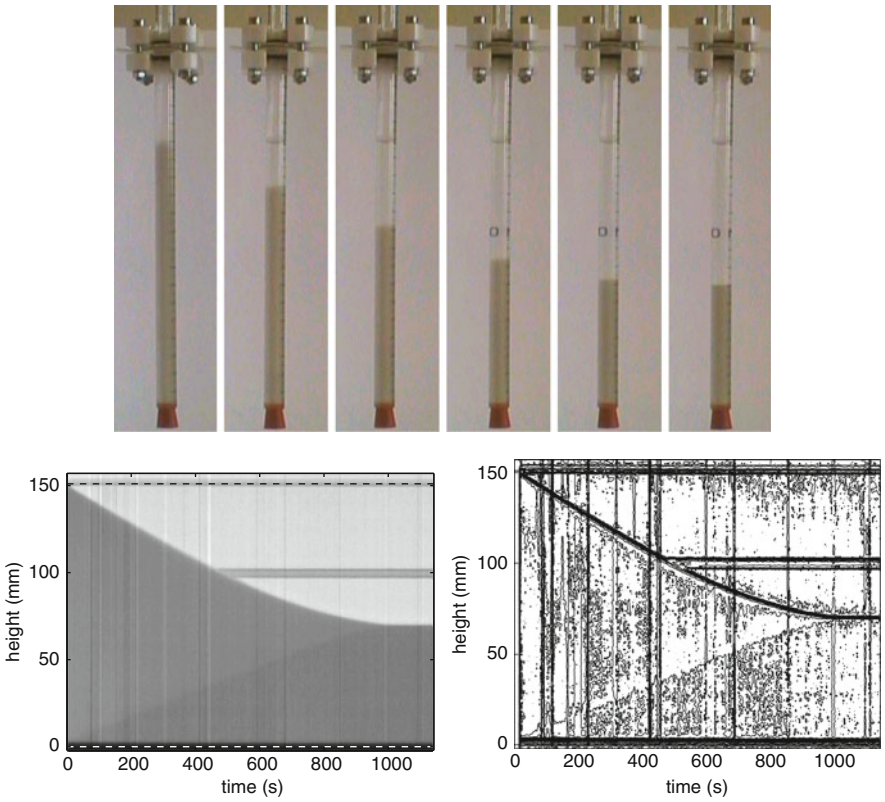


Fig. 7.1 The time evolution of a batch sedimentation test with 20–32 μm silica in water and the initial homogeneous concentration 623 g l^{-1}. A matrix of grey levels is shown below to the *left* and the contours of the matrix to the *right*. Note that the rising sediment-suspension discontinuity from the bottom becomes clearly visible with simple image processing. The polydispersity is probably one reason why this discontinuity is not a straight line

7.2.2 Continuous Sedimentation in Wastewater Treatment

In diverse industrial processes there are flowing particle-liquid suspensions and it is of importance to separate the two components continuously in time. The aim may be to obtain a concentrated slurry from a dilute suspension. This is called *thickening*. If the main aim is to remove solid particles from a suspension to obtain a clear liquid, the process is called *clarification*. The process *continuous sedimentation* involves both clarification and thickening under the continuous removal of liquid and slurry and takes place in a clarifier-thickener unit. Other names of such a sedimentation tank are gravity thickener, clarifier, or just *settler*. In most wastewater treatment plants one can partly find a primary settler for the removal of course particles, partly a secondary settler within the *activated sludge process* (ASP), see Fig. 7.2.

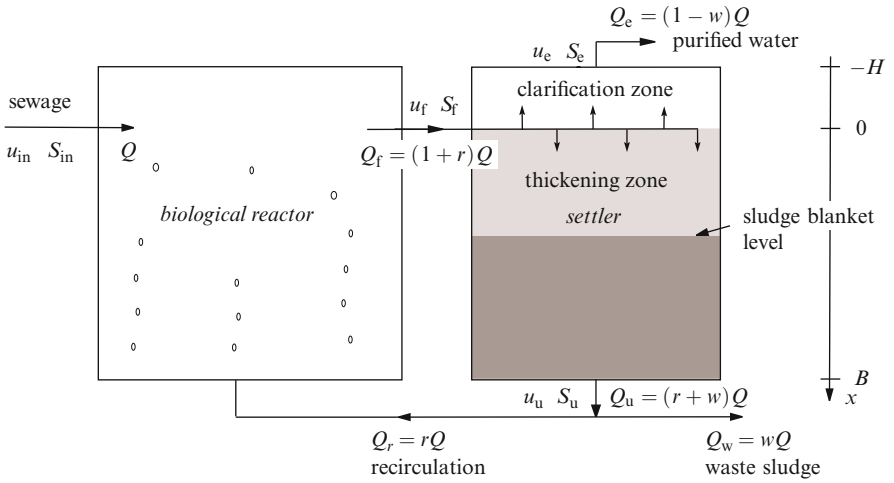


Fig. 7.2 The activated sludge process (ASP) consists of a biological reactor and a sedimentation tank (settler). The settler consists of a clarification zone corresponding to the interval $(-H, 0)$ and a thickening zone in the interval $(0, B)$. The indices stand for f = feed, e = effluent, u = underflow, r = recycle and w = waste

The incoming sewage consists mostly of organic material and nutrients, the *substrate* or soluble material, represented by the S-variables in Fig. 7.2. In the biological reactor, this is consumed and decomposed by microorganisms, the *biomass* or particulate material, represented by u in Fig. 7.2. The substrate and biomass consist each of several components. In the reduced-order model in Sect. 7.4, we consider however only one substrate and one biomass component, i.e. all variables in Fig. 7.2 are then scalars. The biological reactor often consists of several compartments, classified according to the form of the available oxygen (anaerobic, aerobic and anoxic). The incoming biomass concentration u_{in} is usually small or negligible. The main objective of the ASP is to keep the effluent substrate concentrations less than prescribed reference values.

In the secondary settler, the flocculated biomass particles settle slowly, since they have only a slightly higher density than the water. This settler is crucial in a wastewater treatment plant and the most difficult subprocess to control. Besides clarification and thickening, a third purpose of the settler is to be a *buffer of mass* in the ASP, since most of the underflow of the settler is recycled to the biological reactor.

Input variables to the ASP are the volumetric flow rate Q and the substrate (S_{in}) and biomass (u_{in}) concentrations. *Outputs* are the effluent and the underflow concentrations (u_e, S_e, u_u, S_u). The underflow rate of the settler is $Q_u = (r+w)Q$, where the recycle and waste ratios are defined by

$$r := \frac{Q_r}{Q}, \qquad w := \frac{Q_w}{Q}.$$

These two ratios are the main control parameters for the entire ASP. The waste ratio w has to be nonzero since the microorganisms grow in the biological reactor and one purpose of the process is that they should not leave the effluent flow, i.e. u_e should be zero. On the other hand, w should be small to minimize the amount of waste sludge for practical, economical and environmental reasons. A large value of r may cause turbulence and thereby unwanted disturbances in the settler. Orders of magnitude are usually $r \approx 1$ and $w \lesssim 0.01$. The supply of oxygen and carbon to the biological reactor are normally used for control of the biological processes. For the simple model in Sect. 7.4, we assume that the biological reactor consists of only one aerobic reactor and that the supply of oxygen is sufficiently high.

Under normal operating conditions there is a large concentration discontinuity in the thickening zone, which position is called the *sludge blanket level* (or the sediment level in other applications). As for the settler, the main control problem is to maintain the sludge blanket level as the inputs Q, S_{in} and u_{in} vary with time. Then the effluent particulate concentration remains low, preferably zero, and the underflow concentration high.

Even for an idealized one-dimensional settler (with constant cross-sectional area, inlet and outlets at points), there are still two reasons for the nonlinearities:

- *Hindered sedimentation*: The settling speed of the particles relative to the liquid is a nonlinear function of the concentration (Kynch's assumption). This leads to a nonlinear hyperbolic conservation law, see (7.3) in Sect. 7.3.1.
- *Compression*: The effective solid stress for the network of particles, which is formed above a critical concentration, is a nonlinear function of the concentration. Together with Kynch's assumption this leads to a strongly degenerate parabolic PDE, see (7.4) in Sect. 7.3.1.

These physical nonlinear phenomena imply non-uniqueness of solutions of the modelling equation. In addition, we have the following complication:

- *Convective flows:* The flows of the inlet and outlets of the settler imply that there are space-discontinuous coefficients of the model PDE, see Sect. 7.3.3. Even for a soluble material that follows the water in the settler and has a linear model equation (since it undergoes neither sedimentation nor compression), the space-discontinuous coefficients cause non-uniqueness of solutions, see Sect. 7.4.1.

These facts are the main reasons for the problems of modelling and controlling the continuous sedimentation process in a stand-alone settler. This is the case in, for example, the mineral industry.

7.2.3 Goals of Applied Mathematics Research

Concerning a stand-alone settler, we may state the following goals of the applied mathematics research:

Modelling I: equations. Make simplifying assumptions and create a mathematical model, in this case a PDE, which comes out from a physical law and constitutive assumptions.
Well-posedness. Establish the well-posedness of the model, i.e. existence and uniqueness of solutions. Because of the presence of shock waves in the solution of the PDE, the stability of such is related to the choice of uniqueness condition.
Numerical methods and simulation tools. Develop reliable numerical methods to be used for simulation of different scenarios. By reliable is meant that the numerical approximate solutions converge to the exact solution of the PDE model as the size of the mesh, used for the numerical computations, tends to zero. This is often neglected in the applied fields, where the need for simulation models has forced ad hoc assumptions and adjustments in the numerical methods. This is particularly dangerous for solutions having discontinuities.
Manual control. Formulate control objectives mathematically and establish corresponding control strategies, i.e. define the control variables as functions of the input data and the present state of the system. This requires a categorization of the behaviour of the nonlinear system for all types of input data with respect to stationary solutions, step responses and limitations for control.
Automatic control. Obtain automatic control of the model process and ultimately of the real process.
Modelling II: inverse problems. Develop methods to determine the constitutive relations from measured data. Specially designed experiments can be evolved and yield information in addition to real process data. This leads to the inverse problem of estimating parameters in the PDE. This is particularly important for biological sludges, for which the settling and compression properties may change with time. The ultimate goal is to perform estimations in real time from on-line data in order to achieve automatic control.

For the coupled settler within the ASP, the situation becomes even more complicated. Some additional features, which should be added to each of the goals above, are then the following:

- *The processes within the biological reactor*: There are standard nonlinear ODE models, which are considered to be satisfactory although the identification of parameters is difficult. A conventional model consists of 13 ODEs, see Henze et al. [116].
- *The link from the reactor to the settler*: The 13 unknowns of the ODE system correspond to different particulate and soluble materials. As the sludge leaves the biological reactor and enters the settler, it is normally assumed that the biological reactions cease and the particulate material has formed larger flocs. The percentages of the components of the flocs depend on time as they enter the settler. It is the concentration of the flocs that is the scalar feed concentration u_f to the settler in Fig. 7.2. Inside the settler, the percentages are governed by a system of PDEs, see (7.16) for the hyperbolic case. The analogous situation holds for the soluble material with the simplifying fact that the convective fluxes do not contain any nonlinear sedimentation component (cf. Sect. 7.4).

- *The recirculation from the settler to the reactor*: The underflow concentrations u_u and S_u are parts of the solutions of the PDEs and inputs to the ODEs. Furthermore, the control parameter r influences all equations.

7.3 A Selective Overview of Previous Results

The need for modelling continuous sedimentation arose as soon as the process was invented in 1905 by Dorr [88]. For a historical perspective of the twentieth century, we refer to Concha and Bürger [57].

7.3.1 Modelling Batch Sedimentation

Batch sedimentation is performed in a closed column, where the suspension settles without any applied bulk flows, see Fig. 7.1. For a constant cross-sectional area and with $v_s \geq 0$ denoting the downwards settling speed of particles, the conservation of mass leads to the conservation law

$$u_t + (uv_s)_x = 0 \quad \text{or} \quad u_t - (uv_s)_z = 0, \tag{7.2}$$

where x is a downwards and z an upwards pointing axis. (In continuous sedimentation, the spatial axis is often defined downwards and in batch sedimentation upwards, cf. Fig. 7.3). The *first-order model of batch sedimentation* was presented in 1952 in the pioneering work by Kynch [135]. Kynch's constitutive assumption is that the settling speed v_s depends only on the local concentration u. A common assumption is that v_s is a decreasing function since the hindrance to free settling of a single particle increases with the concentration. Hence, the batch-settling flux function $f_b(u) = uv_s(u)$ is nonlinear, see Fig. 7.3 (left). Equation (7.2) then becomes

$$u_t + f_b(u)_x = 0 \quad \text{or} \quad u_t - f_b(u)_z = 0. \tag{7.3}$$

Kynch's model can also be obtained as a special one-dimensional case of a multi-dimensional sedimentation model, see Bürger et al. [21] and the second-order model below. Kynch's assumption turns out as a result of a constitutive assumption in a linear momentum equation: a postulated nonlinear concentration-dependent factor relates the solid-liquid interaction force to the solid-liquid relative velocity.

Ahead of the formalization of entropy solutions and based more on physical insight, Kynch used the method of characteristics to construct solutions. The solution of (7.3) is generally built up by patches where the solution is piecewise smooth and such that a concentration value u_0 is propagated along straight lines – the characteristics – of slope $f_b'(u_0)$, see Fig. 7.3 (right). This can be seen by considering

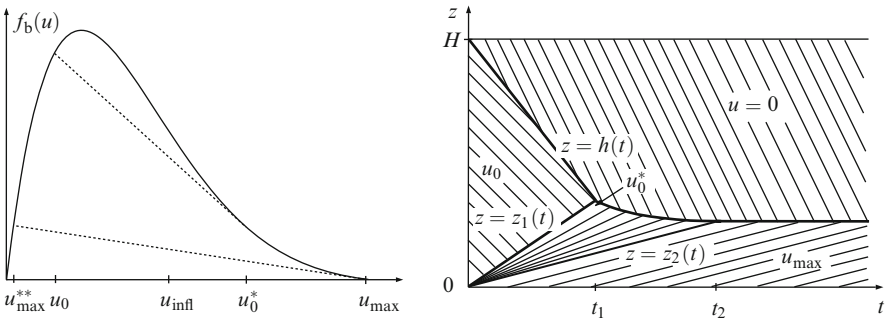


Fig. 7.3 *Left*: A batch-settling flux curve with one inflection point u_{infl}. The mappings $u_0 \mapsto u_0^*$ and $u_{\max} \mapsto u_{\max}^{**}$ are related to tangents to the curve. *Right*: A schematic solution of a standard batch-settling test with the initial homogeneous concentration $u_0 \in (u_{\max}^{**}, u_{\text{infl}})$. *Thin lines* are characteristics. Besides the point $(u_0, f_b(u_0))$ on the flux curve, the interval of estimation in this case is $[u_0^*, u_{\max}]$

a contour $x = x(t)$ that satisfies $u\big(x(t), t\big) = u_0$. Differentiation yields

$$0 = \frac{d}{dt} u\big(x(t), t\big) = u_t + x'(t) u_x.$$

Comparing with (7.3), we see that $x'(t) = f_b'(u) = f_b'(u_0)$ holds. As characteristics with different concentration intersect, a discontinuity appears. This may occur even for smooth initial data. The speed of a discontinuity is governed by the jump condition, which is a direct consequence of the weak formulation of (7.3). Discontinuous solutions may, however, not be unique for given initial data. A physically relevant discontinuity, a shock wave, satisfies the entropy inequality by Oleinik [150]. This ensures a unique solution for given initial data. The construction of solutions of (7.3) in the case of standard batch-settling tests can be found in the book by Bustos et al. [44, Chap. 7]. Kynch's paper was the most important one for the modelling and analysis of sedimentation in the twentieth century. His constitutive assumption has been the most widely used, despite the fact that it is only valid for ideal incompressible particles, such as monosized glass beads.

For most industrial suspensions, the settling particles form a compressible network above a critical concentration. This phenomenon can be captured by another constitutive relationship in addition to the one by Kynch, namely the *effective solid stress* σ_e, or compressive yield stress, as a function of the concentration, see [8,19,21,40,41,58,64,118,136]. A common assumption is that σ_e and its derivative should satisfy

$$\sigma_e(u) \text{ and } \sigma_e'(u) \begin{cases} = 0, & u \le u_c, \\ > 0, & u > u_c, \end{cases}$$

where u_c is the critical concentration above which the particles are in constant contact with each other. In particular, we highlight the sedimentation-consolidation theory by Bürger et al. [21], who carefully derived the following expression for the settling speed of the particles:

$$v_s = \frac{f_b(u)}{u}\left(1 - \frac{\rho_s}{u g_a \Delta\rho}\frac{\partial \sigma_e(u)}{\partial x}\right).$$

Substituting this into (7.2), the following model PDE is obtained, which we may call the *second-order model of batch sedimentation*:

$$u_t + f_b(u)_x = D(u)_{xx} \quad \text{or} \quad u_t - f_b(u)_z = D(u)_{zz}. \tag{7.4}$$

Here, the diffusion function is

$$D(u) := \int_0^u d(v)\,dv \quad \text{with} \quad d(u) := \frac{\rho_s f_b(u)\sigma_e'(u)}{u g_a \Delta\rho}, \tag{7.5}$$

where ρ_s is the density of the solids, $\Delta\rho$ is the density difference between solids and liquid and g_a is the acceleration of gravity. Equation (7.4) is strongly degenerate parabolic, since $D(u)$ disappears for u less than the critical concentration. Then the equation becomes the hyperbolic one (7.3), and we note that the convective flux f_b is precisely the one by Kynch. Particularly thorough analyses of (7.4) have been presented by Bürger et al. [17, 18, 22, 23] and Karlsen et al. [94, 127, 128]. There are several published results that show good agreement between this model and industrial suspensions, e.g. [20, 66, 105]. In Fig. 7.4, numerical simulations of standard batch tests are shown without and with diffusion. For the numerical convective flux, Godunov's method [111] has been used, and for the diffusion term, central differencing.

7.3.2 *Inverse Problems of Batch Sedimentation*

The inverse problem of determining the two constitutive relations (Kynch's settling velocity function and the effective solid stress function) from experimental data is of vital importance for the reliability of the PDE models in both batch and continuous sedimentation. A common approach is to assume explicit expressions for these two functions, depending on a number of parameters, see e.g. [13, 20, 59, 66, 105]. A common way of estimating the parameters is to minimize an objective functional that measures some suitably chosen L^2 distance between a numerical approximate solution of the PDE and measured data. These publications indicate that the inverse problem is both ill-posed and badly conditioned, and that more research is needed.

The batch-settling flux function has usually at least one inflection point, which implies that the determination of the whole function is more delicate. The most

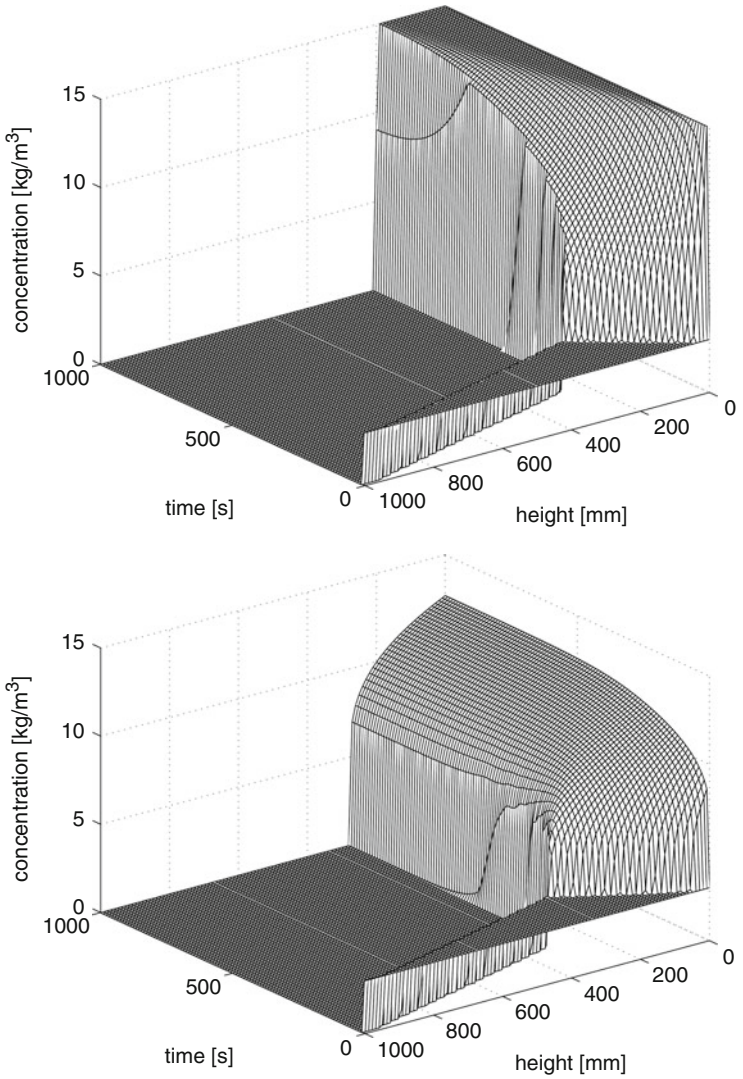


Fig. 7.4 Simulations of batch sedimentation tests without (*upper*) and with (*lower*) diffusion. The initial concentration is $u_0 = 3\,\mathrm{kg\,m^{-3}}$ and the flux function is $f_\mathrm{b}(u) = 10u(e^{-0.35u} - e^{-0.35u_\mathrm{max}})\,\mathrm{m\,h^{-1}}$ with $u_\mathrm{max} = 15\,\mathrm{kg\,m^{-3}}$. The effective solid stress functions used are for the *upper* figure: $\sigma_\mathrm{e} \equiv 0$ and for the *lower*: $\sigma_\mathrm{e}(u) = 0$ for $u \le u_\mathrm{c}$; $= (u - u_\mathrm{c})^4/4$ for $u > u_\mathrm{c}$, where $u_\mathrm{c} = 4\,\mathrm{kg\,m^{-3}}$. In (7.5), the following constants have been used: $\rho_\mathrm{s} = 1,050\,\mathrm{kg\,m^{-3}}$ (biomass density), $\Delta\rho = 52\,\mathrm{kg\,m^{-3}}$ and $g_\mathrm{a} = 9.81\,\mathrm{m\,s^{-2}}$

common approach has been the graphical method by Kynch [135], with which a portion of the flux curve to the right of the (smallest) inflection point can be estimated from only one experiment. This has been used by e.g. [16, 87, 97, 98, 101–104, 168, 179, 183]. It is only Lester et al. [138] that present an explicit formula for

this part of the batch-settling flux function in terms of measurable variables. Their formula involves an integral over the measured settling velocities.

It is, however, possible to derive a simpler explicit formula for this part of the flux function by using Kynch's own arguments. Consider Fig. 7.3 (right). For $t_1 \le t \le t_2$, denote the concentration just below the interface $z = h(t)$ by

$$u_h(t) := u\big(h(t) - 0, t\big), \quad t_1 \le t \le t_2.$$

This increases from $u_h(t_1) = u_0^*$ to $u_h(t_2) = u_{\max}$, cf. Fig. 7.3 (left). We need two facts that relate the interface height $h(t)$ with the batch flux function and the concentration $u_h(t)$. The first one is that the characteristics in the expansion wave (between z_1 and z_2 in Fig. 7.3 (right)) all go through the origin and their slopes satisfy (cf. Sect. 7.3.1)

$$\frac{h(t)}{t} = -f_{\mathrm{b}}'\big(u_h(t)\big), \quad t_1 \le t \le t_2. \tag{7.6}$$

The second one is that the downward speed of the interface is equal to the settling speed of the uppermost particles:

$$-h'(t) = v_{\mathrm{s}}\big(u_h(t)\big) = \frac{f_{\mathrm{b}}\big(u_h(t)\big)}{u_h(t)}, \quad t_1 \le t \le t_2. \tag{7.7}$$

Consider now a fixed time point $t \in (t_1, t_2)$ and the characteristic in the expansion wave that goes from the origin to the point $(t, h(t))$. Along this characteristic, the concentration has the constant value $u_h(t)$ and the speed of the characteristic upwards is $-f_{\mathrm{b}}'(u_h(t)) > 0$. All particles pass this characteristic with the constant relative speed $v_{\mathrm{s}}(u_h(t)) - f_{\mathrm{b}}'(u_h(t))$ and, hence, constant flux

$$u_h(t)\left[v_{\mathrm{s}}\big(u_h(t)\big) - f_{\mathrm{b}}'\big(u_h(t)\big)\right]$$

(mass per unit area and unit time). The total mass per unit area of all particles is Hu_0. During the time period $[0, t]$ all particles have passed this characteristic. This fact, together with (7.6) and (7.7), yields

$$\begin{aligned} Hu_0 &= u_h(t)\left[v_{\mathrm{s}}\big(u_h(t)\big) - f_{\mathrm{b}}'\big(u_h(t)\big)\right] t \\ &= u_h(t)\left[-h'(t) + \frac{h(t)}{t}\right] t = u_h(t)\left[h(t) - th'(t)\right]. \end{aligned} \tag{7.8}$$

This corresponds to Kynch [135, (15)]. Letting $\eta(t) := h(t) - th'(t)$, (7.7) and (7.8) constitute a parametrization of a part of the batch-settling flux function expressed in measurable variables:

$$\begin{cases} u = \dfrac{Hu_0}{\eta(t)} \\ f_{\mathrm{b}}(u) = -\dfrac{Hu_0}{\eta(t)} h'(t) \end{cases} \qquad t_1 \le t \le t_2. \tag{7.9}$$

This formula can also be obtained from the hyperbolic PDE theory, see Diehl [79]. It is easy to see that η is invertible. Hence, an explicit representation of the batch-settling flux function in measurable variables is

$$f_{\mathrm{b}}(u) = -uh'\left(\eta^{-1}\left(\frac{Hu_0}{u}\right)\right), \quad u_0^* \le u \le u_{\max}, \quad \text{where} \quad u_0^* = \frac{Hu_0}{\eta(t_1)}. \tag{7.10}$$

It is remarkable that despite the fact that Kynch's paper has been cited several hundred times, the formulae (7.9) and (7.10) have not been discovered before.

For the flux curve on the left of the inflection point, only a single point has been obtained for each standard batch test in all references above. With a new type of batch test, launched by the author in [79], it is possible to estimate a large part of the flux function containing the maximum point in only one test. Explicit formulae analogous to (7.9) are also given in [79]. Further studies on this topic already utilize (7.9), see Grassia et al. [112].

7.3.3 Modelling Continuous Sedimentation

7.3.3.1 Ad Hoc Simulation Models

The urgent need for dynamic models in the wastewater research community has led to empirically developed simulation models. In order to fit the model to some set of experimental data, various ad hoc assumptions have been used. There are several published such models, e.g. [1,53,62,65,90–92,110,114,115,133,149,154,155,159,160,167,175,178,181,184–187,189]. The most commonly used one is the so-called Takács model, see Takács et al. [167], which in fact is the model by Vitasovic [180] with a special settling velocity function.

With sufficiently many parameters in a model it is possible to fit simulations to some set of data. However, if a new set of data requires substantially different values of the parameters, then such a simulation model does not catch the fundamental physics properly and is not reliable. Another example of unreliable behaviour of ad hoc models is that the approximate solutions is qualitatively different for different mesh sizes of the grid used for the numerical computations. The fundamental problems of the Takács model and comparisons with the author's simulation model, which is based on PDE theory, see [74, 84], can be found in Jeppsson and Diehl [120].

7.3.3.2 Models Based on PDEs

Kynch's model has provided a platform in the water research field called the *solids-flux theory*. From this, many conclusions have been drawn mostly by graphical constructions by using the batch-settling flux curve f_b. Fundamental such results for obtaining concentrations in steady-state operation were presented by Jernqvist [122–124]. Unfortunately, his results seem not to have reached other researchers. Similar developments, but not as extensive, were presented in the 1960s onwards with concepts such as the operating line, the limiting flux and the state point (pivot point, feed point) in so-called operating charts, see [76,92] and the references therein. The results presented in this context have been obtained by direct physical considerations, without utilizing the method of characteristics to construct solutions as Kynch did.

In the thickening zone of the settler, the downward flux or particles (mass per unit time and area) is a superposition of the batch settling flux and the volumetric flux of the suspension. With a constant cross-sectional area A, the hyperbolic model (7.3) then becomes

$$u_t + f\big(u, Q_\mathrm{u}(t)\big)_x = 0 \tag{7.11}$$

$$\text{where} \quad f\big(u, Q_\mathrm{u}(t)\big) := f_\mathrm{b}(u) + \frac{Q_\mathrm{u}(t)}{A}u = \left(v_\mathrm{s}(u) + \frac{Q_\mathrm{u}(t)}{A}\right)u.$$

Construction of solutions by the method of characteristics, implying some analysis of the sludge blanket and the prediction of the underflow concentration u_u, was made by Petty [158], Bustos et al. [42,43,45,46] and Diehl et al. [67,86]. The reason for the restriction to the thickening zone was the lack of mathematical results for the inclusion of the inlet and outlets of the settler.

Modelling the inlet as a point source at $x = 0$ and the flux of the suspension in the two outlet pipes by convective transport terms, one arrives at the following hyperbolic PDE:

$$u_t + F(u,x,t)_x = \frac{Q_\mathrm{f}(t)}{A}u_\mathrm{f}(t)\delta(x), \quad x \in \mathbb{R}, \tag{7.12}$$

where the total flux function is

$$F(u,x,t) := \begin{cases} -\frac{Q_\mathrm{e}(t)}{A}u, & x < -H, \\ f_\mathrm{b}(u) - \frac{Q_\mathrm{e}(t)}{A}u =: g\big(u, Q_\mathrm{e}(t)\big), & -H < x < 0, \\ f_\mathrm{b}(u) + \frac{Q_\mathrm{u}(t)}{A}u =: f\big(u, Q_\mathrm{u}(t)\big), & 0 < x < B, \\ \frac{Q_\mathrm{u}(t)}{A}u, & x > B. \end{cases} \tag{7.13}$$

Equations (7.12) and (7.13) constitute the *first-order model of the settler*. It was presented and analyzed by the author in [71], which was based on the work in

[68, 70]. Independently, Chancelier et al. [50] presented the same equations, but smoothed the spatial discontinuities of the flux function and the source term, so that the standard existence and uniqueness theory could be used. However, one advantage of first-order hyperbolic equations is the possibility of constructing exact solutions. In [68, 70, 72, 85], the author utilized this fact and developed an entropy condition called Condition Γ, which made it possible to give a satisfactory treatment of the inlet and outlets directly. Under some regularity assumptions, construction of unique solutions of (7.12) for given initial data was possible.

Independently, Gimse and Risebro [108] presented an entropy condition, which with a specific interpretation is equivalent to Condition Γ. Their interest in conservation laws with discontinuous flux function originated from the modelling of two-phase flow in heterogeneous porous media, in particular geological discontinuities. With these two applications as starting points, the uniqueness issue has been widely studied, see e.g. [4, 7, 9, 29, 30, 38, 70, 72, 83, 85, 109, 113, 126–128, 130–132, 145–147, 156, 163, 170, 171]. For the hyperbolic settler problem (7.12) with time independent volumetric flows and feed concentration, the existence and uniqueness of global solutions were established by Bürger et al. [29].

The construction of solutions together with Condition Γ made it possible to provide answers to many issues raised in the engineering literature, see e.g. [74, 75]. In particular, the solids-flux theory with its graphical constructions and operating charts can be completely described and extended in the direction of the control, see [76–78, 80–82]. A main ingredient of these results is the possibility of constructing exact steady-state solutions and transitions between such.

Taking into account also the compression phenomenon, one arrives instead at the degenerate parabolic PDE (7.1), the *second-order model of the settler*. Although such an equation can be found in many references, the major breakthrough for this model was presented by Bürger et al. [30]. They treated the well-posedness issues for the model and presented numerical methods for reliable simulations. Further utilization of this model concerning steady-state, control and capacity calculations is provided by Bürger and Narváez [33], and for the application to wastewater treatment by De Clercq et al. [66].

As for the construction of steady-state solutions, the entropy condition at the spatial discontinuities is of major importance. The condition utilized in [30, 33] is the Kružkov-type entropy condition by Karlsen et al. [128]. The latter studied strongly degenerate parabolic equations of the type

$$u_t + \tilde{f}\big(\boldsymbol{\gamma}(x), u\big)_x = D(u)_{xx}, \tag{7.14}$$

where $\boldsymbol{\gamma}(x)$ is a piecewise smooth and bounded vector-valued function with a finite number of discontinuities. Uniqueness was shown via L^1 stability of weak entropy solutions, provided that an additional 'crossing condition' holds. The crossing condition is an assumption on the flux function $\tilde{f}$ at each spatial discontinuity, see also Bürger, Karlsen et al. [29, 126].

In [83], the author considered the equation

$$u_t + \left(\tilde{F}(u,x) - \tilde{D}(u,x)_x\right)_x = 0, \tag{7.15}$$

with the discontinuous flux and diffusion functions

$$\tilde{F}(u,x) = \begin{cases} f_{\mathrm{l}}(u), & x < 0, \\ f_{\mathrm{r}}(u), & x > 0, \end{cases} \quad \text{and} \quad \tilde{D}(u,x) = \begin{cases} D_{\mathrm{l}}(u), & x < 0, \\ D_{\mathrm{r}}(u), & x > 0. \end{cases}$$

An entropy condition at $x = 0$ was presented, which is a reformulation and straight generalization of the previous Condition Γ. The uniqueness results of Karlsen et al. [128] were generalized in the following ways. It was concluded that Condition Γ implies the Kružkov-type condition used in [128]. By modifying the proof of L^1 stability in [128], uniqueness was proved partly without the additional crossing condition, partly with the weaker assumption that $u \in L^\infty$ instead of $u \in L^1 \cap L^\infty$. The latter is important in the application to continuous sedimentation, where the solution does not tend to zero far away (the concentration in the lower outlet pipe is usually high). An additional advantage of Condition Γ is its simple geometrical interpretation, which facilitates the construction of stationary solutions.

As for the existence of solutions and reliable numerical methods, we refer to Bürger, Karlsen and their collaborators [24, 29, 30, 126–128].

Without going into details, it should be mentioned that different types of entropy conditions have been developed for conservation laws with discontinuous flux, see [7, 31, 38, 146]. Condition Γ was justified by viscous profile analyses in [72, 85]. Within the setting of a more general theory covering several types of solution concepts, Andreianov et al. [4, 5] recently concluded that Condition Γ should be recognized as the correct admissibility criterion for the vanishing viscosity limits for this type of equation.

7.3.4 Extensions and Related Problems

7.3.4.1 Varying Cross-Sectional Area

Taking into account a cross-sectional area that varies with depth does not add any substantially new feature concerning the analysis, except for the fact that the characteristics of the corresponding hyperbolic equation will not be straight lines. The cross-sectional area may also be discontinuous at the feed inlet, see [26–28, 30, 74, 75].

7.3.4.2 Polydisperse Sedimentation

In many applications, the suspensions consist of particles of different sizes. For the modelling of sedimentation of polydisperse suspensions we mention only the recent

works by Bürger et al. [35, 36, 39] and the references therein. They have in [32, 36] also modelled the case of a sink term in the hyperbolic equation and demonstrated how it can be used for the classification of polydisperse suspensions.

7.3.4.3 Other Applications

Conservation laws with discontinuous coefficients appear also in other applications, such as the modelling of *two-phase flow* in heterogeneous media [89, 109, 125], *traffic flow* with abruptly changing surface conditions or number of lanes [25, 34, 37, 148], traffic flow on road networks [56], packet flow on *telecommunication networks* [60, 61, 144], *ion etching* in the fabrication of semiconductor devices [162, 169] and *shape-from-shading* problems (reconstruction of a 3D surface from a 2D image) [152, 153].

7.3.5 Controlling the Settler

As for the control of the settler, only a few approaches can be found in the literature before 2008; empirical ones in [11, 55, 129, 142] and more advanced ones in [12, 33, 45, 50, 51, 74, 86, 166]. They all deal with the problem of controlling the sludge blanket level under the assumption that normal operating conditions are maintained.

For the first-order model of the settler (7.12), all stationary solutions have been charted, step responses investigated, control objectives formulated and limitations for control of dynamic solutions established in the series of papers [76–78, 81]. This work of line lead the author to develop the first published regulator for control of the settler, see [80]. An analogous analysis has started with the aim of obtaining a similar regulator for the second-order model (7.1).

7.3.6 Modelling and Controlling the ASP

Despite the fact that it has long been generally accepted that the biological reactor can be modelled by a set of ODEs and the settler modelled by a PDE, no rigorous analysis of the coupled system can be found in the literature. The main reason for this has been the lack of sufficient knowledge of the first- and second-order PDE models of the settler. Nevertheless, the need for controlling the ASP has lead to different approaches. Numerous control strategies have been proposed and the focus has been laid on the processes in the biological reactor. There exist a few standard regulators in addition to more complex ones based on feed-forward and model-based predictive control.

For large biological reactors, Lee et al. [137] argue that the concentration of each component also depends on one spatial variable, and model each component by a linear convection-diffusion PDE that takes into account the plug flow and diffusion.

7.3.6.1 Approaches Using Ad Hoc Settler Models

One of the first attempts to simulate the ASP was presented in 1978 by Attir and Denn [6], who used two equations for the biological reactor (substrate and biomass) and an ad hoc numerical settler implementation. Since then, the settler has often been assumed either to satisfy a very simple model, such as a single ODE, or to behave in an assumed optimal way, such as always in steady state with a sludge blanket in the thickening zone. Some recent references with different objectives and methods of analysis are [14,48,52,93,106,107,119,134,141,161,174,176,177,188]. Some recent approaches using simulation models for the full set of equations can be found in [2, 3, 10, 47, 49, 54, 63, 95, 96, 99, 100, 117, 121, 137, 143, 151, 165, 172, 173, 182]. In most of these, the settler is simulated with a numerical scheme that has no established connection to the PDE, such as the Takács model (which produces incorrect approximate solutions, see [120]). A recent example of two-dimensional simulations of the settler coupled with a simple completely mixed reactor model is described by Patziger et al. [157]. The impact of phenomena like turbulence is then possible to study. However, their conclusions, which are based on simulations, are only rules of thumb on how to adjust the recycle ratio r.

7.3.6.2 A First-Order Simulation Model for the ASP Based on PDEs for the Settler

In the ASP, there are several biological components of the particulate material. As the sludge leaves the biological reactor and is fed into the settler, it is common to assume that the biological activities have ceased and that the particulate material is in the form of larger particles or flocs. Introducing the percentage vector $\mathbf{p}_\mathrm{f}(t)$ of these components, the feed concentrations are contained in the vector $\mathbf{p}_\mathrm{f}(t)u_\mathrm{f}(t)$, where u_f is the total (scalar) concentration of the flocs that are fed to the settler. Ignoring the compression phenomenon, the percentage vector within the settler $\mathbf{p}(x,t)$ satisfies the system of PDEs

$$\frac{\partial(\mathbf{p}u)}{\partial t} + \frac{\partial}{\partial x}\big(\mathbf{p}F(u,x,t)\big) = \frac{Q_\mathrm{f}(t)}{A}\mathbf{p}_\mathrm{f}(t)u_\mathrm{f}(t)\delta(x). \tag{7.16}$$

Note that the sum of all equations in (7.16) yields (7.12). The system (7.16) has been analyzed with respect to existence and uniqueness by the author in [73], where a numerical scheme for approximate solutions can also be found. These results made it possible to present the *first-order simulation model for the ASP*, see Diehl and Jeppsson [84]. Except for (7.12) and (7.16), the model consists of 13 ODEs for the

biological reactor. The difference between this model and all the ad hoc simulation models referred to above is that the numerical method is derived from the exact solutions of the PDEs, although no convergence results have been presented.

7.3.6.3 Reduced-Order Models

When it comes to analysis of equations, a reduced model with a few ODEs for the biological reactor is a natural starting point. An obvious first approach is to establish the steady states of the ASP and how these depend on the influent variables and the control parameters. Sheintuch [164] presents a steady-state analysis of the reactor-settler interaction using two equations for the biological reactor (substrate and biomass). For the steady states of the settler, Sheintuch uses the result of Lev et al. [139]. One shortcoming is that Lev et al. assume that all steady-state solutions are constant in the clarification and thickening zones, respectively, and thereby miss the most interesting one with a sludge blanket within the thickening zone. We comment more on [139] in Sect. 7.4. Kumar et al. [176, 177] perform steady-state analyses using three ODEs for the biological reactor (biomass and two substrates) and a limiting flux condition for the settler. However, as is done in many references above regarding ad hoc simulations of the ASP, they assume that the settler always satisfies an optimal condition no matter what the influent volumetric flow rate Q is. Since this is not true, the results are not reliable. In the next section, an analysis of a similar model of the ASP is presented. The reactor is modelled by two equations and, in contrast to all previous works, the available results for PDE solutions for the settler are utilized.

7.4 Fundamental Properties of a Reduced First-Order Model of the ASP

The recent results on the first-order model (7.12) by the author in [76–78,81] make it possible to proceed and investigate a simple first-order model of the ASP. Some new results on the steady-state solutions of such a model are presented in this section.

7.4.1 A Reduced First-Order Model of the ASP

We shall combine the first-order settler model (7.12) for the particulate biomass and a similar PDE for the soluble substrate with a reduced-order model for the biological reactor consisting of two ODEs for the two components. We assume that there is only one aerobic reactor, which is completely mixed and has the volume V.

The biomass consumes the substrate and a constitutive relation for this is needed. The growth rate of biomass, denoted by μ [time^{-1}], increases with the dissolved oxygen concentration up to a critical level, after which it is approximately constant. We assume that the oxygen concentration is always sufficiently high and thus ignore a possible dependence. As a function of the substrate concentration, we assume that $\mu \in C^2$ and that it satisfies the following properties, where $\hat{\mu}$ is a constant:

$$\begin{aligned} &\mu(0) = 0, && \lim_{S\to\infty} \mu(S) = \hat{\mu}, \\ &\mu'(S) > 0, \quad S > 0, && \mu''(S) < 0, \quad S > 0. \end{aligned} \tag{7.17}$$

The most common example of such a constitutive relation is the Monod relation

$$\mu(S) = \hat{\mu}\frac{S}{K+S}, \tag{7.18}$$

where K is the half-saturation coefficient.

Let $S(x,t)$ denote the substrate concentration in the settler. In the same way as for the particulate concentration $u(x,t)$, it is convenient to extend the bounded interval $(-H, B)$ for the settler to the whole real line and define

$$\begin{aligned} S_{\mathrm{e}}(t) &:= S(-H-0,t), & S_{\mathrm{u}}(t) &:= S(B+0,t), \\ u_{\mathrm{e}}(t) &:= u(-H-0,t), & u_{\mathrm{u}}(t) &:= u(B+0,t). \end{aligned}$$

The convective flux function for the soluble substrate contains no batch settling flux term:

$$F^{\mathrm{s}}(S,x,r,w,Q) := \begin{cases} -\frac{(1-w)Q}{A}S, & x < 0, \\ \frac{(r+w)Q}{A}S, & x > 0. \end{cases}$$

The mass balances for the two species in the reactor and settler yield the following model equations, cf. Fig. 7.2,

$$V\frac{dS_{\mathrm{f}}}{dt} = QS_{\mathrm{in}} + rQS_{\mathrm{u}} - (1+r)QS_{\mathrm{f}} - V\frac{\mu(S_{\mathrm{f}})}{Y}u_{\mathrm{f}}, \tag{7.19}$$

$$V\frac{du_{\mathrm{f}}}{dt} = Qu_{\mathrm{in}} + rQu_{\mathrm{u}} - (1+r)Qu_{\mathrm{f}} + V\big(\mu(S_{\mathrm{f}}) - b\big)u_{\mathrm{f}}, \tag{7.20}$$

$$A\frac{\partial S}{\partial t} + A\frac{\partial}{\partial x}\big(F^{\mathrm{s}}(S,x,r,w,Q)\big) = (1+r)QS_{\mathrm{f}}\delta(x), \tag{7.21}$$

$$A\frac{\partial u}{\partial t} + A\frac{\partial}{\partial x}\big(F(u,x,r,w,Q)\big) = (1+r)Qu_{\mathrm{f}}\delta(x), \tag{7.22}$$

Here, $Y > 0$ is a constant yield factor relating substrate usage to organism growth, b is the death rate of biomass, which is assumed to be a constant less than the

maximum growth constant $\hat{\mu}$. (7.22) is the first-order equation (7.12) for the biomass within the settler and the flux function F is given by (7.13), although we have in (7.22) written out the dependence on r, w and Q instead of time t as in (7.13). For the model system (7.19)–(7.22), the inputs are $Q(t)$, $S_{\text{in}}(t)$, $u_{\text{in}}(t)$, 'state variables' are $S_{\text{f}}(t)$, $u_{\text{f}}(t)$, $S(x,t)$, $u(x,t)$, the outputs are $S_{\text{e}}(t)$, $S_{\text{u}}(t)$, $u_{\text{e}}(t)$, $u_{\text{u}}(t)$, and the control variables are $r(t)$ and $w(t)$.

The main control objective of such an ASP model is to minimize the effluent substrate concentration S_{e} subject to the constraints that (1) no biomass is leaves through the effluent, i.e. $u_{\text{e}} = 0$, (2) there is a sludge blanket level within the thickening zone.

Equation (7.21) for the substrate in the settler means a pure linear convection of liquid with the speed $(1-w)Q/A$ upwards in the clarification zone and $(r+w)Q/A$ downwards in the thickening zone. However, the source term implies that there are infinitely many possible solutions for given initial data. With $S_{\pm} := S(0\pm,t)$, the mass in- and outflow at $x = 0$ yields

$$(1+r)QS_{\text{f}} = (r+w)QS_{+} + (1-w)QS_{-} \tag{7.23}$$

and this equation does not determine the concentrations $S_{\pm}$ uniquely. The natural, physically relevant and unique solution, which satisfies Condition Γ (see [70] for details), is continuous at $x = 0$, i.e. $S_{-} = S_{+}$. Equation (7.23) then yields

$$S(0,t) = S_{\pm} = S_{\text{f}}(t). \tag{7.24}$$

For time independent Q, r and w, the following relations hold (recall that H is the height of the clarification zone and B the depth of the thickening zone):

$$S_{\text{e}}(t) = S_{\text{f}}\left(t - \frac{HA}{(1-w)Q}\right), \qquad S_{\text{u}}(t) = S_{\text{f}}\left(t - \frac{BA}{(r+w)Q}\right). \tag{7.25}$$

Aiming at controlling the ASP (cf. Sect. 7.2.3), it is of key importance to establish all steady states as functions of the input and control variables.

7.4.2 The Steady States of the Settler

All steady-state solutions of (7.22) have been charted by the author in [76] with respect to both their dependence on the feed inputs and the control parameter $Q_{\text{u}} = (r+w)Q$. It is convenient to use the following two feed inputs: the feed concentration u_{f} and the feed flux $\Phi_{\text{f}} := (1+r)Qu_{\text{f}}$. For fixed Q_{u}, the results can then be described by means of *operating charts*, concentration-flux charts, see Fig. 7.5, which are based on the flux function in the thickening zone $f(\cdot, Q_{\text{u}})$. Regarding the notation in this figure, we refer to [76] for strict definitions. We assume that the twice continuously differentiable batch-settling flux function f_{b} has

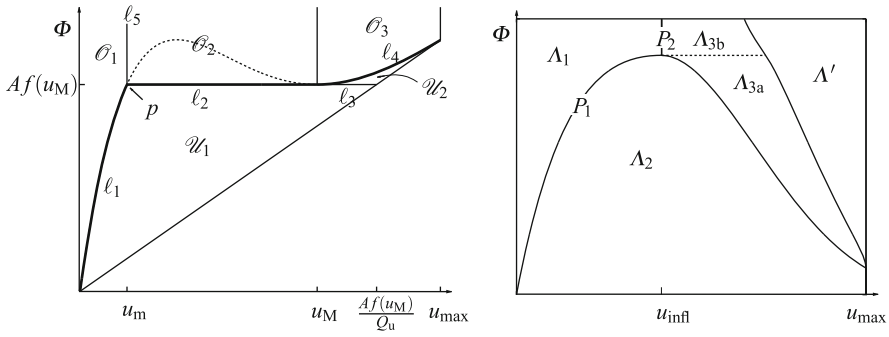


Fig. 7.5 *Left:* The steady-state operating chart. The dotted curve is a part of the graph of the flux function $f(\cdot, Q_u)$. The thick graph is the limiting flux curve $\Phi_{lim}(\cdot, Q_u)$. If the feed point (u_f, Φ_f) lies on this curve, the settler is critically loaded in steady state, which means that it works at its maximum capacity (the excess flux $E = 0$). A feed point below this graph means that the settler is underloaded ($E < 0$), and above means overloaded ($E > 0$). Each region corresponds to a specific steady state which is unique, except on the limiting flux curve (and on ℓ_3 and ℓ_5), where the location of a discontinuity in the thickening and/or the clarification zone is not uniquely determined. Note that the regions in this chart all depend on Q_u. *Right:* The control chart with respect to the steady states. The regions in this chart are fixed (given the batch settling flux f_b). For example, Λ_2 is the region defined by all points of $\ell_2(Q_u)$ for all $Q_u \geq 0$

precisely one inflection point u_{infl}, see Fig. 7.3 (left). Two important concentrations related to the flux function $f(\cdot, Q_u)$ are the following. If the volumetric flow rate Q_u is not too high, there is a local minimum point $u_M(Q_u)$ on the right of u_{infl}, see Fig. 7.5 (left). For low values of Q_u, we may have $u_M(Q_u) = u_{max}$. The concentration $u_m(Q_u)$ is then defined as the lower one with the same flux value. These two values are precisely those found above and below the sludge blanket in the most desired steady-state solution. The settler is then said to work in *optimal operation in steady state*. For details, including a generalized definition of optimal operation for a dynamic solution, we refer to [76–81].

For the purpose of the present paper, the unit on the flux axis has been chosen to mass per time unit, denoted by Φ, whereas the unit of the flux function f (mass per time and cross-sectional area unit) has been used in the previous papers. The *limiting flux function* is defined as (Chancelier et al. [50]):

$$\begin{aligned}\Phi_{lim}(u, Q_u) &:= \min_{u \leq v \leq u_{max}} Af(v, Q_u) \\ &= \begin{cases} Af(u, Q_u), & u \in [0, u_m(Q_u)] \cup \big[u_M(Q_u), u_{max}\big], \\ Af\big(u_M(Q_u), Q_u\big), & u \in \big(u_m(Q_u), u_M(Q_u)\big). \end{cases}\end{aligned} \tag{7.26}$$

Figure 7.5 (left) shows the *steady-state operating chart*, in which the location of the *feed point* (u_f, Φ_f) determines the type of steady-state solution. The same chart, however with less additional information, was earlier presented by Lev et al. [139]. They came to that conclusion based on physical considerations on mass balances,

discussions on the continuity equation with and without a small diffusion term and on some assumptions on the solution of these equations, e.g. on monotonicity, stability and that the concentration is constant within the clarification and thickening zone (for the hyperbolic equation), respectively. The reason for the agreement with the chart in Fig. 7.5 (left) is that their assumptions were correct for the solutions corresponding to the interior of the five regions ($\mathscr{U}_i$ and $\mathscr{O}_i$). As was shown in [76], the solutions corresponding to the boundaries of these five regions (the lines ℓ_i) may have discontinuities within the clarification and thickening zones, and this fact is crucial for the operation and control of the settler, since the sludge blanket is such a discontinuity. Every steady-state solution is piecewise constant and non-decreasing with depth; see [76, Table 1] for a complete table accompanying Fig. 7.5 (left). Of particular interest are the steady-state solutions as $(u_\mathrm{f}, \Phi_\mathrm{f}) \in p \cup \ell_2 \cup \ell_3$, since then the state of optimal operation is possible.

Consider a stand-alone settler and a given feed point $(u_\mathrm{f}, \Phi_\mathrm{f})$ in the *control chart*, Fig. 7.5 (right). The *excess flux* is then defined as (cf. [76])

$$E(u_\mathrm{f}, \Phi_\mathrm{f}, Q_\mathrm{u}) := \Phi_\mathrm{f} - \Phi_\mathrm{lim}(u_\mathrm{f}, Q_\mathrm{u}) \quad \text{(stand-alone settler)}.$$

Then there is a unique value $\tilde{Q}_\mathrm{u}$ and a unique graph $\Phi_\mathrm{lim}(\cdot, \tilde{Q}_\mathrm{u})$ that passes through the feed point, see [76, Theorem 2]. With this unique value $\tilde{Q}_\mathrm{u}$ of the control parameter for the settler, defined implicitly by the equation $E(u_\mathrm{f}, \Phi_\mathrm{f}, \tilde{Q}_\mathrm{u}) = 0$, the settler is *critically loaded* in steady state. If $(u_\mathrm{f}, \Phi_\mathrm{f}) \in P_1 \cup \Lambda_2 \cup \Lambda_{3\mathrm{a}}$ (see Fig. 7.5, right), then there is a possibility for control actions such that the settler can be put in the optimal-operation state with a sludge blanket in the thickening zone. There is a slight difference between the two concepts 'critically loaded' and 'optimal operation' and we refer to the series [76–81] for the details.

For a coupled settler within the ASP, the situation is more complicated because of the recirculation. The feed flux $\Phi_\mathrm{f} = (1 + r)Qu_\mathrm{f}$ then depends on the control parameter r. For the settler within the ASP, we define the *excess flux* as

$$E(u_\mathrm{f}, r, w, Q) := (1 + r)Qu_\mathrm{f} - \Phi_\mathrm{lim}\big(u_\mathrm{f}, (r + w)Q\big) \quad \text{(settler in ASP)}. \tag{7.27}$$

The fluxes in the clarification and thickening zones, denoted by Φ_cl and Φ_th, respectively, are independent of depth in steady state. The concentration within each of these zones, as a function of depth, is generally piecewise constant. The following holds for the flux in the thickening zone for any steady-state solution [76, Corollary 1]:

$$\Phi_\mathrm{th} = \min\big(\Phi_\mathrm{f}, \Phi_\mathrm{lim}(u_\mathrm{f}, (r + w)Q)\big).$$

This fact is the missing link in previous attempts of analyzing similar ASP models.

7.4.3 The Steady States of the ASP

Because of the linear convective transport of the substrate in the settler (7.25) and the continuity (7.24), the stationary substrate concentration is the same in the reactor and the settler, i.e.

$$S = S_{\mathrm{f}} = S_{\mathrm{e}} = S_{\mathrm{u}} \quad \text{in steady state.}$$

The steady-state equations for the ASP model are

$$Q(S_{\mathrm{in}} - S) - V\frac{\mu(S)}{Y}u_{\mathrm{f}} = 0, \tag{7.28}$$

$$Qu_{\mathrm{in}} + rQu_{\mathrm{u}} - (1+r)Qu_{\mathrm{f}} + V\big(\mu(S) - b\big)u_{\mathrm{f}} = 0, \tag{7.29}$$

$$\Phi_{\mathrm{f}} = (1+r)Qu_{\mathrm{f}}, \tag{7.30}$$

$$\Phi_{\mathrm{th}} = \min\big(\Phi_{\mathrm{f}}, \Phi_{\mathrm{lim}}(u_{\mathrm{f}}, (r+w)Q)\big), \tag{7.31}$$

$$\Phi_{\mathrm{th}} = (r+w)Qu_{\mathrm{u}}, \tag{7.32}$$

$$\Phi_{\mathrm{f}} = \Phi_{\mathrm{cl}} + \Phi_{\mathrm{th}}, \tag{7.33}$$

$$\Phi_{\mathrm{cl}} = (1-w)Qu_{\mathrm{e}}. \tag{7.34}$$

Given fixed values of the three inputs Q, S_{in} and u_{in}, and the two control parameters r and w, (7.28)–(7.34) constitute a nonlinear system of equations for the seven variables S, u_{f}, u_{u}, u_{e}, Φ_{f}, Φ_{cl} and Φ_{th}. Note that all these 12 quantities are non-negative. We assume that $Q > 0$, since the case $Q = 0$ is uninteresting. Equations (7.28)–(7.32) and the properties (7.17) of μ imply directly the following:

- $S \le S_{\mathrm{in}}$.
- $S_{\mathrm{in}} = 0 \Leftrightarrow S = 0$ holds. Then there is no need for the ASP at all (although one can find a positive unique solution for the biomass if $u_{\mathrm{in}} > 0$; we leave this case to the reader).
- $S = S_{\mathrm{in}} > 0$ imply that the biomass concentrations are zero in the entire system as well as $u_{\mathrm{in}} = 0$.
- $0 < S < S_{\mathrm{in}} \Leftrightarrow u_{\mathrm{f}} > 0$.

All in all, we assume that $0 < S \le S_{\mathrm{in}}$ and $Q > 0$ hold corresponding to the interesting stationary solutions. Equations (7.28) and (7.30) motivate the following convenient notation

$$U_{\mathrm{f}}(S) := \frac{Y}{\mu(S)\tau}(S_{\mathrm{in}} - S), \quad 0 < S \le S_{\mathrm{in}}, \tag{7.35}$$

$$\Phi_{\mathrm{f}}\big(U_{\mathrm{f}}(S)\big) := (1+r)QU_{\mathrm{f}}(S), \tag{7.36}$$

where

$$\tau := \frac{V}{Q}.$$

This is the hydraulic retention time for an uncoupled bioreactor. Then the system of (7.28)–(7.34) can be rewritten equivalently as

$$u_{\mathrm{f}} = U_{\mathrm{f}}(S), \tag{7.37}$$

$$\begin{aligned} L(S) &:= \Phi_{\mathrm{f}}\big(U_{\mathrm{f}}(S)\big)\left(1+\frac{b\tau}{1+r}\right) \\ &\quad -\frac{r}{r+w}\min\Big(\Phi_{\mathrm{f}}\big(U_{\mathrm{f}}(S)\big), \Phi_{\mathrm{lim}}\big(U_{\mathrm{f}}(S), (r+w)Q\big)\Big) + YQS \\ &= (YS_{\mathrm{in}} + u_{\mathrm{in}})Q, \end{aligned} \tag{7.38}$$

$$\Phi_{\mathrm{f}} = \Phi_{\mathrm{f}}\big(U_{\mathrm{f}}(S)\big), \tag{7.39}$$

$$\Phi_{\mathrm{th}} = \min\big(\Phi_{\mathrm{f}}, \Phi_{\mathrm{lim}}(u_{\mathrm{f}}, (r+w)Q)\big), \tag{7.40}$$

$$u_{\mathrm{u}} = \frac{\Phi_{\mathrm{th}}}{(r+w)Q}, \tag{7.41}$$

$$\Phi_{\mathrm{cl}} = \Phi_{\mathrm{f}} - \Phi_{\mathrm{th}}, \tag{7.42}$$

$$u_{\mathrm{e}} = \frac{\Phi_{\mathrm{cl}}}{(1-w)Q}. \tag{7.43}$$

In the second equation the left-hand side is a function of S denoted by L. We shall show below that this equation has a unique solution. Then the rest of the variables are determined explicitly by the other equations.

In the figures below (Figs. 7.6–7.9) we exemplify different functions and have used the following parameter values and constants of the system (7.37)–(7.43) and the growth rate function (7.18):

$$\begin{aligned} &Q = 1,000\,\mathrm{m^3\,h^{-1}}, && S_{\mathrm{in}} = 0.2\,\mathrm{kg\,m^{-3}}, && u_{\mathrm{in}} = 0.1\,\mathrm{kg\,m^{-3}}, \\ &V = 2,000\,\mathrm{m^3}, && A = 1,500\,\mathrm{m^2}, && Y = 0.7, \\ &b = 0.01\,\mathrm{h^{-1}} && \hat{\mu} = 0.1\,\mathrm{h^{-1}} && K = 0.15\,\mathrm{kg\,m^{-3}}, \\ &r = 1, && w = 0.01. \end{aligned}$$

The settling velocity function v_{s} can be found in Fig. 7.6.

Lemma 1. *Assume that* $0 \le v_{\mathrm{s}} \in C^2$ *is a decreasing function with* $v_{\mathrm{s}}(u_{\max}) = 0$, *the batch settling flux function* $f_{\mathrm{b}}(u) = v_{\mathrm{s}}(u)u$ *has precisely one inflection point* $u_{\mathrm{infl}} \in (0, u_{\max})$ *(cf. Fig. 7.3 (left)) and that the parameters* $Q > 0$, $r > 0$ *and* $0 \le w < 1$ *are fixed. If*

$$v_{\mathrm{s}}(0) > q_{\mathrm{e}} := \frac{Q_{\mathrm{e}}}{A} = \frac{(1-w)Q}{A},$$

then the following equation with the excess flux function (7.27)

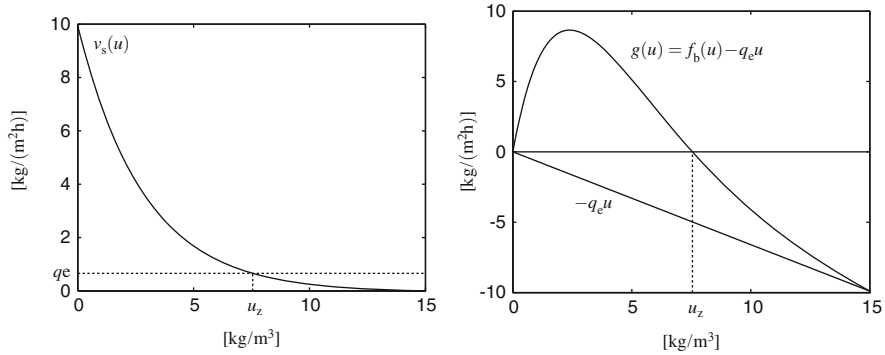


Fig. 7.6 *Left:* The settling velocity function chosen is $v_s(u) = 10(e^{-0.35u} - e^{-0.35u_{max}})\ \mathrm{m\,h^{-1}}$ with $u_{max} = 15\,\mathrm{kg\,m^{-3}}$. *Right:* The flux function g in the clarification zone and its positive zero $u_z \approx 7.55\,\mathrm{kg\,m^{-3}}$

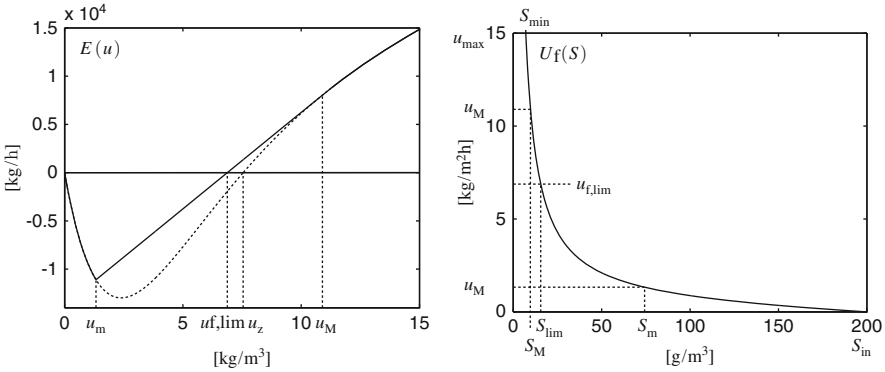


Fig. 7.7 *Left*: The excess flux as a function of u given by (7.27) with its zero $u_{f,lim} \approx 6.88\,\mathrm{kg\,m^{-3}}$. The *dotted graph* is $-Ag$. *Right*: Graph of the function $U_f(S)$ defined by (7.35). The value of S corresponding to a steady-state solution with a critically loaded settler is $S_{lim} \approx 15.5\,\mathrm{g\,m^{-3}}$

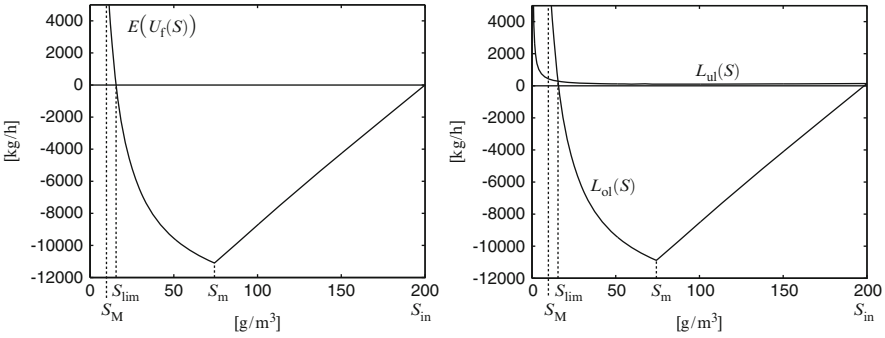


Fig. 7.8 *Left*: The excess flux as a function of S, with the properties (7.52). *Right*: The two functions L_{ol} and L_{ol}

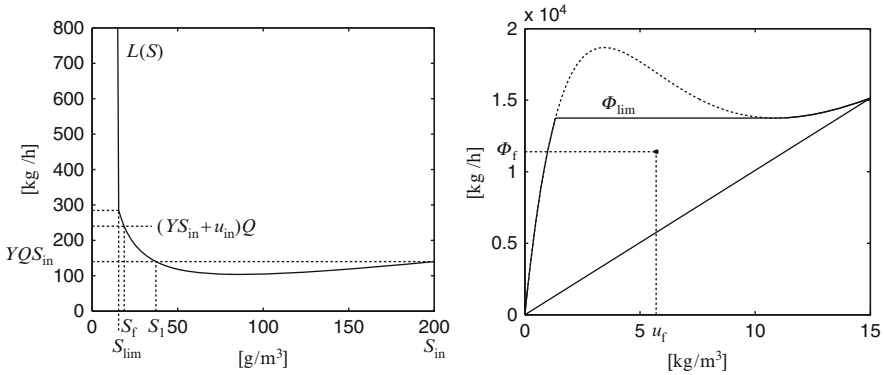


Fig. 7.9 *Left*: The left-hand side of (7.38) as a function of S. The concentration S_l satisfies $L(S_l) = L(S_{in}) = YQS_{in} = 140\,\mathrm{kg\,h^{-1}}$. The right-hand side of (7.38) is $(YS_{in} + u_{in})Q = 240\,\mathrm{kg\,h^{-1}}$ and the steady-state solution is $S = S_f \approx 18.8\,\mathrm{g\,m^{-3}}$. *Right*: The location of the feed point below the limiting flux curve. The settler is underloaded

$$E(u, r, w, Q) = 0 \quad \Longleftrightarrow \quad (1 + r)Qu = \Phi_{\lim}\big(u, (r + w)Q\big) \tag{7.44}$$

has a unique solution $u_{f,\lim} \in (0, u_{\max})$. *Furthermore, the following properties hold for* $u \in (0, u_{\max})$*:*

$$E(u) \lessgtr 0 \quad \Longleftrightarrow \quad u \lessgtr u_{f,\lim},$$
$$E(u) \text{ is increasing for } u > u_{f,\lim}. \tag{7.45}$$

Proof. Recall that $g(u, Q_e) = f_b(u) - q_e u = (v_s(u) - q_e)u$ is the flux function in the clarification zone, where $q_e := Q_e/A = (1 - w)Q/A$ is the upward bulk speed. Since we have assumed that the settling speed v_s is a decreasing function and $v_s(0) > q_e$ holds, we can define u_z as the unique positive solution of the equation $v_s(u) = q_e$, that is, u_z is the largest zero of $g(\cdot, Q_e)$, see Fig. 7.6. It follows partly that $0 < u_z < u_{\max}$, partly that (for $0 < u < u_{\max}$)

$$u \lessgtr u_z \quad \Longleftrightarrow \quad v_s(u) \gtrless q_e \quad \Longleftrightarrow \quad g(u) \gtrless 0. \tag{7.46}$$

Suppressing the parameters r, w and Q, the excess flux (7.27) can be written

$$\begin{aligned} E(u) &= \Phi_f(u) - \Phi_{\lim}(u) \\ &= \begin{cases} (1 + r)Qu - Af(u) = -Ag(u), & u \in [0, u_m] \cup [u_M, u_{\max}], \\ (1 + r)Qu - Af(u_M), & u \in (u_m, u_M). \end{cases} \end{aligned} \tag{7.47}$$

This is a continuous function that satisfies $E(0) = 0$, see Fig. 5 (left). By virtue of (7.46) we can conclude that there exists a unique positive zero of E, which we

denote by $u_{\mathrm{f,lim}}$. If $u_z \in (0, u_m] \cup [u_M, u_{\max})$, then $u_{\mathrm{f,lim}} = u_z$ holds. If $u_z \in (u_m, u_M)$, then $(u_m, u_M) \ni u_{\mathrm{f,lim}} < u_z$ holds. Furthermore, the properties (7.45) also follow. □

There is a unique relation between the biomass and the substrate concentrations in the reactor given by $u_f = U_f(S)$, see (7.35) and (7.37). A graph of this function is shown in Fig. 5 (right). Differentiation gives

$$U_f'(S) = -\frac{Y}{\mu(S)^2\tau}\big(\mu(S) + (S_{in} - S)\mu'(S)\big),$$

$$U_f''(S) = \frac{Y}{\mu(S)^3\tau}\Big(2\mu(S)\mu'(S) + (S_{in} - S)\big(2\mu'(S)^2 - \mu''(S)\mu(S)\big)\Big).$$

The properties (7.17) of μ imply the following:

$$\begin{aligned} &\lim_{S\to 0+} U_f(S) = \infty, \qquad U_f(S_{in}) = 0, \\ &U_f'(S) < 0 \quad \text{and} \quad U_f''(S) > 0 \quad \text{for} \quad 0 < S < S_{in}. \end{aligned} \tag{7.48}$$

Given the constants u_m, u_M, $u_{\mathrm{f,lim}}$ and $u_{\max}$, we can define S_m, S_M, S_{lim} and $S_{\min}$ each as the unique solution (in the interval $(0, S_{in}]$) of each of the following equations:

$$\begin{aligned} U_f(S_m) &= u_m, \qquad U_f(S_{\mathrm{lim}}) = u_{\mathrm{f,lim}}, \\ U_f(S_M) &= u_M, \qquad U_f(S_{\min}) = u_{\max}. \end{aligned} \tag{7.49}$$

Theorem 2. *Assume that $v_s \in C^2[0, u_{\max}]$ is a decreasing function with $v_s(u_{\max}) = 0$, that f_b has precisely one inflection point $u_{\mathrm{infl}} \in (0, u_{\max})$, that $\mu \in C^2$ satisfies (7.17) and that the parameters $Q > 0$, $S_{in} > 0$, $u_{in} \geq 0$, $r > 0$ and $0 \leq w < 1$ are fixed and satisfy*

$$\begin{aligned} v_s(0) > q_e = \frac{Q_e}{A} &= \frac{(1-w)Q}{A}, \\ YS_{in} + u_{in} \leq YS_{\min} &+ u_{\max}(1 + b\tau). \\ S_{\mathrm{lim}} &\leq \check{S}, \end{aligned} \tag{7.50}$$

where $\check{S}$ is the unique minimum of L_{ul} defined by (7.53). If $u_{in} > 0$, then the system (7.37)–(7.43) has a unique non-negative solution that satisfies $S \in [S_{\min}, S_{in})$, $u_f, u_u \in (0, u_{\max}]$ and $u_e \in [0, u_{\max}]$. If $u_{in} = 0$ and

$$\mu(S_{in}) > \frac{w(1+r)}{\tau(r+w)} + b \tag{7.51}$$

holds, then there exist two solutions; partly a non-negative solution that satisfies $S \in [S_{\min}, S_{in})$, $u_f, u_u \in (0, u_{\max}]$ and $u_e \in [0, u_{\max}]$; partly the trivial solution with

$S = S_{\text{in}}$ and all biomass concentrations equal to zero. If $u_{\text{in}} = 0$ and (7.51) is violated, then only the trivial solution exists.

Remark 3. The reason for the presence of inequality (7.50) is only to assure that the feed concentration of the steady-state solution satisfies $u_{\text{f}} \le u_{\max}$. With our example values of the parameters, the inequality (7.50) reads $0.14 + 0.1 \le 0.0049 + 15(1 + 0.02)$ [kg m^{-3}].

Proof. Equation (7.37) means that there is a unique correspondence between u_{f} and S. The same holds between the feed flux Φ_{f} and S via (7.36). We shall now prove that (7.38) has a unique solution S. Then the other equations explicitly determine the remaining variables. We start by examining the minimum term of (7.38), which is the flux in the thickening zone Φ_{th}, see (7.40), as a function of $u_{\text{f}} = U_{\text{f}}(S)$. Note that the subtraction between Φ_{f} and $\Phi_{\lim}$ in the minimum term equals the excess flux (7.27). Recall Lemma 1 and the unique value $u_{\text{f,lim}}$ and its corresponding $S_{\lim}$ defined by (7.49). Since $U_{\text{f}}(S)$ is decreasing we can write the properties (7.45) in the following way for $S \in (0, S_{\text{in}})$:

$$\begin{aligned} &E\big(U_{\text{f}}(S)\big) = \Phi_{\text{f}}\big(U_{\text{f}}(S)\big) - \Phi_{\lim}\big(U_{\text{f}}(S)\big) \lesseqgtr 0 \quad \Longleftrightarrow \quad S \gtreqless S_{\lim}\,, \\ &E\big(U_{\text{f}}(S)\big) \text{ is decreasing for } S < S_{\lim}, \end{aligned} \tag{7.52}$$

cf. Fig. 6.

Hence, we can write

$$L(S) = \begin{cases} L_{\text{ol}}(S), & 0 < S < S_{\lim}, \\ L_{\text{ul}}(S), & S_{\lim} \le S \le S_{\text{in}}, \end{cases}$$

where 'ol' and 'ul' refer to an overloaded and underloaded settler, respectively, and where

$$\begin{aligned} L_{\text{ol}}(S) &:= QU_{\text{f}}(S)(1 + r + b\tau) - \frac{r}{r + w}\Phi_{\lim}\big(U_{\text{f}}(S)\big) + YQS, \\ L_{\text{ul}}(S) &:= QU_{\text{f}}(S)\left(\frac{w(1 + r)}{r + w} + b\tau\right) + YQS, \end{aligned} \tag{7.53}$$

see Figs. 6 (right) and 7.9 (left). The properties (7.48) imply the following properties of L_{ul}:

$$\begin{aligned} &L_{\text{ul}}(S) \to \infty, \quad S \to 0+, \\ &L_{\text{ul}}(S_{\text{in}}) = YQS_{\text{in}} = L(S_{\text{in}}), \\ &L''_{\text{ul}}(S) = QU''_{\text{f}}(S)\left(\frac{w(1 + r)}{r + w} + b\tau\right) > 0, \quad 0 < S < S_{\text{in}}. \end{aligned}$$

This means that L_{ul} is either decreasing on the entire interval $(0, S_{\text{in}}]$ or unimodal with a local minimum within $(0, S_{\text{in}})$. Noting that

$$L_{\mathrm{ol}}(S) - L_{\mathrm{ul}}(S) = \frac{r}{r+w}\left(\Phi_{\mathrm{f}}\big(U_{\mathrm{f}}(S)\big) - \Phi_{\lim}\big(U_{\mathrm{f}}(S)\big)\right) = \frac{r}{r+w}E\big(U_{\mathrm{f}}(S)\big), \tag{7.54}$$

we can write

$$L(S) = \begin{cases} L_{\mathrm{ul}}(S) + \frac{r}{r+w}E\big(U_{\mathrm{f}}(S)\big), & 0 < S < S_{\lim}, \\ L_{\mathrm{ul}}(S), & S_{\lim} \le S \le S_{\mathrm{in}}, \end{cases}$$

By (7.52), we can now conclude that L has the same property as L_{ul}, namely, it is either decreasing on the entire interval $(0, S_{\mathrm{in}}]$ or unimodal with a global minimum point within $(0, S_{\mathrm{in}})$. The right-hand side of (7.38) satisfies $(YS_{\mathrm{in}} + u_{\mathrm{in}})Q \ge YQS_{\mathrm{in}} = L(S_{\mathrm{in}})$, with strict inequality iff $u_{\mathrm{in}} > 0$. If $u_{\mathrm{in}} > 0$, then (7.38) has a unique solution $S_{\mathrm{f}} \in (0, S_{\mathrm{in}})$ (cf. Fig. 7.9, left). If $u_{\mathrm{in}} = 0$, the trivial solution with $S = S_{\mathrm{in}}$ and all biomass concentrations equal to zero is one possibility. Note that condition (7.51) is equivalent to

$$L'_{\mathrm{ul}}(S_{\mathrm{in}}) = YQ\left(1 - \frac{1}{\mu(S_{\mathrm{in}})}\left(\frac{w(1+r)}{\tau(r+w)} + b\right)\right) > 0.$$

If $u_{\mathrm{in}} = 0$ and $L'_{\mathrm{ul}}(S_{\mathrm{in}}) > 0$, then L_{ul} is unimodal and there exist two solutions. If $u_{\mathrm{in}} = 0$ and $L'_{\mathrm{ul}}(S_{\mathrm{in}}) \le 0$, then L_{ul} is decreasing on $(0, S_{\mathrm{in}}]$ and the only solution is the trivial one with zero biomass concentrations. It remains to prove that the non-trivial solution satisfies $S_{\mathrm{f}} \ge S_{\min}$, which is equivalent to $u_{\mathrm{f}} \le u_{\max}$. Then all biomass concentrations in the settler and the outlets are less than or equal to $u_{\max}$ (see [71, Theorem 6.2]). Suppose first that $S_{\mathrm{f}} \ge S_{\lim}$. Since U_{f} is decreasing, we have

$$U_{\mathrm{f}}(S_{\lim}) = u_{\mathrm{f},\lim} < u_{\max} = U_{\mathrm{f}}(S_{\min}) \quad \Longleftrightarrow \quad S_{\lim} > S_{\min},$$

hence $S_{\mathrm{f}} > S_{\min}$ holds. Suppose now that $S_{\mathrm{f}} < S_{\lim}$ holds. The properties on L implies that it is decreasing on $(0, S_{\mathrm{f}}]$. Then we have the following equivalence:

$$S_{\mathrm{f}} \ge S_{\min} \quad \Longleftrightarrow \quad L(S_{\mathrm{f}}) \le L(S_{\min}), \tag{7.55}$$

where we want to prove the former inequality. Since the left-hand side of the latter inequality is $L(S_{\mathrm{f}}) = Q(YS_{\mathrm{in}} + u_{\mathrm{in}})$ and the right-hand side is

$$\begin{aligned} L(S_{\min}) = L_{\mathrm{ol}}(S_{\min}) &= Qu_{\max}(1 + r + b\tau) - \frac{r}{r+w}\Phi_{\lim}(u_{\max}) + YQS_{\min} \\ &= Qu_{\max}(1 + r + b\tau) - \frac{r}{r+w}\big(A\underbrace{f_{\mathrm{b}}(u_{\max})}_{=0} + (r+w)Qu_{\max}\big) + YQS_{\min} \\ &= Q\big(u_{\max}(1 + b\tau) + YS_{\min}\big), \end{aligned}$$

we can conclude that the inequality (7.55) is equivalent to (7.50). □

The values of the steady-state solution in the example, supported by Figs. 5–7.9, are given by

$$S = S_{\mathrm{f}} \approx 18.8\,\mathrm{g\,m^{-3}}, \qquad u_{\mathrm{f}} \approx 5.70\,\mathrm{kg\,m^{-3}}, \quad \Phi_{\mathrm{lim}} \approx 13{,}754\,\mathrm{kg\,h^{-1}},$$
$$\Phi_{\mathrm{f}} = \Phi_{\mathrm{th}} \approx 11{,}399\,\mathrm{kg\,h^{-1}}, \quad u_{\mathrm{u}} \approx 11.3\,\mathrm{kg\,m^{-3}}, \quad \Phi_{\mathrm{cl}} = 0\,\mathrm{kg\,h^{-1}}.$$

The location of the feed point can be seen in Fig. 7.9 (right).

In the description of the ASP in Sect. 7.2.2, it was argued that although w should be small, it has to be nonzero because of the growth of the biomass. This is in accordance with the corresponding stationary solution as $w = 0$. From (7.29)–(7.33) we can infer that $w = 0$ yields

$$0 \le \Phi_{\mathrm{cl}} = V\big(\mu(S) - b\big)u_{\mathrm{f}} + Q u_{\mathrm{in}}$$

and we cannot expect the right-hand side to be zero, since (7.28) determines the relation between S and u_{f}. Hence, $\Phi_{\mathrm{cl}} > 0$ holds generally and there is an overflow of biomass.

7.5 Conclusions

A main physical observation of the continuous-sedimentation process is that under optimal operating conditions there is a concentration discontinuity, the sediment level or sludge blanket level, within the thickening zone. The need for controlling the location of this discontinuity, under time-varying input concentration and flow, has been a driving force for fundamental experimental and theoretical research with published results in different fields during a century. This type of "shock-wave behaviour" can also be found in other processes, for example, traffic flow along a motorway. An additional complication arise as a concentration shock wave moves to the inlet or outlets of the sedimentation tank, or a vehicle density shock wave moves to a motorway entrance. For a proper description of such phenomena, fundamental research on nonlinear PDEs with discontinuous coefficients has been inevitable. It started in the 1990s and comprises well-posedness (existence, uniqueness, stability), numerical analysis, automatic control and inverse problems. The author's contributions in relation to others have been discussed in Sect. 7.3. No detailed results are given except for the following curiosity, that provides an example of the importance of applied mathematics.

The most well-known publication in the sedimentation history is the one by Kynch [135] in 1952. It contains a widely used graphical method for estimating a part of the batch settling flux function. Utilizing basic knowledge of first-order hyperbolic PDEs, the author presented in [79] simple explicit formulae as a solution of this problem. In Sect. 7.3.2, it is demonstrated that these formulae can actually be obtained by Kynch's own arguments. It is remarkable that despite Kynch's paper

has been cited in several hundreds of publications, these formulae have not been discovered before.

The fact that the continuous-sedimentation process can be seen as a "rich problem" is further emphasized by the fact that the process is a critical part of the complex activated sludge process, which can be found in most wastewater treatment plants. The simplest possible model of such a system contains two ODEs and two PDEs. Although similar models can be found in the literature, there is no utilization of PDE theory. Instead, further assumption have been used with questionable results, such as non-uniqueness of stationary solutions.

In Sect. 7.4, a simple model of an ASP is provided by the system of equations (7.19)–(7.22). Under normal operating conditions, the existence and uniqueness of a stationary solution is established in Theorem 2. The result is obtained only for the first-order hyperbolic model. Thus, most of the issues in applied mathematics mentioned in Sect. 7.2.3, from well-posedness to inverse problems, are still to be explored even for a simple model for the ASP.

Acknowledgements I am grateful to Raimund Bürger and Sebastian Farås for valuable comments on the manuscript.

References

1. Abusam, A., Keesman, K.J.: Dynamic modeling of sludge compaction and consolidation processes in wastewater secondary settling tanks. Water Environ. Res. **81**(1), 51–56 (2009)
2. Ahnert, M., Traenckner, J., Guenther, N., Hoeft, S., Krebs, P.: Model-based comparison of two ways to enhance WWTP capacity under stormwater conditions. Water Sci. Tech. **60**(7), 1875–1883 (2009)
3. Alasino, N., Mussati, M.C., Scenna, N.: Wastewater treatment plant synthesis and design. Ind. Eng. Chem. Res. **46**(23), 7497–7512 (2007)
4. Andreianov, B., Karlsen, K.H., Risebro, N.H.: On vanishing viscosity approximation of conservation laws with discontinuous flux. Netw. Heterog. Media **5**, 617–633 (2010)
5. Andreianov, B., Karlsen, K.H., Risebro, N.H.: A theory of L^1-dissipative solvers for scalar conservation laws with discontinuous flux. Arch. Ration. Mech. Anal. 1–60 (2011)
6. Attir, U., Denn, M.M.: Dynamics and control of the activated sludge wastewater process. AIChE J. **24**, 693–698 (1978)
7. Audusse, E., Perthame, B.: Uniqueness for scalar conservation laws with discontinuous flux via adapted entropies. Proc. Roy. Soc. Edinburgh A **135**, 253–265 (2005)
8. Auzerais, F.M., Jackson, R., Russel, W.B.: The resolution of shocks and the effects of compressible sediments in transient settling. J. Fluid Mech. **195**, 437–462 (1988)
9. Bachmann, F., Vovelle, J.: Existence and uniqueness of entropy solution of scalar conservation law with a flux function involving discontinuous coefficients. Comm. Partial Differ. Equat. **31**, 371–395 (2006)
10. Balku, S., Berber, R.: Dynamics of an activated sludge process with nitrification and denitrification: Start-up simulation and optimization using evolutionary algorithm. Comput. Chem. Eng. **30**(3), 490–499 (2006)
11. Balslev, P., Nickelsen, C., Lynggaard-Jensen, A.: On-line flux-theory based control of secondary clarifiers. Water Sci. Tech. **30**(2), 209–218 (1994)

12. Barton, N.G., Li, C.-H., Spencer, J.: Control of a surface of discontinuity in continuous thickeners. J. Austral. Math. Soc. Ser. B **33**, 269–289 (1992)
13. Berres, S., Bürger, R., Coronel, A., Sepulveda, M.: Numerical identification of parameters for a flocculated suspension from concentration measurements during batch centrifugation. Chem. Eng. J. **111**, 91–103 (2005)
14. Boulkroune, B., Darouach, M., Zasadzinski, M., Gille, S.: State and unknown input estimation for nonlinear singular systems: application to the reduced model of the activated sludge process. In: 16th Mediterranean Conference on Control and Automation, pp. 1399–1404 (2008)
15. Bressan, A.: Hyperbolic Systems of Conservation Laws: The One-Dimensional Cauchy Problem. Oxford University Press, London (2000)
16. Bueno, J.L., Coca, J., Cuesta, E., A.G., Lavin, Velasco, G.: Sedimentation of coal slurries: A procedure for the determination of the flocculated-solid flux curve useful for the design of continuous settling tanks. Powder Tech. **63**, 133–140 (1990)
17. Bürger, R., Wendland, W.L.: Entropy boundary and jump conditions in the theory of sedimentation of with compression. Math. Meth. Appl. Sci. **21**, 865–882 (1998)
18. Bürger, R., Wendland, W.L.: Existence, uniqueness, and stability of generalized solutions of an initial-boundary value problem for a degenerating quasilinear parabolic equation. J. Math. Anal. Appl. **218**, 207–239 (1998)
19. Bürger, R., Bustos, M.C., Concha, F.: Settling velocities of particulate systems: 9. Phenomenological theory of sedimentation processes: Numerical simulation of the transient behaviour of flocculated suspensions in an ideal batch or continuous thickener. Int. J. Miner. Process. **55**, 267–282 (1999)
20. Bürger, R., Concha, F., Tiller, F.M.: Applications of the phenomenological theory to several published experimental cases of sedimentation processes. Chem. Eng. J. **80**, 105–117 (2000)
21. Bürger, R., Wendland, W.L., Concha, F.: Model equations for gravitational sedimentation-consolidation processes. Z. Angew. Math. Mech. **80**, 79–92 (2000)
22. Bürger, R., Evje, S., Karlsen, K.H., Lie, K.-A.: Numerical methods for the simulation of the settling of flocculated suspensions. Chem. Eng. J. **80**, 91–104 (2000)
23. Bürger, R., Evje, S., Karlsen, K.H.: On strongly degenerate convection-diffusion problems modeling sedimentation-consolidation processes. J. Math. Anal. Appl. **247**, 517–556 (2000)
24. Bürger, R., Karlsen, K.H., Klingenberg, C., Risebro, N.H.: A front tracking approach to a model of continuous sedimentation in ideal clarifier-thickener units. Nonl. Anal. Real World Appl. **4**, 457–481 (2003)
25. Bürger, R., Karlsen, K.H.: On a diffusively corrected kinematic-wave traffic model with changing road surface conditions. Math. Model. Meth. Appl. Sci. **13**, 1767–1799 (2003)
26. Bürger, R., Karlsen, K.H., Risebro, N.H., Towers, J.D.: On a model for continuous sedimentation in vessels with discontinuously varying cross-sectional area. In: Hou, T.Y., Tadmor, E. (eds.) Hyperbolic Problems: Theory, Numerics, Applications. Proceedings of the Ninth International Conference on Hyperbolic Problems Held in CalTech, Pasadena, 25–29 March 2002, pp. 397–406. Springer, Berlin (2003)
27. Bürger, R., Damasceno, J.J.R., Karlsen, K.H.: A mathematical model for batch and continuous thickening of flocculated suspensions in vessels with varying cross-section. Int. J. Miner. Process. **73**, 183–208 (2004)
28. Bürger, R., Karlsen, K.H., Risebro, N.H., Towers, J.D.: Monotone difference approximations for the simulation of clarifier-thickener units. Comput. Vis. Sci. **6**, 83–91 (2004)
29. Bürger, R., Karlsen, K.H., Risebro, N.H., Towers, J.D.: Well-posedness in BV_t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. Numer. Math. **97**, 25–65 (2004)
30. Bürger, R., Karlsen, K.H., Towers, J.D.: A model of continuous sedimentation of flocculated suspensions in clarifier-thickener units. SIAM J. Appl. Math. **65**, 882–940 (2005)
31. Bürger, R., Karlsen, K.H., Mishra, S., Towers, J.D.: On conservation laws with discontinuous flux. In: Wang, Y., Hutter, K. (eds.) Trends in Applications of Mathematics to Mechanics, pp. 75–84. Shaker, Aachen (2005)

32. Bürger, R., Garcìa, A., Karlsen, K.H., Towers, J.D.: On an extended clarifier-thickener model with singular source and sink terms. Eur. J. Appl. Math. **17**(6), 257–292 (2006)
33. Bürger, R. Narváez, A.: Steady-state, control, and capacity calculations for flocculated suspensions in clarifier-thickeners. Int. J. Mineral Process. **84**(1–4), 274–298 (2007)
34. Bürger, R., Garcìa, A., Karlsen, K.H., Towers, J.D.: A family of numerical schemes for kinematic flows with discontinuous flux. J. Eng. Math. **60**(3), 387–425 (2008)
35. Bürger, R., Garcìa, A., Kunik, M.: A generalized kinetic model of sedimentation of polydisperse suspensions with a continuous particle size distribution. Math. Model. Meth. Appl. Sci. **18**, 1741–1785 (2008)
36. Bürger, R., Garcìa, A., Karlsen, K.H., Towers, J.D.: A kinematic model of continuous separation and classification of polydisperse suspensions. Comput. Chem. Eng. **32**(6), 1173–1194 (2008)
37. Bürger, R., Garcìa, A., Karlsen, K.H., Towers, J.D.: Difference schemes, entropy solutions, and speedup impulse for an inhomogeneous kinematic traffic flow model. Netw. Heterog. Media **3**, 1–41 (2008)
38. Bürger, R., Karlsen, K.H., Towers, J.D.: An Engquist-Osher-type scheme for conservation laws with discontinuous flux adapted to flux connections. SIAM J. Num. Anal. **47**(3), 1684–1712 (2009)
39. Bürger, R., Donat, R., Mulet, P., Vega, C.A.: Hyperbolicity analysis of polydisperse sedimentation models via a secular equation for the flux Jacobian. SIAM J. Appl. Math. **70**, 2186–2213 (2010)
40. Buscall, R.: The sedimentation of concentrated colloidal suspensions. Colloid. Surface. **43**, 33–53 (1990)
41. Buscall,R. White, L.R.: The consolidation of concentrated suspensions. 1. The theory of sedimentation. J. Chem. Soc. Faraday Trans. 1 **83**, 873–891 (1987)
42. Bustos, M.C., Concha, F.: Boundary conditions for the continuous sedimentation of ideal suspensions. AIChE J. **38**(7), 1135–1138 (1992)
43. Bustos, M.C., Concha, F.: Settling velocities of particulate systems. 7. Kynch sedimentation process: Continuous thickening. Int. J. Minern. Process. **34**, 33–51 (1992)
44. Bustos, M.C., Concha, F., Bürger, R., Tory, E.M.: Sedimentation and Thickening: Phenomenological Foundation and Mathematical Theory. Kluwer, Dordrecht (1999)
45. Bustos, M.C., Paiva, F., Wendland, W.: Control of continuous sedimentation as an initial and boundary value problem. Math. Meth. Appl. Sci. **12**, 533–548 (1990)
46. Bustos, M.C., Concha, F., Wendland, W.: Global weak solutions to the problem of continuous sedimentation of an ideal suspension. Math. Meth. Appl. Sci. **13**, 1–22 (1990)
47. Cadet, C., Beteau, J.F., Carlos Hernandez, S.: Multicriteria control strategy for cost/quality compromise in wastewater treatment plants. Contr. Eng. Pract. **12**(3), 335–347 (2004)
48. Chachuat, B., Roche, N., Latifi, M.A.: Optimal aeration control of industrial alternating activated sludge plants. Biochem. Eng. J. **23**(3), 277–289 (2005)
49. Chai, Q., Lie, B.: Predictive control of an intermittently aerated activated sludge process. 2008 American Control Conference, pp. 2209–2214 (2008)
50. Chancelier, J.-Ph., Cohen de Lara, M., Pacard, F.: Analysis of a conservation PDE with discontinuous flux: A model of settler. SIAM J. Appl. Math. **54**(4), 954–995 (1994)
51. Chancelier, J.-Ph., Cohen de Lara, M., Joannis, C., Pacard, F.: New insight in dynamic modelling of a secondary settler – II. Dynamical analysis. Water Res. **31**(8), 1857–1866 (1997)
52. Charef, A., Ghauch, A., Martin-Bouyer, M.: An adaptive and predictive control strategy for an activated sludge process. Bioproc. Biosyst. Eng. **23**, 529–534 (2000)
53. Chatellier, P., Audic, J.M.: A new model for wastewater treatment plant clarifier simulation. Water Res. **34**(2), 690–693 (2000)
54. Chen, J., Gao, Y., Zhang, Z., Shi, D., Xi, D.: Integrated modeling and simulation of activated sludge process. 2006 Chinese Control Conference, pp. 1382–1386 (2006)
55. Clauss, F., Hélaine, D., Balavoine, C., Martin, G.: Controlling the settling of activated sludge in pulp and paper wastewater treatment plants. Water Sci. Tech. **40**, 223–229 (1999)

56. Coclite, G.M., Garavello, M., Piccoli, B.: Traffic flow on a road network. SIAM J. Math. Anal. **36**, 1862–1886 (2005)
57. Concha, F., Bürger, R.: A century of research in sedimentation and thickening. KONA Powder Part. **20**, 38–70 (2002)
58. Concha, F., Bustos, M.C., Barrentios, A.: Phenomenological theory of sedimentation. In: Tory, E. (ed.) Sedimentation of Small Particles in a Viscous Fluid, pp. 51–96. Computational Mechanics Publications, Southampton (1996)
59. Coronel, A., James, F., Sepúlveda, M.: Numerical identification of parameters for a model of sedimentation processes. Inverse Probl. **19**(4), 951–972 (2003)
60. D'Apice, C., Manzo, R., Piccoli, B.: Packet flow on telecommunication networks. SIAM J. Math. Anal. **38**, 717–740 (2006)
61. D'Apice, C., Manzo, R., Piccoli, B.: A fluid dynamic model for telecommunication networks with sources and destinations. SIAM J. Appl. Math. **68**, 981–1003 (2008)
62. David, R., Saucez, P., Vasel, J.-L., Wouwer, A.V.: Modeling and numerical simulation of secondary settlers: A method of lines strategy. Water Res. **43**, 319–330 (2009)
63. David, R., Vasel, J.-L., Wouwer, A.V.: Settler dynamic modeling and MATLAB simulation of the activated sludge process. Chem. Eng. J. **146**, 174–183 (2009)
64. Davis, K.E., Russel, W.B.: An asymptotic description of transient settling and ultrafiltration of colloidal dispersions. Phys. Fluids A **1**(1), 82–100 (1989)
65. De Clercq, J., Devisscher, M., Boonen, I., Vanrolleghem, P.A., Defrancq, J.: A new one-dimensional clarifier model – verification using full-scale experimental data. Water Sci. Tech. **47**, 105–112 (2003)
66. De Clercq, J., Nopens, I., Defrancq, J., Vanrolleghem, P.A.: Extending and calibrating a mechanistic hindered and compression settling model for activated sludge using in-depth batch experiments. Water Res. **42**, 781–791 (2008)
67. Diehl, S.: Shock behaviour of sedimentation in wastewater treatment. Master's thesis, Department of Mathematics, Lund Institute of Technology (1988)
68. Diehl, S.: Scalar conservation laws with source term and discontinuous flux function. Licentiate's thesis, Department of Mathematics, Lund Institute of Technology (1992)
69. Diehl, S.: Conservation Laws with Application to Continuous Sedimentation. PhD thesis, Lund University (1995). ISBN 91-628-1632-2
70. Diehl, S.: On scalar conservation laws with point source and discontinuous flux function. SIAM J. Math. Anal. **26**(6), 1425–1451 (1995)
71. Diehl, S.: A conservation law with point source and discontinuous flux function modelling continuous sedimentation. SIAM J. Appl. Math. **56**(2), 388–419 (1996)
72. Diehl, S.: Scalar conservation laws with discontinuous flux function: I. The viscous profile condition. Comm. Math. Phys. **176**, 23–44 (1996)
73. Diehl, S.: Continuous sedimentation of multi-component particles. Math. Meth. Appl. Sci. **20**, 1345–1364 (1997)
74. Diehl, S.: Dynamic and steady-state behaviour of continuous sedimentation. SIAM J. Appl. Math. **57**(4), 991–1018 (1997)
75. Diehl, S.: On boundary conditions and solutions for ideal clarifier-thickener units. Chem. Eng. J. **80**, 119–133 (2000)
76. Diehl, S.: Operating charts for continuous sedimentation I: Control of steady states. J. Eng. Math. **41**, 117–144 (2001)
77. Diehl, S.: Operating charts for continuous sedimentation II: Step responses. J. Eng. Math. **53**, 139–185 (2005)
78. Diehl, S.: Operating charts for continuous sedimentation III: Control of step inputs. J. Eng. Math. **54**, 225–259 (2006)
79. Diehl, S.: Estimation of the batch-settling flux function for an ideal suspension from only two experiments. Chem. Eng. Sci. **62**, 4589–4601 (2007)
80. Diehl, S.: A regulator for continuous sedimentation in ideal clarifier-thickener units. J. Eng. Math. **60**, 265–291 (2008)

81. Diehl, S.: Operating charts for continuous sedimentation IV: Limitations for control of dynamic behaviour. J. Eng. Math. **60**, 249–264 (2008)
82. Diehl, S.: The solids-flux theory – confirmation and extension by using partial differential equations. Water Res. **42**, 4976–4988 (2008) http://dx.doi.org/10.1016/j.watres.2008.09.005
83. Diehl, S.: A uniqueness condition for nonlinear convection-diffusion equations with discontinuous coefficients. J. Hyperbolic Differ. Equat. **6**, 127–159 (2009)
84. Diehl, S., Jeppsson, U.: A model of the settler coupled to the biological reactor. Water Res. **32**(2), 331–342 (1998)
85. Diehl, S., Wallin, N.-O.: Scalar conservation laws with discontinuous flux function: II. On the stability of the viscous profiles. Comm. Math. Phys. **176**, 45–71 (1996)
86. Diehl, S., Sparr, G., Olsson, G.: Analytical and numerical description of the settling process in the activated sludge operation. In: Briggs, R. (ed.) Instrumentation, Control and Automation of Water and Wastewater Treatment and Transport Systems, IAWPRC, pp. 471–478. Pergamon Press, NY (1990)
87. Diplas, P., Papanicolaou, A.N.: Batch analysis of slurries in zone settling regime. J. Environ. Eng. **123**(7), 659–667 (1997)
88. Dorr, J.V.N.: The use of hydrometallurgical apparatus in chemical engineering. J. Ind. Eng. Chem. **7**, 119–130 (1915)
89. Van Duijn, C.J., Molenaar, J., De Neef, M.J.: The effect of capillary forces on immiscible two-phase flow in heterogeneous porous media. Int. J. Multiphase Flow **22**, 150 (1996)
90. Dupont, R., Dahl, C.: A one-dimensional model for a secondary settling tank including density current and short-circuiting. Water Sci. Tech. **31**(2), 215–224 (1995)
91. Dupont, R., Henze, M.: Modelling of the secondary clarifier combined with the activated sludge model no. 1. Water Sci. Tech. **25**(6), 285–300 (1992)
92. Ekama, G.A., Barnard, J.L., Günthert, F.W., Krebs, P., McCorquodale, J.A., Parker, D.S., Wahlberg, E.J.: Secondary Settling Tanks: Theory, Modelling, Design and Operation. IAWQ scientific and technical report no. 6, 1997
93. Ekman, M.: Bilinear black-box identification and mpc of the activated sludge process. J. Process Contr. **18**(7–8), 643–653 (2008)
94. Evje, S., Karlsen, K.H.: Monotone difference approximations of BV solutions to degenerate convection-diffusion equations. SIAM J. Numer. Anal. **37**, 1838–1860 (2000)
95. Ferrer, J., Seco, A., Serralta, J.: Desass: A software tool for designing, simulating and optimising WWTPs. Environ. Model. Software **23**(1), 19–27 (2008)
96. Fikar, M., Chachuat, B., Latifi, M.A.: Optimal operation of alternating activated sludge processes. Contr. Eng. Pract. **13**(7), 853–861 (2005)
97. Fitch, B.: Kynch theory and compression zones. AIChE J. **29**(6), 940–947 (1983)
98. Fitch, B.: Thickening theories – an analysis. AIChE J. **39**(1), 27–36 (1993)
99. Flores-Alsina, X., Rodriguez-Roda, I., Sin, G., Gernaey, K.V.: Multi-criteria evaluation of wastewater treatment plant control strategies under uncertainty. Water Res. **42**(17), 4485–4497 (2008)
100. Flores-Tlacuahuac, A., Hernandez Esparza, M., Lopez-Negrete de la Fuente, R.: Bifurcation Behavior of a Large Scale Waste Water Treatment Plant. Ind. Eng. Chem. Res. **48**(5), 2605–2615 (2009)
101. Font, R.: Compression zone effect in batch sedimentation. AIChE J. **34**(2), 229–238 (1988)
102. Font, R., Garcia, P., Perez, M.: Analysis of the variation of the upper discontinuity in sedimentation batch test. Sep. Sci. Technol. **33**, 1487–1510 (1998)
103. Font, R., Laveda, M.L.: Semi-batch test of sedimentation. Application to design. Chem. Eng. J. **80**, 157–165 (2000)
104. Font, R., Perez, M., Pastor, C.: Permeability values from batch tests of sedimentation. Ind. Eng. Chem. Res. **33**, 2859–2867 (1994)
105. Garrido, P., Bürger, R., Concha, F.: Settling velocities of particulate systems: 11. Comparison of the phenomenological sedimentation-consolidation model with published experimental results. Int. J. Miner. Process. **60**, 213–227 (2000)

106. Georgieva, P., Ilchmann, A.: Adaptive lambda-tracking control of activated sludge processes. Int. J. Contr. **12**, 1247–1259 (2001)
107. Georgieva, P.G., Feyo De Azevedo, S.: Robust control design of an activated sludge process. Int. J. Robust Nonlin. Contr. **9**, 949–967 (1999)
108. Gimse, T., Risebro, N.H.: Riemann problems with a discontinuous flux function. In: Engquist, B., Gustavsson, B. (ed.) Third International Conference on Hyperbolic Problems, Theory, Numerical Methods and Applications, vol. I, pp. 488–502 (1990)
109. Gimse, T., Risebro, N.H.: Solution of the Cauchy problem for a conservation law with a discontinuous flux function. SIAM J. Math. Anal. **23**(3), 635–648 (1992)
110. Giokas, D.L., Kim, Y., Paraskevas, P.A., Paleologos, E.K., Lekkas, T.D.: A simple empirical model for activated sludge thickening in secondary clarifiers. Water Res. **36**, 3245–3252 (2002)
111. Godunov, S.K.: A finite difference method for the numerical computations of discontinuous solutions of the equations of fluid dynamics. Mat. Sb. **47**, 271–306 (1959) (In Russian)
112. Grassia, P., Usher, S.P., Scales, P.J.: A simplified parameter extraction technique using batch settling data to estimate suspension material properties in dewatering applications. Chem. Eng. Sci. **63**(7), 1971–1986 (2008)
113. Greenberg, J.M., Leroux, A.Y., Baraille, R., Noussair, A.: Analysis and approximation of conservation laws with source terms. SIAM J. Num. Anal. **34**, 1980–2007 (1997)
114. Hamilton, J., Jain, R., Antoniou, P., Svoronos, S.A., Koopman, B., Lyberatos, G.: Modeling and pilot-scale experimental verification for predenitrification process. J. Environ. Eng. **118**, 38–55 (1992)
115. Härtel, L., Pöpel, H.J.: A dynamic secondary clarifier model including processes of sludge thickening. Water Sci. Tech. **25**(6), 267–284 (1992)
116. Henze, M., Grady, C.P.L., Gujer, W., Marais, G.v.R., Matsuo, T.: Activated sludge model no. 1. Technical Report 1, IAWQ, London, UK, 1987
117. Holenda, B., Domokos, E., Redey, A., Fazakas, J.: Dissolved oxygen control of the activated sludge wastewater treatment process using model predictive control. Comput. Chem. Eng. **32**(6), 1278–1286 (2008)
118. Howells, I., Landman, K.A., Panjkov, A., Sirakoff, C., White, L.R.: Time-dependent batch settling of flocculated suspensions. Appl. Math. Modell. **14**(2), 77–86 (1990)
119. Ito, H.: A dissipative approach to control of biological wastewater treatment plants based on entire nonlinear process models. In: Proceedings of the 2004 American Control Conference, vol. 6, pp. 5489–5495. IEEE, NY (2004)
120. Jeppsson, U., Diehl, S.: An evaluation of a dynamic model of the secondary clarifier. Water Sci. Tech. **34**(5–6), 19–26 (1996)
121. Jeppsson, U., Pons, M.-N., Nopens, I., Alex, J., Copp, J., Gernaey, K.V., Rosen, C., Steyer, J.-P., Vanrolleghem, P.A.: Benchmark simulation model no 2 – general protocol and exploratory case studies. Water Sci. Tech. **56**, 67–78 (2007)
122. Jernqvist, Å.: Experimental and theoretical studies of thickeners. Part 1. Derivation of basic equations and basic graphical constructions. Svensk Papperstidning **68**, 506–511 (1965)
123. Jernqvist, Å.: Experimental and theoretical studies of thickeners. Part 2. Graphical calculation of thickener capacity. Svensk Papperstidning **68**, 545–548 (1965)
124. Jernqvist, Å.: Experimental and theoretical studies of thickeners. Part 3. Concentration distribution of the steady and unsteady state operation of thickeners. Svensk Papperstidning **68**, 578–582 (1965)
125. Kaasschieter, E.F.: Solving the Buckley-Leverett equation with gravity in a heterogeneous porous medium. Comput. Geosci. **3**, 23–48 (1999)
126. Karlsen, K.H., Towers, J.D.: Convergence of the Lax-Friedrichs scheme and stability for conservation laws with a discontinuous space-time dependent flux. Chinese Ann. Math. Ser. B. **25**(3), 287–318 (2004)
127. Karlsen, K.H., Risebro, N.H., Towers, J.D.: Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient. IMA J. Numer. Anal. **22**(4), 623–664 (2002)

128. Karlsen, K.H., Risebro, N.H., Towers, J.D.: L^1 stability for entropy solutions of nonlinear degenerate parabolic convection-diffusion equations with discontinuous coefficients. Trans. Royal Norwegian Society Sci. Letters (Skr. K. Nor. Vidensk. Selsk.) **3**, 49 (2003)
129. Keinath, T.M.: Operational dynamics and control of secondary clarifiers. J. Water Pollut. Control Fed. **57**(7), 770–776 (1985)
130. Klausen, R.A., Risebro, N.H.: Stability of conservation laws with discontinuous coefficients. J. Differ. Equat. **157**, 41–60 (1999)
131. Klingenberg, C., Risebro, N.H.: Convex conservation laws with discontinuous coefficients. Existence, uniqueness and asymptotic behavior. Comm. Part. Differ. Equat. **20**, 1959–1990 (1995)
132. Klingenberg, C., Risebro, N.H.: Stability of a resonant system of conservation laws modeling polymer flow with gravitation. J. Differ. Equat. **170**, 344–380 (2001)
133. Koehne, M., Hoen, K., Schuhen, M.: Modelling and simulation of final clarifiers in wastewater treatment plants. Math. Comput. Simul. **39**(5–6), 609–616 (1995)
134. Koumboulis, F.N., Kouvakas, N.D., King, R.E., Stathaki, A.: Two-stage robust control of substrate concentration for an activated sludge process. ISA Trans. **47**(3), 267–278 (2008)
135. Kynch, G.J.: A theory of sedimentation. Trans. Faraday Soc. **48**, 166–176 (1952)
136. Landman, K.A., White, L.R.: Solid/liquid separation of flocculated suspensions. Adv. Colloid Interface Sci. **51**, 175–246 (1994)
137. Lee, T.T., Wang, F.Y., Newell, R.B.: Advances in distributed parameter approach to the dynamics and control of activated sludge processes for wastewater treatment. Water Res. **40**(5), 853–869 (2006)
138. Lester, D.R., Usher, S.P., Scales, P.J.: Estimation of the hindered settling function $R(\phi)$ from batch-settling tests. AIChE J. **51**, 1158–1168 (2005)
139. Lev, O., Rubin, E., Sheintuch, M.: Steady state analysis of a continuous clarifier-thickener system. AIChE J. **32**(9), 1516–1525 (1986)
140. LeVeque, R.J.: Numerical Methods for Conservation Laws. Birkhäuser, Boston (1992)
141. Liu, C., Qiao, J., Zhang, F.: The control of wastewater treatment process based on fuzzy neural network. The Sixth World Congress on Intelligent Control and Automation. WCICA 2006, vol. 2, pp. 9347–9351 (2006)
142. Lynggaard-Jensen, A., Andreasen, P., Husum, F., Nygaard, M., Kaltoft, J., Landgren, L., Moller, F., Brodersen, E.: Increased performance of secondary clarifiers using dynamic distribution of minimum return sludge rates. Water Sci. Tech. **60**(9), 2439–2445 (2009)
143. Ma, Y., Peng, Y., Wang, S.: New automatic control strategies for sludge recycling and wastage for the optimum operation of predenitrification processes. J. Chem. Tech. Biotechnol. **81**(1), 41–47 (2006)
144. Marigo, A.: Optimal traffic distribution and priority coefficients for telecommunication networks. Networks Heterogeneous Media **1**, 315–336 (2006)
145. Mishra, A.S., Gowda, G.D.V.: Godunov-type methods for conservation laws with a flux function discontinuous in space. SIAM J. Numer. Anal. 42, 179–208 (2004)
146. Mishra, A.S., Gowda, G.D.V.: Optimal entropy solutions for conservation laws with discontinuous flux-functions. J. Hyperbolic Differ. Equat. **2**, 783–837 (2005)
147. Mishra, A.S., Gowda, G.D.V.: Convergence of Godunov type methods for a conservation law with a spatially varying discontinuous flux function. Math. Comp. **76**, 1219–1242 (2007)
148. Mochon, S.: An analysis of the traffic on highways with changing surface conditions. Math. Model. **9**(1), 1–11 (1987)
149. Nocoń, W.: Mathematical modelling of distributed feed in continuous sedimentation. Simulat. Model. Pract. Theor. **14**(5), 493–505 (2006)
150. Oleinik, O.A.: Uniqueness and stability of the generalized solution of the Cauchy problem for a quasi-linear equation. Uspekhi Mat. Nauk **14**, 165–170 (1959); Am. Math. Soc. Trans. Ser. 2 **33**, 285–290 (1964)
151. Olsson, G., Nielsen, M.K., Yuan, Z., Lynggaard-Jensen, A., Steyer, J.-P.: Instrumentation, Control and Automation in Wastewater Systems. International Water Association, London (2005)

152. Ostrov, D.N.: Viscosity solutions and convergence of monotone schemes for synthetic aperture radar shape-from-shading equations with discontinuous intensities. SIAM J. Appl. Math. **59**(6), 2060–2085 (1999)
153. Ostrov, D.N.: Solutions of Hamilton-Jacobi equations and scalar conservation laws with discontinuous space-time dependence. J. Differ. Equat. **182**(1), 51–77 (2002)
154. Otterpohl, R., Freund, M.: Dynamic models for clarifiers of activated sludge plants with dry and wet weather flows. Water Sci. Tech. **26**(5–6), 1391–1400 (1992)
155. Ozinsky, A.E., Ekama, G.A., Reddy, B.D.: Mathematical simulation of dynamic behaviour of secondary settling tanks. Technical Report W85. Department of Civil Engineering, University of Cape Town, South Africa, 1994
156. Panov, E.Yu.: On existence and uniqueness of entropy solutions to the Cauchy problem for a conservation law with discontinuous flux. J. Hyperbolic Differ. Equat. **6**(3), 525–549 (2009)
157. Patziger, M., Kainz, H., Hunze, M., Józsa, J.: Analysing sludge balance in activated sludge systems with a novel mass transport model. Water Sci. Tech. **57**, 1413–1419 (2008)
158. Petty, C.A.: Continuous sedimentation of a suspension with a nonconvex flux law. Chem. Eng. Sci. **30**, 1451–1458 (1975)
159. Plośz, B.Gy., Weiss, M., Printemps, C., Essemiani, K., Meinhold, J.: One-dimensional modelling of the secondary clarifier – factors affecting simulation in the clarification zone and the assessment of the thickening flow dependence. Water Res. **41**, 3359–3371 (2007)
160. Queinnec, I., Dochain, D.: Modelling and simulation of the steady-state of secondary settlers in wastewater treatment plants. Water Sci. Tech. **43**, 39–46 (2001)
161. Rauh, A., Kletting, M., Aschemann, H., Hofer, E.P.: Robust controller design for bounded state and control variables and uncertain parameters using interval methods. International Conference on Control and Automation. ICCA '05., vol. 2, pp, 777–782 (2005)
162. Ross, D.S.: Two new moving boundary problems for scalar conservation laws. Comm. Pure Appl. Math. **41**, 725–737 (1988)
163. Seguin, N., Vovelle, J.: Analysis and approximation of a scalar conservation law with a flux function with discontinuous coefficients. Math. Model. Meth. Appl. Sci. **13**, 221–257 (2003)
164. Sheintuch, M.: Steady state modeling of reactor-settler interaction. Water Res. **21**(12), 1463–1472 (1987)
165. Shen, W., Chen, X., Pons, M.N., Corriou, J.P.: Model predictive control for wastewater treatment process with feedforward compensation. Chem. Eng. J. **155**(1–2), 161–174 (2009)
166. Stehfest, H.: An operational dynamic model of the final clarifier. Trans. Inst. Meas. Contr.**6**(3), 160–164 (1984)
167. Takács, I., Patry, G.G., Nolasco, D.: A dynamic model of the clarification-thickening process. Water Res. **25**(10), 1263–1271 (1991)
168. Talmage, W.P., Fitch, E.B.: Determining thickener unit areas. Ind. Eng. Chem. **47**(1), 38–41 (1955)
169. Terracina, A.: A free boundary problem for scalar conservation laws. SIAM J. Math. Anal. **30**(5), 985–1010 (1999)
170. Towers, J.D.: Convergence of a difference scheme for conservation laws with a discontinuous flux. SIAM J. Numer. Anal. **38**, 681–698 (2000)
171. Towers, J.D.: A difference scheme for conservation laws with a discontinuous flux: The nonconvex case. SIAM J. Numer. Anal. **39**, 1197–1218 (2001)
172. Tränckner, J., Franz, T., Seggelke, K., Krebs, P.: Dynamic optimisation of WWTP inflow to reduce total emission. Water Sci. Tech. **56**(10), 11–19 (2007)
173. Traoré, A., Grieu, S., Thiery, F., Polit, M., Colprim, J.: Control of sludge height in a secondary settler using fuzzy algorithms. Comput. Chem. Eng. **30**(8), 1235–1243 (2006)
174. Tzoneva, R.: Method for real time optimal control of the activated sludge process. 2007 Mediterranean Conference on Control and Automation, pp. 1–6 (2007)
175. Vaccari, D.A., Uchrin, C.G.: Modeling and simulation of compressive gravity thickening of activated sludge. J. Environ. Sci. Health **A24**(6), 645–674 (1989)

176. Vasudeva Kumar, M., Sree Rama Raju, V., Pushpavanam, S., Kienle, A.: Effect of the minimum flux condition in the settler on the nonlinear behavior of the activated sludge process. Ind. Eng. Chem. Res. **45**(17), 5996–6006 (2006)
177. Vasudeva Kumar, M., Zeyer, K.P., Kienle, A., Pushpavanam, S.: Conceptual analysis of the effect of kinetics on the stability and multiplicity of a coupled bioreactor-separator system using a cybernetic modeling approach. Ind. Eng. Chem. Res. **48**(24), 10962–10975 (2009)
178. Verdickt, L., Smets, I., Van Impe, J.: Sensitivity analysis of a one-dimensional convection-diffusion model for secondary settling tanks. Chem. Eng. Comm. **192**, 1567–1585 (2005)
179. Vesilind, P.A.: Treatment and Disposal of Wastewater Sludges. Ann Arbor Science Publishers, Ann Arbor, Michigan, 236 p. (1974)
180. Vitasovic, Z.Z.: An Integrated Control Strategy for the Activated Sludge Process. PhD thesis, Rice University, TX (1985)
181. Vitasovic, Z.Z.: Continuous settler operation: A dynamic model. In: Patry, G.G., Chapman, D. (eds.) Dynamic Modelling and Expert Systems in Wastewater Engineering, pp. 59–81. Lewis, MI (1989)
182. Wahab, N.A., Katebi, R., Balderud, J.: Multivariable pid control design for activated sludge process with nitrification and denitrification. Biochem. Eng. J. **45**(3), 239–248 (2009)
183. Waters, A.G., Galvin, K.P.: Theory and application of thickener design. Filtrat. Separ. **28**, 110–116 (1991)
184. Watts, R.W., Svoronos, S.A., Koopman, B.: One-dimensional clarifer model with sludge blanket heights. J. Environ. Eng. **122**(12), 1094–1100 (1996)
185. Watts, R.W., Svoronos, S.A., Koopman, B.: One-dimensional modeling of secondary clarifiers using a concentration and feed velocity-dependent dispersion coefficient. Water Res. **30**(9), 2112–2124 (1996)
186. Wett, B.: A straight interpretation of the solids flux theory for a three-layer sedimentation model. Water Res. **36**, 2949–2958 (2002)
187. Zeidan, A., Rohani, S., Bassi, Z.: Dynamic and steady-state sedimentation of polydisperse suspension and prediction of outlets particle-size distribution. Chem. Eng. Sci. **59**, 2619–2632 (2004)
188. Zhao, L., Chai, T., Cong, Q.: Hybrid dynamic model of anoxic-aeration biological wastewater treatment plant. 6th World Congress on Intelligent Control and Automation, vol. 1, pp. 4781–4785 (2006)
189. Zheng, Y., Bagley, D.M.: Dynamic model for zone settling and compression in gravity thickeners. J. Environ. Eng. **124**(10), 953–958 (1998)

Chapter 8
Scaling, Wavelets, Image Compression, and Encoding

Palle E.T. Jorgensen and Myung-Sin Song

Abstract In this paper we develop a family of multi-scale algorithms with the use of filter functions in higher dimension.

While our primary application is to images, i.e., processes in two dimensions, we prove our theorems in a more general context, allowing dimension 3 and higher.

The key tool for our algorithms is the use of tensor products of representations of certain algebras, the Cuntz algebras O_N, from the theory of algebras of operators in Hilbert space. Our main result offers a matrix algorithm for computing coefficients for images or signals in specific resolution subspaces. A special feature with our matrix operations is that they involve products and iteration of slanted matrices. Slanted matrices, while large, have many zeros, i.e., are sparse. We prove that as the operations increase the degree of sparseness of the matrices increase. As a result, only a few terms in the expansions will be needed for achieving a good approximation to the image which is being processed. Our expansions are local in a strong sense.

An additional advantage with the use of representations of the algebras O_N, and tensor products is that we get easy formulas for generating all the choices of matrices going into our algorithms.

P.E.T. Jorgensen (✉)
Department of Mathematics, The University of Iowa, Iowa City, IA52242, USA
e-mail: jorgen@math.uiowa.edu

M.-S. Song
Department of Mathematics and Statistics, Southern Illinois University Edwardsville, Edwardsville, IL62026, USA
e-mail: msong@siue.edu

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_8,

8.1 Introduction

The paper is organized as follows: first motivation, history, and discussion of our applications. Since a number of our techniques use operator theory, we have a separate section with these results. We feel they are of independent interest, but we have developed them here tailored to use for image processing.

A key tool in our algorithms is the use of slanted matrices, and tensor products of representations. This material receives separate sections. The significance of the slanted property is that matrix products of slanted matrices become increasingly more sparse (i.e., the resulting matrices have wide patterns of zeros) which makes computations fast.

Motivation and Applications. We consider interconnections between three subjects which are not customarily thought to have much to do with one another: (1) the theory of stochastic processes, (2) wavelets, and (3) sub-band filters (in the sense of signal processing).

While connections between (2) and (3) have gained recent prominence, see for example [9], applications of these ideas to stochastic integration is of more recent vintage. Nonetheless, there is always an element of noise in the processing of signals with systems of filters. But this has not yet been modeled with stochastic processes, and it hasn't previously been clear which processes do the best job.

Recall however that the notion of low-pass and high-pass filters derives in part from probability theory. Here high and low refers to frequency bands, but there may well be more than two bands (not just high and low, but a finite range of bands). The idea behind this is that signals can be decomposed according to their frequencies, with each term in the decomposition corresponding to a range of a chosen frequency interval, for example high and low. Sub-band filtering amounts to an assignment of filter functions which accomplish this: each of the filters will then block signals in one band, and passes the others. This is known to allow for transmission of the signal over a medium, for example wireless. It was discovered recently (see [9]), perhaps surprisingly, that the functions which give good filters in this context serve a different purpose as well: they offer the parameters which account for families of wavelet bases, for example families of bases functions in the Hilbert space $L^2(\mathbb{R})$. Indeed the simplest quadrature-mirror filter is known to produce the Haar wavelet basis in $L^2(\mathbb{R})$.

It is further shown in [9] that both principles (2) and (3) are governed by families of representations of one of the Cuntz algebras $\mathcal{O}_N$, with the number N in the subscript equal to the number of sub-bands in the particular model. So for the Haar case, $N = 2$.

A main purpose in this paper is pointing out that fractional Brownian motion (fBm) may be understood with the data in (2) and (3), and as a result that fBm may be understood with the use of a family of representations of $\mathcal{O}_N$; albeit a quite different family of representations from those used in [9].

A second purpose we wish to accomplish is to show that the operators and representations we use in one dimension can be put together in a tensor product

construction and then account for those filters which allow for processing of digital images. Here we think of both black and white, in which case we will be using a single matrix of pixel numbers. In the case of color images, the same idea applies, but then we will rather be using three matrices accounting for exposure of each of the three primary colors. If one particular family F of representations of one of the Cuntz algebras $\mathcal{O}_N$ is used in 1D, then for 2D (images) we show that the relevant families of representations of $\mathcal{O}_N$ are obtained from F with the use of tensor product of pairs of representations, each one chosen from F.

8.1.1 Interdisciplinary Dimensions

While there is a variety of interdisciplinary dimensions, math (harmonic, numerical, functional, ...), computer science, engineering, physics, image science, we will offer some pointers here some pointers to engineering, signal and image processing.

Engineering departments teach courses in digital images, as witnessed by such standard texts as [11]. Since 1997, there was a separate track of advances involving color images and wireless communication, the color appearance of surfaces in the world as a property that allows us to identify objects, see e.g., [17,29], and [7,8,30].

From statistical inference we mention [28]. Color imaging devices such as scanners, cameras, and printers, see [26]. And on the more theoretical side, CS, [10].

8.2 History and Motivation

Here we discuss such tools from engineering as sub-band filters, and their manifold variations. They are ubiquitous in the processing of multi-scale data, such as arises in wavelet expansions.

A powerful tool in the processing of signals or images, in fact going back to the earliest algorithms, is that of subdividing the data into sub-bands of frequencies. In the simplest case of speech signals, this may involve a sub-division into the low frequency range, and the high. An underlying assumption for this is that the data admits such a selection of a total *finite* range for the set of all frequencies. If such a finite frequency interval can be chosen we talk about bandlimited analog signals. Now depending of the choice of analysis and synthesis method to be used, the suitability of bandlimited signals may vary. Shannon proved that once a frequency band B has been chosen, then there is a Fourier representation for the signals, as time-series, with frequencies in B which allow reconstruction from a discrete set of samples in the time variable, sampled in an arithmetic progression at a suitable rate, the Nyquist rate. The theorem is commonly called the Shannon sampling theorem, and is also known as Nyquist–Shannon–Kotelnikov, Whittaker–Shannon, WKS, etc., sampling theorem, as well as the Cardinal Theorem of Interpolation Theory.

While this motivates a variety of later analogues to digital (A/D) tools, the current techniques have gone far beyond Fourier analysis.

Here we will focus on tools based on such multi-scale which are popular in wavelet analysis. But the basic idea of dividing the total space of data into subspaces corresponding to bands, typically frequency bands, will be preserved. In passing from speech to images, we will be aiming at the processes underlying the processing of digital images, i.e., images arising as a matrix (or checkerboard) of pixel numbers, or in case of color-images a linked system of three checkerboards, with the three representing the primary colors.

While finite or infinite families of nested subspaces are ubiquitous in mathematics, and have been popular in Hilbert space theory for generations (at least since the 1930s), this idea was revived in a different guise in 1986 by Stéphane Mallat. It has since found a variety of application to multiscale processes such as analysis on fractals. In its adaptation to wavelets, the idea is now referred to as the multiresolution method.

What made the idea especially popular in the wavelet community was that it offered a skeleton on which various discrete algorithms in applied mathematics could be attached and turned into wavelet constructions in harmonic analysis. In fact what we now call multiresolutions have come to signify a crucial link between the world of discrete wavelet algorithms, which are popular in computational mathematics and in engineering (signal/image processing, data mining, etc.) on the one side, and on the other side continuous wavelet bases in function spaces, especially in $L^2(\mathbb{R}^d)$. Further, the multiresolution idea closely mimics how fractals are analyzed with the use of finite function systems.

8.3 Operator Theory

Those key ideas multi-scale, wavelets, image processing, and the operator theory involved discussed about, and used inside our paper are covered in a number of references. Especially relevant are the following [2–4, 9, 16, 19–21], but the reader will find additional important reference lists in the books [9] and [19]. A key idea in the analysis we present here is to select a Hilbert space which will represent the total space of data; it may be $L^2(\mathbb{R}^d)$ for a suitable chosen d for dimension; or it may be anyone of a carefully selected closed subspaces. A representation of a function in d variables into an expansion relative to an orthogonal basis (or a frame basis) corresponds to a subdivision of the total space into one-dimensional subspaces. But to get started one must typically select a fixed subspace which represent a resolution of the data (or image) under consideration, then the further subdivision into closed subspaces can be accomplished with a scaling: The result is a family of closed subspaces, each representing a detail. There will then be a system of isometries which account for these subspaces, and there will be a scaling operation which makes (mathematically) precise the scale-similarity of the data in different detail-components of the total decomposition.

In wavelets, the scaling is by a number, for example 2, for dyadic wavelets. In this case, there will be two frequency bands. If the scale is instead by a positive integer $N > 2$, then there will be N natural frequency bands.

For images, or higher dimensional date, it is then natural to use an invertible $d \times d$ matrix (over the integers), say A, to model a choice of scaling. Scaling will then be modeled with powers of A, i.e., with A^j as j ranges over the integers $\mathbb{Z}$. In this case, the number of frequency bands will be $N := |det(A)|$.

Here we review a harmonic analysis of isometries in Hilbert space. Our results are developed with view to applications to multi-scale processes. In the Hilbert space framework, this takes the form of an exhaustion of closed subspaces, a monotone family of closed subspaces where one arises from the other by an application of a scaling operator.

But in mathematics, or more precisely in operator theory, the underlying idea dates back to work of John von Neumann, Norbert Wiener, and Herman Wold, where nested and closed subspaces in Hilbert space were used extensively in an axiomatic approach to stationary processes, especially for time series. Wold proved that any (stationary) time series can be decomposed into two different parts: The first (deterministic) part can be exactly described by a linear combination of its own past, while the second part is the opposite extreme; it is *unitary*, in the language of von Neumann.

von Neumann's version of the same theorem is a pillar in operator theory. It states that every isometry in a Hilbert space $\mathscr{H}$ is the unique sum of a shift isometry and a unitary operator, i.e., the initial Hilbert space $\mathscr{H}$ splits canonically as an orthogonal sum of two subspaces $\mathscr{H}_s$ and $\mathscr{H}_u$ in $\mathscr{H}$, one which carries the shift operator, and the other $\mathscr{H}_u$ the unitary part. The shift isometry is defined from a nested scale of closed spaces V_n, such that the intersection of these spaces is $\mathscr{H}_u$. Specifically,

$$\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset V_2 \subset \cdots \subset V_n \subset V_{n+1} \subset \cdots$$

$$\bigwedge_n V_n = \mathscr{H}_u, \text{ and } \bigvee_n V_n = \mathscr{H}.$$

However, Stéphane Mallat was motivated by the notion of scales of resolutions in the sense of optics. This in turn is based on a certain "artificial-intelligence" approach to vision and optics, developed earlier by David Marr at MIT, an approach which imitates the mechanism of vision in the human eye.

The connection from these developments in the 1980s back to von Neumann is this: Each of the closed subspaces V_n corresponds to a level of resolution in such a way that a larger subspace represents a finer resolution. Resolutions are relative, not absolute! In this view, the relative complement of the smaller (or coarser) subspace in larger space then represents the visual detail which is added in passing from a blurred image to a finer one, i.e., to a finer visual resolution.

Subsequently, this view became popular in the wavelet community, as it offered a repository for the fundamental father and the mother functions, also called the scaling function φ, and the wavelet function ψ (see details below). Via a system of translation and scaling operators, these functions then generate nested subspaces,

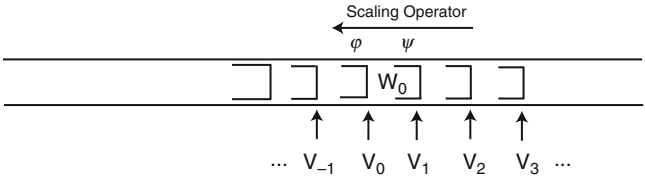


Fig. 8.1 Multiresolution. $L^2(\mathbb{R}^d)$-version (continuous); $\varphi \in V_0$, $\psi \in W_0$

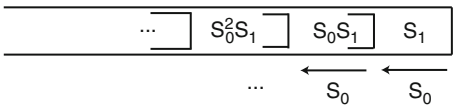


Fig. 8.2 Multiresolution. $l^2(\mathbb{Z})$-version (discrete); $\varphi \in V_0$, $\psi \in W_0$

and we recover the scaling identities which initialize the appropriate algorithms. This is now called the family of pyramid algorithms in wavelet analysis. The approach itself is called the multiresolution approach (MRA) to wavelets. And in the meantime various generalizations (GMRAs) have emerged.

In all of this, there was a second "accident" at play: As it turned out, pyramid algorithms in wavelet analysis now lend themselves via multiresolutions, or nested scales of closed subspaces, to an analysis based on frequency bands. Here we refer to bands of frequencies as they have already been used for a long time in signal processing.

One reason for the success in varied disciplines of the same geometric idea is perhaps that it is closely modeled on how we historically have represented numbers in the positional number system. Analogies to the Euclidean algorithm seem especially compelling.

8.3.1 Multiresolutions

Haar's work in 1909–1910 holds implicitly the key idea which got wavelet mathematics started later with Yves Meyer, Ingrid Daubechies, Stéphane Mallat, and others (see [19] for a resent bibliograph)–namely the idea of a multiresolution. See Figs. 8.1 and 8.2 for details.

This refers to $f_v(x) = \sum_{k\in\mathbb{Z}} v_k \varphi(\cdot - k)$, as a representation by numbers $v = (v_k) \in l^2(\mathbb{Z})$. The dyadic scaling operator as a unitary operator in $L^2(\mathbb{R})$, $U = \frac{1}{\sqrt{2}} f(\frac{x}{2})$.

There are three different ways of representation, function, sequence, each represented by a link as follows:

$$L^2(\mathbb{R}) \underset{\sim}{\supseteq} l^2 \simeq L^2(0,1) : f \longleftrightarrow v \longleftrightarrow V$$

$\sum_k v_k z^k = v(z)$, $l^2 \simeq L^2(\mathbb{T})$, $v = (v_k) \in l^2(\mathbb{Z})$.

Representation of Fourier series:

$$v_k = \langle z^k, v\rangle_{L^2(\mathbb{T})},$$

$$v_k = \int_0^1 e^{-i2\pi kt} v(t)dt,$$

$$\sum_{k\in\mathbb{Z}} |v_k|^2 = \int_0^1 |v(t)|^2 dt.$$

Generating functions (engineering term):

$$H(z) = \sum_{k\in\mathbb{Z}} h_k z^k \quad \text{where } z = e^{i2\pi t}, \quad \text{and } z^k = e^{i2\pi kt}$$

An equivalent form of the dyadic scaling operator U in $L^2(\mathbb{R})$ is as follows: $U : V^\varphi \to V^\varphi$, so U restricts to an isometry in the subspace V^φ.

$$\begin{aligned}(U\varphi)(x) &= \frac{1}{\sqrt{2}}\varphi\left(\frac{x}{2}\right)\\ &= \sqrt{2}\sum_j h_j \varphi(x-j).\end{aligned}$$

Set $\varphi_k = \varphi(\cdot - k)$; then we have

$$\begin{aligned}(U\varphi_k)(x) &= \frac{1}{\sqrt{2}}\varphi_k\left(\frac{x}{2}\right)\\ &= \frac{1}{\sqrt{2}}\varphi\left(\frac{x}{2}-k\right)\\ &= \frac{1}{\sqrt{2}}\varphi\left(\frac{x-2k}{2}\right)\\ &= \sqrt{2}\sum_j h_j \varphi(x-2k-j)\\ &= \sqrt{2}\sum_j h_j \varphi_{2k+j}.\end{aligned}$$

It follows that

$$\begin{aligned}U\sum_k v_k\varphi_k &= \sqrt{2}\sum_l\sum_k v_k h_{2k-l}\\ &= \sqrt{2}\sum_l h_{2k-l}\varphi(x-l) \quad \text{if we let } l = 2k-j.\end{aligned}$$

As a result, we get:

$$(M_H v)_l = \sqrt{2}\sum_{k\in\mathbb{Z}} h_{2k+l} v_k;$$

and

$$(M_H \delta_p)_l = \sqrt{2}\sum_{k\in\mathbb{Z}} h_{2k+l}\delta_{p,k} = h_{2p+l}.$$

Lemma 1. *Setting* $W : f_v \to \sum_{k\in\mathbb{Z}} v_k \varphi(\cdot - k) \in V^\varphi$, *and* $(S_0 f_v)(z) := H(z) f_v(z^2)$, *we then get the intertwining identity* $WS_0 = UW$ *on* V^φ.

Proof.

$$\begin{aligned}(WS_0 f_v)(x) &= (W f_{S_0 v})(x)\\ &= \sum_{k\in\mathbb{Z}} (S_0 v)_k \varphi(x-k)\\ &= \frac{1}{\sqrt{2}}\sum_{k\in\mathbb{Z}} v_k \varphi\left(\frac{x}{2} - k\right)\\ &= \frac{1}{\sqrt{2}} f_v(\frac{x}{2}) = (UW f_v)(x).\end{aligned}$$

□

Note that the isometry $S_0 = S_H$ and $S_1 = S_G$ are $\mathcal{O}_2$ Cuntz algebra.

$$\sqrt{2}\begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \\ \cdots & h_{2p-l} & h_{2p-l+2} & \cdots & \\ \cdots & h_{2p-l-1} & h_{2p-l+1} & \cdots & \\ & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

$S_0^j S_1$ are isometries for $j = 0, 1, 2, \cdots$ that generate $\mathcal{O}_\infty$,

$$Q_j = S_0^j S_1 (S_0^j S_1)^* = S_0^j S_1 S_1^* S_0^{*j} \underset{S_1 S_1^* = 1 - S_0 S_0^*}{=} \underset{P_j}{S_0^j S_0^{*j}} - \underset{P_{j+1}}{S_0^{j+1} S_0^{*j+1}},$$

$$\sum_{j=0}^{n} Q_j = \sum_{j=0}^{n} (P_j - P_{j+1}) = I - P_{n+1} \underset{N\to\infty}{\to} I$$

since

$$P_{n+1} = S_0^{n+1} S_0^{*n+1} \to 0 \quad \text{by pure isometry lemma in [19].}$$

Recall an isometry S in a Hilbert space $\mathcal{H}$ is a shift if and only if $\lim_{N\to\infty} S^N S^{*N} = 0$. L^2 is same as resolution space $\simeq V^\varphi$. The operators Are as follows: $Q_0 = S_1 S_1^*$, $Q_1 = S_0 S_1 S_1^* S_0^* = S_0 S_1 (S_0 S_1)^*$. Note that $Q_0 = I - P_0$, and $Q_j = P_j - P_{j+1}$.

$$\psi_{j,k}(x) = 2^{-j/2}\psi(2^{-j}x + k),\ j = 0, 1, 2, \cdots$$

$$\sum\sum c_{j,k}\psi_{j,k} = \sum\sum f_v(Q_j v)_k = \sum\sum (S_0^j S_1 S_1^* S_0^{*j} v)_k \psi_{j,k}$$

If we now take the adjoint of these matrices, these correspond to the isometries.

$$S_0^* \sim M_H^* \sim F_0,$$

$$S_1^* \sim M_G^* \sim F_1.$$

Theorem 2. *The wavelet representation is given by slanted matrices as follows:* $\sum\sum (F_1 F_0^j v)_k \psi_{j,k} = f_v$

$$c_{j,k} = \langle \psi_{j,k}, f \rangle = \int \psi_{j,k}(x) f(x) dx.$$

In this figure we have the following configuration of closed subspaces: $\cdots \subset V_{-1} \subset V_0 \subset V_1 \subset \cdots$, $V_0 + W_0 = V_1$. The word "multiresolution" suggests a connection to optics from physics. So that should have been a hint to mathematicians to take a closer look at trends in signal and image processing! Moreover, even staying within mathematics, it turns out that as a general notion this same idea of a "multiresolution" has long roots in mathematics, even in such modern and pure areas as operator theory and Hilbert-space geometry. Looking even closer at these interconnections, we can now recognize scales of subspaces (so-called multiresolutions) in classical algorithmic construction of orthogonal bases in inner-product spaces, now taught in lots of mathematics courses under the name of the Gram–Schmidt algorithm. Indeed, a closer look at good old Gram–Schmidt reveals that it is a matrix algorithm, hence new mathematical tools involving non-commutativity!

If the signal to be analyzed is an image, then why not select a fixed but suitable *resolution* (or a subspace of signals corresponding to a selected resolution), and then do the computations there? The selection of a fixed "resolution" is dictated by practical concerns. That idea was key in turning computation of wavelet coefficients into iterated matrix algorithms. As the matrix operations get large, the computation is carried out in a variety of paths arising from big matrix products. The dichotomy, continuous vs. discrete, is quite familiar to engineers. The industrial engineers typically work with huge volumes of numbers.

Numbers! – Why wavelets? What matters to engineers is not really the wavelets, but the fact that special wavelet functions serve as an efficient way to encode large data sets– encode for computations. And the wavelet algorithms are computational. Encoding numbers into pictures, images, or graphs of functions comes later, perhaps at the very end of the computation. But without the graphics, we would not understand any of this as well as we do now. The same can be said for the many issues that relate to the mathematical concept of self-similarity, as we know it from fractals, and more generally from recursive algorithms.

8.4 The Discrete Versus Continuous Wavelet Algorithms

8.4.1 The Discrete Wavelet Transform

If one stays with function spaces, it is then popular to pick the d-dimensional Lebesgue measure on $\mathbb{R}^d$, $d = 1, 2, \cdots$, and pass to the Hilbert space $L^2(\mathbb{R}^d)$ of all square integrable functions on $\mathbb{R}^d$, referring to d-dimensional Lebesgue measure. A wavelet basis refers to a family of basis functions for $L^2(\mathbb{R}^d)$ generated from a finite set of normalized functions ψ_i, the index i chosen from a fixed and finite index set I, and from two operations, one called scaling, and the other translation. The scaling is typically specified by a d by d matrix over the integers $\mathbb{Z}$ such that all the eigenvalues in modulus are bigger than one, lie outside the closed unit disk in the complex plane. The d-lattice is denoted $\mathbb{Z}^d$, and the translations will be by vectors selected from $\mathbb{Z}^d$. We say that we have a wavelet basis if the triple indexed family

$$\psi_{i,j,k}(x) := |detA|^{j/2}\psi_i(A^j x + k)$$

forms an orthonormal basis (ONB) for $L^2(\mathbb{R}^d)$ as i varies in I, $j \in \mathbb{Z}$, and $k \in \mathbb{R}^d$. The word "orthonormal" for a family F of vectors in a Hilbert space $\mathscr{H}$ refers to the norm and the inner product in $\mathscr{H}$: The vectors in an orthonormal family F are assumed to have norm one, and to be mutually orthogonal. If the family is also total (i.e., the vectors in F span a subspace which is dense in $\mathscr{H}$), we say that F is an orthonormal basis (ONB).

While there are other popular wavelet bases, for example frame bases, and dual bases (see e.g., [6, 15] and the papers cited there), the ONBs are the most agreeable at least from the mathematical point of view.

That there are bases of this kind is not at all clear, and the subject of wavelets in this continuous context has gained much from its connections to the discrete world of signal- and image processing.

Here we shall outline some of these connections with an emphasis on the mathematical context. So we will be stressing the theory of Hilbert space, and bounded linear operators acting in Hilbert space $\mathscr{H}$, both individual operators, and families of operators which form algebras.

As was noticed recently the operators which specify particular subband algorithms from the discrete world of signal- processing turn out to satisfy relations that were found (or rediscovered independently) in the theory of operator algebras, and which go under the name of Cuntz algebras, denoted $\mathscr{O}_N$ if n is the number of bands. For additional details, see e.g., [19].

In symbols the C^*-algebra has generators $(S_i)_{i=0}^{N-1}$, and the relations are

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1} \quad \text{(see Fig. 8.3)} \tag{8.1}$$

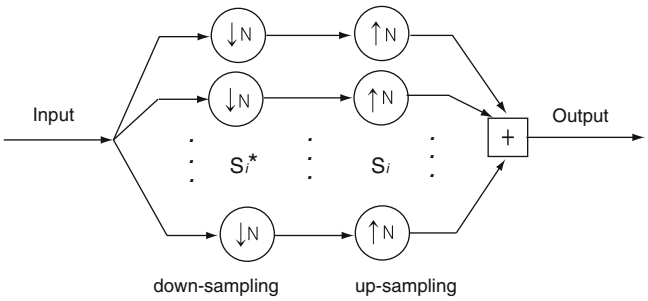


Fig. 8.3 Perfect reconstruction in a subband filtering as used in signal and image processing

(where **1** is the identity element in $\mathcal{O}_N$) and

$$\sum_{i=0}^{N-1} S_i S_i^* = \mathbf{1}, \text{ and } S_i^* S_j = \delta_{i,j}\mathbf{1}. \tag{8.2}$$

In a representation on a Hilbert space, say $\mathcal{H}$, the symbols S_i turn into bounded operators, also denoted S_i, and the identity element **1** turns into the identity operator I in $\mathcal{H}$, i.e., the operator $I : h \to h$, for $h \in \mathcal{H}$. In operator language, the two formulas (8.1) and (8.2) state that each S_i is an isometry in $\mathcal{H}$, and that the respective ranges $S_i\mathcal{H}$ are mutually orthogonal, i.e., $S_i\mathcal{H} \perp S_j\mathcal{H}$ for $i \neq j$. Introducing the projections $P_i = S_i S_i^*$, we get $P_i P_j = \delta_{i,j} P_i$, and

$$\sum_{i=0}^{N-1} P_i = I.$$

Example 8.1. Fix $N \in \mathbb{Z}_+$. Then the easiest representation of $\mathcal{O}_N$ is the following: Let $\mathcal{H} := l^2(\mathbb{Z}_{\geq 0})$, where $\Gamma := \mathbb{Z}_{\geq 0} = \{0\} \cup \mathbb{N} = \{0, 1, 2, \cdots\}$.

Set $\mathbb{Z}_N = \{0, 1, \cdots, N-1\} = \mathbb{Z}/N\mathbb{Z} =$ the cyclic group of order N.

We shall denote canonical ONB in $\mathcal{H} = l^2(\Gamma)$ by $|x\rangle = \delta_x$ with Dirac's formalism. For $i \in \mathbb{Z}_N$ set $S_i = \rho(s_i)$, given by

$$S_i|x\rangle = |Nx + i\rangle, \quad x \in \Gamma \tag{8.3}$$

then

$$S_i^*|x\rangle = \begin{cases} |\frac{x-i}{N}\rangle & \text{if } x - i \equiv 0 \mod N \\ 0 & \text{otherwise.} \end{cases}$$

The reader may easily verify the two relations in (8.2) by hand.

For the use of $\text{Rep}(O_N, \mathcal{H})$ in signal/image processing, more complicated formulas than (8.3) are needed for the operators $S_i = \rho(s_i)$.

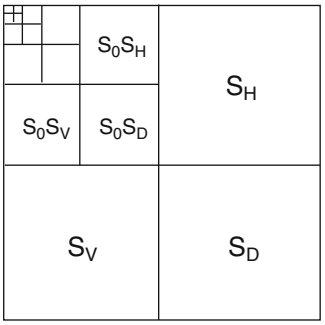
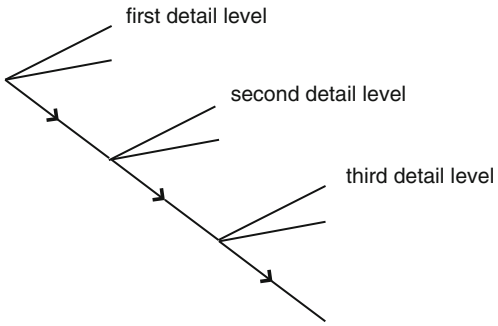


Fig. 8.4 The subdivided squares represent the use of the pyramid subdivision algorithm to image processing, as it is used on pixel squares. At each subdivision step the top left-hand square represents averages of nearby pixel numbers, averages taken with respect to the chosen low-pass filter; while the three directions, horizontal, vertical, and diagonal represent detail differences, with the three represented by separate bands and filters. So in this model, there are four bands, and they may be realized by a tensor product construction applied to dyadic filters in the separate x- and the y-directions in the plane. For the discrete WT used in image-processing, we use iteration of four isometries $S_0, S_H, S_V, and S_D$ with mutually orthogonal ranges, and satisfying the following sum-rule $S_0S_0^* + S_HS_H^* + S_VS_V^* + S_DS_D^* = I$, with I denoting the identity operator in an appropriate l^2-space

In the engineering literature this takes the form of programming diagrams:
If the process of Fig. 8.3 is repeated, we arrive at the discrete wavelet transform



or stated in the form of images ($n = 5$).

But to get successful subband filters, we must employ a more subtle family of representations than those of (8.3) in Example 8.1. We now turn to the study of those representations (Figs. 8.4 and 8.5).

Selecting a resolution subspace $V_0 = closure\ span\{\varphi(\cdot - k)|k \in \mathbb{Z}\}$, we arrive at a wavelet subdivision $\{\psi_{j,k}|j \geq 0, k \in \mathbb{Z}\}$, where $\psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$, and the continuous expansion $f = \sum_{j,k}\langle\psi_{j,k}|f\rangle\psi_{j,k}$ or the discrete analogue derived from the isometries, $i = 1, 2, \cdots, N-1$, $S_0^k S_i$ for $k = 0, 1, 2, \cdots$; called the discrete wavelet transform.

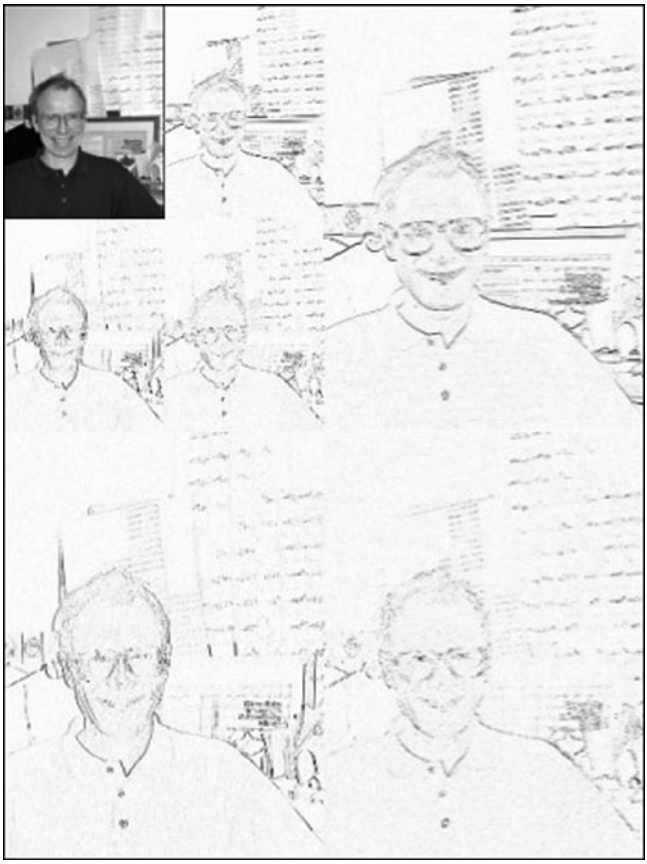


Fig. 8.5 An example of Fig. 8.4: $n = 2$ Jorgensen. Jorgensen's picture after level 2 wavelet decomposition. Notice that The average, horizontal, diagonal and vertical details are captured clockwise. Also, the average detail is decomposed one more time since level 2 decomposition was done

8.4.1.1 Notational Convention

In algorithms, the letter N is popular, and often used for counting more than one thing.

In the present contest of the Discete Wavelet Algorithm (DWA) or DWT, we count two things, "the number of times a picture is decomposed via subdivision". We have used n for this. The other related but different number N is the number of subbands, $N = 2$ for the dyadic DWT, and $N = 4$ for the image DWT. The image-processing WT in our present context is the tensor product of the 1-D dyadic WT, so $2 \times 2 = 4$. Caution: Not all DWAs arise as tensor products of $N = 2$ models. The wavelets coming from tensor products are called separable. When a

particular image-processing scheme is used for generating continuous wavelets it is not transparent if we are looking at a separable or inseparable wavelet!

To clarify the distinction, it is helpful to look at the representations of the Cuntz relations by operators in Hilbert space. We are dealing with representations of the two distinct algebras $\mathcal{O}_2$, and $\mathcal{O}_4$; two frequency subbands vs 4 subbands. Note that the Cuntz $\mathcal{O}_2$, and $\mathcal{O}_4$ are given axiomatic, or purely symbolically. It is only when subband filters are chosen that we get representations. This also means that the choice of N is made initially; and the same N is used in different runs of the programs. In contrast, the number of times a picture is decomposed varies from one experiment to the next!

Summary: $N = 2$ for the dyadic DWT: The operators in the representation are S_0, S_1. One average operator, and one detail operator. The detail operator S_1 "counts" local detail variations.

Image-processing. Then $N = 4$ is fixed as we run different images in the DWT: The operators are now: S_0, S_H, S_V, S_D. One average operator, and three detail operator for local detail variations in the three directions in the plane.

8.4.1.2 Increasing the Dimension

In wavelet theory, [13] there is a tradition for reserving φ for the father function and ψ for the mother function. A 1-level wavelet transform of an $N \times M$ image can be represented as

$$\mathbf{f} \mapsto \begin{pmatrix} \mathbf{a}^1 & | & \mathbf{h}^1 \\ -- & & -- \\ \mathbf{v}^1 & | & \mathbf{d}^1 \end{pmatrix} \tag{8.4}$$

where the subimages $\mathbf{h}^1, \mathbf{d}^1, \mathbf{a}^1$ and $\mathbf{v}^1$ each have the dimension of $N/2$ by $M/2$.

$$\begin{aligned}
\mathbf{a}^1 &= V_m^1 \otimes V_n^1 : \varphi^A(x,y) = \varphi(x)\varphi(y) = \sum_i \sum_j h_i h_j \varphi(2x-i)\varphi(2y-j) \\
\mathbf{h}^1 &= V_m^1 \otimes W_n^1 : \psi^H(x,y) = \psi(x)\varphi(y) = \sum_i \sum_j g_i h_j \varphi(2x-i)\varphi(2y-j) \\
\mathbf{v}^1 &= W_m^1 \otimes V_n^1 : \psi^V(x,y) = \varphi(x)\psi(y) = \sum_i \sum_j h_i g_j \varphi(2x-i)\varphi(2y-j) \\
\mathbf{d}^1 &= W_m^1 \otimes W_n^1 : \psi^D(x,y) = \psi(x)\psi(y) = \sum_i \sum_j g_i g_j \varphi(2x-i)\varphi(2y-j)
\end{aligned} \tag{8.5}$$

where φ is the father function and ψ is the mother function in sense of wavelet, V space denotes the average space and the W spaces are the difference space from multiresolution analysis (MRA) [13].

We now introduce operators T_H and T_G in l^2 such that the expression on the RHS in (8.5) becomes $T_H \otimes T_H$, $T_G \otimes T_H$, $T_H \otimes T_G$ and $T_G \otimes T_G$, respectively.

We use the following representation of the two wavelet functions φ (father function), and ψ (mother function). A choice of filter coefficients (h_j) and (g_j) is made.

In the formulas, we have the following two indexed number systems $\mathbf{a} := (h_i)$ and $\mathbf{d} := (g_i)$, $\mathbf{a}$ is for averages, and $\mathbf{d}$ is for local differences. They are really the input for the DWT. But they also are the key link between the two transforms, the discrete and continuous. The link is made up of the following scaling identities:

$$\varphi(x) = 2\sum_{i\in\mathbb{Z}} h_i \varphi(2x - i);$$

$$\psi(x) = 2\sum_{i\in\mathbb{Z}} g_i \varphi(2x - i);$$

and (low-pass normalization) $\sum_{i\in\mathbb{Z}} h_i = 1$. The scalars (h_i) may be real or complex; they may be finite or infinite in number. If there are four of them, it is called the "four tap", etc. The finite case is best for computations since it corresponds to compactly supported functions. This means that the two functions φ and ψ will vanish outside some finite interval on a real line.

The two number systems are further subjected to orthgonality relations, of which

$$\sum_{i\in\mathbb{Z}} \bar{h}_i h_{i+2k} = \frac{1}{2}\delta_{0,k} \tag{8.6}$$

is the best known.

The systems h and g are both low-pass and high-pass filter coefficients. In (8.5), $\mathbf{a}^1$ denotes the first averaged image, which consists of average intensity values of the original image. Note that only φ function, V space and h coefficients are used here. Similarly, $\mathbf{h}^1$ denotes the first detail image of horizontal components, which consists of intensity difference along the vertical axis of the original image. Note that φ function is used on y and ψ function on x, W space for x values and V space for y values; and both h and g coefficients are used accordingly. The data $\mathbf{v}^1$ denotes the first detail image of vertical components, which consists of intensity difference along the horizontal axis of the original image. Note that φ function is used on x and ψ function on y, W space for y values and V space for x values; and both h and g coefficients are used accordingly. Finally, $\mathbf{d}^1$ denotes the first detail image of diagonal components, which consists of intensity difference along the diagonal axis of the original image. The original image is reconstructed from the decomposed image by taking the sum of the averaged image and the detail images and scaling by a scaling factor. It could be noted that only ψ function, W space and g coefficients are used here. See [34, 37].

This decomposition not only limits to one step but it can be done again and again on the averaged detail depending on the size of the image. Once it stops at certain level, quantization (see [33, 36]) is done on the image. This quantization step may be lossy or lossless. Then the lossless entropy encoding is done on the decomposed and quantized image.

The relevance of the system of identities (8.6) may be summarized as follows. Set

$$m_0(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} h_k z^k \text{ for all } z \in \mathbb{T};$$

$$g_k := (-1)^k \bar{h}_{1-k} \text{ for all } k \in \mathbb{Z};$$

$$m_1(z) := \frac{1}{2}\sum_{k\in\mathbb{Z}} g_k z^k; \text{ and}$$

$$(S_j f)(z) = \sqrt{2} m_j(z) f(z^2), \text{ for } j = 0, 1, \ f \in L^2(\mathbb{T}), \ z \in \mathbb{T}.$$

Then the following conditions are equivalent:

(a) The system of equations (8.6) is satisfied.
(b) The operators S_0 and S_1 satisfy the Cuntz relations.
(c) We have perfect reconstruction in the subband system of Fig. 8.3.

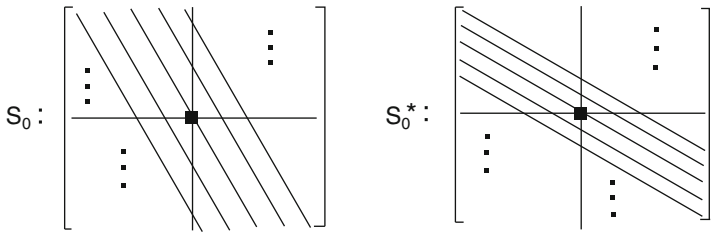
Note that the two operators S_0 and S_1 have equivalent matrix representations. Recall that by Parseval's formula we have $L^2(\mathbb{T}) \simeq l^2(\mathbb{Z})$. So representing S_0 instead as an $\infty \times \infty$ matrix acting on column vectors $x = (x_j)_{j\in\mathbb{Z}}$ we get

$$(S_0 x)_i = \sqrt{2}\sum_{j\in\mathbb{Z}} h_{i-2j} x_j$$

and for the adjoint operator $F_0 := S_0^*$, we get the matrix representation

$$(F_0 x)_i = \frac{1}{\sqrt{2}}\sum_{j\in\mathbb{Z}} \bar{h}_{j-2i} x_j$$

with the overbar signifying complex conjugation. This is computational significance to the two matrix representations, both the matrix for S_0, and for $F_0 := S_0^*$, is slanted. However, the slanting of one is the mirror-image of the other, i.e.,



8.4.1.3 Significance of Slanting

The slanted matrix representations refers to the corresponding operators in L^2. In general operators in Hilbert function spaces have many matrix representations, one for each orthonormal basis (ONB), but here we are concerned with the ONB consisting of the Fourier frequencies z^j, $j \in \mathbb{Z}$. So in our matrix representations for the S operators and their adjoints we will be acting on column vectors, each infinite column representing a vector in the sequence space l^2. A vector in l^2 is said to be of finite size if it has only a finite set of non-zero entries.

It is the matrix F_0 that is effective for iterated matrix computation. Reason: When a column vector x of a fixed size, say 2 s is multiplied, or acted on by F_0, the result is a vector y of half the size, i.e., of size s. So $y = F_0 x$. If we use F_0 and F_1 together on x, then we get two vectors, each of size s, the other one $z = F_1 x$, and we can form the combined column vector of y and z; stacking y on top of z. In our application, y represents averages, while z represents local differences: Hence the wavelet algorithm.

$$\begin{bmatrix} \vdots \\ y_{-1} \\ y_0 \\ y_1 \\ \vdots \\ -- \\ \vdots \\ z_{-1} \\ z_0 \\ z_1 \\ \vdots \end{bmatrix} = \begin{bmatrix} F_0 \\ -- \\ F_1 \end{bmatrix} \begin{bmatrix} \vdots \\ x_{-2} \\ x_{-1} \\ x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix},$$

$$y = F_0 x,$$

$$z = F_1 x.$$

8.4.2 Entropy Encoding

In this section we discuss the encoding aspect of our algorithms. While the theory here dates back to the start of information theory, see [24, 25, 27, 31, 32] and the references cited there, its adaptation to advances in technology have been amazingly successful, see e.g., [24, 25].

An important part of digital imaging is the choice of encoding, for example the encoding of the letters in the alphabet, a, b, c, etc. As a rough principle, one selects the shortest code for the most frequently occurring letter. But to do this, both of these notions must be quantified.

It is clearly of relevance for efficiency, speed, and error detection. As it turns out, probabilities and entropy are helpful. Indeed the way Shannon quantified information a reduction in entropy by an amount A costs A units of information. We have discussed this part of the theory in more detail in [35], and here we offer just an example for illustration of the main points.

There are various entropy encoding schemes being used, and one example of it is Shannon-Fano entropy encoding. In Shannon-Fano entropy encoding, for each data on an image, i.e., pixel, a set of probabilities p_i is computed, where $\sum_{i=1}^{n} p_i = 1$. The entropy of this set gives the measure of how much choice is involved, in the selection of the pixel value of average.

Definition 3. Shannon's entropy $E(p_1, p_2, \ldots, p_n)$ which satisfy the following:

- E is a continuous function of p_i.
- E should be steadily increasing function of n.
- If the choice is made in k successive stages, then E = sum of the entropies of choices at each stage, with weights corresponding to the probabilities of the stages.

$E = -k \sum_{i=1}^{n} p_i \log p_i$. k controls the units of the entropy, which is "bits." logs are taken base 2. [5, 32]

Shannon-Fano entropy encoding is done according to the probabilities of data and the method is as follows:

- The data is listed with their probabilities in decreasing order of their probabilities.
- The list is divided into two parts that has roughly equal probability.
- Start the code for those data in the first part with a 0 bit and for those in the second part with a 1.
- Continue recursively until each subdivision contains just one data. [5, 32]

An example on a text: letters-to-codes: may be better to depict how the mechanism works. Suppose we have a text with letters a, e, f, q, r with the following probability distribution:

Letter	Probability
a	0.3
e	0.2
f	0.2
q	0.2
r	0.1

Then applying the Shannon-Fano entropy encoding scheme on the above table gives us the following assignment.

Letter	Probability	code
a	0.3	00
e	0.2	01
f	0.2	100
q	0.2	101
r	0.1	110

Note that instead of using 8-bits to represent a letter, 2 or 3-bits are being used to represent the letters in this case.

The following is an elementary example of Shannon-Fano entropy encoding

Letter	Probability	code
a	0.3	00
e	0.2	01
f	0.2	100
q	0.2	101
r	0.1	110

While this is an oversimplification, it is nonetheless a key idea used in more realistic algorithms:

- In a given text, list all letters in decreasing order of their probabilities.
- Divide the list into two parts with approximately equal probability (i.e., by the median, the total probability of each part is approximately 0.5).
- For the letters in the first part, start the code with a 0 bit, and for those in the second part with a 1.
- Recursively continue until each subdivision is left with just one letter [5].

Note that the divisions of the list are following a binary tree-rule. The initial important uses of encoding were to texts and to signals. Much more recent uses are to a host of big data sets such as data on the color images. These uses are: quantization, entropy encoding. As a result, the mathematics of encoding has seen a recent revival.

While entropy encoding is popular in engineering, [14, 33, 36], the choices made in signal processing are often more by trial and error than by theory. Reviewing the literature, we found that the mathematical foundation of the current use of entropy in encoding deserves closer attention.

8.5 Slanted Matrix Representations, Frame Estimates, and Computations

We will be using finite and infinite slanted matrices, and we prove two results about their tensor products, and their significance in representation theory. The significance of the slanted property is that matrix products of slanted matrices

become increasingly more sparse (i.e., the resulting matrices have wide patterns of zeros) which makes computations fast. In the application this means that an image, or a more general problem from applied mathematics may be synthesized faster with the use of scale similar orthogonal bases, such as wavelet bases in $L^2(\mathbb{R}^d)$.

In this section we prove mathematical theorems supporting the applications outlined above: The operators in Fig 8.1 have slanted matrix representations determined by the masking sequences (h_n) and (g_n), and with the slanting changing from one operator S to the corresponding adjoint operator S^*. We then show how frame estimates are preserved under filtering with our S-systems, i.e., with the slanted matrices that realize the Cuntz relations in (a) and (b) above. The slanted matrix representation is what make computations fast. The slanting is such that repeated matrix operations in the processing make for more sparse matrices, and hence for a smaller number of computational steps in digital program operations for image processing.

We begin by introducing the Cuntz operators S . The two operators come from the two masking sequences (h_n) and (g_n), also called filter-coefficients, also called low-pass and high-pass filters.

Definition 4. If $(h_n)_{n\in\mathbb{Z}}$ is a double infinite sequence of complex numbers, i.e., $h_n \in \mathbb{C}$, for all $n \in \mathbb{Z}$; set

$$(S_0 x)(m) = \sqrt{2}\sum_{n\in\mathbb{Z}} h_{m-2n}\, x(n) \tag{8.7}$$

and adjoint

$$(S_0^* x)(m) = \sqrt{2}\sum_{n\in\mathbb{Z}} \bar{h}_{n-2m}\, x(n);\ \text{for all } m \in \mathbb{Z}. \tag{8.8}$$

Then

(a) The $\infty \times \infty$ matrix representations (8.7) and (8.8) have the following slanted forms
(b) The set of non-zero numbers in $(h_n)_{n\in\mathbb{Z}}$ is finite if and only if the two matrices in the figure are *banded* (Fig 8.6).
(c) Relative to the inner product

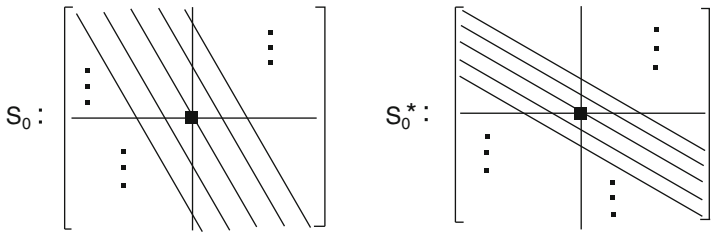


Fig. 8.6 S_0 and S_0^*

$$\langle x|y\rangle_{l^2} := \sum_{n\in\digamma} \bar{x}_n y_n \text{ in } l^2$$

(i.e., conjugate-linear in the first variable), the operator S_0 is *isometric* if and only if

$$\sum_{n\in\digamma} \bar{h}_n h_{n+2p} = \frac{1}{2}\delta_{0,p}, \text{ for all } p \in \mathbb{Z}. \tag{8.9}$$

(d) If (8.9) holds, and

$$(S_1 x)(m) = \sqrt{2}\sum_{n\in\mathbb{Z}} g_{m-2n}x(n), \tag{8.10}$$

then

$$S_0 S_0^* + S_1 S_1^* = I_{l^2} \tag{8.11}$$

$$S_k^* S_l = \delta_{k,l} I_{l^2} \text{ for all } k, l \in \{0, 1\} \tag{8.12}$$

(the Cuntz relations) holds for

$$g_n := (-1)^n \bar{g}_{1-n},\ n \in \mathbb{Z}.$$

Proof. By Parseval's identity and Fourier transforms, we have the *isometric* isomorphism $l^2(\mathbb{Z}) \simeq L^2(\mathbb{T})$ where $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$ is equipped with Haar measure.

Hence the assertions (a)-(d) may be checked instead in the following function representation:

$$f(z) = \sum_{n\in\digamma} x(n)z^n, \tag{8.13}$$

$$m_0(z) = \sum_{n\in\digamma} h_n z^n, \tag{8.14}$$

$$m_1(z) = \sum_{n\in\digamma} g_n z^n; \tag{8.15}$$

setting

$$(S_j f)(z) = m_j(z) f(z^2), \text{ for all } z \in \mathbb{T}, \text{ for all } f \in L^2(\mathbb{T}),\ j = 0, 1. \tag{8.16}$$

In this form, the reader may check that conditions (a)-(d) are equivalent to the following unitary principle: For almost every $z \in \mathbb{T}$ (relative to Haar measure), we have that the 2×2 matrix

$$U(z) = \begin{pmatrix} m_0(z) & m_0(-z) \\ m_1(z) & m_1(-z) \end{pmatrix} \tag{8.17}$$

is unitary; i.e., that $U(z)^* = U(z)^{-1}$, almost every $z \in \mathbb{T}$, where $(U^*)_{k,l} := \bar{U}_{l,k}$ denotes the adjoint matrix. □

8.5.1 Warning

Note that the tensor product of two matrix-functions (8.17) does not have the same form. Nonetheless, there is a more indirect way of creating new multiresolution-wavelet filters from old ones with the use of tensor product; see details below.

Suppose A is an index set and $(v_\alpha)_{\alpha\in A} \subset l^2(\mathbb{Z})$ is a system of vectors subject to the frame bound ($B < \infty$)

$$\sum_{\alpha\in A} |\langle v_\alpha|x\rangle_{l^2}|^2 \le B\|x\|_{l^2}^2, \text{ for all } x \in l^2. \tag{8.18}$$

Set

$$w_{j,k} := S_0^j S_1 v_\alpha,\ j \in \mathbb{N}_0 = \{0, 1, 2, \ldots\}, \alpha \in A. \tag{8.19}$$

If the unitarity condition (8.17) in the lemma is satisfied, then the induced system (8.19) satisfies the same frame bound (8.18).

Proof. Introducing Dirac's notation for rank-one operators:

$$|u\rangle\langle v||x\rangle = \langle v|x\rangle|u\rangle, \tag{8.20}$$

we see that (8.18) is equivalent to the following operator estimate

$$\sum_{\alpha\in A} |v_\alpha\rangle\langle v_\alpha| \le BI_{l^2} \tag{8.21}$$

where we use the standard ordering of the Hermitian operators, alias matrices.

For the system $(w_{j,\alpha})_{(j,\alpha)\in\mathbb{N}_0\times A}$ in (8.19), we therefore get

$$\begin{aligned}\sum_{(j,\alpha)\in\mathbb{N}_0\times A} |w_{j,\alpha}\rangle\langle w_{j,\alpha}| &= \sum_j \sum_\alpha S_0^j S_1 |v_\alpha\rangle\langle v_\alpha| S_1^* S_0^{*j} \\ &\underset{by(8.21)}{\le} B\sum_j S_0^j S_1 S_1^* S_0^{*j} \le B,\end{aligned}$$

Since

$$\sum_j S_0^j S_1 S_1^* S_0^{*j} = I - S_0^{n+1} S_0^{*n+1} \le I \quad \text{for all } n.$$

But the RHS in the last expression is the limit of the finite partial sums

$$\sum_{j=0}^N S_0^j S_1 S_1^* S_0^{*j} = \sum_{j=0}^N S_0^j (I - S_0 S_0^*) S_0^{*j} \tag{8.22}$$

$$= I_{l^2} - S_0^{N+1} S_0^{*N+1} \tag{8.23}$$

$$\le I_{l^2} \text{ since} \tag{8.24}$$

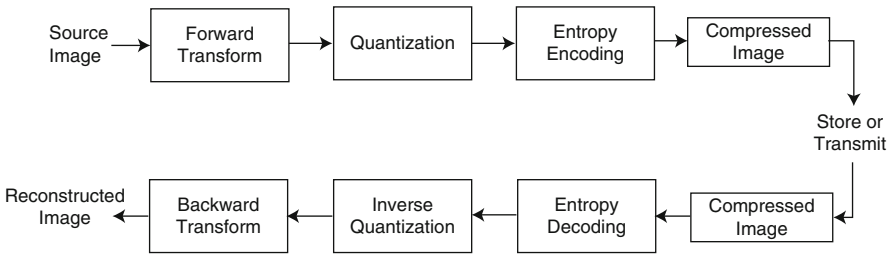


Fig. 8.7 Outline of the wavelet image compression process [33]

$P_{N+1} := S_0^{N+1} S_0^{*N+1}$ is a projection for all $N \in \mathbb{N}_0$. In fact

$$\cdots \leq P_{N+1} \leq P_N \leq \cdots \leq P_1 = S_0 S_0^*;$$

and P_1 denotes the projection onto $S_1 l^2$. □

8.5.1.1 Digital Image Compression

In [23], we showed that use of Karhunen–Loève's theorem enables us to pick the best basis for encoding, thus to minimize the entropy and error, to better represent an image for optimal storage or transmission. Here, optimal means it uses least memory space to represent the data; i.e., instead of using 16 bits, use 11 bits. So the best basis found would allow us to better represent the digital image with less storage memory.

The particular orthonomal bases (ONBs) and frames which we use come from the operator theoretic context of the Karhunen–Loève theorem [1]. In approximation problems involving a stochastic component (for example noise removal in time-series or data resulting from image processing) one typically ends up with correlation kernels; in some cases as frame kernels.

Summary of the mathematics used in the various steps of the image compression flow chart in Fig. 8.7:

- At the beginning of the diagram (source image) we will typically input a digital image. Of the possible forward transforms that apply to images, this proposal uses the discrete wavelet transforms.
- The quantization step refers to the following: the output from the transform will be quantized in the sense of economy of processing; for instance, with the thresholding method. After the wavelet forward transform, the image is decomposed into different details at different levels; the thresholding method will eliminate some of these details selectively resulting in lossy compression. In our recent paper [23], we initiated the use of thresholding for exactly this purpose.

- In our approach to image compression the encoding step does the following: with the quantization, the process was lossy where the step is irreversible. With the entropy encoding, if we started off with an 16 bit image we find a better representation meaning we use fewer bits to present the pixel values. This step is lossless. Entropy encoding has been used for long time in information theory, and it has found a recent revival in data compression schemes. The idea is to assign codes to symbols in order to match code lengths with the frequencies of occurrence of the symbols. By entropy we refer to a priori probability distributions of symbols. Entropy encoders compress data by replacing equal length codes with codes where the length of each code would be determined by quantitative measures of entropy. Therefore, the most frequent symbols will be assigned the shortest code. Hence the economy.

  There are number of other entropy encoding schemes, Shannon-Fano encoding, Huffman coding, arithmetic coding, etc. [5, 12, 38, 39]. These alternative encoding schemes have been successful in signal compression and in simplest static codes, but in the most recent uses of discrete algorithm on images the entropy encoding proved to have practical advantages. The lossless data compression has in common with the discrete wavelet synthesis that exact data can be reconstructed from the compressed versions. Hence, the processing is reversible. Lossless compression is used for example in zip file forward and gzip in Unix/Linux. For images, the formats png and gif also use lossless compression. Examples are executable programs and source codes.

  In carrying out compression, one generates the statistical output from the input data, and the next step maps the input data into bit-streams, achieving economy by mapping data into shorter output codes.
- The result of the three steps is the compressed information which could be stored in the memory or transmitted through internet. The technology in both ECW (Enhanced Compressed Wavelet technology) and JPEG2000 allow for fast loading and display of large image files. JPEG2000 is wavelet-based image compression [18, 19]. There is a different brand new compressing scheme called Enhanced Compressed Wavelet technology (ECW). Both JPEG2000 and ECW are used in reconstruction as well.

8.5.2 Reconstruction

The methods described above apply to reconstruction questions from signal processing. Below we describe three important instances of this:

1. In entropy decoding, the fewer bits assigned to represent the digital image are mapped back to the original number of bits without losing any information.
2. In inverse quantization, there is not much of recovery to be obtained if thresholding was used for that was a lossy compression. For other quantization methods, this may yield some recovery.

3. The backward transform is the wavelet inverse transform where the decomposed image is reconstructed back into an image that is of the same size as the original image in dimension but maybe of smaller size in memory.

Let $\mathbb{T} = \mathbb{R}/\mathbb{Z} \simeq \{z \in \mathbb{C}; |z| = 1\}$. Let A be a $d \times d$ matrix over $\mathbb{Z}$ such that $\|\lambda\| > 1$ for all $\lambda \in \text{spec}(A)$. Then the order of the group $A^{-1}\mathbb{Z}^d/\mathbb{Z}^d$ is $|det A| := N$. Consider $A^{-1}\mathbb{Z}^d/\mathbb{Z}^d$ as a subgroup in $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$.

We shall need the following bijection correspondence:

Definition 5. For $z \in \mathbb{T}^d$, $z_j = e^{i2\pi\theta_j}$, $1 \le j \le d$, set

$$z^A = (e^{i2\pi\eta_j})^d, \tag{8.25}$$

where

$$A\theta = \eta, \tag{8.26}$$

i.e., $\sum_{j=1}^d A_{kj}\theta_j = \eta_k$, $1 \le k \le d$. Then for $z \in \mathbb{T}^d$, the set $\{w \in \mathbb{T}^d; w^A = z\}$ is in bijection correspondence with the finite group $A^{-1}\mathbb{Z}^d/\mathbb{Z}^d$.

Definition 6. Let $\mathscr{U}_N(\mathbb{C})$ be the group of all unitary $N \times N$ matrices, and let $\mathscr{U}_N^A(\mathbb{T}^d)$ be the group of all measurable function

$$\mathbb{T}^d \ni z \mapsto U(z) \in U_N(\mathbb{C}). \tag{8.27}$$

Let $\mathscr{M}_N^A(\mathbb{T}^d)$ be the multiresolution functions,

$$\mathbb{T}^d \ni z \mapsto M(z) = (m_j(z))_1^N \in \mathbb{C}^N; \tag{8.28}$$

i.e., satisfying

$$\frac{1}{N} \sum_{w \in \mathbb{T}^d, w^A = z} \overline{m_j(w)} m_k(w) = \delta_{j,k}. \tag{8.29}$$

Example 8.2. Let $\{k_j\}_{j=1}^N$ be a selection of representatives for the group $\mathbb{Z}^d/A\mathbb{Z}^d$; then

$$M_0(z) = (z^{k_j})_{j=1}^N \tag{8.30}$$

is in $\mathscr{M}_N^A(\mathbb{T}^d)$.

For $z \in \mathbb{T}^d$ and $k \in \mathbb{Z}^d$ we write

$$z^k = (z_1^{k_1} z_2^{k_2} \cdots z_d^{k_d}) \quad \text{multinomial.} \tag{8.31}$$

Lemma 7. *There is a bijection between $\mathscr{U}_N^A(\mathbb{T}^d)$ and $\mathscr{M}_N^A(\mathbb{T}^d)$ given as follows:*

(i) If $u \in \mathscr{U}_N^A(\mathbb{T}^d)$ set

$$M_U(z) = U(z^A) M_0(z) \tag{8.32}$$

where M_0 is the function in (8.30).

(ii) If $M \in \mathscr{M}_N^A(\mathbb{T}^d)$, *set*

$$U_M(z) = (U_{i,j}(z))_{i,j=1}^N \tag{8.33}$$

with

$$U_{i,j}(z) = \frac{1}{N} \sum_{w \in \mathbb{T}^d, w^A = z} \overline{M_j(w)} M_i(w). \tag{8.34}$$

Proof. Case (i). Given $U \in \mathscr{U}_N^A(\mathbb{T}^d)$, we compute

$$\begin{aligned}
\frac{1}{N} \sum_{w^A=z} \overline{(M_U)_i(w)}(M_U)_j(w) &\underset{\text{by (8.32)}}{=} \frac{1}{N} \sum_{w^A=z} \overline{(U(z)(w^k)_i} (U(z)(w^k)_j \\
&= \frac{1}{N} \sum_{w^A=z} \overline{w^{k_i}} w^{k_j} \\
&= \frac{1}{N} \sum_{w^A=z} w^{k_j - k_i} \\
&= \delta_{i,j}
\end{aligned}$$

where we have used the usual Fourier duality for the two finite groups

$$A^{-1}\mathbb{Z}^d/\mathbb{Z}^d \simeq \mathbb{Z}^d/A\mathbb{Z}^d; \tag{8.35}$$

i.e., a finite Fourier transform.

Case (ii). Given $M \in \mathscr{M}_N^A(\mathbb{T}^d)$, we compute $U_{i,j}(z)$ according to (8.34). The claim is that $U(z) = (U_{i,j}(z))$ is in $U_N(\mathbb{C})$, i.e.,

$$\sum_{l=1}^N \overline{U_{l,j}(z)} U_{l,j}(z) = \delta_{i,j} \qquad \text{for all } z \in \mathbb{T}^d. \tag{8.36}$$

Proof. Proof of (8.36):

$$\begin{aligned}
&\sum_{l=1}^N \overline{U_{l,j}(z)} U_{l,j}(z) \\
&\quad \underset{\text{by (8.34)}}{=} \frac{1}{N^2} \sum_l \sum_w \sum_{w'} M_i(w) \overline{M_l(w) M_j(w')} M_l(w') \\
&\quad = \frac{1}{N} \sum_w \sum_{w'} \delta_{w,w'} M_i(w) \overline{M_j(w')}
\end{aligned}$$

$$= \frac{1}{N} \sum_{w} M_i(w) \overline{M_j(w)}$$

$$\underset{\text{by (8.30)}}{=} \delta_{i,j}.$$

□

This completes the proof of (8.36) and therefore the Lemma. In the summations we consider independent pairs $w, w' \in \mathbb{T}^d$ satisfying $w^A = w'^A = z$. So each ranges over a set of cardinality $N = |detA|$. □

In the simplest case of $N = 2$, $d = 1$, we have just two frequency bands, are two filter functions

$$M = \begin{pmatrix} m_0 \\ m_1 \end{pmatrix}$$

see (8.28), (8.29). In that case we select the two points ± 1 in bt. In additive notation, these represents the two elements in $\frac{1}{2}\mathbb{Z}/\mathbb{Z}$ viewed as a subgroup of $\mathbb{T} \simeq \mathbb{R}/\mathbb{Z}$. The band-conditions are then $M_0(1) = 1$ and $M_1(-1) = 1$, in addition notation (Fig. 8.8):

Multi-band: high-pass/low-pass. If $N > 2$, there will be more than two frequency bands. They can be written with the use of duality for the finite group $A^{-1}\mathbb{Z}^d/\mathbb{Z}^d$ in (8.36). Recall $|detA| = N$, and the group is cyclic of order N. The matrix for its Fourier transform matrix is denoted H_N with H Hadamard.

$$H_N = \frac{1}{\sqrt{N}} \left(e^{i\frac{2\pi jk}{N}} \right)_{j,k \in \mathbb{Z}_N}. \tag{8.37}$$

Lemma 8. *If $U \in \mathscr{U}_N^A(\mathbb{T}^d)$ and $M_U = U(z^A)M_0(z)$ is the multiresolution for the lemma then*

$$\mathbb{T}^d \ni z \mapsto H_N M_U(z) \tag{8.38}$$

satisfying the multi-band frequency pass condition.

Proof. Immediate. □

Lemma 9. *Let M be a multi-band frequency pass function, see (8.38), and assume continuity at the zero-multi frequency, i.e., at the unit element in $\mathbb{T}^d$ (with multiplicative notation).*

(a) Then there are functions $(\psi_j)_{j=0}^{N-1}$ in $L^2(\mathbb{R}^d)$ such that $\psi_0 = \varphi$ satisfies

$$\widehat{\varphi}(\theta) = \prod_{k=1}^{\infty} m_0(A^{-1}\theta) \tag{8.39}$$

$$\widehat{\psi_j}(\theta) = m_j(A^{-1}\theta)\widehat{\varphi}(A^{-1}\theta), \quad j = 1, \cdots, N-1, \quad \text{for all } \theta \in \mathbb{R}^d \tag{8.40}$$

where we have used addition notation, i.e., $\theta \in \mathbb{R}^d$ is written as a column vector, and $A^{-1}\theta$ is a matrix action.

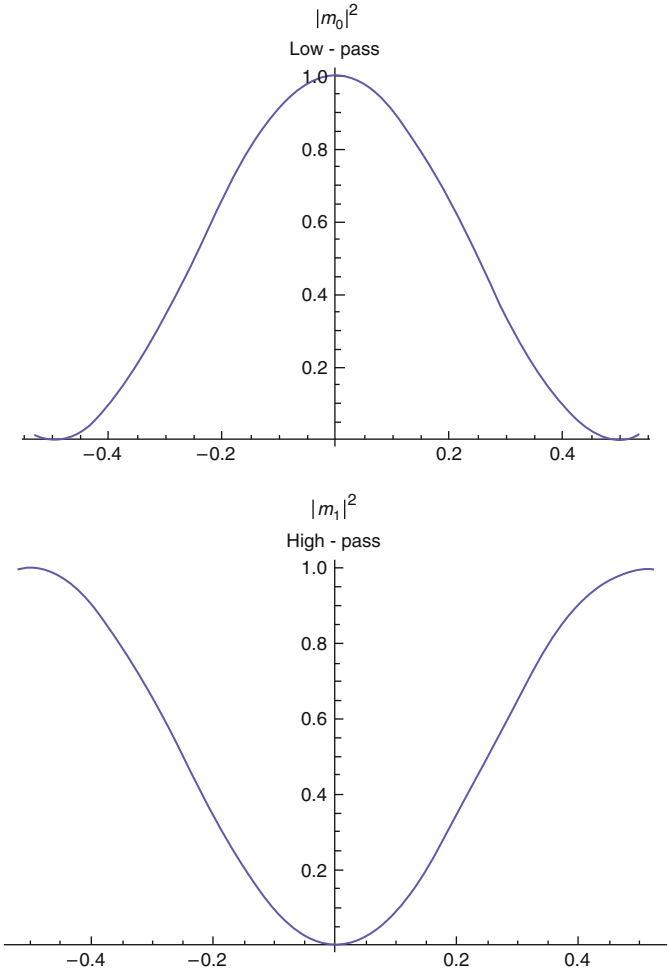


Fig. 8.8 The two filters of probability distributions

(b) With the tripple index set $j = 1, 2, \cdots, N-1$, $k \in \mathbb{Z}$, *and* $l \in \mathbb{Z}^d$, *we get the system*

$$\psi_{j,k,l}(x) = |detA|^{k/2}\psi_j(A^k x - l). \tag{8.41}$$

(c) While the system in (8.41) in general does not form an orthonormal basis (ONB) in $L^2(\mathbb{R}^d)$, *it satisfies the following Parseval-fram condition: For every* $f \in L^2(\mathbb{R}^d)$, *we have*

$$\int_{\mathbb{R}^d} |f(x)|^2 dx = \sum_{1 \le j < N} \sum_{k \in \mathbb{Z}} \sum_{l \in \mathbb{Z}^d} |\langle \psi_{j,k,l}, f \rangle_{L^2(\mathbb{R}^d)}|^2 \tag{8.42}$$

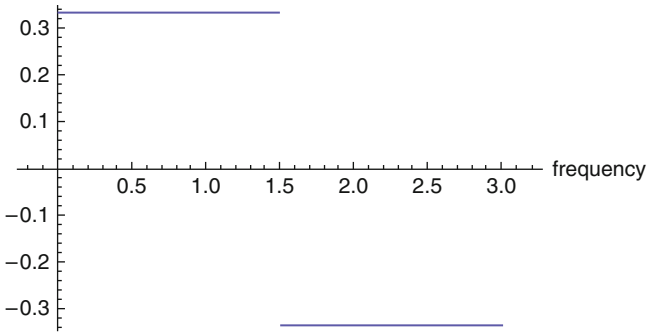


Fig. 8.9 Streched Haar Wavelet

with the expression $\langle \cdot, \cdot \rangle_{L^2(\mathbb{R}^d)}$ *on the RHS in (8.42) representing the usual* $L^2(\mathbb{R}^d)$ *inner product*

$$\langle \psi, f \rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} \overline{\psi(x)} f(x) dx. \tag{8.43}$$

Proof. The essential idea is contained in [9], and the remaining details are left to the reader. □

Remark 10. The condition in (8.42) (Parseval-frame) is weaker than ONB, referring to function $(\psi_{j,k,l})$ in (8.41). Rather than asking for an ONB in $L^2(\mathbb{R}^d)$, we seek instead a Parseval-frame. But there is a variety of explicit additional conditions on the given wavelet filter from (8.39) and (8.40) which imply that $(\psi_{j,k,l})$ is in fact an ONB.

In the language of frequency bands, the low pass filter should not pass "false" frequencies. The simplest example of a non ONB Parseval wavelet is the stretched Haar wavelet (Fig. 8.9)

$$\psi(x) = \begin{cases} \frac{1}{3} & \text{if } 0 \le x < \frac{3}{2} \\ \frac{-1}{3} & \text{if } \frac{3}{2} \le x < 3 \\ 0 & \text{if } x \in \mathbb{R} \setminus [0, 3]. \end{cases}$$

For additional details, see [9].

Lemma 11. *Let*

$$M(z) = \sum_{k \in \mathbb{Z}^d} h_k z^k, \quad \textit{with} \quad z = e^{i 2\pi \theta}, \quad \theta \in \mathbb{R}^d, \tag{8.44}$$

$$M(\theta) = \sum_{k \in \mathbb{Z}^d} h_k e^{i 2\pi k \cdot \theta}, \quad k \cdot \theta = \sum_{j=1}^{d} k_j \theta_j, \tag{8.45}$$

be the Fourier expansion of a filter function note (8.45) is the same as (8.44), but written in the additional form of Fourier analysis for $\mathbb{T}^d$. *Then the function* φ *in (8.39) satisfies*

$$\varphi(x) = |detA| \sum_{k\in\mathbb{Z}^d} h_k \varphi(A^t x - k) \tag{8.46}$$

where A^t *denotes the transpose matrix to* A.

Proof. The result follows from an $L^2(\mathbb{R}^d)$-Fourier transform applied to both sides in formular (8.46). □

Lemma 12. *Let* M *be a low-pass filter function, and let* φ *be a scaling function, see (8.46), or equivalently (8.39). Then the operator*

$$T_\varphi : l^2(\mathbb{Z}^d) \to L^2(\mathbb{R}^d)$$

given by

$$T_\varphi((\xi)_k(x)) = \sum_{k\in\mathbb{Z}^d} (\xi)_k \varphi(x-k)$$

is isometric, i.e., we have

$$\sum_{k\in\mathbb{Z}^d} |\xi_k|^2 = \int_{\mathbb{R}^d} |\sum_{k\in\mathbb{Z}^d} \xi_k \varphi(x-k)|^2 dx.$$

Proof. See [9]. □

8.6 Consistency of Tensor Product for Filters and for Functions on $\mathbb{R}^d$

There are two operations we shall need to perform on infinite slanted matrices, matrix product and tensor product. The first is for the computation of the coefficients in expansions with scale similar orthogonal bases, such as wavelet bases. We prove that when these operations are performed, the resulting new slanted matrices become increasingly more sparse (i.e., the resulting matrices have wide patterns of zeros) which makes computations fast.

Since, by Lemma 11, the filter functions coming from the group $\mathscr{U}_N^A(\mathbb{T}^d)$, it follows that the corresponding filter functions $\mathscr{M}_N^A(\mathbb{T}^d)$ are also closed under tensor product.

8.6.1 Creating Multiresolutions and Wavelets

Using the lemmas, we may create new filters and new multiresolutions from old. In outline, the process is as follows:

- projections.
  ↓
- unitary matrices.
  ↓
- unitary matrix-functions.
  ↓
- filter functions.
  ↓
- tensor product.
  ↓
- scaling functions.
  ↓
- multi-resolutions.
  ↓
- wavelet functions (higher dimensions with tensor product).

Definition 13. Let $\mathscr{H}_i$, $i = 1, 2$ be Hilbert spaces, and let T_i, $i = 1, 2$, be linear operators in the respective spaces. By the tensor product $T_1 \otimes T_2$ we mean the operator $T_1 \otimes T_2$ in $\mathscr{H}_1 \otimes \mathscr{H}_2$ given by

$$(T_1 \otimes T_2)(u_1 \otimes u_2) = (T_1 u_1) \otimes (T_2 u_2), \quad \text{for all } u_i \in \mathscr{H}_i, \quad i = 1, 2. \tag{8.47}$$

An operator P in a Hilbert space $\mathscr{H}$ is called a *projection* if $P = P^* = P^2$. An operator U is called *unitary* if and only if $U^* = U^{-1}$, i.e., $UU^* = U^*U = I$ where I denotes the identity operator.

Easy Facts:

1. The tenstor product of two projections is a projection.
2. The tensor product of two unitary operators is a unitary operator.
3. If P is a projection in a Hilbert space $\mathscr{H}$, then the function

$$U(z) := zP + (I - P) = zP + P^{\perp}, \quad (z \in \mathbb{T}) \tag{8.48}$$

maps $\mathbb{T}$ into the group of all unitary operators.

Proof. of (iii). The conclusion follows from representing (8.48) in the following operator-block matrix form

$$U(z) = \begin{bmatrix} z & 0 \\ 0 & 1 \end{bmatrix}_{P^{\perp}}^{P} \tag{8.49}$$

Note that the RHS in (8.49) is unitary if and only if $|z| = 1$. □

Definition 14. 1. Let A be a $d \times d$ matrix on $\mathbb{Z}$ such that

$$\operatorname{spec}(A) \subset \{\lambda \in \mathbb{C}; |\lambda| > 1\}.$$

Let $\mathbb{Z}^d \overset{\pi_A}{\to} \mathbb{Z}^d/A\mathbb{Z}^d =: Q_A$ be the natural quotient mapping. If $\rho_A : \mathbb{Z}^d/A\mathbb{Z}^d \to \mathbb{Z}^d$ satisfying $\pi_A \circ \rho_A = id$, we saw that

$$\mathbb{T}^d \ni z \mapsto \left(z^{\rho_A(q)}\right)_{q\in Q_A}$$

is a multiresolution filter on $\mathbb{T}^d$.

2. If B is a $e \times e$ matric over $\mathbb{Z}$ such that $\text{spec}(B) \subset \{\lambda \in \mathbb{C}; |\lambda| > 1\}$, we introduce $Q_B = \mathbb{Z}^e/B\mathbb{Z}^e$ and ρ_B as in (1) by analogy. We then get a $d + e$ multiresolution filter

$$\mathbb{T}^d \times \mathbb{T}^e \ni (z, w) \mapsto \left(z^{\rho_A(q)} w^{\rho_B(r)}\right)_{q\in Q_A, r\in Q_B}.$$

3. If $N_A := |detA|$, and $N_B := |detB|$, then the filter in (2) takes values in $bc^{N_A+N_B}$.

Corollary 15. *The families of multiresolution filters $\mathscr{M}^A_{N_A}(\mathbb{T}^d)$ is closed under tensor product, i.e.,*

$$\mathscr{M}^A_{N_A}(\mathbb{T}^d) \otimes \mathscr{M}^B_{N_B}(\mathbb{T}^e) \subset \mathscr{M}^{A\otimes B}_{N_A\cdot N_B}(\mathbb{T}^{d+e}).$$

Proof. By the lemma, we only need to observed that the unitary matrix-functions $\mathbb{T}^d \ni z \mapsto U_A(z) \in U_{N_A}(\mathbb{C})$ are closed under tensor product, i.e.,

$$\mathbb{T}^d \otimes \mathbb{T}^e \ni (z, w) \mapsto U_A(z) \otimes U_B(w) \in U_{N_A\cdot N_B}(\mathbb{C}).$$

□

8.6.2 Tensor Products: Applications and Examples

Lexicographical of basis vectors, for example $\mathbb{C}^2 \otimes \mathbb{C}^2 \to (11), (12), (21), (22)$. Scaling matrices:

$$I \otimes B \longrightarrow \begin{pmatrix} B & 0 \\ 0 & B \end{pmatrix}.$$

Fourier Hadamard transform matrices:

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \longrightarrow \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

$$H_2 \otimes H_2 \neq H_4.$$

Compare with the matrix F_4 for the Fourier transform on $\mathbb{Z}_4$:

$$H_4 : \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & i & -1 & -i \\ 1 & -1 & 1 & -1 \\ 1 & -i & -1 & i \end{pmatrix}.$$

8.6.3 *Unitary Matrix-Functionsi $U_p(z)$:*

$$\mathbb{T} \ni z \longmapsto zP + P^{\perp} = \begin{pmatrix} zP & 0 \\ 0 & I - P \end{pmatrix}. \tag{8.50}$$

Here the function $U_P(z)$ is constructed from a fixed projection P in a Hilbert space $\mathscr{H}$, and we set $P^{\perp} = I_{\mathscr{H}} - P$ the projection onto the orthogonal complement of $P\mathscr{H}$ in $\mathscr{H}$.

Then for $(z, w) \in \mathbb{T}^2 = \mathbb{T} \times \mathbb{T}$ we have:

$$U_P(z) \times U_P(w) = \begin{pmatrix} zwP \otimes P & 0 \\ 0 & I_{\mathscr{H}\otimes\mathscr{H}} - P \otimes P \end{pmatrix}.$$

Giving (8.50) the matrix representation

$$\begin{pmatrix} z & 0 \\ 0 & 1 \end{pmatrix}$$

we get

$$U_P(z) \otimes U_P(w) = \begin{pmatrix} zw & 0 & 0 & 0 \\ 0 & z & 0 & 0 \\ 0 & 0 & w & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}. \tag{8.51}$$

For the slanted matrices

$$F = \begin{pmatrix} h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & h_0 & h_1 & h_2 & h_3 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & h_0 & h_1 & h_2 & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & h_0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}$$

we see that sparsity increases when both matrix multiplication and tensor product $\otimes$ is applied. If for example the two matrices F and F' have slanting degree 2, then the matrix product FF' has slanting degree 4. An $n-$fold product of slanted matrices of slant 2 is again slanted but of slanting degree 2^n.

The situation for tensor product is more subtle.

Example 8.3. Set

$$F = \begin{pmatrix} a & b & 0 \\ 0 & 0 & a \\ 0 & 0 & 0 \end{pmatrix},$$

and

$$F' = \begin{pmatrix} \alpha & \beta & 0 \\ 0 & 0 & \alpha \\ 0 & 0 & 0 \end{pmatrix}.$$

Then $F \otimes F'$ has the matrix representation

$$F \otimes F' \sim \begin{pmatrix} aF' & bF' & 0 \\ 0 & 0 & aF' \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} a\alpha & a\beta & 0 & b\alpha & b\beta & 0 & 0 & 0 & 0 \\ 0 & 0 & a\alpha & 0 & 0 & b\alpha & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & a\alpha & a\beta & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a\alpha \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

8.6.3.1 Slanted Matrices

Definition 16. An (infinite) matrix $F = (a_{j,k})$, $j, k \in \mathbb{Z}$, is said to be slanted of degree d if there is a function f on $\mathbb{Z}$ (depending only on F) such that

$$a_{j,k} = f_F(k - dj) \tag{8.52}$$

holds for all $j, k \in \mathbb{Z}$. We write $deg_s(F) = d$.

Lemma 17. *Let F and F' be slanted matrices, with F of slanting degree d and F' of degree e. Then the matrix product $G = FF'$ is again a slanted matrix, and for the degree we have*

$$deg_s(FF') \geq deg_s(F) \cdot deg_s(F'). \tag{8.53}$$

Proof. 1. By matrix-multiplicative. For the entries in $G = FF'$,

$$G = (c_{i,j}), \quad i, j \in \mathbb{Z},$$

$$c_{i,j} = \sum_{k\in\mathbb{Z}} a_{i,k}a'_{k,j} = \sum_{k\in\mathbb{Z}} f_F(k - di)\, f_{F'}(j - ek).$$

Hence the matrix entries $c_{i,j+(d\cdot e)i}$ are represented by a function g in j, i.e.,

$$c_{i,j+(d\cdot e)i} = g(j), \quad \text{for all } i, j \in \mathbb{Z}.$$

2. By generating functions. Because of the assumptions on the entries in the matrices $F = (a_{i,j})$ and $F' = (a'_{i,j})$ the generating functions (alias, frequency response functions) are in fact Fourier series.

   As a result, a slanted matrix F represents a bounded operator T_F in $L^2(\mathbb{T}) \simeq l^2(\mathbb{Z})$. An easy computation shows that F is slanted of degree d if and only if

$$T_F M_{z^d} = M_z T, \tag{8.54}$$

   where

$$(M_z\xi)(z) = z\xi(z), \quad \text{for all } z \in \mathbb{T};$$

   and

$$(M_z d\,\xi)(z) = z^d\xi(z), \quad \xi \in L^2(\mathbb{T});$$

   i.e., $M_{z^d} = (M_z)^d$, $d \in \mathbb{Z}_+$.

   As a result, we have

$$T_F M_{z^d} = M_z T_F, \quad \text{and} \tag{8.55}$$

$$T_{F'} M_{z^e} = M_z T_{F'}. \quad \text{Since} \tag{8.56}$$

$$T_G = T_F T_{F'} \quad \text{we get} \tag{8.57}$$

$$\begin{aligned} M_z T_G &\underset{by(8.57)}{=} M_z T_F T_{F'} \\ &\underset{by(8.55)}{=} T_F M_{z^d} T_{F'} \\ &\underset{by(8.56)}{=} T_F T_{F'} M_{(z^d)^e} \\ &= T_G M_{z^{d\cdot e}}, \end{aligned}$$

   proving the assertion in the lemma.

□

The proof of the following result is analogous to that in Lemma 17 above.

Lemma 18. *Let F and F' be slanted matrices with $\mathbb{Z}$ as index set for rows and columns, or possibly subset of $\mathbb{Z}$. If $deg_s(F) = d$, and $deg_s(F') = e$, then $F \otimes F'$ is slanted relative to the index set $\mathbb{Z}^2$ by vector degree (d, e). Setting*

$$(F \otimes F')_{(i,j),(k,l)} = F_{i,k} F'_{j,l} \tag{8.58}$$

there is a function g on $\mathbb{Z}^2$ *such that*

$$(F \otimes F')_{(i,j),(k,l)} = g(k - di, l - ej) \quad \textit{for all } (i,j) \in \mathbb{Z}^2 \quad \textit{and all } (k,l) \in \mathbb{Z}^2.$$

8.7 Rep($\mathscr{O}_N, \mathscr{H}$)

Here we show that the algorithms developed in the previous two sections may found from certain representations of an algebra in a family indexed by a positive integer N, called the Cuntz algebras $\mathscr{O}_N$.

It will be important to make a distribution between $\mathscr{O}_N$ as a C^*-algebra, and its representation by operators in some Hilbert space. As we show distinct representation of $\mathscr{O}_N$ yield distinct algorithms, distinct wavelets, and distinct matrix computations.

It is known that $\mathscr{O}_N$ is the unique (up to isomorphism) C^*-algebra on generated $(s_i)_i \in \mathbb{Z}_N$, and relations

$$s_i^* s_j = \delta_{i,j} 1, \quad \sum_{i \in \mathbb{Z}_N} s_i s_i^* = 1. \tag{8.59}$$

Here s_i in (8.59) is a symbolic expression and 1 denotes the unit-element in the C^*-algebra generated by $\{s_i; i \in \mathbb{Z}_N\}$. Hence a representation ρ of $\mathscr{O}_N$ acting on a Hilbert space $\mathscr{H}$, $\rho \in \text{Rep}(\mathscr{O}_N, \mathscr{H})$, is a system of operators $S_i = S_i^{(\rho)} = \rho(s_i)$ satisfying the same relations (8.59), but with 1 replaced by $I_{\mathscr{H}} = \rho(1) =$ the identity operator in $\mathscr{H}$.

From (8.59) it follows that $\mathscr{O}_N \otimes \mathscr{O}_M = \mathscr{O}_{NM}$. Hence it follows from an analysis of tensor product of representation that not all

$$\rho \in \text{Rep}(\mathscr{O}_N, \mathscr{H})$$

is a tensor product of a pair of representations, one of $\mathscr{O}_N$ and the second of $\mathscr{O}_M$.

Acknowledgements The first named author was supported in part by a grant from the US NSF. Also first named author acknowledges partial support of the Swedish Foundation for International Cooperation in Research and Higher education (STINT) during his visits to Lund University.

References

1. Ash, R.B.: Information Theory. Dover Publications Inc., New York (1990) Corrected reprint of the 1965 original.
2. Aldroubi, A., Cabrelli, C., Hardin, D., Molter, U.: Optimal Shift Invariant Spaces and Their Parseval Frame Generators. Appl. Comput. Harmon. Anal. **23**(2), 273–283 (2007)

3. Albeverio, S., Jorgensen, P.E.T., Paolucci, A.M.: Multiresolution Wavelet Analysis of Integer Scale Bessel Functions. J. Math. Phys. **48**(7), 073516 (2007)
4. Ball, J.A., Vinnikov, V.: Functional Models for representations of the Cuntz algebra. Operator theory, systems theory and scattering theory: Multidimensional generalizations. Oper. Theory Adv. Appl. **157**, 1–60, Birkhäuser, Basel (2005)
5. Cleary, J.G., Witten, I.H., Bell, T.C.: Text Compression. Prentice Hall, Englewood Cliffs (1990)
6. Baggett, L., Jorgensen, P.E.T., Merrill, K., Packer, J.: A non-MRA C^r frame wavelet with rapid decay. Acta Appl. Math. (2005)
7. Bose, T.: Digital Signal and Image Processing. Wiley, New York (2003)
8. Bose, T., Chen, M.-Q., Thamvichai, R.: Stability of the 2-D Givone-Roesser model with periodic coefficients: IEEE Trans. Circuits Syst. I. Regul. Pap. **54**(3), 566–578 (2007)
9. Bratelli, O., Jorgensen, P.E.T.: Wavelets Through a Looking Glass: The World of the Spectrum. Birkhäuser, Basel (2002)
10. Burger, W.: Principles of Digital Image Processing: Fundamental Techniques, Springer (2009)
11. Burdick, H.E.: Digital Imaging: Theory and Applications. Mcgraw-Hill, New York (1997)
12. John, G., Cleary, J.G., Witten, I.H.: A comparison of enumerative and adaptive codes. IEEE Trans. Inform. Theory **30**(2, part 2), 306–315 (1984)
13. Daubechies, I.: Ten lectures on wavelets, volume 61 of CBMS-NSF Regional Conference Series in Applied Mathematics. (1992)
14. Donoho, D.L., Vetterli, M., DeVore, R.A., Daubechies, I.: Data compression and harmonic analysis. IEEE Trans. Inform. Theory **44**(6), 2435–2476 (1998)
15. Dutkay, D.E., Roysland, K.: Covariant representations for matrix-valued transfer operators. arXiv:math/0701453 (2007)
16. Hutchinson, J.E.: Fractals and self-similarity. Indiana Univ. Math. J. **30**(5), 713–747 (1981)
17. Green, P., MacDonald, L.: Colour Engineering: Achieving Device Independent Colour. Wiley, New York (2002)
18. Jaffard, S., Meyer, Y., Ryan, R.D.: Wavelets Tools for science & technology.. Society for Industrial and Applied Mathematics (SIAM). Philadelphia, PA, revised edition (2001)
19. Jorgensen, P.E.T.: Analysis and probability: wavelets, signals, fractals. Graduate Texts in Mathematics, vol. 234. Springer, New York (2006)
20. Jorgensen, P.E.T., Kornelson, K., Shuman, K.: Harmonic analysis of iterated function systems with overlap. J. Math. Phys. **48**(8), 083511 (2007)
21. Jorgensen, P.E.T., Mohari, A.: Localized bases in $L^2(0, 1)$ and their use in the analysis of Brownian motion. J. Approx. Theory **151**(1), 20–41 (2008)
22. Jorgensen, P.E.T., Song, M.-S.: Comparison of discrete and continuous wavelet transforms. Springer Encyclopedia of Complexity and Systems Science (2007)
23. Jorgensen, P.E.T., Song, M.-S.: Entropy encoding, hilbert space, and karhunen-loève transforms. J. Math. Phys. **48**(10), 103503 (2007)
24. MacKay, D.J.C.: Information Theory, Inference and Learning Algorithms. Cambridge University Press, New York (2003)
25. Keyl, M.: Fundamentals of Quantum Information Theory. Phys. Rep. **369**(5), 431–548 (2002)
26. Rastislav L., Plataniotis, K.N.: Color Image Processing: Methods and Applications, 1 edn. CRC, Boca Raton, FL (2006)
27. Roman, S.: Introduction to Coding and Information Theory: Undergraduate Texts in Mathematics. Springer, New York (1997)
28. Lukac, R., Plataniotis, K.N., Venetsanopoulos, A.N.: Bayer pattern demosaicking using local-correlation approach. Computational science–ICCS 2004. Part IV, Lecture Notes in Comput. Sci. **3039**, 26–33, Berlin, Springer (2004)
29. MacDonald, L. (Editor), Luo, M.R. (Editor): Colour Image Science: Exploiting Digital Media, 1 edn. Wiley, New York (2002)
30. Russ, J.C.: The image processing handbook, 5th edn. CRC, Boca Raton, FL (2007)
31. Salomon, D.: Data compression. The complete reference. 4th edn. With contributions by Giovanni Motta and David Bryant. Springer, London (2007)

32. Shannon C.E., Weaver W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana and Chicago (1998)
33. Skodras, A., Christopoulos, C., Ebrahimi, T.: Jpeg 2000 still image compression standard" ieee signal processing magazine. IEEE Signal process. Mag. **18**, 36–58 (2001)
34. Song, M.-S.: Wavelet image compression. In Operator theory, operator algebras, and applications, **414** Contemp. Math., 41–73. Amer. Math. Soc., Providence, RI, (2006)
35. Song, M.-S.: Entropy encoding in wavelet image compression. Representations, Wavelets and Frames A Celebration of the Mathematical Work of Lawrence Baggett, 293–311 (2008)
36. Usevitch, B.E.: A tutorial on modern lossy wavelet image compression: Foundations of jpeg 2000. IEEE Signal Process. Mag. **18**, 22–35 (2001)
37. Walker, J.S.: A Primer on Wavelets and Their Scientific Applications. Chapman & Hall, CRC (1999)
38. Witten, I.H.: Adaptive text mining: inferring structure from sequences. J. Discrete Algorithms **2**(2), 137–159 (2004)
39. Witten, I.H., Neal, R.M., Cleary, J.G.: Arithmetic coding for data compression. Comm. ACM **30**(6), 520–540 (1987)

Chapter 9
Wavelet Representations and Their Commutant

Dorin Ervin Dutkay and Sergei Silvestrov

Abstract We study the reducibility of the wavelet representation associated to various QMF filters, including those associated to Cantor sets. We show there are connections between this problem, the harmonic analysis of transfer operators and the ergodic properties of shifts on solenoids. We prove that if the QMF filter does not have constant absolute value, then the wavelet representations is reducible.

9.1 Introduction

Wavelets are functions that generate orthonormal bases under certain actions of translation and dilation operators. They have the advantage over Fourier series that they are better localized. By definition, a wavelet is a function $\psi \in L^2(\mathbb{R})$ with the property that

$$\{2^{j/2}\psi(2^j \cdot -k) : j, k \in \mathbb{Z}\} \tag{9.1}$$

is an orthonormal basis for $L^2(\mathbb{R})$. We refer to Daubechies' wonderful book [Dau92] for details.

We can rephrase this in terms of two unitary operators, the dilation operator U and the translation operator T on $L^2(\mathbb{R})$,

D.E. Dutkay
Department of Mathematics, University of Central Florida, 4000 Central Florida Blvd., P.O. Box 161364, Orlando, FL 32816-1364, USA
e-mail: ddutkay@mail.ucf.edu

S. Silvestrov (✉)
Centre for Mathematical Sciences, Lund University, Box 118, 22100 Lund, Sweden

Division of Applied Mathematics, The School of Education, Culture and Communication, Mälardalen University, Box 883, 72123 Västerås, Sweden
e-mail: sergei.silvestrov@math.lth.se; sergei.silvestrov@mdh.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_9,

$$Uf(x) = \frac{1}{\sqrt{2}} f\left(\frac{x}{2}\right), \quad Tf(x) = f(x-1), \quad (x \in \mathbb{R}, f \in L^2(\mathbb{R})) \tag{9.2}$$

The family defined in (9.1) is

$$\left\{U^j T^k \psi : j, k \in \mathbb{Z}\right\}.$$

One of the main techniques of constructing wavelets, is by a *multiresolution analysis*. By this we mean a sequence $(V_n)_{n\in\mathbb{Z}}$ of closed subspaces of $L^2(\mathbb{R})$ with the following properties:

1. $V_n \subset V_{n+1}$ for all $n \in \mathbb{Z}$
2. $UV_{n+1} = V_n$ for all $n \in \mathbb{Z}$
3. $\cup_n V_n$ is dense in $L^2(\mathbb{R})$ and $\cap_n V_n = \{0\}$
4. There exists a function $\varphi \in L^2(\mathbb{R})$ called *the scaling function*, such that $\{T^k\varphi : k \in \mathbb{Z}\}$ is an orthonormal basis for V_0

The subspaces V_n correspond to various resolution levels. Once a multiresolution analysis is given, the wavelet can be found in the *detail space*: $W_0 := V_1 \ominus V_0$. It is a function ψ with the property that $\{T^k\psi : k \in \mathbb{Z}\}$ is an orthonormal basis for W_0.

The multiresolution is constructed easily from the scaling function φ. Since $U\varphi$ is in $V_{-1} \subset V_0$, it can be written as a combination of translates of φ, and this gives the *scaling equation*:

$$U\varphi = \sum_{k\in\mathbb{Z}} a_k T^k \varphi. \tag{9.3}$$

Since T is a unitary operator, it has the spectrum contained in the unit circle $\mathbb{T}$; one can define a representation π of $L^\infty(\mathbb{T})$ on $L^2(\mathbb{R})$, by applying Borel functonal calculus

$$\pi(f) = f(T), \quad (f \in L^\infty(\mathbb{R})),$$

which means that for polynomials

$$\pi\left(\sum_k \alpha_k z^k\right) = \sum_k \alpha_k T^k.$$

The representation satisfies the *covariance relation*:

$$U\pi(f)U^{-1} = \pi(f(z^2)), \quad (f \in L^\infty(\mathbb{T})).$$

Using this representation, the scaling equation can be rewritten as

$$U\varphi = \pi(m_0)\varphi,$$

where

$$m_0(z) = \sum_{k\in\mathbb{Z}} a_k z^k, \quad (z \in \mathbb{T}).$$

The function m_0 is called *the low-pass* filter, and it is the starting point for the construction of the multiresolution analysis.

Since one is aiming at scaling functions whose translates are orthogonal, a necessary condition on m_0 is the *quadrature mirror filter (QMF) condition*:

$$\frac{1}{2}\sum_{w^2=z}|m_0(w)|^2 = 1, \quad (z \in \mathbb{T}).$$

Wavelet representations were introduced in [Jor01, Dut02, DJ07b] in an attempt to apply the multiresolution techniques of wavelet theory [Dau92] to a larger class of problems where self-similarity, or refinement is the central phenomenon. They were used to construct wavelet bases and multiresolutions on fractal measures and Cantor sets [DJ06] or on solenoids [Dut06].

Wavelet representations can be defined axiomatically as follows: let X be a compact metric space and let $r : X \to X$ be a Borel measurable function which is onto and finite-to-one, i.e., $0 < \#r^{-1}(x) < \infty$ for all $x \in X$. Let μ be a *strongly invariant measure* on X, i.e.

$$\int_X f\, d\mu = \int_X \frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} f(y)\, d\mu(x), \quad (f \in L^\infty(X)) \tag{9.4}$$

Let $m_0 \in L^\infty(X)$ be a *QMF filter*, i.e.,

$$\frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} |m_0(y)|^2 = 1 \text{ for } \mu\text{-a.e. } x \in X \tag{9.5}$$

Theorem 1. *[DJ07b] There exists a Hilbert space $\mathscr{H}$, a unitary operator U on $\mathscr{H}$, a representation π of $L^\infty(X)$ on $\mathscr{H}$ and an element φ of $\mathscr{H}$ such that:*

1. *(Covariance) $U\pi(f)U^{-1} = \pi(f \circ r)$ for all $f \in L^\infty(X)$.*
2. *(Scaling equation) $U\varphi = \pi(m_0)\varphi$*
3. *(Orthogonality) $\langle \pi(f)\varphi\,,\, \varphi\rangle = \int f\, d\mu$ for all $f \in L^\infty(X)$.*
4. *(Density) $\{U^{-n}\pi(f)\varphi \mid n \in \mathbb{N}, f \in L^\infty(X)\}$ is dense in $\mathscr{H}$.*

Moreover they are unique up to isomorphism.

Definition 2. We say that $(\mathscr{H}, U, \pi, \varphi)$ in Theorem 1 is the *wavelet representation* associated to m_0.

Our main focus will be the reducibility of the wavelet representations.

The most familiar wavelet representation is the classical one on $L^2(\mathbb{R})$, where as we described above, U is the operator of dilation by two and π is obtained by applying the Borel functional calculus to the translation operator T, i.e. $\pi(f) = f(T)$ for f bounded function on $\mathbb{T}$ – the unit circle. This representation is associated to the map $r(z) = z^2$ on $\mathbb{T}$, the measure μ is just the Haar measure on the circle, and m_0 can be any low-pass QMF filter which produces an orthogonal scaling function

(see [Dau92]). For example, one can take the Haar filter $m_0(z) = (1+z)/\sqrt{2}$ which produces the Haar scaling function φ.

This representation is reducible; its commutant was computed in [HL00] and the direct integral decomposition was presented in [LPT01].

Some low-pass filters, such as the stretched Haar filter $m_0(z) = (1+z^3)/\sqrt{2}$ give rise to non-orthogonal scaling functions. In this case super-wavelets appear, and the wavelet representation is realized on a direct sum of finitely many copies of $L^2(\mathbb{R})$. See [BDP05]. This representation is also reducible and its direct integral decomposition is similar to the one for $L^2(\mathbb{R})$. See [BDP05, Dut06].

When one takes the QMF filter $m_0 = 1$ the situation is very different. As shown in [Dut06], the representation can be realized on a solenoid and in this case it is irreducible. The result holds even for more general maps r, if they are ergodic (see [DLS09]).

The general theory of the decomposition of wavelet representations into irreducible components was given in [Dut06], but there is a large class of examples where it is not known whether these representations are irreducible or not.

One interesting example, introduced in [DJ07a], is the following: take the map $r(z) = z^3$ on the unit circle $\mathbb{T}$ with the Haar measure μ. Consider the QMF filter $m_0(z) = (1+z^2)/\sqrt{2}$. The wavelet representation associated to this data is strongly connected to the middle-third Cantor set. It can be realized as follows:

Let $\mathbf{C}$ be the middle-third Cantor set. Let

$$\mathscr{R} := \bigcup \left\{ \mathbf{C} + \frac{k}{3^n} \,|\, k, n \in \mathbb{Z} \right\}.$$

Let $\mathscr{H}^s$ be the Hausdorff measure of dimension $s := \log_3 2$, i.e., the Hausdorff dimension of the Cantor set. Restrict $\mathscr{H}^s$ to the set $\mathscr{R}$. Consider the Hilbert space $\mathscr{H} := L^2(\mathscr{R}, \mathscr{H}^s)$. Define the unitary operators on $\mathscr{H}$:

$$Uf(x) = \frac{1}{\sqrt{2}} f\left(\frac{x}{3}\right), \quad Tf(x) = f(x-1)$$

and define the representation π of $L^\infty(\mathbb{T})$ on $\mathscr{H}$, by applying Borel functional calculus to the operator T: $\pi(f) = f(T)$ for $f \in L^\infty(X)$.

The scaling function is defined as the characteristic function of the Cantor set $\varphi := \chi_{\mathbf{C}}$.

Then $(\mathscr{H}, U, \pi, \varphi)$ is the wavelet representation associated to the QMF filter $m_0(z) = (1+z^2)/\sqrt{2}$. To make this more intuitive, note that the Cantor set has the property that when it is dilated by three, the result consists of the Cantor set and a translation of it by two. In other words, the characteristic function of the Cantor set satisfies the scaling equation

$$U\chi_{\mathbf{C}} = \frac{1}{\sqrt{2}} \left(\chi_{\mathbf{C}} + T^2 \chi_{\mathbf{C}}\right).$$

In [DMP08], d'Andrea, Merrill and Packer construct a wavelet representation whose scaling function is the Sierpinski gasket. See Example 13 below. They also

present some numerical experiments showing how this multiresolution behaves under the usual wavelet compression algorithm.

In [BLP+10, LR07, LR06] the wavelet representations are given a more operator theoretic flavour. A groupoid approach is offered in [IM08]. Investigations of general multiresolution theories are presented in [BFMP09b, BFMP09a, BLP+10, BLM+09].

9.2 Solenoids

We recall some facts from [DJ07b]. The wavelet representation can be realized on a solenoid as follows: Let

$$X_\infty := \left\{(x_0, x_1, \dots) \in X^{\mathbb{N}} \mid r(x_{n+1}) = x_n \text{ for all } n \geq 0\right\} \tag{9.6}$$

We call X_∞ the *solenoid* associated to the map r.

On X_∞ consider the σ-algebra generated by cylinder sets. Let $r_\infty : X_\infty \to X_\infty$

$$r_\infty(x_0, x_1, \dots) = (r(x_0), x_0, x_1, \dots) \text{ for all } (x_0, x_1, \dots) \in X_\infty \tag{9.7}$$

Then r_∞ is a measurable automorphism on X_∞.

Define $\theta_0 : X_\infty \to X$,

$$\theta_0(x_0, x_1, \dots) = x_0. \tag{9.8}$$

The measure μ_∞ on X_∞ will be defined by constructing some path measures P_x on the fibers $\Omega_x := \{(x_0, x_1, \dots) \in X_\infty \mid x_0 = x\}$.

Let

$$c(x) := \#r^{-1}(r(x)), \quad W(x) = |m_0(x)|^2/c(x), \quad (x \in X).$$

Then

$$\sum_{r(y)=x} W(y) = 1, \quad (x \in X) \tag{9.9}$$

$W(y)$ can be thought of as the transition probability from $x = r(y)$ to one of its roots y.

For $x \in X$, the path measure P_x on Ω_x is defined on cylinder sets by

$$P_x(\{(x_n)_{n\geq 0} \in \Omega_x \mid x_1 = z_1, \dots, x_n = z_n\}) = W(z_1) \dots W(z_n) \tag{9.10}$$

for any $z_1, \dots, z_n \in X$. This value can be interpreted as the probability of the random walk to go from x to z_n through the points $x_1, \dots, x_n$.

Next, define the measure μ_∞ on X_∞ by

$$\int f \, d\mu_\infty = \int_X \int_{\Omega_x} f(x, x_1, \dots) \, dP_x(x, x_1, \dots) \, d\mu(x) \tag{9.11}$$

for bounded measurable functions on X_∞.

Consider now the Hilbert space $\mathscr{H} := L^2(\mu_\infty)$. Define the operator

$$U\xi = (m_0 \circ \theta_0)\, \xi \circ r_\infty, \quad (\xi \in L^2(X_\infty, \mu_\infty)) \tag{9.12}$$

Define the representation of $L^\infty(X)$ on $\mathscr{H}$

$$\pi(f)\xi = (f \circ \theta_0)\, \xi, \quad (f \in L^\infty(X), \xi \in L^2(X_\infty, \mu_\infty)) \tag{9.13}$$

Let $\varphi = 1$ the constant function 1.

Theorem 3. *[DJ07b] Suppose m_0 is non-singular, i.e., $\mu(\{x \in X \mid m_0(x) = 0\}) = 0$. Then the data $(\mathscr{H}, U, \pi, \varphi)$ forms the wavelet representation associated to m_0.*

We are interested in the reducibility of wavelet representations, this involves the study of the commutant of the representation; it was shown in [DJ07b] that there are several equivalent ways to formulate this problem:

Theorem 4. *[DJ07b] Suppose m_0 is non-singular. Then there is a one-to-one correspondence between the following data:*

1. *Operators S in the commutant of $\{U, \pi\}$.*
2. *Cocycles, i.e., functions $f \in L^\infty(X_\infty, \mu_\infty)$ such that $f \circ r_\infty = f$, μ_∞-a.e.*
3. *Harmonic functions $h \in L^\infty(X)$ for the transfer operator R_{m_0}, i.e., $R_{m_0} h = h$, where*

$$R_{m_0} f(x) = \frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} |m_0(y)|^2 f(y).$$

The correspondence $1 \leftrightarrow 2$ is given by $S = M_f$ where M_f is the multiplication operator $M_f \xi = f\xi$, $\xi \in L^2(X_\infty, \mu_\infty)$. The correspondence from two to three is given by

$$h(x) = \int_{\Omega_x} f(x, x_1, \dots)\, dP_x(x, x_1, \dots).$$

The correspondence from three to two is given by

$$f(x, x_1, \dots) = \lim_{n \to \infty} h(x_n), \text{ for } \mu_\infty\text{-a.e. } (x, x_1, \dots) \text{ in } X_\infty.$$

9.3 Reducible Wavelet Representations

Using the correspondences in Theorem 4, the reducibility problem can be then formulated in several ways:

Theorem 5. *[DLS09] Suppose m_0 is non-singular. The following affirmations are equivalent:*

1. *The wavelet representation is irreducible, i.e., the commutant* $\{U, \pi\}'$ *is trivial.*
2. *The automorphism* r_∞ *on* (X_∞, μ_∞) *is ergodic.*
3. *The only bounded measurable harmonic functions for the transfer operator* R_{m_0} *are the constants.*
4. *There are no non-constant fixed points of the transfer operator* $h \in L^p(X, \mu)$, *for some* $p > 1$ *with the property that*

$$\sup_{n\in\mathbb{N}} \int_X |m_0^{(n)}(x)|^2 |h(x)|^p \, d\mu(x) < \infty \tag{9.14}$$

where

$$m_0^{(n)}(x) = m_0(x)m_0(r(x)) \dots m_0(r^{n-1}(x)), \quad (x \in X). \tag{9.15}$$

5. *If* $\varphi' \in L^2(X_\infty, \mu_\infty)$, *satisfies the same scaling equation as* φ, *i.e.,* $U\varphi' = \pi(m_0)\varphi'$, *then* φ' *is a constant multiple of* φ.

Next, we show that under some mild assumptions, the wavelet representations are reducible.

Theorem 6. *Suppose* $r : (X, \mu) \to (X, \mu)$ *is ergodic. Assume* $|m_0|$ *is not constant* 1 μ*-a.e., non-singular, i.e.,* $\mu(m_0(x) = 0) = 0$, *and* $\log |m_0|^2$ *is in* $L^1(X)$. *Then the wavelet representation* $(\mathscr{H}, U, \pi, \varphi)$ *is reducible.*

Proof. From the QMF relation and the strong invariance of μ we have

$$\int_X |m_0|^2 \, d\mu = \int_X \frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} |m_0(y)|^2 \, d\mu = 1.$$

By Jensen's inequality we have

$$a := \int_X \log |m_0|^2 \, d\mu \le \log \int_X |m_0|^2 \, d\mu = 0.$$

Since log is strictly concave, and $|m_0|^2$ is not constant μ-a.e., it follows that the inequality is strict, and $a < 0$.

Since r is ergodic, applying Birkoff's ergodic theorem, we obtain that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} \log |m_0 \circ r^k|^2 = \int_X \log |m_0|^2 \, d\mu = a, \ \mu - \text{ a.e.}$$

This implies that

$$\lim_{n\to\infty} \left(|m_0(x)m_0(r(x)) \dots m_0(r^{n-1}(x))|^2\right)^{1/n} = e^a < 1, \ \mu - \text{a.e.}$$

Take b with $e^a < b < 1$.

By Egorov's theorem, there exists a measurable set A_0, with $\mu_\infty(A_0) > 0$, such that $(|m_0(x)m_0(r(x))\dots m_0(r^{n-1}(x))|^2)^{1/n}$ converges uniformly to e^a on A_0. This implies that there exists an n_0 such for all $m \geq n_0$:

$$\left(|m_0(x)m_0(r(x))\dots m_0(r^{m-1}(x))|^2\right)^{1/m} \leq b \text{ for } x \in A_0$$

so

$$|m_0(x)m_0(r(x))\dots m_0(r^{m-1}(x))|^2 \leq b^m, \text{ for } m \geq n_0 \text{ and all } x \in A_0. \tag{9.16}$$

Next, given $m \in \mathbb{N}$, we compute the probability of a sequence $(z_n)_{n\in\mathbb{N}} \in X_\infty$ to have $z_m \in A_0$. We have, using the strong invariance of μ:

$$\begin{aligned} P(z_m \in A_0) &= \mu_\infty\left(\{(z_n)_n \mid z_m \in A_0\}\right) = \int_{X_\infty} \chi_{A_0} \circ \theta_m \, d\mu_\infty \\ &= \int_X \frac{1}{\#r^{-m}(z_0)} \sum_{r(z_1)=z_0,\dots,r(z_m)=z_{m-1}} |m_0(z_1)|^2 \dots |m_0(z_m)|^2 \chi_{A_0}(z_m)\, d\mu(z_0) \\ &= \int_X |m_0(z_m)m_0(r(z_m))\dots m_0(r^{m-1}(z_m))|^2 \chi_{A_0}(z_m)\, d\mu(z_m) \\ &= \int_X |m_0(x)m_0(r(x))\dots m_0(r^{m-1}(x))|^2 \chi_{A_0}(x)\, d\mu(x). \end{aligned}$$

Then

$$\sum_{m=1}^{\infty} P(z_m \in A_0) = \sum_{m\geq 1} \int_X |m_0(x)m_0(r(x))\dots m_0(r^{m-1}(x))|^2 \chi_{A_0}\, d\mu(x) < \infty$$

and we used (9.16) in the last inequality.

Now we can use Borel–Cantelli's lemma, to conclude that the probability that $z_m \in A_0$ infinitely often is zero. Thus, for μ_∞-a.e. $z := (z_n)_n$, there exists k_z (depending on the point) such that $z_n \notin A_0$ for $n \geq k_z$.

Suppose now the representation is irreducible. Then r_∞ is ergodic on (X_∞, μ_∞). So r_∞^{-1} is too. Using Birkhoff's ergodic theorem it follows that, μ_∞-a.e.,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} (\chi_{A_0} \circ \theta_0) \circ r_\infty^{-k} = \int_{X_\infty} \chi_{A_0} \circ \theta_0 \, d\mu_\infty = \mu(A_0) > 0 \tag{9.17}$$

But

$$\left[(\chi_{A_0} \circ \theta_0) \circ r_\infty^{-k}\right](z_n)_n = \chi_{A_0}(z_k) = 0, \text{ for } k \geq k_z.$$

Therefore the sum on the left of (9.17) is bounded by k_z so the limit is zero, a contradiction. Thus the representation has to be reducible. □

Using the results from [DLS09], we obtain that there are non-trivial solutions to refinement equations and non-trivial fixed points for transfer operators:

Corollary 7. *Let m_0 be as in Theorem 6 and let $(\mathscr{H}, U, \pi, \varphi)$ be the associated wavelet representation. Then*

1. *There exist solutions $\varphi' \in \mathscr{H}$ for the scaling equation $U\varphi' = \pi(m_0)\varphi'$ which are not constant multiples of φ.*
2. *There exist non-constant, bounded fixed points for the transfer operator*

$$R_{m_0} f(x) = \frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} |m_0(y)|^2 f(y), \quad (f \in L^\infty(X), x \in X).$$

Remark 8. As shown in [DJ07b], operators in the commutant of $\{U, \pi\}$ are multiplication operators M_g, with $g \in L^\infty(X_\infty, r_\infty)$ and $g = g \circ r_\infty$. Therefore, if $\mathcal{K}$ is a subspace which is invariant for U and $\pi(f)$ for all $f \in L^\infty(X)$, then the orthogonal projection onto $\mathcal{K}$ is an operator in the commutant and so it corresponds to a multiplication by a characteristic function χ_A, where A is an invariant set for r_∞, i.e., $A = r_\infty^{-1}(A) = r_\infty(A)$, μ_∞-a.e., and $\mathcal{K} = L^2(A, \mu_\infty)$.

In conclusion the study of invariant spaces for the wavelet representation $\{U, \pi\}$ is equivalent to the study of the invariant sets for the dynamical system r_∞ on (X_∞, μ_∞).

Proposition 9. *Under the assumptions of Theorem 6, there are no finite-dimensional invariant subspaces for the wavelet representation.*

Proof. We reason by contradiction. Suppose $\mathcal{K}$ is a finite-dimensional invariant subspaces. Then, as in remark 8, this will correspond to a set A invariant under r_∞, $\mathcal{K} = L^2(A, \mu_\infty)$. But if $\mathcal{K}$ is finite dimensional then A must contain only atoms. Let $(z_n)_{n\in\mathbb{N}}$ be such an atom. We have

$$0 < \mu_\infty((z_n)_{n\in\mathbb{N}}) = \mu(z_0) P_{z_0}((z_n)_{n\in\mathbb{N}}),$$

so z_0 is an atom for μ. Since μ is strongly invariant for μ, it follows that it is also invariant for μ. Then $\mu(r(z_0)) = \mu(r^{-1}(r(z_0))) \geq \mu(z_0)$. By induction, $\mu(r^{n+1}(z_0)) \geq \mu(r^n(z_0))$. Since $\mu(X) < \infty$ and $\mu(z_0) > 0$ this implies that at for some $n \in \mathbb{N}$ and $p > 0$ we have $r^{n+p}(z_0) = r^n(z_0)$. We relabel $r^n(z_0)$ by z_0 so we have $r^p(z_0) = z_0$ and $\mu(z_0) > 0$.

Since μ is invariant for r we have $\mu(z_0) \leq \mu(r^{-p}(z_0)) = \mu(z_0)$, and this shows that all the points in $r^{-p}(z_0)$ except z_0 have measure μ zero. The same can be said for $r(z_0), \ldots, r^{p-1}(z_0)$. But then $C := \{z_0, r(z_0), \ldots, r^{p-1}(z_0)\}$ is invariant for r, μ-a.e., and has positive measure. Since r is ergodic this shows that $C = X$, μ-a.e., and so we can consider that $\#r^{-1}(x) = 1$ for μ-a.e. $x \in X$. And then the QMF condition implies that $|m_0| = 1$ μ-a.e., which contradicts the assumptions in the hypothesis. □

9.4 Examples

Example 10. Consider the map $r(z) = z^2$ on the unit circle $\mathbb{T} = \{z \in \mathbb{C} : |z| = 1\}$. Let μ be the Haar measure on $\mathbb{T}$. Let $m_0(z) = \frac{1}{\sqrt{2}}(1+z)$ be the Haar low-pass filter, or any filter that generates an orthonormal scaling function in $L^2(\mathbb{R})$ (see [Dau92]). Then the wavelet representation associated to m_0 can be realized on the Hilbert space $L^2(\mathbb{R})$. The dilation operator is

$$U\xi(x) = \frac{1}{\sqrt{2}}\xi\left(\frac{x}{2}\right), \quad (x \in \mathbb{R}, \xi \in L^2(\mathbb{R}))$$

The representation π of $L^\infty(\mathbb{T})$ is constructed by applying Borel functional calculus to the translation operator

$$\begin{aligned} T\xi(x) &= \xi(x-1), \quad (x \in \mathbb{R}, f \in L^2(\mathbb{R})), \\ \pi(f) &= f(T), \quad (f \in L^\infty(\mathbb{R})), \end{aligned}$$

in particular

$$\pi\left(\sum_{k\in\mathbb{Z}} a_k z^k\right) = \sum_{k\in\mathbb{Z}} a_k T^k,$$

for any finitely supported sequence of complex numbers $(a_k)_{k\in\mathbb{Z}}$.

The Fourier transform of the scaling function is given by an infinite product ([Dau92]):

$$\widehat{\varphi}(x) = \prod_{n=1}^{\infty} m_0\left(\frac{x}{2^n}\right), \quad (x \in \mathbb{R}).$$

The commutant of this wavelet representation can be explicitly computed (see [HL00]): let $\mathscr{F}$ be the Fourier transform. An operator A is in the commutant $\{U, \pi\}'$ of the wavelet representation if and only if its Fourier transform $\widehat{A} := \mathscr{F}A\mathscr{F}^{-1}$ is a multiplication operator by a bounded, dilation invariant function, i.e., $\widehat{A} = M_f$, with $f \in L^\infty(\mathbb{R})$, $f(2x) = f(x)$, for a.e. $x \in \mathbb{R}$. Here

$$M_f\xi = f\xi, \quad (\xi \in L^2(\mathbb{R})).$$

Thus, invariant subspaces correspond, through the Fourier transform, to sets which are invariant under dilation by two.

The measure μ_∞ on the solenoid $\mathbb{T}_\infty$ can also be computed, see [Dut06]. It is supported on the embedding of $\mathbb{R}$ in the solenoid $\mathbb{T}_\infty$. The path measures P_x are in this case atomic.

The direct integral decomposition of the wavelet representation was described [LPT01].

For the low-pass filters that generate non-orthogonal scaling function, such as the stretched Haar filter $m_0(z) = \frac{1}{\sqrt{2}}(1+z^3)$, the wavelet representation can be realized in a finite sum of copies of $L^2(\mathbb{R})$. These filters correspond to super-wavelets, and the computation of the commutant, of the measure μ_∞ and the direct integral decomposition of the wavelet representation can be found in [BDP05, Dut06].

Example 11. Let $r(z) = z^N$, $N \in \mathbb{N}$, $N \geq 2$ on the unit circle $\mathbb{T}$ and let $m_0(z) = 1$ for all $z \in \mathbb{T}$. In this case (see [Dut06]) the wavelet representation can be realized on the solenoid $\mathbb{T}_\infty$ and the measure μ_∞ is just the Haar measure on the solenoid $\mathbb{T}_\infty$, and the operators U, π are defined above in the proof of Theorem 6. For this particular wavelet representation the commutant is trivial, so the representation is *irreducible*. It is interesting to see that, by Theorem 6, just any small perturbation of the constant function $m_0 = 1$ will generate a *reducible* wavelet representation.

Example 12. We turn now to the example on the Cantor set. Let $r(z) = z^3$ on $\mathbb{T}$ with the Haar measure, and $m_0(z) = \frac{1}{\sqrt{2}}(1+z^2)$. As we explained in the introduction, this low-pass filter generates a wavelet representation involving the middle third Cantor set. See [DJ06] for details. We know that $r(z) = z^3$ is an ergodic map and it is easy to see that the function m_0 satisfies the hypotheses of Theorem 6. Actually, an application of Jensen's formula to the analytic function m_0^2 shows that

$$\int_{\mathbb{T}} \log |m_0|^2 \, d\mu = -2\pi \log 2.$$

Thus, by Theorem 6, it follows that this wavelet representation is reducible. However, the problem of *constructing* the operators in the commutant of the wavelet representation remains open for analysis.

Example 13. Consider the wavelet representation constructed by d'Andrea, Merrill and Packer in [DMP08]. It is associated to the map r on $\mathbb{T}^2$ with Haar measure, $r(z_{1,2}) = (z_1^2, z_2^2)$ and the QMF filter

$$m_0(z_1, z_2) = \frac{1}{\sqrt{3}}(1 + z_1 + z_2).$$

The point of choosing this filter is to obtain the scaling equation for the Sierpinski gasket $\mathscr{S}$ with vertices $(0,0)$, $(1,0)$ and $(0,1)$. This set has the property

$$2\mathscr{S} = \mathscr{S} \cup (\mathscr{S} + (1,0)) \cup (\mathscr{S} + (0,1)).$$

Define the set

$$\mathscr{R}_{\mathscr{S}} := \cup_{j=-\infty}^{\infty} \cup_{(m,n)\in\mathbb{Z}^2} [A^j(\mathscr{S} + (m,n)].$$

Let $\mathscr{H}$ be the Hausdorff measure of dimension $\frac{\log 3}{\log 2}$ restricted to the set $\mathscr{R}_{\mathscr{S}}$. The Hilbert space we work on is $L^2(\mathscr{R}_{\mathscr{S}}, \mathscr{H})$.

Define the dilation operator by

$$Uf(x) = \frac{1}{\sqrt{3}} f\left(\frac{1}{2}x\right), \quad (x \in \mathscr{R}_{\mathscr{S}}, f \in L^2(\mathscr{R}_{\mathscr{S}}, \mathscr{H})).$$

Define the translation operators

$$T_{(m,n)} f(x, y) = f(x - m, y - n), \quad ((x, y) \in \mathscr{R}_{\mathscr{S}}, f \in L^2(\mathscr{R}_{\mathscr{S}}, \mathscr{H})).$$

The representation π of $L^\infty(\mathbb{T}^2)$ is defined by Borel functional calculus applied to the translation operators $T_{(m,n)}$.

Then the characteristic function of the Sierpinski gasket $\varphi = \chi_{\mathscr{S}}$ satisfies the scaling equation

$$U\varphi = \pi(m_0)\varphi.$$

It is shown in [DMP08] that the wavelet representation associated to m_0 is $(L^2(\mathscr{R}_{\mathscr{S}}, \mathscr{H}), U, \pi, \varphi)$.

It is well know that the map r is ergodic. The filter m_0 is non-singular and its absolute value is not constant. To see that $\log |m_0|$ is in $L^1(\mathbb{T}^2)$, note that $1 + z_1 + z_2 = 0$ only if $(z_1, z_2) \in \{(e^{2\pi i/3}, e^{4\pi i/3}), (e^{4\pi i/3}, e^{2\pi i/3})\}$. Then by Theorem 6, the wavelet representation associated to the Sierpinski gasket is reducible.

There is a case not covered by Theorem 6, namely the case when $|m_0| = 1$. In this case the situation can be different, the representation can be irreducible:

Theorem 14. *Let $m_0 = 1$ and let $(L^2(X_\infty, \mu_\infty), U, \pi, \varphi)$ be the associated wavelet representation. The following affirmations are equivalent:*

1. *The automorphism r_∞ on (X_∞, μ_∞) is ergodic.*
2. *The wavelet representation is irreducible.*
3. *The only bounded functions which are fixed points for the transfer operator R_1, i.e.,*
$$R_1 h(x) := \frac{1}{\#r^{-1}(x)} \sum_{r(y)=x} h(y) = h(x)$$
are the constant functions.
4. *The only $L^2(X, \mu)$-functions which are fixed points for the transfer operator R_1, are the constants.*
5. *The endomorphism r on (X, μ) is ergodic.*

Acknowledgements This research was supported in part by The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Swedish Research Council, The Swedish Royal Academy of Sciences and The Crafoord Foundation. The second author also is grateful to Institut Mittag-Leffler for support of his participation in Quantum Information Theory program in Autumn 2010, which has been beneficial for completion of this work.

References

[BDP05] Bildea, S., Dutkay, D.E., Picioroaga, G.: MRA super-wavelets. New York J. Math. **11**, 1–19 (electronic) (2005)

[BFMP09a] Baggett, L.W., Furst, V., Merrill, K.D., Packer, J.A.: Generalized filters, the low-pass condition, and connections to multiresolution analyses. J. Funct. Anal. **257**(9), 2760–2779 (2009)

[BFMP09b] Baggett, L. W., Furst, V., Merrill, K. D., Packer, J. A.: Classification of generalized multiresolution analyses. J. Funct. Anal. 258, no. 12, 4210-4228 (2010)

[BLM+09] Baggett, L.W., Larsen, N.S., Merrill, K.D., Packer, J.A., Raeburn, I.: Generalized multiresolution analyses with given multiplicity functions. J. Fourier Anal. Appl. **15**(5), 616–633 (2009)

[BLP+10] Baggett, L.W., Larsen, N.S., Packer, J.A., Raeburn, I., Ramsay, A.: Direct limits, multiresolution analyses, and wavelets. J. Funct. Anal. **258**(8), 2714–2738 (2010)

[Dau92] Daubechies, I.: Ten lectures on wavelets, volume 61 of CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, (1992)

[DJ06] Dutkay, D.E., Jorgensen, P.E.T.: Wavelets on fractals. Rev. Mat. Iberoam. **22**(1), 131–180 (2006)

[DJ07a] Dutkay, D.E., Jorgensen, P.E.T.: Fourier frequencies in affine iterated function systems. J. Funct. Anal. **247**(1), 110–137 (2007)

[DJ07b] Dutkay, D.E., Jorgensen, P.E.T.: Martingales, endomorphisms, and covariant systems of operators in Hilbert space. J. Oper. Theory **58**(2), 269–310 (2007)

[DLS09] Dutkay, D. E., Larson, D. R., Silvestrov, S.: Irreducible wavelet representations and ergodic automorphisms on solenoids. Oper. Matrices 5, no. 2, 201–219 (2011)

[DMP08] d'Andrea, J., Merrill, K., Packer, J.: Fractal wavelets of Dutkay-Jorgensen type for the Sierpinski gasket space Frames and operator theory in analysis and signal processing, Contemp. Math. **451**, 69–88 (2008)

[Dut02] Dutkay, D.E.: Harmonic analysis of signed Ruelle transfer operators. J. Math. Anal. Appl. **273**(2), 590–617 (2002)

[Dut06] Dutkay, D.E.: Low-pass filters and representations of the Baumslag Solitar group. Trans. Amer. Math. Soc. **358**(12), 5271–5291 (electronic) (2006)

[HL00] Han, D., Larson, D.R.: Frames, bases and group representations. Mem. Amer. Math. Soc. **147**(697), x+94 (2000)

[IM08] Ionescu, M., Muhly, P.S.: Groupoid methods in wavelet analysis. In: Group representations, ergodic theory, and mathematical physics: a tribute to George W. Mackey, volume 449 of Contemp. Math., pages 193–208. Amer. Math. Soc., Providence, RI (2008)

[Jor01] Jorgensen, P.E.T.: Ruelle operators: functions which are harmonic with respect to a transfer operator. Mem. Am. Math. Soc. **152**(720), viii+60 (2001)

[LPT01] Lim, L.-H., Packer, J.A., Taylor, K.F.: A direct integral decomposition of the wavelet representation. Proc. Am. Math. Soc. **129**(10), 3057–3067 (electronic) (2001)

[LR06] Larsen, N.S., Raeburn, I.: From filters to wavelets via direct limits. In: Operator theory, operator algebras, and applications, volume 414 of Contemp. Math., pages 35–40. Amer. Math. Soc., Providence, RI (2006)

[LR07] Larsen, N.S., Raeburn, I.: Projective multi-resolution analyses arising from direct limits of Hilbert modules. Math. Scand. **100**(2), 317–360 (2007)

Chapter 10
Multidimensional Inequalities of Hardy and (Limit) Pólya-Knopp Types

Maria Johansson and Lars-Erik Persson

Abstract In this review paper we complement the classical two-dimensional Hardy-type inequality by E. Sawyer (see MR87f:42052) in various ways. In particular, ideas and results from three recent Ph.D. theses are unified and presented in this frame. Also some complementary new results are proved and some open questions are raised.

10.1 Introduction

In this paper we complement the classical two-dimensional Hardy-type inequality by E. Sawyer (see Theorem 1 below) in various ways. In this case three different (independent) conditions are used to characterize the Hardy inequality for the case $1 < p \leq q < \infty$. In this review article we have unified and extracted ideas and results from the Ph.D. theses [33] and [34] (see also [12]), which are closely related to this important result of E. Sawyer. Also some complementary results and ideas are included and discussed in this frame.

Following the ideas introduced by A. Wedestig [34] (see also [35]) in Sect. 10.2 we state and prove the fact that in the case when the weight on the right hand side is of product type, then only one condition is necessary and sufficient for this Hardy-type inequality to hold. In Sect. 10.3 we state and prove that as a limiting case (when $p \to \infty$) we get a characterization also of the corresponding two-dimensional

M. Johansson (✉)
Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden
e-mail: maria.l.johansson@ltu.se

L.-E. Persson
Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden

Narvik University College, P.O. Box 385, N 8505, Narvik, Norway
e-mail: larserik@sm.luth.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_10,

Pólya-Knopp inequality. We also include a second proof of independent interest. By using induction the results in these Sections can be given also in n-dimensional settings. Instead of going on in this direction we now turn to another technique used by E. Ushakova in her Ph.D. thesis [33] (see also [29]). More exactly, in Sect. 10.4 we prove that n-dimensional Hardy-type inequalities in the case $1 < p \le q < \infty$ can be characterized by using just one condition both when the right hand side weight is of product type and also when the left hand side is. Some similar results for the case $1 < q < p < \infty$ are proved in Sect. 10.5 but with some additional restrictions on the weights. In Sect. 10.6 we prove some corresponding n-dimensional limit Pólya-Knopp type inequalities. Finally, Sect. 10.7 is reserved for some further results and remarks. In particular, we also discuss the relations with the Ph.D. thesis by M. Johansson [12] and some open questions are raised.

10.2 On Sawyer's Characterization of the Two-Dimensional Hardy Inequality

The following remarkable result was proved by E.T. Sawyer in [30, Theorem 1].

Theorem 1. *Let $1 < p \le q < \infty$ and let u and v be weight functions on $\mathbb{R}^2_+$. Then*

$$\left(\int_0^\infty\int_0^\infty\left(\int_0^{x_1}\int_0^{x_2} f(t_1,t_2)dt_1dt_2\right)^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}} \tag{10.1}$$

$$\le C\left(\int_0^\infty\int_0^\infty f\,(x_1,x_2)^p\, v(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}$$

holds for all positive and measurable functions f on $\mathbb{R}^2_+$ if and only if

$$\sup_{y_1,y_2>0}\left(\int_{y_1}^\infty\int_{y_2}^\infty u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}\left(\int_0^{y_1}\int_0^{y_2} v(x_1,x_2)^{1-p'}dx_1dx_2\right)^{\frac{1}{p'}} = A_1 < \infty, \tag{10.2}$$

and

$$\sup_{y_1,y_2>0}\frac{\left(\int_0^{y_1}\int_0^{y_2}\left(\int_0^{x_1}\int_0^{x_2} v(t_1,t_2)^{1-p'}dt_1dt_2\right)^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}}{\left(\int_0^{y_1}\int_0^{y_2} v(x_1,x_2)^{1-p'}dx_1dx_2\right)^{\frac{1}{p}}} = A_2 < \infty, \tag{10.3}$$

and

$$\sup_{y_1,y_2>0} \frac{\left(\int_{y_1}^{\infty}\int_{y_2}^{\infty}\left(\int_{x_1}^{\infty}\int_{x_2}^{\infty} u(t_1,t_2)\,dt_1dt_2\right)^{p'} v(x_1,x_2)^{1-p'}dx_1dx_2\right)^{\frac{1}{p'}}}{\left(\int_{y_1}^{\infty}\int_{y_2}^{\infty} u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q'}}} = A_3 < \infty. \tag{10.4}$$

Note that in this two-dimensional case all three conditions (10.2)–(10.4) are needed in order to characterize the weighted Hardy inequality. However, in our next Theorem we shall point out the fact that in one dimension the situation is simpler. More exactly, in this case the conditions corresponding to (10.2)–(10.4) are in fact equivalent and each of them can be used to characterize the corresponding one-dimensional Hardy inequality (see also [3] and [31]).

Theorem 2. *Let $1 < p \le q < \infty$, and let u and v be weight functions on $\mathbb{R}_+$. Then each of the following conditions are necessary and sufficient for the inequality*

$$\left(\int_0^{\infty}\left(\int_0^{x} f(t)dt\right)^{q} u(x)dx\right)^{\frac{1}{q}} \le C\left(\int_0^{\infty} f^{p}(x)v(x)dx\right)^{\frac{1}{p}} \tag{10.5}$$

to hold for all positive and measurable functions on $\mathbb{R}_+$.
(a) The Muckenhoupt condition,

$$A_M = \sup_{x>0}\left(\int_x^{\infty} u(t)dt\right)^{\frac{1}{q}}\left(\int_0^{x} v(t)^{1-p'}dt\right)^{\frac{1}{p'}} < \infty. \tag{10.6}$$

Moreover, the best constant C in (10.5) can be estimated as follows:

$$A_M \le C \le \left(1+\frac{q}{p'}\right)^{\frac{1}{q}}\left(1+\frac{p'}{q}\right)^{\frac{1}{p'}} A_M.$$

(b) The condition of L. E Persson and V. D. Stepanov,

$$A_{PS} = \sup_{x>0} V(x)^{-\frac{1}{p}}\left(\int_0^{x} u(t)V(t)^{q}dt\right)^{\frac{1}{q}} < \infty, \qquad V(x) = \int_0^{x} v(t)^{1-p'}dt. \tag{10.7}$$

Moreover, the best constant C in (10.5) satisfies the following estimates:

$$A_{PS} \le C \le p' A_{PS}.$$

(c) The condition (c.f. (10.4)),

$$A_{PS}^{*}=\sup_{x>0}\left(\int_{x}^{\infty}\left(\int_{t}^{\infty}u(s)ds\right)^{p'}v^{1-p'}(t)dt\right)^{\frac{1}{p'}}\left(\int_{x}^{\infty}u(t)dt\right)^{-\frac{1}{q'}}<\infty. \tag{10.8}$$

Moreover, the best constant C *in (10.5) can be estimated sa follows:*

$$A_{PS}^{*}\leq C\leq qA_{PS}^{*}. \tag{10.9}$$

Remark 3. A direct proof of Theorem 2 can be found in [34]. For the original proofs of a) and b) see [19] and [26], respectively, and (10.8) is just the dual condition of (10.7). A much more general result where (10.6)–(10.8) is replaced by 4 different scales of conditions is proved in [8]. For more results, historical remarks and references in the one dimensional case see the recent book [17] and historical article [16].

Moreover, it was recently discovered by A. Wedestig in her Ph.D. thesis [34] (see also [35]) that if the weight on the right hand side is of product type, then, in fact, 10.1 can be characterized by just one condition (or, more generally, just one of infinite possible conditions). More exactly our main result in this Section reads:

Theorem 4. *Let* $1<p\leq q<\infty$, $s_1,s_2\in(1,p)$ *and let* u *be a weight functions on* $\mathbb{R}_{+}^{2}$ *and let* v_1 *and* v_2 *be weight functions on* $\mathbb{R}_{+}$. *Then the inequality*

$$\left(\int_{0}^{\infty}\int_{0}^{\infty}\left(\int_{0}^{x_1}\int_{0}^{x_2}f(t_1,t_2)dt_1dt_2\right)^{q}u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}\leq$$

$$C\left(\int_{0}^{\infty}\int_{0}^{\infty}f^{p}(x_1,x_2)v_1(x_1)v_2(x_2)\,dx_1dx_2\right)^{\frac{1}{p}} \tag{10.10}$$

holds for all measurable functions $f\geq 0$ *if*

$$A_W(s_1,s_2)=\sup_{t_1,t_2>0}V_1(t_1)^{\frac{s_1-1}{p}}V_2(t_2)^{\frac{s_2-1}{p}}\times$$

$$\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty}u(x_1,x_2)V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)}V_2(x_2)^{q\left(\frac{p-s_2}{p}\right)}dx_1dx_2\right)^{\frac{1}{q}}<\infty, \tag{10.11}$$

where $V_1(t_1)=\int_{0}^{t_1}v_1(x_1)^{1-p'}dx_1$ *and* $V_2(t_2)=\int_{0}^{t_2}v_2(x_2)^{1-p'}dx_2$.

Moreover, if C is the best possible constant in (10.10), then

$$\sup_{1<s_1,s_2<p}\left(\frac{\left(\frac{p}{p-s_1}\right)^p}{\left(\frac{p}{p-s_1}\right)^p+\frac{1}{s_1-1}}\right)^{\frac{1}{p}}\left(\frac{\left(\frac{p}{p-s_2}\right)^p}{\left(\frac{p}{p-s_2}\right)^p+\frac{1}{s_2-1}}\right)^{\frac{1}{p}}A_W(s_1,s_2)\le C \tag{10.12}$$

$$\le \inf_{1<s_1,s_2<p} A_W(s_1,s_2)\left(\frac{p-1}{p-s_1}\right)^{\frac{1}{p'}}\left(\frac{p-1}{p-s_2}\right)^{\frac{1}{p'}}.$$

Proof. Let $f^p(x_1,x_2)v_1(x)v_2(x_2)=g(x_1,x_2)$ in (10.10). Then (10.10) is equivalent to the inequality

$$\left(\int_0^\infty\int_0^\infty\left(\int_0^{x_1}\int_0^{x_1} g(t_1,t_2)^{\frac{1}{p}}v_1(t_1)^{-\frac{1}{p}}v_2(t_2)^{-\frac{1}{p}}dt_1dt_2\right)^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}\le$$

$$C\left(\int_0^\infty\int_0^\infty g(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}. \tag{10.13}$$

Assume that (10.11) holds. By applying Hölder's inequality, the fact that $\frac{d}{dt_1}V_1(t_1)=v_1(t_1)^{1-p'}=v_1(t_1)^{-\frac{p'}{p}}$, $\frac{d}{dt_2}V_2(t_2)=v_2(t_2)^{1-p'}=v_2(t_2)^{-\frac{p'}{p}}$ and Minkowski's inequality we have

$$\left(\int_0^\infty\int_0^\infty\left(\int_0^{x_1}\int_0^{x_2} g(t_1,t_2)^{\frac{1}{p}}v_1(t_1)^{-\frac{1}{p}}v_2(t_2)^{-\frac{1}{p}}dt_1dt_2\right)^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}$$

$$=\left(\int_0^\infty\int_0^\infty\left(\int_0^{x_1}\int_0^{x_2} g(t_1,t_2)^{\frac{1}{p}}V_1(t_1)^{\frac{s_1-1}{p}}V_2(t_2)^{\frac{s_2-1}{p}}\times\right.\right.$$

$$\left.\left.V_1(t_1)^{-\frac{s_1-1}{p}}v_1(t_1)^{-\frac{1}{p}}V_2(t_2)^{-\frac{s_2-1}{p}}v_2(t_2)^{-\frac{1}{p}}dt_1dt_2\right)^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}$$

$$\le\left(\int_0^\infty\int_0^\infty\left(\int_0^{x_1}\int_0^{x2} g(t_1,t_2)V_1(t_1)^{s_1-1}V_2(t_2)^{s_2-1}dt_1dt_2\right)^{\frac{q}{p}}\times\right.$$

$$\left(\int_0^{x_1} V_1(t_1)^{-\frac{(s_1-1)p'}{p}} v_1(t_1)^{-\frac{p'}{p}}\,dt_1\right)^{\frac{q}{p'}} \times$$

$$\left(\int_0^{x_2} V_2(t_2)^{-\frac{(s_2-1)p'}{p}} v_2(t_2)^{-\frac{p'}{p}}\,dt_2\right)^{\frac{q}{p'}} u(x_1,x_2)dx_1dx_2\Bigg)^{\frac{1}{q}}$$

$$= \left(\frac{p-1}{p-s_1}\right)^{\frac{1}{p'}} \left(\frac{p-1}{p-s_2}\right)^{\frac{1}{p'}} \left(\int_0^{\infty}\int_0^{\infty}\left(\int_0^{x_1}\int_0^{x_2} g(t_1,t_2)V_1(t_1)^{s_1-1} \times\right.\right.$$

$$\left.\left. V_2(t_2)^{s_2-1}dt_1dt_2\right)^{\frac{q}{p}} V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)} V_2(x_2)^{q\left(\frac{p-s_2}{p}\right)} u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}$$

$$\le \left(\frac{p-1}{p-s_1}\right)^{\frac{1}{p'}} \left(\frac{p-1}{p-s_2}\right)^{\frac{1}{p'}} \left(\int_0^{\infty}\int_0^{\infty} g(t_1,t_2)V_1(t_1)^{s_1-1}V_2(t_2)^{s_2-1} \times\right.$$

$$\left.\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)} V_2(x_2)^{q\left(\frac{p-s_2}{p}\right)} u(x_1,x_2)dx_1dx_2\right)^{\frac{p}{q}} dt_1dt_2\right)^{\frac{1}{p}}$$

$$\le \left(\frac{p-1}{p-s_1}\right)^{\frac{1}{p'}} \left(\frac{p-1}{p-s_2}\right)^{\frac{1}{p'}} A_W(s_1,s_2)\left(\int_0^{\infty}\int_0^{\infty} g(t_1,t_2)dt_1dt_2\right)^{\frac{1}{p}}.$$

Hence (10.13) and, thus, (10.10) holds with a constant satisfying the right hand side inequality in (10.12).

Now we assume that (10.10) and, thus, (10.13) holds and choose the test function

$$\begin{aligned} g(x_1,x_2) = &\\ &\left(\frac{p}{p-s_1}\right)^p \left(\frac{p}{p-s_2}\right)^p \times\\ &V_1(t_1)^{-s_1}v_1(x_1)^{1-p'}V_2(t_2)^{-s_2}v_2(x_2)^{1-p'}\chi_{(0,t_1)}(x_1)\chi_{(0,t_2)}(x_2)\\ &+\left(\frac{p}{p-s_1}\right)^p V_1(t_1)^{-s_1}v_1(x_1)^{1-p'}V_2(x_2)^{-s_2}v_2(x_2)^{1-p'}\chi_{(0,t_1)}(x_1)\chi_{(t_2,\infty)}(x_2)\\ &+\left(\frac{p}{p-s_2}\right)^p V_1(x_1)^{-s_1}v_1(x_1)^{1-p'}V_2(t_2)^{-s_2}v_2(x_2)^{1-p'}\chi_{(t_1,\infty)}(x_1)\chi_{(0,t_2)}(x_2)\\ &+V_1(x_1)^{-s_1}v_1(x_1)^{1-p'}V_2(x_2)^{-s_2}v_2(x_2)^{1-p'}\chi_{(t_1,\infty)}(x_1)\chi_{(t_2,\infty)}(x_2), \end{aligned} \tag{10.14}$$

where t_1, t_2 are fixed numbers > 0. Then the integral on the right hand side of (10.13) can be estimated as follows:

$$\left(\int_0^\infty\int_0^\infty g(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}} =$$

$$\left(\int_0^{t_1}\left(\frac{p}{p-s_1}\right)^p V_1(t_1)^{-s_1}v_1(x_1)^{1-p'}dx_1\int_0^{t_2}\left(\frac{p}{p-s_2}\right)^p V_2(t_2)^{-s_2}v_2(x_2)^{1-p'}dx_2\right.$$

$$+\int_0^{t_1}\left(\frac{p}{p-s_1}\right)^p V_1(t_1)^{-s_1}v_1(x_1)^{1-p'}dx_1\int_{t_2}^\infty V_2(x_2)^{-s_{21}}v_2(x_2)^{1-p'}dx_2 \qquad (10.15)$$

$$+\int_{t_1}^\infty V_1(x_1)^{-s_1}v_1(x_1)^{1-p'}dx_1\int_0^{t_2}\left(\frac{p}{p-s_2}\right)^p V_2(t_2)^{-s_2}v_2(x_2)^{1-p'}dx_2 \qquad (10.16)$$

$$\left.+\int_{t_1}^\infty V_1(x_1)^{-s_1}v_1(x_1)^{1-p'}dx_1\int_{t_2}^\infty V_2(x_2)^{-s_2}v_2(x_2)^{1-p'}dx_2\right)^{\frac{1}{p}}$$

$$\le\left(\left(\frac{p}{p-s_1}\right)^p+\frac{1}{s_1-1}\right)^{\frac{1}{p}}\left(\left(\frac{p}{p-s_2}\right)^p+\frac{1}{s_2-1}\right)^{\frac{1}{p}}V_1(t_1)^{\frac{1-s_1}{p}}V_2(t_2)^{\frac{1-s_2}{p}}.$$

Moreover, the left hand side of (10.13) is greater than

$$\int_{t_1}^\infty\int_{t_2}^\infty\left[\left(\left(\int_0^{t_1}\frac{p}{p-s_1}V_1(t_1)^{-\frac{s_1}{p}}v_1(y_1)^{1-p'}dy_1\right)\right.\right.$$

$$\left.\left.\times\left(\int_0^{t_2}\frac{p}{p-s_2}V_2(t_2)^{-\frac{s_2}{p}}v_2(y_2)^{1-p'}dy_2\right)\right)\right.$$

$$+\left(\int_0^{t_1}\frac{p}{p-s_1}V_1(t_1)^{-\frac{s_1}{p}}v_1(y_1)^{1-p'}dy_1\right)\left(\int_{t_2}^{x_2}V_2(y_1)^{-\frac{s_2}{p}}v_2(y_2)^{1-p'}dy_2\right)$$

$$+\left(\int_{t_1}^{x_1}V_1(y_1)^{-\frac{s_1}{p}}v_1(y_1)^{1-p'}dy_1\right)\left(\int_0^{t_2}\frac{p}{p-s_2}V_2(t_2)^{-\frac{s_2}{p}}v_2(y_2)^{1-p'}dy_2\right)$$

$$+\left(\left(\int_{t_1}^{x_1} V_1(y_1)^{-\frac{s_1}{p}} v_1(y_1)^{1-p'} dy_1\right)\right.$$

$$\left.\times\left(\int_{t_2}^{x_2} V_2(y_1)^{-\frac{s_2}{p}} v_2(y_2)^{1-p'} dy_2\right)\right)\Bigg]^q u(x_1,x_2)dx_1dx_2\Bigg)^{\frac{1}{q}}$$

$$= \ldots = \frac{p}{p-s_1}\frac{p}{p-s_2}\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} u(x_1,x_2)V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)}V_2(x_2)^{q\left(\frac{p-s_2}{p}\right)}dx_1dx_2\right)^{\frac{1}{q}}.$$

Hence, (10.13) implies that

$$= \frac{p}{p-s_1}\frac{p}{p-s_2}\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} u(x_1,x_2)V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)}V_2(x_2)^{q\left(\frac{p-s_2}{p}\right)}dx_1dx_2\right)^{\frac{1}{q}}$$

$$\le C\left(\left(\frac{p}{p-s_1}\right)^p + \frac{1}{s_1-1}\right)^{\frac{1}{p}}\left(\left(\frac{p}{p-s_2}\right)^p + \frac{1}{s_2-1}\right)^{\frac{1}{p}} V_1(t_1)^{\frac{1-s_1}{p}}V_2(t_2)^{\frac{1-s_2}{p}},$$

i.e. that

$$\left(\frac{\left(\frac{p}{p-s_1}\right)^p}{\left(\frac{p}{p-s_1}\right)^p + \frac{1}{s_1-1}}\right)^{\frac{1}{p}}\left(\frac{\left(\frac{p}{p-s_2}\right)^p}{\left(\frac{p}{p-s_2}\right)^p + \frac{1}{s_2-1}}\right)^{\frac{1}{p}} V_1(t_1)^{\frac{s_1-1}{p}}V_2(t_2)^{\frac{s_2-1}{p}}\times$$

$$\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} u(x_1,x_2)V_1(x_1)^{q\left(\frac{p-s_1}{p}\right)}V(x_2)^{q\left(\frac{p-s_2}{p}\right)}dx_1dx_2\right)^{\frac{1}{q}} \le C.$$

We conclude that (10.11) and the left hand side of the estimate of (10.12) hold. The proof is complete. □

10.3 The (Limit) Two-Dimensional Pólya-Knopp Type Inequality

The main result in this section is just the following natural limit result of Theorem 4:

Theorem 5. *Let $0 < p \le q < \infty$ and let u and v be strictly positive and measurable functions on $\mathbb{R}^2_+$. Then*

$$\left(\int_0^\infty\int_0^\infty\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log f(y_1,y_2)dy_1dy_2\right)\right]^q u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}} \tag{10.17}$$
$$\le C\left(\int_0^\infty\int_0^\infty f^p(x_1,x_2)v(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}$$

for all positive and measurable functions on $[0,\infty]\times[0,\infty]$ *if and only if*

$$D_W(s_1,s_2):=\sup_{\substack{y_1\in(0,b_1)\\ y_2\in(0,b_2)}} y_1^{\frac{s_1-1}{p}}y_2^{\frac{s_2-1}{p}}\left(\int_{y_1}^\infty\int_{y_2}^\infty x_1^{-\frac{s_1q}{p}}x_2^{-\frac{s_2q}{p}}w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}<\infty, \tag{10.18}$$

where $s_1,s_2>1$ *and*

$$w(x_1,x_2)=\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log\frac{1}{v(t_1,t_2)}dt_1dt_2\right)\right]^{\frac{q}{p}}u(x_1,x_2) \tag{10.19}$$

and the best possible constant C *in (10.17) can be estimated in the following way:*

$$\sup_{s_1,s_2>1}\left(\frac{e^{s_1}(s_1-1)}{e^{s_1}(s_1-1)+1}\right)^{\frac{1}{p}}\left(\frac{e^{s_2}(s_2-1)}{e^{s_2}(s_2-1)+1}\right)^{\frac{1}{p}}D_W(s_1,s_2)\le C \tag{10.20}$$
$$\le\inf_{s_1,s_2>1}e^{\frac{s_1+s_2-2}{p}}D_W(s_1,s_2).$$

Remark 6. For the case $p=q=1$ a similar result was recently proved by H. P. Heinig, R. Kerman and M. Krbec [9] but without the estimates of the operator norm (= the best constant C) in (10.17) pointed out in (10.20) here.

Remark 7. It is easy to see that the inequality

$$\int_0^\infty e\left(\frac{1}{x}\int_0^x\ln f(t)dt\right)dx\le e\left(\int_0^\infty f(x)dx\right)$$

may be regarded as a limit case (as $p\to\infty$) of the original Hardy's inequality

$$\int_0^\infty\left(\frac{1}{x}\int_0^x f(t)dt\right)^p dx\le\left(\frac{p}{p-1}\right)^p\int_0^\infty f^p(x)dx.$$

This inequality is sometimes referred to as Knopp's inequality but it was obvious known to Pólya before (see [16] and the references given there), so nowadays it is usually referred to as the Pólya-Knopp inequality. Therefore it is natural to regard (10.17) as a two-dimensional Pólya-Knopp type inequality.

Proof. If we in the inequality (10.17) replace $f^p(x_1,x_2)v(x_1,x_2)$ with $f^p(x_1,x_2)$ and let $w(x_1,x_2)$ be defined as in (10.19), then (10.17) is equivalent to

$$\left(\int_0^\infty\int_0^\infty\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log f(y_1,y_2)dy_1dy_2\right)\right]^q w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}$$
$$\le C\left(\int_0^\infty\int_0^\infty f^p(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}.$$

Further, by using Theorem 4 with the special weights $u(x_1,x_2)=w(x_1,x_2)x_1^{-q}x_2^{-q}$ and $v_1(x_1)=v_2(x_2)=1$ we have that

$$\left(\int_0^\infty\int_0^\infty\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}f(t_1,t_2)dt_1dt_2\right)^q w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}} \tag{10.21}$$
$$\le C\left(\int_0^\infty\int_0^\infty f^p(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}$$

holds for all $f\ge 0$ if and only if

$$A_W(s_1,s_2)=\sup_{t_1,t_2>0}t_1^{\frac{s_1-1}{p}}t_2^{\frac{s_2-1}{p}}\left(\int_{t_1}^\infty\int_{t_2}^\infty w(x_1,x_2)x_1^{-s_1\frac{q}{p}}x_2^{-s_2\frac{q}{p}}dx_1dx_2\right)^{\frac{1}{q}}<\infty. \tag{10.22}$$

We note that $A_W(s_1,s_2)$ coincides with the constant $D_W(s_1,s_2)=D_W(s_1,s_2,q,p)$ defined by (10.18) and (10.19). Moreover, if C is the best possible constant in (10.21), then

$$\sup_{1<s_1,s_2<p}\left(\frac{\left(\frac{p}{p-s_1}\right)^p}{\left(\frac{p}{p-s_1}\right)^p+\frac{1}{s_1-1}}\right)^{\frac{1}{p}}\left(\frac{\left(\frac{p}{p-s_2}\right)^p}{\left(\frac{p}{p-s_2}\right)^p+\frac{1}{s_2-1}}\right)^{\frac{1}{p}}D_W(s_1,s_2)\le C$$
$$\le\inf_{1<s_1,s_2<p}D_W(s_1,s_2)\left(\frac{p-1}{p-s_1}\right)^{\frac{1}{p'}}\left(\frac{p-1}{p-s_2}\right)^{\frac{1}{p'}}. \tag{10.23}$$

Now, if we replace f in (10.21) with f^{α}, $0 < \alpha < p$ and after that replace p with $\frac{p}{\alpha}$ and q with $\frac{q}{\alpha}$ in (10.21)–(10.23), then we find that, for $1 < s_1, s_2 < \frac{p}{\alpha}$,

$$\left(\int_0^{\infty}\int_0^{\infty}\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2} f^{\alpha}(t_1,t_2)dt_1dt_2\right)^{q} w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}} \tag{10.24}$$
$$\leq C_{\alpha}\left(\int_0^{\infty}\int_0^{\infty} f^{p}(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}$$

holds for all $f \geq 0$ if and only if $D_W(s_1,s_2,\frac{q}{\alpha},\frac{p}{\alpha}) = D_W^{\alpha}(s_1,s_2,q,p) < \infty$. Moreover, if C_{α} is the best possible constant in (10.24), then

$$\sup_{1<s_1,s_2<p}\left(\frac{\left(\frac{p}{p-\alpha s_1}\right)^{\frac{p}{\alpha}}}{\left(\frac{p}{p-\alpha s_1}\right)^{p}+\frac{1}{s_1-1}}\right)^{\frac{1}{p}}\left(\frac{\left(\frac{p}{p-\alpha s_2}\right)^{\frac{p}{\alpha}}}{\left(\frac{p}{p-\alpha s_2}\right)^{p}+\frac{1}{s_2-1}}\right)^{\frac{1}{p}} D_W^{\alpha}(s_1,s_2,q,p) \leq C_{\alpha}$$
$$\leq \inf_{1<s_1,s_2<p} D_W^{\alpha}(s_1,s_2,q,p)\left(\frac{p-\alpha}{p-\alpha s_1}\right)^{\frac{p-\alpha}{\alpha p}}\left(\frac{p-\alpha}{p-\alpha s_2}\right)^{\frac{p-\alpha}{\alpha p}}. \tag{10.25}$$

We also note that

$$\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2} f^{\alpha}(t_1,t_2)dt_1dt_2\right)^{\frac{1}{\alpha}} \downarrow e^{\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\ln f(t_1,t_2)dt_1dt_2}, \text{ as } \alpha \to 0_+.$$

We conclude that (10.18) holds exactly when $\limsup_{\alpha\to 0_+} C_{\alpha} < \infty$ and this holds, according to (10.25), exactly when (10.22) holds. Moreover, when $\alpha \to 0_+$ (10.25) implies that (10.20) holds. For the lower estimate we apply the testfunction

$$g(x_1,x_2) = g_0(x_1,x_2) = t_1^{-1}t_2^{-1}\chi_{(0,t_1)}(x_1)\chi_{(0,t_2)}(x_2) + \tag{10.26}$$
$$t_1^{-1}\chi_{(0,t_1)}(x_1)\frac{e^{-s_2}t_2^{s_2-1}}{x_2^{s_2}}\chi_{(t_2,\infty)}(x_2) +$$
$$\frac{e^{-s_1}t_1^{s_1-1}}{x_1^{s_1}}\chi_{(t_1,\infty)}(x_1)t_2^{-1}\chi_{(0,t_2)}(x_2) +$$
$$\frac{e^{-(s_1+s_2)}t_1^{s_1-1}t_2^{s_2-1}}{x_1^{s_1}x_2^{s_2}}\chi_{(t_1,\infty)}(x_1)\chi_{(t_2,\infty)}(x_2).$$

The proof is complete. □

Remark 8. This proof shows that the Pólya-Knopp inequality characterized in Theorem 5 may be regarded as a natural limiting inequality of the (Sawyer type) Hardy inequality characterized in Theorem 4.

We will finish this Section by presenting an alternative proof of Theorem 5 which is independent of Theorem 4 but heavily depending of the following well-known two dimensional version of the Minkowski integral inequality:

Lemma 9. *Let $r > 1$, $-\infty \le a_1 < b_1 \le \infty$, $-\infty \le a_2 < b_2 \le \infty$ and let Φ and Ψ be positive measurable functions on $[a_1, b_1] \times [a_2, b_2]$. Then*

$$\int_{a_1}^{b_1}\int_{a_2}^{b_2} \Phi(x_1,x_2)\left(\int_{a_1}^{x_1}\int_{a_2}^{x_2}\Psi(y_1,y_2)dy_1dy_2\right)^r dx_1dx_2 \tag{10.24}$$

$$\le \int_{a_1}^{b_1}\int_{a_1}^{b_2}\Psi(y_1,y_2)\left(\int_{y_1}^{b_1}\int_{y_1}^{b_2}\Phi(x_1,x_2)dx_1d_2\right)^{\frac{1}{r}} dy_1dy_2,$$

For an elementary proof of Lemma 9 see e.g. [34, p. 41]

Alternative proof of Theorem 5. Let $g(x_1,x_2) = f^p(x_1,x_2)v(x_1,x_2)$ in (10.17). Then we see that (10.17) is equivalent to the inequality

$$\left(\int_0^{b_1}\int_0^{b_2}\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g(y_1,y_2)dy_1dy_2\right)\right]^{\frac{q}{p}}\times\right. \tag{10.25}$$

$$\left.\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log\frac{1}{v(t_1,t_2)}dt_1dt_2\right)\right]^{\frac{q}{p}}u(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}$$

$$\le C\left(\int_0^{b_1}\int_0^{b_2}g(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}.$$

If $w(x_1,x_2)$ is defined by (10.19), then we can equivalently write (10.25) as

$$\left(\int_0^{b_1}\int_0^{b_2}\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g(y_1,y_2)dy_1dy_2\right)\right]^{\frac{q}{p}}w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}} \tag{10.26}$$

$$\le C\left(\int_0^{b_1}\int_0^{b_2}g(x_1,x_2)dx_1dx_2\right)^{\frac{1}{p}}.$$

Let $y_1 = x_1 t_1$ and $y_2 = x_2 t_2$ and (10.26) becomes

$$\left(\int_0^{b_1} \int_0^{b_2} \left[e \left(\int_0^1 \int_0^1 \log g(x_1 t_1, x_2 t_2) dt_1 dt_2 \right) \right]^{\frac{q}{p}} w(x_1, x_2) dx_1 dx_2 \right)^{\frac{1}{q}}$$

$$\leq C \left(\int_0^{b_1} \int_0^{b_2} g(x_1, x_2) dx_1 dx_2 \right)^{\frac{1}{p}}. \tag{10.27}$$

By using that

$$\left(e \int_0^1 \int_0^1 \log t_1^{s_1-1} t_2^{s_2-1} dt_1 dt_2 \right)^{\frac{q}{p}} = e^{-(s_1+s_2-2)\frac{q}{p}}$$

and Jensen's inequality, the left hand side of (10.27) becomes

$$e^{\frac{s_1+s_2-2}{p}} \left(\int_0^{b_1} \int_0^{b_2} \left[e \int_0^1 \int_0^1 \log \left(t_1^{s_1-1} t_2^{s_2-1} g(x_1 t_1, x_2 t_2) \right) dt_1 dt_2 \right]^{\frac{q}{p}} w(x_1, x_2) dx_1 dx_2 \right)^{\frac{1}{q}}$$

$$\leq e^{\frac{s_1+s_2-2}{p}} \left(\int_0^{b_1} \int_0^{b_2} \left[\int_0^1 \int_0^1 t_1^{s_1-1} t_2^{s_2-1} g(x_1 t_1, x_2 t_2) dt_1 dt_2 \right]^{\frac{q}{p}} w(x_1, x_2) dx_1 dx_2 \right)^{\frac{1}{q}}$$

$$= e^{\frac{s_1+s_2-2}{p}} \left(\int_0^{b_1} \int_0^{b_2} \left[\int_0^{x_1} \int_0^{x_2} y_1^{s_1-1} y_2^{s_2-1} g(y_1, y_2) dy_1 dy_2 \right]^{\frac{q}{p}} \frac{w(x_1, x_2)}{x_1^{s_1 \frac{q}{p}} x_2^{s_2 \frac{q}{p}}} dx_1 dx_2 \right)^{\frac{1}{q}}.$$

Therefore, by also using Proposition 9 with $r = \frac{q}{p}$ for $p < q$ and Fubini's theorem for $p = q$, we find that the left hand side in (10.27) can be estimated as follows:

$$\leq e^{\frac{s_1+s_2-2}{p}}$$

$$\times \left(\int_0^{b_1} \int_0^{b_2} y_1^{s_1-1} y_2^{s_2-1} g(y_1, y_2) \left(\int_{y_1}^{b_1} \int_{y_2}^{b_2} x_1^{-\frac{s_1 q}{p}} x_2^{-\frac{s_2 q}{p}} w(x_1, x_2) dx_1 dx_2 \right)^{\frac{p}{q}} dy_1 dy_2 \right)^{\frac{1}{p}}$$

$$\leq e^{\frac{s_1+s_2-2}{p}} D_W(s_1, s_2) \left(\int_0^{b_1} \int_0^{b_2} g(y_1, y_2) dy_1 dy_2 \right)^{\frac{1}{p}}.$$

Hence, (10.27) and, thus, (10.17) holds with a constant C satisfying the right hand side estimate in (10.20).

Now, assume that (10.17) and, thus, (10.26) holds. For fixed t_1 and t_2, $0 < t_1 < b_1, 0 < t_2 < b_2$, we choose the test function 10.26 .Then for the right side of (10.26) it yields that

$$\begin{aligned}
&\left(\int_0^{b_1}\int_0^{b_2} g_0(y_1,y_2)\,dy_1dy_2\right)^{\frac{1}{p}}\\
&= \left(\int_0^{t_1}\int_0^{t_2} t_1^{-1}t_2^{-1}dy_1dy_2 + \int_0^{t_1}\int_{t_2}^{b_2} t_1^{-1}\frac{e^{-s_2}t_2^{s_2-1}}{y_2^{s_2}}dy_1dy_2\right.\\
&\quad \left. + \int_{t_1}^{b_1}\int_0^{t_2} t_2^{-1}\frac{e^{-s_1}t_1^{s_1-1}}{y_1^{s_1}}dy_1dy_2 + \int_{t_1}^{b_1}\int_{t_2}^{b_2}\frac{e^{-(s_1+s_2)}t_1^{s_1-1}t_2^{s_1-1}}{y_1^{s_1}y_2^{s_2}}dy_1dy_2\right)^{\frac{1}{p}}\\
&= \left(1+\frac{e^{-s_2}}{s_2-1}\left(1-\left(\frac{t_2}{b_2}\right)^{s_2-1}\right)+\frac{e^{-s_1}}{s_1-1}\left(1-\left(\frac{t_1}{b_1}\right)^{s_1-1}\right)\right.\\
&\quad \left.+\frac{e^{-s_1}e^{-s_2}}{(s_1-1)(s_2-1)}\left(1-\left(\frac{t_1}{b_1}\right)^{s_1-1}\right)\left(1-\left(\frac{t_2}{b_2}\right)^{s_2-1}\right)\right)^{\frac{1}{p}}\\
&\le \left(1+\frac{e^{-s_2}}{s_2-1}+\frac{e^{-s_1}}{s_1-1}+\frac{e^{-s_1}e^{-s_2}}{(s_1-1)(s_2-1)}\right)^{\frac{1}{p}},
\end{aligned}$$

i.e.,

$$\left(\int_0^{b_1}\int_0^{b_2} g_0(y_1,y_2)\,dy_1dy_2\right)^{\frac{1}{p}} \le \left(\frac{e^{s_1}(s_1-1)+1}{e^{s_1}(s_1-1)}\right)^{\frac{1}{p}}\left(\frac{e^{s_2}(s_2-1)+1}{e^{s_2}(s_2-1)}\right)^{\frac{1}{p}}. \tag{10.28}$$

Moreover, for the left hand side in (10.26) we have

$$\left(\int_0^{b_1}\int_0^{b_2} w(x_1,x_2)\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g(y_1,y_2)dy_1dy_2\right)\right]^{\frac{q}{p}}dx_1dx_2\right)^{\frac{1}{q}} \ge \tag{10.29}$$

$$\left(\int_{t_1}^{b_1}\int_{t_2}^{b_2} w(x_1,x_2)\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g(y_1,y_2)dy_1dy_2\right)\right]^{\frac{q}{p}}dx_1dx_2\right)^{\frac{1}{q}}.$$

With the function $g_0(y_1, y_2)$ we get that

$$e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g_0(y_1,y_2)dy_1dy_2\right)=e\left(I_1+I_2+I_3+I_4\right),$$

where

$$I_1=\frac{1}{x_1x_2}\int_0^{t_1}\int_0^{t_2}\log\left(t_1^{-1}t_2^{-1}\right)dy_1dy_2=-\frac{t_1t_2}{x_1x_2}\log t_1-\frac{t_1t_2}{x_1x_2}\log t_2,$$

$$I_2=\frac{1}{x_1x_2}\int_0^{t_1}\int_{t_2}^{x_2}\log\left(t_1^{-1}\frac{e^{-s_2}t_2^{s_2-1}}{y_2^{s_2}}\right)dy_1dy_2=$$
$$-\frac{t_1}{x_1}\log t_1+\frac{t_1t_2}{x_1x_2}\log t_1+(s_2-1)\frac{t_1}{x_1}\log t_2+\frac{t_1t_2}{x_1x_2}\log t_2-s_2\frac{t_1}{x_1}\log x_2,$$

$$I_3=\frac{1}{x_1x_2}\int_{t_1}^{x_1}\int_0^{t_2}\log\left(t_2^{-1}\frac{e^{-s_1}t_1^{s_1-1}}{y_1^{s_1}}\right)dy_1dy_2=$$
$$-\frac{t_2}{x_2}\log t_2+\frac{t_1t_2}{x_1x_2}\log t_2+(s_1-1)\frac{t_2}{x_2}\log t_1+$$
$$\frac{t_1t_2}{x_1x_2}\log t_1-s_1\frac{t_2}{x_2}\log x_1,$$

and

$$I_4=(s_1-1)\log t_1-(s_1-1)\frac{t_2}{x_2}\log t_1+\frac{t_1t_2}{x_1x_2}\log t_1$$
$$+(s_2-1)\log t_2-(s_2-1)\frac{t_1}{x_1}\log t_2+\frac{t_1t_2}{x_1x_2}\log t_2$$
$$-s_1\log x_1+\frac{t_1}{x_1}\log t_1+s_1\frac{t_2}{x_2}\log x_1$$
$$-s_2\log x_2+\frac{t_2}{x_2}\log t_2+s_2\frac{t_1}{x_1}\log x_2.$$

Now we see that

$$I_1+I_2+I_3+I_4=\log\left(\frac{t_1^{(s_1-1)}t_2^{(s_2-1)}}{x_1^{s_1}x_2^{s_2}}\right)$$

so that, by (10.29),

$$\left(\int_{t_1}^{b_1}\int_{t_2}^{b_2} w(x_1,x_2)\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log g_0(y_1,y_2)dy_1dy_2\right)\right]^{\frac{q}{p}}dx_1dx_2\right)^{\frac{1}{q}}$$

$$=\left(\int_{t_1}^{b_1}\int_{t_2}^{b_2} w(x_1,x_2)\left[\frac{t_1^{(s_1-1)}t_2^{(s_2-1)}}{x_1^{s_1}x_2^{s_2}}\right]^{\frac{q}{p}}dx_1dx_2\right)^{\frac{1}{q}}$$

Hence, by (10.26) and (10.28),

$$t_1^{\frac{s_1-1}{p}}t_2^{\frac{s_2-1}{p}}\left(\int_{t_1}^{b_1}\int_{t_2}^{b_2} x_1^{-\frac{q}{p}s_1}x_2^{-\frac{q}{p}s_2}w(x_1,x_2)dx_1dx_2\right)^{\frac{1}{q}}\le$$

$$C\left(\frac{e^{s_1}(s_1-1)+1}{e^{s_1}(s_1-1)}\right)^{\frac{1}{p}}\left(\frac{e^{s_2}(s_2-1)+1}{e^{s_2}(s_2-1)}\right)^{\frac{1}{p}},$$

i.e.

$$\left(\frac{e^{s_1}(s_1-1)}{e^{s_1}(s_1-1)+1}\right)^{\frac{1}{p}}\left(\frac{e^{s_2}(s_2-1)}{e^{s_2}(s_2-1)+1}\right)^{\frac{1}{p}}D_W(s_1,s_2)\le C.$$

We conclude that (10.18) and the left hand side inequality of (10.20) hold. The proof is complete. □

Corollary 10. *Let $0<p\le q<\infty$, and $\alpha_1,\alpha_2,\beta_1,\beta_2\in\mathbb{R}$. Then*

$$\left(\int_0^{\infty}\int_0^{\infty}\left[e\left(\frac{1}{x_1x_2}\int_0^{x_1}\int_0^{x_2}\log f(y_1,y_2)dy_1dy_2\right)\right]^{q}x_1^{\alpha_1}x_2^{\alpha_2}dx_1dx_2\right)^{\frac{1}{q}} \tag{10.30}$$

$$\le C\left(\int_0^{\infty}\int_0^{\infty} f^p(x_1,x_2)x_1^{\beta_1}x_2^{\beta_2}dx_1dx_2\right)^{\frac{1}{p}}$$

holds for all positive and measurable functions f on $\mathbb{R}^2_+$ with a finite constant C if and only if

$$\frac{\alpha_1+1}{q}=\frac{\beta_1+1}{p},$$

and

$$\frac{\alpha_2+1}{q}=\frac{\beta_2+1}{p},$$

and the best constant C in (10.30) can be estimated as follows:

$$\sup_{s_1,s_2>1}\left(\frac{e^{s_1}(s_1-1)}{e^{s_1}(s_1-1)+1}\cdot\frac{e^{s_2}(s_2-1)}{e^{s_2}(s_2-1)+1}\right)^{\frac{1}{p}}\left(\frac{1}{s_1-1}\cdot\frac{1}{s_2-1}\right)^{\frac{1}{q}}\times$$

$$e^{\frac{\beta_1+\beta_2}{p}}\left(\frac{p}{q}\right)^{\frac{2}{q}}\le C\le e^{\frac{\beta_1+\beta_2}{p}+\frac{2}{q}}. \tag{10.31}$$

Proof. Apply Theorem 5 with the weights $u(x_1,x_2)=x_1^{\alpha_1}x_2^{\alpha_2}$ and $v(x_1,x_2)=x_1^{\beta_1}x_2^{\beta_2}$. The left hand side estimate of (10.31) follows directly and for the right hand side we can use the optimal values $s_1=1+\frac{p}{q}$ and $s_2=1+\frac{p}{q}$ found by B. Opic and P. Gurka [25] and obtain

$$\begin{aligned}C &\le \inf_{s_1,s_2>1}\left(\frac{1}{s_1-1}\cdot\frac{1}{s_2-1}\right)^{\frac{1}{q}}e^{\frac{\beta_1+\beta_2+s_1+s_2-2}{p}}\left(\frac{p}{q}\right)^{\frac{2}{q}}\\ &= \inf_{s_1>1}\left(\frac{1}{s_1-1}\cdot\right)^{\frac{1}{q}}e^{\frac{\beta_1+s_1-1}{p}}\inf_{s_2>1}\left(\frac{1}{s_2-1}\right)^{\frac{1}{q}}e^{\frac{\beta_2+s_2-1}{p}}\left(\frac{p}{q}\right)^{\frac{2}{q}}\\ &= e^{\frac{\beta_1+\beta_2}{p}+\frac{2}{q}}.\end{aligned}$$

The proof is complete. □

Remark 11. If $p=q$, then the inequality (10.30) is sharp with the constant $C=e^{\frac{\beta_1+\beta_2+2}{p}}$, see Theorem 2.2 in [10].

Remark 12. By using the techniques and results in this section and induction all results can be formulated and proved also in an n-dimensional setting (see [34]). However, in our next Sections we will prove some closely related results by using another technique, which was recently presented in another Ph.D. thesis by E. Ushakova [33].

10.4 The Multi-dimensional Case $1<p\le q<\infty$

In this and the next sections we deal with the inequality

$$\left(\int_{\mathbb{R}^n_+}(H_nf)^q(\mathbf{x})w(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{q}}\le C\left(\int_{\mathbb{R}^n_+}f^p(\mathbf{y})v(\mathbf{y})d\mathbf{y}\right)^{\frac{1}{p}}, \tag{10.32}$$

where

$$(H_nf)(\mathbf{x})=\int_0^{x_1}\ldots\int_0^{x_n}f(t_1\ldots t_n)dt_1\ldots dt_n,\qquad \mathbf{x}:=(\mathbf{x}_1\ldots x_n)\in\mathbb{R}^n_+$$

and one of the two weight functions v and w is of product type, that is where

$$v(\mathbf{y}) = v(y_1, \dots, y_n) = v_1(y_1) \dots v_n(y_n) \tag{10.33}$$

or

$$w(\mathbf{x}) = w(x_1, \dots, x_n) = w_1(x_1) \dots w_n(x_n). \tag{10.34}$$

Conditions (10.33) and (10.34) are satisfied, for instance, by a power function of n variables.

In this Section we obtain new necessary and sufficient conditions for the validity of (10.32) in the case $1 < p \le q < \infty$ and when (10.33) is satisfied. The same problem is considered here with the assumption (10.34). Our estimates are n-dimensional analogies of well known criteria for the one-dimensional integral Hardy inequality (see [8], [18] and [27]).

In the next preliminary Lemmas we state some necessary conditions for the inequality (10.32) to hold in the case $1 < p \le q < \infty$ without any restrictions on the weight functions w and v. These Lemmas are useful in our proofs later on but also of independent interest because they indicate the problem to extend Theorem 1 to the n-dimensional case.

Lemma 13. *Let $1 < p \le q < \infty$ and assume that the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with a finite constant C, which is independent on f. Then*

$$\sup_{\substack{t_i>0 \\ i=1,\dots,n}} W(t_1, \dots, t_n)^{\frac{1}{q}} V(t_1, \dots, t_n)^{\frac{1}{p'}} < \infty, \tag{10.35}$$

where

$$W(t_1, \dots, t_n) := W(\mathbf{t}) = \int_{t_1}^{\infty} \dots \int_{t_n}^{\infty} w(\mathbf{x}) d\mathbf{x}$$

and

$$V(t_1, \dots, t_n) := V(\mathbf{t}) = \int_0^{t_1} \dots \int_0^{t_n} v(\mathbf{y})^{1-p'} d\mathbf{y}.$$

Proof. For $\mathbf{t} = (t_1, \dots, t_n)$ such that $t_i > 0$, $i = 1, \dots, n$, we take a test function

$$f_{\mathbf{t}}(\mathbf{y}) := \chi_{[0,t_1]}(y_1) \dots \chi_{[0,t_n]}(y_n) v(\mathbf{y})^{1-p'} \tag{10.36}$$

and put it into the inequality (10.32). Then we have that

$$C \ge \frac{\left(\int_{\mathbb{R}^n_+} \left(\int_0^{x_1} \dots \int_0^{x_n} f_{\mathbf{t}}(\mathbf{y}) d\mathbf{y} \right)^q w(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{q}}}{\left(\int_{\mathbb{R}^n_+} f_{\mathbf{t}}^{\,p}(\mathbf{y}) v(\mathbf{y}) d\mathbf{y} \right)^{\frac{1}{p}}}$$

$$\geq \frac{\left(\int_{t_1}^{\infty}\dots\int_{t_n}^{\infty} w(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{q}}\left(\int_0^{t_1}\dots\int_0^{t_n} v(\mathbf{y})^{1-p'}d\mathbf{y}\right)}{\left(\int_0^{t_1}\dots\int_0^{t_n} v(\mathbf{y})^{1-p'}d\mathbf{y}\right)^{\frac{1}{p}}} = W(\mathbf{t})^{\frac{1}{q}}V(\mathbf{t})^{\frac{1}{p'}}.$$

Thus, (10.35) follows by taking the supremum over all $t_i > 0$, $i = 1,\dots,n$. □

Lemma 14. *Let $1 < p \leq q < \infty$ and suppose that the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C independent on f. Then*

$$\sup_{\substack{t_i>0\\ i=1,\dots,n}} V(t_1,\dots,t_n)^{-\frac{1}{p}}\left(\int_0^{t_1}\dots\int_0^{t_n} w(\mathbf{x})V(\mathbf{x})^q d\mathbf{x}\right)^{\frac{1}{q}} < \infty. \tag{10.37}$$

Proof. This statement follows evidently by substituting into the inequality (10.32) the function $f_{\mathbf{t}}(\mathbf{y})$ (see (10.36)) for $\mathbf{t} = (t_1,\dots,t_n)$ such that $t_i > 0$, $i = 1,\dots,n$. □

Lemma 15. *Let $1 < p \leq q < \infty$ and assume that the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C independent on f. Then*

$$\sup_{\substack{t_i>0\\ i=1,\dots,n}} W(t_1,\dots,t_n)^{-\frac{1}{q'}}\left(\int_{t_1}^{\infty}\dots\int_{t_n}^{\infty} v(\mathbf{x})^{1-p'}W(\mathbf{x})^{p'}d\mathbf{x}\right)^{\frac{1}{p'}} < \infty. \tag{10.38}$$

Proof. By duality the inequality (10.32) is equivalent to the inequality

$$\left(\int_{\mathbb{R}^n_+}\left(H_n^* g\right)^{p'}(\mathbf{x})v^{1-p'}(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{p'}} \leq C\left(\int_{\mathbb{R}^n_+} g^{q'}(\mathbf{y})w^{1-q'}(\mathbf{y})d\mathbf{y}\right)^{\frac{1}{q'}} \tag{10.39}$$

with the dual operator H_n^* defined by

$$(H_n^* g)(\mathbf{x}) := \int_{x_1}^{\infty}\dots\int_{x_n}^{\infty} g(\mathbf{y})d\mathbf{y}, \qquad x_1,\dots,x_n > 0. \tag{10.40}$$

Now (10.38) follows by substituting into the inequality (10.39) the function

$$g_{\mathbf{t}}(\mathbf{y}) := \chi_{[t_1,\infty)}(y_1)\dots\chi_{[t_n,\infty)}(y_n)w(\mathbf{y})$$

for $\mathbf{t} = (t_1,\dots,t_n)$ such that $t_i > 0$, $i = 1,\dots,n$, and taking supremum. □

Remark 16. Note that for $n = 2$ the statements of Lemmas 13 – 15 follow from Theorem 1.

The first main theorem in this Section reads:

Theorem 17. *Let $1 < p \le q < \infty$ and the weight function v be of product type (10.33). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $A_{M_n} < \infty$, where*

$$A_{M_n} := \sup_{\substack{t_i>0\\ i=1,\dots,n}} W(t_1,\dots,t_n)^{\frac{1}{q}} V_1(t_1)^{\frac{1}{p'}} \dots V_n(t_n)^{\frac{1}{p'}} \tag{10.41}$$

and

$$V_i(t_i) := \int_0^{t_i} v_i(x_i)^{1-p'}\,dx_i, \qquad i = 1,\dots,n.$$

Moreover, $C \approx A_{M_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.

Proof. The necessary part of the proof follows from Lemma 13 while the sufficiency can be obtained from the n-dimensional extension of Theorem 1 and from the following Lemma: □

Lemma 18. *Let*

$$A_{W_n} := A_{W_n}(s_1,\dots,s_n) := \sup_{\substack{t_i>0\\ i=1,\dots,n}} V_1(t_1)^{\frac{s_1-1}{p}} \dots V_n(t_n)^{\frac{s_n-1}{p}} \times \left(\int_{t_1}^{\infty}\dots\int_{t_n}^{\infty} w(\mathbf{x}) V_1(x_1)^{\frac{q(p-s_1)}{p}} \dots V_n(x_n)^{\frac{q(p-s_n)}{p}}\,d\mathbf{x}\right)^{\frac{1}{q}},$$

where $s_i \in (1,p)$, $i = 1,\dots,n$. Then

$$A_{W_n} \ll A_{M_n}. \tag{10.42}$$

Proof. Let $n = 2$ and $s_1 = s_2 = \frac{1+p}{2}$. Then

$$A_{W_2} = \sup_{t_1,t_2>0} V_1(t_1)^{\frac{1}{2p'}} V_2(t_2)^{\frac{1}{2p'}} \left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} w(x_1,x_2) V_1(x_1)^{\frac{q}{2p'}} V_2(x_2)^{\frac{q}{2p'}}\,dx_2dx_1\right)^{\frac{1}{q}}.$$

Since

$$V_i(x_i)^{\frac{q}{2p'}} = \frac{q}{2p'}\int_0^{x_i} v_i(y_i)^{1-p'} V_i(y_i)^{\frac{q}{2p'}-1} dy_i, \qquad i = 1,2,$$

we have that

$$\begin{aligned} &V_1(x_1)^{\frac{q}{2p'}} V_2(x_2)^{\frac{q}{2p'}} \\ &\approx \left(\left[\int_0^{t_1} + \int_{t_1}^{x_1}\right] v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1\right) \\ &\quad\times \left(\left[\int_0^{t_2} + \int_{t_2}^{x_2}\right] v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2\right) \end{aligned}$$

$$
\begin{aligned}
&= \int_0^{t_1} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1 \int_0^{t_2} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2 \\
&\quad + \int_{t_1}^{x_1} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1 \int_{t_2}^{x_2} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2 \\
&\quad + \int_0^{t_1} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1 \int_{t_2}^{x_2} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2 \\
&\quad + \int_{t_1}^{x_1} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1 \int_0^{t_2} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2 \\
&=: I_{11} + I_{22} + I_{12} + I_{21}.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) V_1(x_1)^{\frac{q}{2p'}} V_2(x_2)^{\frac{q}{2p'}} dx_2 dx_1 \\
&\quad \approx \int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) \left[I_{11} + I_{22} + I_{12} + I_{21} \right] dx_2 dx_1 \\
&\quad =: J_{11} + J_{22} + J_{12} + J_{21}.
\end{aligned}
$$

Clearly it yields that

$$
V_1(t_1)^{\frac{1}{2p'}} V_2(t_2)^{\frac{1}{2p'}} \left[J_{11} \right]^{\frac{1}{q}} \ll A_{M_2}.
$$

Further

$$
\begin{aligned}
J_{22} &= \int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) \left(\int_{t_1}^{x_1} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} dy_1 \right) \\
&\quad \times \left(\int_{t_2}^{x_2} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} dy_2 \right) dx_2 dx_1 \\
&= \int_{t_1}^{\infty} \int_{t_2}^{\infty} v_1(y_1)^{1-p'} V_1(y_1)^{\frac{q}{2p'}-1} v_2(y_2)^{1-p'} V_2(y_2)^{\frac{q}{2p'}-1} \\
&\quad \times \left(\int_{y_1}^{\infty} \int_{y_2}^{\infty} w(x_1, x_2) dx_2 dx_1 \right) dy_2 dy_1 \\
&\le A_{M_2}^q \int_{t_1}^{\infty} V_1(y_1)^{-\frac{q}{2p'}-1} v_1(y_1)^{1-p'} dy_1 \int_{t_2}^{\infty} V_2(y_2)^{-\frac{q}{2p'}-1} v_2(y_2)^{1-p'} dy_2 \\
&\ll A_{M_2}^q V_1(t_1)^{-\frac{q}{2p'}} V_2(t_2)^{-\frac{q}{2p'}}.
\end{aligned} \tag{10.43}
$$

Hence,

$$V_1(t_1)^{\frac{1}{2p'}} V_2(t_2)^{\frac{1}{2p'}} [J_{11}]^{\frac{1}{q}} \ll A_{M_2}.$$

The terms with J_{12} and J_{21} are estimated analogously. The method works for any $n > 2$ by induction and the proof is complete. □

Remark 19. The condition $A_{M_n} < \infty$ may be regarded as a natural end point of the conditions given in the n-dimensional version of Theorem 4 and also as a natural generalization of the usual Muckenhoupt-Bradley condition in one dimension.

The alternative criterion for the Hardy inequality (10.32) to hold with product type weight v satisfying (10.33) in the case $1 < p \le q < \infty$ is stated by the following

Theorem 20. *Let $1 < p \le q < \infty$ and the weight function v be of product type (10.33). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $A_{PS_n} < \infty$, where*

$$A_{PS_n} := \sup_{\substack{t_i>0 \\ i=1,\dots,n}} V_1(t_1)^{-\frac{1}{p}} \dots V_n(t_n)^{-\frac{1}{p}} \times \left(\int_0^{t_1} \dots \int_0^{t_n} w(\mathbf{x}) V_1(x_1)^q \dots V_n(x_n)^q d\mathbf{x} \right)^{\frac{1}{q}}. \tag{10.44}$$

Moreover, $C \approx A_{PS_n}$ with constants of equivalence depending only on the parameters p, q and n.

Proof. The necessary part follows from Lemma 14. The proof of the sufficiency can be obtained from Theorem 17 and the following Lemma 21. □

Lemma 21. *We have*

$$A_{M_n} \ll A_{PS_n}. \tag{10.45}$$

Proof. Let $n = 2$. Suppose first that $V_1(\infty) = V_2(\infty) = \infty$. Then

$$\int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) dx_2 dx_1 = \int_{t_2}^{\infty} \int_{t_1}^{\infty} w(x_1, x_2) V_1(x_1)^q V_1(x_1)^{-q} dx_1 dx_2$$

$$= q \int_{t_2}^{\infty} \int_{t_1}^{\infty} w(x_1, x_2) V_1(x_1)^q \left(\int_{x_1}^{\infty} V_1(y_1)^{-q-1} dV_1(y_1) \right) dx_1 dx_2$$

$$= q \int_{t_2}^{\infty} \int_{t_1}^{\infty} V_1(y_1)^{-q-1} \left(\int_{t_1}^{y_1} w(x_1, x_2) V_1(x_1)^q dx_1 \right) dV_1(y_1) dx_2$$

$$\le q \int_{t_1}^{\infty} V_1(y_1)^{-q-1} \int_{t_2}^{\infty} V_2(x_2)^q V_2(x_2)^{-q} \left(\int_0^{y_1} w(x_1, x_2) V_1(x_1)^q dx_1 \right) dx_2 dV_1(y_1)$$

$$
\begin{aligned}
&\le q^2 \int_{t_1}^{\infty} \int_{t_2}^{\infty} V_1(y_1)^{-q-1} V_2(y_2)^{-q-1} \\
&\times \left(\int_0^{y_2} \int_0^{y_1} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q dx_1 dx_2 \right) d V_2(y_2) d V_1(y_1) \\
&\le q^2 A_{PS_2}^q \int_{t_1}^{\infty} \int_{t_2}^{\infty} V_1(y_1)^{-\frac{q}{p'}-1} V_2(y_2)^{-\frac{q}{p'}-1} d V_2(y_2) d V_1(y_1) \\
&= (p')^2 A_{PS_2}^q V_1(t_1)^{-\frac{q}{p'}} V_2(t_2)^{-\frac{q}{p'}}.
\end{aligned}
$$

Thus, we get that

$$
V_1(t_1)^{\frac{1}{p'}} V_2(t_2)^{\frac{1}{p'}} \left(\int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) dx_2 dx_1 \right)^{\frac{1}{q}} \ll A_{PS_2}.
$$

Further, suppose that $V_i(\infty) < \infty$ for all $i = 1, 2$. Note that

$$
V_i(x_i)^{-q} = V_i(\infty)^{-q} + q \int_{x_i}^{\infty} V_i(y_i)^{-q-1} d V_i(y_i), \qquad i = 1, 2. \tag{10.46}
$$

Therefore,

$$
\begin{aligned}
&\int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) dx_2 dx_1 \\
&= \int_{t_2}^{\infty} \int_{t_1}^{\infty} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q V_1(x_1)^{-q} V_2(x_2)^{-q} dx_1 dx_2 \\
&= V_1(\infty)^{-q} V_2(\infty)^{-q} \int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q dx_2 dx_1 \\
&+ q V_1(\infty)^{-q} \int_{t_1}^{\infty} \int_{t_2}^{\infty} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q \left(\int_{x_2}^{\infty} V_2(y_2)^{-q-1} d V_2(y_2) \right) dx_2 dx_1 \\
&+ q V_2(\infty)^{-q} \int_{t_2}^{\infty} \int_{t_1}^{\infty} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q \left(\int_{x_1}^{\infty} V_1(y_1)^{-q-1} d V_1(y_1) \right) dx_1 dx_2 \\
&+ q^2 \int_{t_2}^{\infty} \int_{t_1}^{\infty} w(x_1, x_2) V_1(x_1)^q V_2(x_2)^q \\
&\times \left(\int_{x_1}^{\infty} \int_{x_2}^{\infty} V_1(y_1)^{-q-1} V_2(y_2)^{-q-1} d V_2(y_2) d V_1(y_1) \right) dx_1 dx_2 \\
&=: J_{11} + J_{12} + J_{21} + J_{22}.
\end{aligned}
$$

Obviously that

$$J_{11} \le V_1(\infty)^{-q} V_2(\infty)^{-q} \int_0^\infty \int_0^\infty w(x_1,x_2) V_1(x_1)^q V_2(x_2)^q dx_2 dx_1$$
$$\le A_{PS_2}^q V_1(\infty)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}}.$$

By changing the order of integration we have that

$$J_{12} \le qV_1(\infty)^{-q} \int_0^\infty \int_{t_2}^\infty w(x_1,x_2) V_1(x_1)^q V_2(x_2)^q$$
$$\times \left(\int_{x_2}^\infty V_2(y_2)^{-q-1} dV_2(y_2) \right) dx_2 dx_1 \le qV_1(\infty)^{-q} \int_0^\infty \int_{t_2}^\infty V_2(y_2)^{-q-1}$$
$$\times \left(\int_0^{y_2} w(x_1,x_2) V_1(x_1)^q V_2(x_2)^q dx_2 \right) dV_2(y_2) dx_1 = qV_1(\infty)^{-q} \int_{t_2}^\infty V_2(y_2)^{-q-1}$$
$$\times \left(\int_0^{y_2} \int_0^\infty w(x_1,x_2) V_1(x_1)^q V_2(x_2)^q dx_2 dx_1 \right) dV_2(y_2)$$
$$\le qA_{PS_2}^q V_1(\infty)^{-\frac{q}{p'}} \int_{t_2}^\infty V_2(y_2)^{-\frac{q}{p'}-1} dV_2(y_2)$$
$$= p' A_{PS_2}^q \left[V_1(\infty)^{-\frac{q}{p'}} V_2(t_2)^{-\frac{q}{p'}} - V_1(\infty)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}} \right].$$

Analogously,

$$J_{21} \le p' A_{PS_2}^q \left[V_1(t_1)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}} - V_1(\infty)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}} \right].$$

By changing the order of integration we have for J_{22} that

$$J_{22} \le q^2 \int_{t_1}^\infty \int_{t_2}^\infty \left(\int_0^{y_1} \int_0^{y_2} w(x_1,x_2) V_1(x_1)^q V_2(x_2)^q dx_2 dx_1 \right)$$
$$\times V_1(y_1)^{-q-1} V_2(y_2)^{-q-1} dV_2(y_2) dV_1(y_1)$$
$$\le (p')^2 A_{PS_2}^q \left[V_1(t_1)^{-\frac{q}{p'}} V_2(t_2)^{-\frac{q}{p'}} - V_1(\infty)^{-\frac{q}{p'}} V_2(t_2)^{-\frac{q}{p'}} \right.$$
$$\left. - V_1(t_1)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}} + V_1(\infty)^{-\frac{q}{p'}} V_2(\infty)^{-\frac{q}{p'}} \right].$$

Therefore, it follows that

$$V_1(t_1)^{\frac{1}{p'}} V_2(t_2)^{\frac{1}{p'}} \left[J_{11} + J_{12} + J_{21} + J_{22} \right]^{\frac{1}{q}} \le (p')^{\frac{2}{q}} A_{PS_2}.$$

Consider now one of the mixed cases when $V_1(\infty) = \infty$ and $V_2(\infty) < \infty$. Write

$$\begin{aligned}
&\int_{t_1}^{\infty}\int_{t_2}^{\infty} w(x_1,x_2)dx_2dx_1 \\
&= \int_{t_2}^{\infty}\int_{t_1}^{\infty} w(x_1,x_2)V_1(x_1)^q V_1(x_1)^{-q}dx_1dx_2 \\
&= q\int_{t_2}^{\infty}\int_{t_1}^{\infty} w(x_1,x_2)V_1(x_1)^q\left(\int_{x_1}^{\infty} V_1(y_1)^{-q-1}dV_1(y_1)\right)dx_1dx_2 \\
&\le q\int_{t_2}^{\infty}\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\left(\int_0^{y_1} w(x_1,x_2)V_1(x_1)^q dx_1\right)dV_1(y_1)dx_2.
\end{aligned}$$

Further, by using (10.46) with $i = 2$ we get that

$$\begin{aligned}
&\int_{t_1}^{\infty}\int_{t_2}^{\infty} w(x_1,x_2)dx_2dx_1 \\
&\le q\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\int_{t_2}^{\infty}\left(\int_0^{y_1} w(x_1,x_2)V_1(x_1)^q dx_1\right)V_2(x_2)^q V_2(x_2)^{-q}dx_2dV_1(y_1) \\
&= qV_2(\infty)^{-q}\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\int_{t_2}^{\infty}\left(\int_0^{y_1} w(x_1,x_2)V_1(x_1)^q dx_1\right)V_2(x_2)^q dx_2dV_1(y_1) \\
&\quad + q^2\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\int_{t_2}^{\infty}\left(\int_0^{y_1} w(x_1,x_2)V_1(x_1)^q dx_1\right)V_2(x_2)^q \\
&\quad \times\left(\int_{x_2}^{\infty} V_2(y_2)^{-q-1}dV_2(y_2)\right)dx_2dV_1(y_1) \\
&\le qV_2(\infty)^{-q}\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\left(\int_0^{\infty}\int_0^{y_1} w(x_1,x_2)V_1(x_1)^q V_2(x_2)^q dx_1dx_2\right)dV_1(y_1) \\
&\quad + q^2\int_{t_1}^{\infty} V_1(y_1)^{-q-1}\int_{t_2}^{\infty} V_2(y_2)^{-q-1} \\
&\quad \times\left(\int_0^{y_1}\int_0^{y_2} w(x_1,x_2)V_1(x_1)^q V_2(x_2)^q dx_2dx_1\right)dV_2(y_2)dV_1(y_1) \\
&\le (p')^2 A_{PS_2}^q V_1(t_1)^{-\frac{q}{p'}}V_2(t_2)^{-\frac{q}{p'}}.
\end{aligned}$$

Hence, it yields that

$$V_1(t_1)^{\frac{1}{p'}}V_2(t_2)^{\frac{1}{p'}}\left(\int_{t_1}^{\infty}\int_{t_2}^{\infty} w(x_1,x_2)dx_2dx_1\right)^{\frac{1}{q}} \le (p')^{\frac{2}{q}}A_{PS_2}.$$

The case $V_1(\infty) < \infty$, $V_2(\infty) = \infty$ can be proved analogously. The proof for $n = 2$ is complete. For any $n > 2$ the statement of Lemma follows by induction and the proof is complete. □

Further we discuss the inequality (10.32) with the left hand side weight function of product type. In particular, in Theorems 25 and 26 we state a Muckenhoupt-type and Persson-Stepanov-type criteria for the inequality (10.32) to hold in the case $1 < p \le q < \infty$ with the left hand side weight w to be of product type (10.34). The proofs of these results are analogous to the proofs of Theorems 17, 20 and based on some statements formulated below. The first of them is dual to n-dimensional extension of Theorem 4 and reads:

Theorem 22. *Let $1 < q' \le p' < \infty$, $s_i \in (1, q')$, $i = 1, \dots, n$, and the weight function w be of product type (10.34) Then the inequality (10.39) holds for all measurable functions g if and only if $A^*_{W_n} < \infty$, where*

$$A^*_{W_n} := A^*_{W_n}(s_1, \dots, s_n) := \sup_{\substack{t_i > 0 \\ i=1,\dots,n}} W_1(t_1)^{\frac{s_1-1}{q'}} \dots W_n(t_n)^{\frac{s_n-1}{q'}}$$

$$\times \left(\int_0^{t_1} \dots \int_0^{t_n} v(\mathbf{x})^{1-p'} W_1(x_1)^{\frac{p'(q'-s_1)}{q'}} \dots W_n(x_n)^{\frac{p'(q'-s_n)}{q'}} d\mathbf{x} \right)^{\frac{1}{p'}}$$

and

$$W(t_i) := \int_{t_i}^{\infty} w_i(x_i) dx_i, \qquad i = 1, \dots, n.$$

Moreover, $C \approx A^*_{W_n}$ with constants of equivalence depending only on the parameters p, q and n.

The following two auxiliary statements are similar to Lemmas 18 and 21, respectively.

Lemma 23. *Let*

$$A^*_{M_n} := \sup_{\substack{t_i > 0 \\ i=1,\dots,n}} W_1(t_1)^{\frac{1}{q}} \dots W_n(t_n)^{\frac{1}{q}} V(t_1, \dots, t_n)^{\frac{1}{p'}}.$$

Then

$$A^*_{W_n} \ll A^*_{M_n}. \tag{10.47}$$

Lemma 24. *Let*

$$A^*_{PS_n} := \sup_{\substack{t_i > 0 \\ i=1,\dots,n}} W_1(t_1)^{-\frac{1}{q'}} \dots W_n(t_n)^{-\frac{1}{q'}}$$

$$\times \left(\int_{t_1}^{\infty} \dots \int_{t_n}^{\infty} v(\mathbf{x})^{1-p'} W_1(x_1)^{p'} \dots W_n(x_n)^{p'} d\mathbf{x} \right)^{\frac{1}{p'}}.$$

Then

$$A^*_{M_n} \ll A^*_{PS_n}. \tag{10.48}$$

Now by passing to the dual inequality (10.39) of (10.32) we can get a Muckenhoupt-type and Persson-Stepanov-type criteria for (10.32) with the left hand side weight w of product type (10.34). The necessity in the proofs of these results follow from Lemmas 13 and 15, while the sufficient parts can be proved in the similar ways as in Theorems 17 and 20 but by using Theorem 22, Lemmas 23 and 24 instead of Theorem 17, Lemmas 18 and 21, respectively.

Theorem 25. *Let $1 < p \le q < \infty$ and the weight function w be of product type (10.34). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $A^*_{M_n} < \infty$. Moreover, $C \approx A^*_{M_n}$ with constants of equivalence depending only on the parameters p, q and n.*

Theorem 26. *Let $1 < p \le q < \infty$ and the weight function w be of product type(10.34). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $A^*_{PS_n} < \infty$. Moreover, $C \approx A^*_{PS_n}$ with constants of equivalence depending only on the parameters p, q and n.*

10.5 The Multi-dimensional Case $1 < q < p < \infty$

In this Section we will prove the similar results as in the previous Section but in the case $1 < q < p < \infty$. Let us introduce the following n-dimensional versions of the Mazya-Rozin and Persson-Stepanov conditions in this case:

$$B_{MR_n} := \left(\int_{\mathbb{R}^n_+} W(\mathbf{t})^{\frac{r}{q}} V_1(t_1)^{\frac{r}{q'}} \dots V_n(t_n)^{\frac{r}{q'}} dV_1(t_1) \dots dV_n(t_n) \right)^{\frac{1}{r}},$$

$$B_{PS_n} := \left(\int_{\mathbb{R}^n_+} \left(\int_0^{t_1} \dots \int_0^{t_n} w(\mathbf{x}) V_1(x_1)^q \dots V_n(x_n)^q d\mathbf{x} \right)^{\frac{r}{q}} \right.$$
$$\left. \times V_1(t_1)^{-\frac{r}{q}} \dots V_n(t_n)^{-\frac{r}{q}} dV_1(t_1) \dots dV_n(t_n) \right)^{\frac{1}{r}}.$$

The following comparison between these constants is useful later on but also of independent interest.

Lemma 27. *We have*

$$B_{PS_n} \ll B_{MR_n}. \tag{10.49}$$

Proof. It yields that

$$\begin{aligned}
&\int_0^{t_1}\dots\int_0^{t_n} w(\mathbf{x})V_1(x_1)^q\dots V_n(x_n)^q\,d\mathbf{x}\\
&\quad= q^n\int_0^{t_1}\dots\int_0^{t_n} w(\mathbf{x})\\
&\quad\times\left(\int_0^{x_1}\dots\int_0^{x_n} V_1(y_1)^{q-1}\dots V_n(y_n)^{q-1}dV_n(y_n)\dots dV_1(y_1)\right)d\mathbf{x}\\
&\quad\le q^n\int_0^{t_1}\dots\int_0^{t_n} W(\mathbf{y})V_1(y_1)^{q-1}\dots V_n(y_n)^{q-1}dV_n(y_n)\dots dV_1(y_1)
\end{aligned}$$

[applying Hölder's inequality with the exponents r/q and p/q]

$$\begin{aligned}
&= q^n\int_0^{t_1}\dots\int_0^{t_n}\left\{W(\mathbf{y})V_1(y_1)^{(q-1)+\frac{q}{2p}}\dots V_n(y_n)^{(q-1)+\frac{q}{2p}}\right\}\\
&\quad\times V_1(y_1)^{-\frac{q}{2p}}\dots V_n(y_n)^{-\frac{q}{2p}}dV_n(y_n)\dots dV_1(y_1)\\
&\quad\le q^n\left(\int_0^{t_1}\dots\int_0^{t_n} W(\mathbf{y})^{\frac{r}{q}}V_1(y_1)^{(q-1+\frac{q}{2p})\frac{r}{q}}\dots V_n(y_n)^{(q-1+\frac{q}{2p})\frac{r}{q}}dV_n(y_n)\dots\right.\\
&\quad\left.\times\,V_1(y_1))^{\frac{q}{r}}\left(\int_0^{t_1}\dots\int_0^{t_n} V_1(y_1)^{-\frac{1}{2}}\dots V_n(y_n)^{-\frac{1}{2}}dV_n(y_n)\dots V_1(y_1)\right)^{\frac{q}{p}}\right.\\
&= q^n 2^{\frac{qn}{p}}\left(\int_0^{t_1}\dots\int_0^{t_n} W(\mathbf{y})^{\frac{r}{q}}V_1(y_1)^{\frac{r}{q'}+\frac{r}{2p}}\dots\right.\\
&\quad\left.\times\,V_n(y_n)^{\frac{r}{q'}+\frac{r}{2p}}dV_n(y_n)\dots dV_1(y_1)\right)^{\frac{q}{r}}V_1(t_1)^{\frac{q}{2p}}\dots V_n(t_n)^{\frac{q}{2p}}.
\end{aligned}$$

Hence, we obtain that

$$\begin{aligned}
B_{PS_n}^r \le q^{\frac{rn}{q}}2^{\frac{rn}{p}}\int_{\mathbb{R}_+^n}&\left(\int_0^{t_1}\dots\int_0^{t_n} W(\mathbf{y})^{\frac{r}{q}}V_1(y_1)^{\frac{r}{q'}+\frac{r}{2p}}\dots\right.\\
&\left.\times\,V_n(y_n)^{\frac{r}{q'}+\frac{r}{2p}}dV_n(y_n)\dots dV_1(y_1)\right)\\
&\times V_1(t_1)^{\frac{r}{2p}-\frac{r}{q}}\dots V_n(t_n)^{\frac{r}{2p}-\frac{r}{q}}dV_1(t_1)\dots dV_n(t_n).
\end{aligned}$$

Therefore, by changing the order of integration, we get that

$$B_{PS_n}^r \le q^{\frac{rn}{q}} 2^{\frac{rn}{p}} \int_{\mathbb{R}_+^n} W(\mathbf{y})^{\frac{r}{q}} V_1(y_1)^{\frac{r}{q'}+\frac{r}{2p}} \dots V_n(y_n)^{\frac{r}{q'}+\frac{r}{2p}}$$
$$\times \left(\int_{y_1}^{\infty} \dots \int_{y_n}^{\infty} V_1(t_1)^{\frac{r}{2p}-\frac{r}{q}} \dots V_n(t_n)^{\frac{r}{2p}-\frac{r}{q}} dV_n(t_n) \dots dV_1(t_1) \right)$$
$$\times dV_1(y_1) \dots dV_n(y_n) \le q^{\frac{rn}{q}} 2^{\frac{rn}{p}} \left(\frac{2p}{r} \right)^n B_{MR_n}^r$$

and the required estimate (10.49) is proved. □

Next we will state a similar comparison between the following dual versions of the constants B_{MR_n} and B_{PS_n}:

$$B_{MR_n}^* := \left(\int_{\mathbb{R}_+^n} V(\mathbf{t})^{\frac{r}{p'}} W_1(t_1)^{\frac{r}{p}} \dots W_n(t_n)^{\frac{r}{p}} d\,[-W_1(t_1)] \dots d\,[-W_n(t_n)] \right)^{\frac{1}{r}},$$

$$B_{PS_n}^* := \left(\int_{\mathbb{R}_+^n} \left(\int_{t_1}^{\infty} \dots \int_{t_n}^{\infty} v(\mathbf{x})^{1-p'} W_1(x_1)^{p'} \dots W_n(x_n)^{p'} d\mathbf{x} \right)^{\frac{r}{p'}} \right.$$
$$\left. \times W_1(t_1)^{-\frac{r}{p'}} \dots W_n(t_n)^{-\frac{r}{p'}} d\,[-W_1(t_1)] \dots d\,[-W_n(t_n)] \right)^{\frac{1}{r}}.$$

Lemma 28. *It yields that*

$$B_{PS_n}^* \ll B_{MR_n}^*. \tag{10.50}$$

Proof. The proof is similar to that of Lemma 27 so we omit the details. □

The following Theorems state necessary and sufficient conditions for the validity of (10.32) in the case $1 < q < p < \infty$ with weights satisfying some of the following additional conditions:

$$V_1(\infty) = \dots = V_n(\infty) = \infty \tag{10.51}$$

or

$$W_1(0) = \dots = W_n(0) = \infty \tag{10.52}$$

Theorem 29. *Let $1 < q < p < \infty$ and $1/r = 1/q - 1/p$. Suppose that the weight function v satisfies the conditions (10.33) and (10.51). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}_+^n$ with some finite constant C, which is independent on f, if and only if $B_{MR_n} < \infty$. Moreover, $C \approx B_{MR_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.*

Proof. Necessity. Suppose that the inequality (10.32) holds with $C < \infty$ and put

$$f(\mathbf{y}) = W(\mathbf{y})^{\frac{r}{pq}} V_1(y_1)^{\frac{r}{pq'}} v_1(y_1)^{1-p'} \dots V_n(y_n)^{\frac{r}{pq'}} v_n(y_n)^{1-p'}.$$

It is easy to see that $\left(\int_{\mathbb{R}^n_+} f^p(\mathbf{x})v(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{p}} = B_{MR_n}^{\frac{r}{p}}$. On the left hand side we have

$$\left(\int_{\mathbb{R}^n_+} (H_n f)^q(\mathbf{x})w(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{q}}$$

$$= \left(\int_{\mathbb{R}^n_+} \left(\int_0^{x_1}\dots\int_0^{x_n} f(\mathbf{t})d\mathbf{t}\right)\left(\int_0^{x_1}\dots\int_0^{x_n} f(\mathbf{y})d\mathbf{y}\right)^{q-1} w(\mathbf{x})d\mathbf{x}\right)^{\frac{1}{q}}$$

$$= \left(\int_{\mathbb{R}^n_+} f(\mathbf{t})\left(\int_{t_1}^{\infty}\dots\int_{t_n}^{\infty}\left(\int_0^{x_1}\dots\int_0^{x_n} f(\mathbf{y})d\mathbf{y}\right)^{q-1} w(\mathbf{x})d\mathbf{x}\right)d\mathbf{t}\right)^{\frac{1}{q}}$$

$$= \left(\int_{\mathbb{R}^n_+} W(\mathbf{t})^{\frac{r}{pq}} V_1(t_1)^{\frac{r}{pq'}}\dots V_n(t_n)^{\frac{r}{pq'}}\left(\int_{t_1}^{\infty}\dots\int_{t_n}^{\infty}\left(\int_0^{x_1}\dots\int_0^{x_n} W(\mathbf{y})^{\frac{r}{pq}}\right.\right.\right.$$

$$\left.\left.\left.\times V_1(y_1)^{\frac{r}{pq'}}\dots V_n(y_n)^{\frac{r}{pq'}} dV_n(y_n)\dots dV_1(y_1)\right)^{q-1} w(\mathbf{x})d\mathbf{x}\right) dV_1(t_1)\dots dV_n(t_n)\right)^{\frac{1}{q}}$$

$$\geq \left(\int_{\mathbb{R}^n_+} W(\mathbf{t})^{\frac{r}{pq}+1} V_1(t_1)^{\frac{r}{pq'}}\dots V_n(t_n)^{\frac{r}{pq'}}\left(\int_0^{t_1}\dots\int_0^{t_n} W(\mathbf{y})^{\frac{r}{pq}}\right.\right.$$

$$\left.\left.\times\ V_1(y_1)^{\frac{r}{pq'}}\dots V_n(y_n)^{\frac{r}{pq'}} dV_n(y_n)\dots dV_1(y_1)\right)^{q-1} dV_1(t_1)\dots dV_n(t_n)\right)^{\frac{1}{q}}$$

[since the function W is non-increasing and $r/pq' + 1 = r/p'q$]

$$\geq \left(\int_{\mathbb{R}^n_+}\left(\int_0^{t_1}\dots\int_0^{t_n} V_1(y_1)^{\frac{r}{pq'}}\dots V_n(y_n)^{\frac{r}{pq'}} dV_n(y_n)\dots dV_1(y_1)\right)^{q-1}\right.$$

$$\left.\times\ W(\mathbf{t})^{\frac{r}{q}} V_1(t_1)^{\frac{r}{pq'}}\dots V_n(t_n)^{\frac{r}{pq'}} dV_1(t_1)\dots dV_n(t_n)\right)^{\frac{1}{q}} = \left(\frac{p'q}{r}\right)^{\frac{n}{q'}} B_{MR_n}^{\frac{r}{q}}$$

and the estimate $B_{MR_n}^{\frac{r}{q}} \ll CB_{MR_n}^{\frac{r}{p}}$ follows. Therefore, $B_{MR_n} \ll C < \infty$.

Sufficiency. Suppose that $B_{MR_n} < \infty$. On the strength of (10.51) we find that

$$\int_{\mathbb{R}^n_+} (H_n f)^q(\mathbf{x})w(\mathbf{x})d\mathbf{x}$$

$$= \int_{\mathbb{R}^n_+} (H_n f)^q(\mathbf{x})V_1(x_1)^q V_1(x_1)^{-q}\dots V_n(x_n)^q V_n(x_n)^{-q} w(\mathbf{x})d\mathbf{x}$$

$$= q^n \int_{\mathbb{R}^n_+} (H_n f)^q (\mathbf{x}) V_1(x_1)^q \dots V_n(x_n)^q$$

$$\times \left(\int_{x_1}^{\infty} \dots \int_{x_n}^{\infty} V_1(y_1)^{-q-1} \dots V_n(y_n)^{-q-1} dV_n(y_n) \dots dV_1(y_1) \right) w(\mathbf{x}) d\mathbf{x}$$

$$\le q^n \int_{\mathbb{R}^n_+} (H_n f)^q (\mathbf{y}) V_1(y_1)^{-q-1} \dots V_n(y_n)^{-q-1}$$

$$\times \left(\int_0^{y_1} \dots \int_0^{y_n} w(\mathbf{x}) V_1(x_1)^q \dots V_n(x_n)^q d\mathbf{x} \right) dV_n(y_n) \dots dV_1(y_1)$$

$$= q^n \int_{\mathbb{R}^n_+} \{(H_n f)^q (\mathbf{y}) V_1(y_1)^{-q} \dots V_n(y_n)^{-q}\} \{V_1(y_1)^{-1} \dots V_n(y_n)^{-1}$$

$$\times \left(\int_0^{y_1} \dots \int_0^{y_n} w(\mathbf{x}) V_1(x_1)^q \dots V_n(x_n)^q d\mathbf{x} \right) \Big\} dV_n(y_n) \dots dV_1(y_1)$$

[by using Hölder's inequality with exponents p/q and r/q]

$$\ll B_{PS_n}^q \left(\int_{\mathbb{R}^n_+} (H_n f)^p (\mathbf{y}) V_1(y_1)^{-p} \dots V_n(y_n)^{-p} dV_n(y_n) \dots dV_1(y_1) \right)^{\frac{q}{p}}.$$

Moreover, according to Theorem 20,

$$\left(\int_{\mathbb{R}^n_+} (H_n f)^p (\mathbf{y}) V_1(y_1)^{-p} \dots V_n(y_n)^{-p} dV_n(y_n) \dots dV_1(y_1) \right)^{\frac{q}{p}}$$

$$\ll \left(\int_{\mathbb{R}^n_+} f^p(\mathbf{x}) v_1(x_1) \dots v_n(x_n) d\mathbf{x} \right)^{\frac{q}{p}}.$$

By combining these inequalities we have that

$$\int_{\mathbb{R}^n_+} (H_n f)^q (\mathbf{x}) w(\mathbf{x}) d\mathbf{x} \ll B_{PS_n}^q \left(\int_{\mathbb{R}^n_+} f^p(\mathbf{x}) v_1(x_1) \dots v_n(x_n) d\mathbf{x} \right)^{\frac{q}{p}}. \quad (10.53)$$

Therefore, in view of Lemma 27, the inequality (10.32) holds and the proof is complete. □

The corresponding result with the constant B_{PS_n} involved reads:

Theorem 30. *Let $1 < q < p < \infty$ and $1/r = 1/q - 1/p$. Suppose that the weight function v satisfies the conditions (10.33) and (10.51). Then the inequality (10.32) holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant $\mathbf{C}$, which is*

independent on f, if and only if $B_{PS_n} < \infty$. Moreover, $C \approx B_{PS_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.

Proof. The necessity follows from Lemma 27 and Theorem 29. The sufficiency is proved by (10.53). □

Remark 31. Note that the sufficient parts of Theorems 29 and 30 in fact hold for all $0 < q < p < \infty$. Moreover, the necessary parts of these Theorems are correct even without assuming that the condition (10.51) is satisfied.

By passing to the dual inequality (10.39) of (10.32) we can in a similar way as above (but now using Lemma 28 instead of Lemma 27) get the following results for the case $1 < q < p < \infty$ with the left hand side weight w of product type (10.34).

Theorem 32. *Let $1 < q < p < \infty$ and $1/r = 1/q - 1/p$. Assume that the weight function w satisfies the conditions* (10.34) *and* (10.52)*. Then the inequality* (10.32) *holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $B^*_{MR_n} < \infty$. Moreover, $C \approx B^*_{MR_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.*

Theorem 33. *Let $1 < q < p < \infty$ and $1/r = 1/q - 1/p$. Suppose that the weight function w satisfies the conditions* (10.34) *and* (10.52)*. Then the inequality* (10.32) *holds for all measurable functions f on $\mathbb{R}^n_+$ with some finite constant C, which is independent on f, if and only if $B^*_{PS_n} < \infty$. Moreover, $C \approx B^*_{PS_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.*

10.6 Multi-dimensional Limit Pólya–Knopp Type Inequalities

In this Section we will apply the results of Theorems 20 and 30. Namely, we will characterize the inequality

$$\left(\int_{\mathbb{R}^n_+} (G_n f)^q(\mathbf{x}) w(\mathbf{x}) d\mathbf{x}\right)^{\frac{1}{q}} \leq C \left(\int_{\mathbb{R}^n_+} f^p(\mathbf{y}) v(\mathbf{y}) d\mathbf{y}\right)^{\frac{1}{p}} \tag{10.54}$$

in the case $0 < p \leq q < \infty$ and give a sufficient condition for (10.54) to hold in the case $0 < q < p < \infty$. Here G_n denotes the n-dimensional geometric mean operator.

According to Jensen's inequality it holds for any $\mathbf{x} \in \mathbb{R}^n_+$ that

$$(G_n f)(\mathbf{x}) \leq \frac{1}{x_1 \dots x_n} (H_n f)(\mathbf{x}), \tag{10.55}$$

where H_n is the usual Hardy operator. This fact allows us to find a upper estimate for the best constant of (10.54) via the inequality (10.32) for the Hardy operator H_n,

which was considered in the previous section and with a product type weight on one side. It is useful to rewrite (10.54) in the following way

$$\left(\int_{\mathbb{R}^n_+} (G_n g)^q(\mathbf{x}) u(\mathbf{x}) d\mathbf{x}\right)^{\frac{1}{q}} \le C \left(\int_{\mathbb{R}^n_+} g^p(\mathbf{x}) d\mathbf{x}\right)^{\frac{1}{p}} \tag{10.56}$$

with $g(\mathbf{x}) = f(\mathbf{x}) v(\mathbf{x})^{1/p}$ and

$$u(\mathbf{x}) := (G_n v)(\mathbf{x})^{-\frac{q}{p}} w(\mathbf{x}). \tag{10.57}$$

Further, for any $0 < s < q$ we put $\tilde{p} := p/s$, $\tilde{q} := q/s$ and after a new substitution $g(\mathbf{x}) = h(\mathbf{x})^{1/s}$ the inequality (10.56) gets the form

$$\left(\int_{\mathbb{R}^n_+} (G_n h)^{\tilde{q}}(\mathbf{x}) u(\mathbf{x}) d\mathbf{x}\right)^{1/\tilde{q}} \le \widetilde{C} \left(\int_{\mathbb{R}^n_+} h^{\tilde{p}}(\mathbf{x}) d\mathbf{x}\right)^{1/\tilde{p}}, \tag{10.58}$$

where $\widetilde{C} = C^s$. Therefore, in view of (10.55) we have that the inequality corresponding to (10.58) for the operator

$$(\widetilde{H}_n h)(\mathbf{x}) := \frac{1}{x_1 \dots x_n} (H_n h)(\mathbf{x}) \tag{10.59}$$

has the form

$$\left(\int_{\mathbb{R}^n_+} (\widetilde{H}_n h)^{\tilde{q}}(\mathbf{x}) u(\mathbf{x}) d\mathbf{x}\right)^{1/\tilde{q}} \le \bar{C} \left(\int_{\mathbb{R}^n_+} h^{\tilde{p}}(\mathbf{x}) d\mathbf{x}\right)^{1/\tilde{p}}. \tag{10.60}$$

This is an inequality for the Hardy operator H_n with $1 < \tilde{p}, \tilde{q} < \infty$, $w(\mathbf{x}) = (x_1 \dots x_n)^{-\tilde{q}} u(\mathbf{x})$ and with the product weight $v(\mathbf{x}) \equiv 1$. Now we are ready to state and prove our results for the inequality (10.54). Our main result for the case $0 < p \le q < \infty$ reads:

Theorem 34. *Let $0 < p \le q < \infty$. Then the inequality (10.54) holds for all positive measurable functions f on $\mathbb{R}^n_+$ if and only if $A_{G_n} < \infty$, where*

$$A_{G_n} := \sup_{\substack{t_i > 0 \\ i=1,\dots,n}} t_1^{-1/p} \dots t_n^{-1/p} \left(\int_0^{t_1} \dots \int_0^{t_n} u(\mathbf{x}) d\mathbf{x}\right)^{1/q} \tag{10.61}$$

with $u(x)$ defined by (10.57). Moreover, $C \approx A_{G_n}$ with constants of equivalence depending only on the parameters p, q and the dimension n.

Proof. Sufficiency. On the strength of (10.55) and Theorem 20 for $1 < \tilde{p} \le \tilde{q} < \infty$ the inequality (10.60) holds if

$$\bar{A}_{G_n} := \sup_{\substack{t_i>0\\ i=1,\dots,n}} t_1^{-1/\tilde{p}} \dots t_n^{-1/\tilde{p}} \left(\int_0^{t_1} \dots \int_0^{t_n} u(\mathbf{x}) d\mathbf{x} \right)^{1/\tilde{q}} < \infty.$$

Note that because of definitions of $\tilde{p}$ and $\tilde{q}$ it yields that

$$\left(\bar{A}_{G_n}\right)^{\frac{1}{s}} = A_{G_n}.$$

Therefore, according to the fact that $C = \widetilde{C}^{1/s}$ it follows that $A_{G_n} < \infty$ is a sufficient condition for the validity of the inequality (10.54) in the case $0 < p \le q < \infty$.

Necessity. Suppose that (10.54) and, thus, (10.56) holds with $C < \infty$. Take a test function

$$g_{\mathbf{t}}(\mathbf{y}) = \chi_{[0,t_1]}(y_1) t_1^{-\frac{1}{p}} \dots \chi_{[0,t_n]}(y_n) t_n^{-\frac{1}{p}}$$

and put it into the inequality (10.56). The function $g_{\mathbf{t}}(\mathbf{y})$ is such that the right hand side of (10.56) is equal to 1. Therefore,

$$C \ge \left(\int_{\mathbb{R}^n_+} (G_n g_{\mathbf{t}})^q (\mathbf{x}) u(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{q}} \ge t_1^{-\frac{1}{p}} \dots t_n^{-\frac{1}{p}} \left(\int_0^{t_1} \dots \int_0^{t_n} u(\mathbf{x}) d\mathbf{x} \right)^{\frac{1}{q}}.$$

Hence, by taking supremum over all t_i, $i = 1, \dots, n$, we have that $A_{G_n} < \infty$ and the proof is complete. □

Remark 35. Our proof above shows that Theorem 34 may be regarded as a limit case of the result in Theorem 20.

Remark 36. Note that for the case $n = 2$ we have here obtained another characterization of (10.54) Than that in Theorem 5. They are both endpoint characterizations of the corresponding scales of Hardy type inequalities with Wedestig and Persson-Stepanov type descriptions, respectively.

Moreover, the inequality (10.55) and Theorem 30 allow us to obtain a sufficient condition for (10.54) to hold in the case $0 < q < p < \infty$. We state this result in the following form:

Theorem 2. *Let $0 < q < p < \infty$. Then the inequality (10.54) holds if $B_{G_n} < \infty$, where*

$$B_{G_n} := \left(\int_{\mathbb{R}^n_+} \left(\int_0^{t_1} \dots \int_0^{t_n} u(\mathbf{x}) d\mathbf{x} \right)^{\frac{r}{q}} t_1^{-\frac{r}{q}} \dots t_n^{-\frac{r}{q}} dt_1 \dots dt_n \right)^{\frac{1}{r}}.$$

Proof. The statement follows from Theorem 30 by using the same arguments as for the proof of a sufficiency part of Theorem 34. □

Remark 3. Note that the condition $B_{G_n} < \infty$ is also necessary for (10.54) to hold in the case $0 < q < p < \infty$ with the additional assumption that the weight function u is of product type. In this case we also have that $C \approx B_{G_n}$, where C is the best constant in (10.54).

10.7 Further Results and Remarks

Theorem 4 and also the limiting results in Sect. 10.3 are based on the Ph.D. thesis [34] by A. Wedestig (see also [35]). Moreover, Sects. 10.4, 10.5 and 10.6 are based on the Ph.D. thesis by Elena Ushakova [33] (see also [29]). Also the Ph.D. theses by S. Barza [1] and M. Johansson [12] have influenced our results and ideas in this paper. In particular, [1] together with the paper [28] were important. Finally, for another type of multidimensional Hardy type inequalities we refer to the review article [5] and for some recent results concerning Hardy type inequalities involving both general measures and scales of conditions we refer to [23] and [24] and the references given in these papers. For some further results we also refer to the new book [15].

In the Ph.D. thesis [12] by M. Johansson and the paper [13] the following result was presented, which actually unifies the result of E. Sawyer [30] (the one dimensional case) and G. Sinnamon [32]:

Theorem 4. *Let* $1 < p \leq q < \infty$ *and let* $u(x)$, $v(x)$ *and* $\varphi(x)$ *be weight functions on* $(0, \infty)$, *where* $\varphi(x)$ *is decreasing. Then the inequality*

$$\left(\int_0^\infty \left(\int_0^x f(t)\, \varphi(t)\, v(t)\, dt \right)^q u(x)\, dx \right)^{\frac{1}{q}} \leq C \left(\int_0^\infty f^p(x)\, v(x)\, dx \right)^{\frac{1}{p}} \tag{10.62}$$

holds for all $C < \infty$ *and decreasing* f *if and only if one of the following conditions holds for some* $s > 0$*:*

$$D_\varphi(s) := \sup_{t>0} \left(\int_t^\infty u(x) \left(\int_0^x v(y)\, \varphi^{p'}(y)\, dy \right)^{q\left(\frac{1}{p'}-s\right)} dx \right)^{\frac{1}{q}} \times \tag{10.63}$$

$$\left(\int_0^t v(x)\, \varphi^{p'}(x)\, dx \right)^s < \infty.$$

$$D_{\varphi}^{*}(s) := \sup_{t>0}\left(\int_0^t v(x)\,\varphi^{p'}(x)\left(\int_x^{\infty} u(y)\,dy\right)^{p'\left(\frac{1}{q}-s\right)} dx\right)^{\frac{1}{p'}} \times \tag{10.64}$$

$$\left(\int_t^{\infty} u(x)\,dx\right)^{s} < \infty.$$

$$E_{\varphi}(s) := \sup_{t>0}\left(\int_0^t u(x)\left(\int_0^x v(y)\,\varphi^{p'}(y)\,dy\right)^{q\left(\frac{1}{p'}+s\right)} dx\right)^{\frac{1}{q}} \times \tag{10.65}$$

$$\left(\int_0^t v(x)\,\varphi^{p'}(x)\,dx\right)^{-s} < \infty.$$

$$E_{\varphi}^{*}(s) := \sup_{t>0}\left(\int_t^{\infty} v(x)\,\varphi^{p'}(x)\left(\int_x^{\infty} u(y)\,dy\right)^{p'\left(\frac{1}{q}+s\right)} dx\right)^{\frac{1}{p'}} \times \tag{10.66}$$

$$\left(\int_t^{\infty} u(x)\,dx\right)^{-s} < \infty.$$

Remark 5. By applying Theorem 4 with $\varphi \equiv 1$ we obtain the before mentioned result by G. Sinnamon. Moreover, by applying Theorem 4 with $\varphi(t) = \frac{1}{v(t)}$ and $u(x)$ replaced by $u(x)\,x^{-q}$ we obtain an alternative to the Sawyer result for the case when $v(t)$ is increasing.

Remark 6. According to E. Sawyer [30, Theorem 1] two conditions are necessary to characterize the one dimensional Hardy inequality (10.62) for decreasing functions in the general case but, in view of Theorem 4, for the special case when $v(x)$ is increasing only one condition is required (but there are infinite many such equivalent conditions). Moreover, in this case each of the conditions above are also equivalent.

Finally, we raise the following open questions connected to this paper:

Open Question 1: Is it possible to prove a natural extension of Theorem 4 to two and more dimensions?

Open Question 2: Characterize the two variables weights v and w ensuring that (10.1) holds for any of the following three remaining cases:

(a) $1 \le q < p < \infty$;
(b) $1 = p \le q < \infty$;
(c) $0 < q \le 1 \le p < \infty$.

Remark 7. The special case with product weights are of interest also in these cases.

Remark 8. The corresponding questions as those above and in Theorem 1 are of interest and still open also for any dimension $N = 3, 4, ...$

Remark 9. For the one-dimensional case Hardy´s inequality for $1 < p \leq q < \infty$ can be characterized by many different conditions (see e.g. Theorem 2), even some scales of conditions conditions (see [8]). Moreover, as seen in this paper for the case when ONE of the weights in (10.1) is of product type then (10.1) can be characterized by just one condition but this condition is not unique.. These considerations lead to the following:

Open Question 3: For the case $1 < p \leq q < \infty$ and when ONE of the weights v and w in (10.1) is of product type: Does there exist some scales of conditions (of the type as those in [2] for the one dimensional case) which, in particular, implies all corresponding two-dimensional results in this paper.

Remark 10. This question is of interest also for the cases (a)–(c) considered in Open Question 2 (concerning case (a) in the one dimensional case see [7]).

Remark 11. Some new Hardy type inequalities derived by mainly using convexity arguments can be found in the papers [4, 6, 11, 14, 20, 21] and [22].

References

1. Barza, S.: Weighted multidimensional integral inequalities and applications, Ph.D. Thesis, Department of Mathematics, Luleå University of Technology (1999)
2. Barza, S., Persson, L.-E., Soria, J.: Multidimensional rearrangements and Lorentz spaces. Acta Math. Hungar. **104**(3), 203–224 (2004)
3. Bloom, S., Kerman, R.: Weighted norm inequalities for operators of Hardy type. Proc. Am. Math. Soc. **113**(1), 135–141 (1991)
4. Cizmesija, A., Oguntuase, J., Persson, L.E.: Multidimensional Hardy-type inequalities via convexity. Bull. Austral. Math. Soc. **77**, 245–260 (2008)
5. Cizmesija, A., Persson, L.E., Wedestig, A.: Weighted integral inequalities for Hardy and geometric mean operators with kernals over cones in $\mathbb{R}^n$. Italian J. Pure Appl. Math. **18**, 89–118 (2005)
6. Essel, E., Oguntuase, J., Persson, L.E., Poopola, B.: Refined Multidiminsional Hardy-type inequalities via superquadraticity. Banach J. Math. Anal. **2**(2), 129–139 (2008)
7. Fefferman, R., Stein, E.M.: Singular integrals on product spaces. Adv. Math. **45**, 117–143 (1982)
8. Gogatishvili, A., Kufner, A., Persson, L.-E., Wedestig, A.: An equivalence theorem for integral conditions related to Hardy's inequality. Real Anal. Exchange **29**(2), 867–880 (2003/04)
9. Heinig, H.P., Kerman, R., Krbec, M.: Weighted exponential inequalities. Georgian Math. J. **8**(1), 69–86 (2001)
10. Jain, P., Hassija, R.: Some Remarks on Two Dimensional Knopp Type Inequalities. Appl. Math. Lett. **16**(4), 459–464 (2003)
11. Jain, P., Persson, L.E., Wedestig, A.: From Hardy to Carleman and general mean-type inequalities. In: Function Spaces and Applications, pp. 117–130. Narosa Publishing House, New Delhi (2000)

12. Johansson, M.: Carleman type inequalities and Hardy type inequalities for monotone functions. PhD. Thesis, Department of Mathematics, Luleå University of Technology (2007)
13. Johansson, M., Persson, L.-E., Wedestig, A.: A new approach to the Sawyer and Sinnamon characterizations of Hardy's inequality for decreasing functions. Georgian Math. J. **15**(2), 295–306 (2008)
14. Kaijser, S., Nikolova, L., Persson, L.E., Wedestig, A.: Hardy-type inequalities via convexity. Math. Inequal. Appl. **3**, 403–417 (2005)
15. Kokilashvili, V., Meshki, A., Persson, L.E.: Weighted Norm Inequalities for Integral Transforms with Product Kernels, Nova Science Publishers, Inc., New York (339 pages), 2010
16. Kufner, A., Maligranda, L., Persson, L.E.: The prehistory of the Hardy inequality. Am. Math. Mon.**113**(8), 715–732 (2006)
17. Kufner, A., Maligranda, L., Persson, L.-E.: The Hardy inequality - About its history and some related results. Vydavatelsky Servis Publishing House, Pilsen (2007)
18. Kufner, A., Persson, L.-E.: Weighted Inequalities Of Hardy Type. World Scientific, New Jersey/ London/ Singapore/ Hong Kong, (357 pages) (2003)
19. Muckenhoupt, B: Hardy's inequality with weights. Stud. Math. **4**, 31–38 (1972)
20. Oguntuase, J., Okpoti, C., Persson, L.E., Alotey, F.: Weighted multidimensional Hardy type inequalities via Jensen's inequality. J. Proc. A. Razmadze Inst. **144**, 91–105 (2007)
21. Oguntuase, J., Okpoti, C., Persson, L.E.: Multidimensional Hardy type inequalities for $p < 0$ and $0 < p < 1$. J. Math. Inequal. **1**(1), 1–11 (2007)
22. Oguntuase, J., Persson, L.E., Essel, E.: Multidimensional Hardy-type inequalities with general kernals. J. Math. Anal. Appl. **348**(1), 411–418 (2008)
23. Okpoti, C., Persson, L.E., Sinnamon, G.: An equivalence theorem for some integral conditions with general measures related to Hardy's inequality. J. Math. Anal. Appl. **326**(1), 398–413 (2007)
24. Okpoti, C., Persson, L.E., Sinnamon, G.: An equivalence theorem for some integral conditions with general measures related to Hardy's inequality II. J. Math. Anal. Appl. **337**(1), 219–230 (2008)
25. Opic, B., Gurka, P.: Weighted inequalities for geometric means. Proc. Am. Math. Soc. **120**(3), 771–779 (1994)
26. Persson, L.-E., Stepanov, V.D.: Weighted integral inequalities with the geometric mean operator. J. Inequal. Appl. **7**, 727–746 (2002)
27. Persson, L.-E., Stepanov, V., Wall, P.: Some scales of equivalent weight characterizations of Hardy's inequality: the case $q < p$. Math. Inequal. Appl. **10**(2), 267–279 (2007)
28. Persson, L.-E., Stepanov, V., Ushakova, E.: Equivalence of Hardy-type inequalities with general measures on the cones of non-negative respective non-increasing functions. Proc. Am. Math. Soc. **134**(8), 2363–2372 (2006)
29. Persson, L.-E., Ushakova, E.: Some multi-dimensional Hardy type integral inequalities. Math. Inequal. Appl. **1**(3), 301–319 (2007)
30. Sawyer, E.: Weighted inequalities for two-dimensional Hardy operator. Stud. Math. **82**(1), 1–16 (1985)
31. Stepanov, V.D.: Two-weight estimates for Riemann-Liouville integrals. (Russian) Izv. Akad. Nauk SSSR Ser. Mat. **54**(3), 645–656 (1990) translation in Math. USSR-Izv. **36**(3), 669–681 (1991)
32. Sinnamon, G.: Hardy's inequality and monotonocity. In: Drábec, P., Rákosnik, J. (eds.) Function Spaces and Nonlinear Analysis, pp. 292–310. Mathematical Institute of the Academy of Sciences of the Czech Republic, Prague (2005)
33. Ushakova, E: Norm inequalities of Hardy and Pólya-Knopp types, PhD. Thesis, Department of Mathematics, Luleå University of Technology (2006)
34. Wedestig, A.: Weighted inequalities of Hardy-type and their limiting inequalities, PhD Thesis, Department of Mathematics, Luleå University of Technology (2003)
35. Wedestig, A.: Weighted inequalities for the Sawyer two-dimensional Hardy operator and its limiting geometric mean operator. J. Inequal. Appl. **4**, 387–394 (2005)

Chapter 11
A Lizorkin Theorem on Fourier Series Multipliers for Strong Regular Systems

Lars-Erik Persson, Lyazzat Sarybekova, and Nazerke Tleukhanova

Abstract A new Fourier series multiplier theorem of Lizorkin type is proved for the case $1 < p < q < \infty$. The result is given for a general strong regular system and, in particular, for the trigonometric system it implies an analogy of the original Lizorkin theorem.

11.1 Introduction

Let $1 \le p \le q \le \infty$. We say that a sequence of complex numbers $\lambda = \{\lambda_k\}_{k\in\mathbb{Z}}$ is a multiplier of a trigonometrical Fourier series from $L_p[0,1]$ to $L_q[0,1]$, if, for every function $f \in L_p[0,1]$, with Fourier series $\sum_{k\in Z} \hat{f}(k)e^{2\pi ikx}$, there exists a function $f_\lambda \in L_q[0,1]$, having a Fourier series which coincides with the series $\sum_{k\in\mathbb{Z}} \lambda_k \hat{f}(k)e^{2\pi ikx}$, such that the operator T_λ, $T_\lambda f = f_\lambda$, is bounded from $L_p[0,1]$ to $L_q[0,1]$.

The set m_p^q of all such multipliers is a normed space with the norm

$$\|\lambda\|_{m_p^q} := \sup_{f\neq 0} \frac{\|f_\lambda\|_{L_q}}{\|f\|_{L_p}}.$$

In the case $p = q$, we put $m_p^p = m_p$. In the present paper, the letters c (c_1, c_2, etc.), denote positive constants that depend on the indicated parameters.

L.-E. Persson (✉)
Department of Mathematics, Luleå University of Technology, SE-97187 Luleå, Sweden

Narvik University College, P.O. Box 385, N 8505, Narvik, Norway
e-mail: larserik@sm.luth.se

L. Sarybekova · N. Tleukhanova
L. N. Gumilyov Eurasian National University, 5 Munaitpasov St, Astana 010008, Kazakhstan
e-mail: lsarybekova@yandex.ru; tleukhanova@rambler.ru

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_11,

A main result in the theory of Fourier series is the following one by Marcinkiewicz [3]:

Theorem 1. *Let $1 < p < \infty$, and let $\lambda = \{\lambda_m\}_{m\in\mathbb{Z}}$ be a sequence of real numbers satisfying the following condition:*

$$F_0(\lambda) = \sup_{m\in\mathbb{N}} \left(\sum_{k=2^m}^{2^{m+1}} |\lambda_k - \lambda_{k+1}| + |\lambda_{-k} - \lambda_{-k-1}| \right) + \sup_{m\in\mathbb{Z}} |\lambda_m| < \infty.$$

Then λ is a Fourier multiplier in $L_p[0, 2\pi)$ and

$$\|\lambda\|_{m_p} \le c F_0(\lambda).$$

The problem of finding sufficient conditions for λ to belong to m_p, such that they essentially depend on p, was solved by Nursultanov [6]. In his paper examples, which illustrate the importance of these conditions, are derived and discussed.

Note that an analogous (Marcinkiewicz) theorem for Fourier transform multipliers was given by Mihlin [4].

Moreover, Lizorkin [5] strengthened and generalized this Mihlin result for the case $1 < p \le q < \infty$:

Theorem 2. *Let $1 < p \le q < \infty$, $A > 0$, and assume that the function $\varphi \in AC^{loc}(\mathbb{R} \setminus \{0\})$ satisfies the following conditions:*

$$\sup_{y\in\mathbb{R}\setminus\{0\}} |y|^{\frac{1}{p}-\frac{1}{q}} |\varphi(y)| \le A,$$

$$\sup_{y\in\mathbb{R}\setminus\{0\}} |y|^{1+\frac{1}{p}-\frac{1}{q}} \left|\varphi'(y)\right| \le A.$$

Then $\varphi \in m_p^q$ and $\|\varphi\|_{m_p^q} \le cA$, where c depends only on p and q.

An exact analogue of Theorem 2 holds also for the Fourier series case which e.g. can be seen as a special case of the result in this paper (see Corollary 4). In fact, in this paper we will in particular prove a generalization of this theorem to the case with Fourier series multipliers for strong regular systems. This system is rather general e.g. all trigonometrical type systems, the Walsh systems and multiplicative systems are regular.

The paper is organized as follows: In Sect. 11.2 we present and discuss our main results. The detailed proofs can be found in Sect. 11.3.

11.2 Main Results

We say that an orthonormal system $\Phi = \{\varphi_k\}_{k\in\mathbb{N}}$ of functions defined on [0,1] is a strong regular system, if there exists a constant $B > 0$ such that for every segment w from $\mathbb{N}$ (finite set of consecutive integers) and $t \in (0, 1]$, it yields that

$$\left(\sum_{k\in w}\varphi_k(\cdot)\varphi(y)\right)^*(t)\le B\min(|w|,1/t),\ y\in[0,1], \tag{11.1}$$

where $\left(\sum_{k\in w}\varphi_k(\cdot)\varphi_k(y)\right)^*(t)$ is the non-increasing rearrangement of the function $D_w(x,y)=\sum_{k\in w}\varphi_k(x)\varphi_k(y)$ by variable x with fixed second variable $y\in(0,1]$, and $|w|$ is the number of elements in w. Since $D_w(x,y)=\sum_{k\in w}\varphi_k(x)\varphi_k(y)$ is symmetric, we have

$$\left(\sum_{k\in w}\varphi_k(x)\varphi_k(\cdot)\right)^*(t)\le B\min(|w|,1/t),\ x\in[0,1].$$

Let $1\le p\le q\le\infty$ and $\Phi=\{\varphi_k\}_{k=1}^{\infty}$ be a strong regular system, $f\in L_p[0,1]$ with Fourier series $\sum\limits_{k\in N}\hat{f}(k)\varphi_k(x)$, and let $\lambda=\{\lambda_k\}_{k\in\mathbb{N}}$ be the sequence of complex numbers.

Let us define the sequence of partial sums $S_n=S_n(f,\lambda,x)$ by

$$S_n(f,\lambda,x)=\sum_{k=1}^{n}\lambda_k\hat{f}(k)\varphi_k(x),\ \ n\in\mathbb{N}.$$

We say that $\lambda=\{\lambda_k\}_{k\in\mathbb{N}}$ is a Fourier series multiplier for the strong regular system Φ from $L_p[0,1]$ to $L_q[0,1]$, if

$$\|\lambda\|_{m_p^q}:=\sup_{n\in\mathbb{N}}\sup_{f\ne0}\frac{\|S_n(f,\lambda,x)\|_{L_q}}{\|f\|_{L_p}}<\infty.$$

In the sequel we always consider this case of general strong regular systems. Our main result reads:

Theorem 3. *Let $1<p<q<\infty$, $0\le\alpha<1-\frac{1}{p}+\frac{1}{q}$, and $\beta=\alpha+\frac{1}{p}-\frac{1}{q}$. Let the sequence of complex numbers $\lambda=\{\lambda_k\}_{k\in\mathbb{N}}$ satisfy the following conditions:*

$$\sup_{k\in\mathbb{N}}k^{\frac{1}{p}-\frac{1}{q}}|\lambda_k|\le A,$$

$$\sup_{k\in\mathbb{N}}k^{1-\alpha}\left(m^{\beta}(\lambda_m-\lambda_{m+1})\right)^*(k)\le A. \tag{11.2}$$

Then $\lambda\in m_p^q$ for each regular system, and

$$\|\lambda\|_{m_p^q}\le cA,$$

where $c>0$ depends only on p,q and α.

The following corollary is a genuine generalization of (the Lizorkin) Theorem 2:

Corollary 4. *Let* $1 < p < q < \infty$, $A > 0$. *If a sequence of complex numbers* $\lambda = \{\lambda_k\}_{k\in\mathbb{N}}$ *satisfies to the following conditions:*

$$\sup_{k\in\mathbb{N}} k^{\frac{1}{p}-\frac{1}{q}} |\lambda_k| \le A,$$

$$\sup_{k\in\mathbb{N}} k^{1+\frac{1}{p}-\frac{1}{q}} |\lambda_k - \lambda_{k+1}| \le A, \tag{11.3}$$

then $\lambda \in m_p^q$ *for each regular system, and*

$$\|\lambda\|_{m_p^q} \le cA,$$

where $c > 0$ *depends on* p, q *and* α.

Remark 5. There exists a sequence λ satisfying the assumptions of Theorem 3, but not satisfying the assumptions in Corollary 4, i.e. there exists a sequence λ such that

$$\sup_{k\in\mathbb{N}} k^{\frac{1}{p}-\frac{1}{q}} |\lambda_k| < \infty,$$

$$\sup_{k\in\mathbb{N}} k^{1-\alpha} \left(m^{\beta}(\lambda_m - \lambda_{m+1})\right)^*(k) < \infty,$$

but

$$\sup_{k\in\mathbb{N}} k^{1+\frac{1}{p}-\frac{1}{q}} |\lambda_k - \lambda_{k+1}| = \infty.$$

The proof of this statement can be found at the end of this paper.

For the proof of Theorem 3 we need the following embedding theorem of independent interest:

Theorem 6. *Let* $1 < p < q \le \infty, 0 < \tau \le \infty$, *and* $0 \le \alpha < 1 - \frac{1}{p} + \frac{1}{q}$. *Then*

$$L_{p,\tau} \hookrightarrow n^{\alpha,\beta,\tau}(L_q),$$

where $\beta = \alpha + \frac{1}{p} - \frac{1}{q}$.

Here $n^{\alpha,\beta,r}(L_q)$ is a version of the net spaces, which was introduced and studied in [1], [7] and [8], defined as follows:

Let $\{\varphi_k(x)\}$ be a strong regular system. For a function $f \in L_1[0,1]$ with Fourier series $\sum_{k\in\mathbb{N}} a_k\varphi_k(x)$ and for any finite set $Q \subset \mathbb{N}$ let us define the sum

$$S_Q(f,x) := \sum_{k\in Q} a_k\varphi_k(x),$$

which is called the Fourier sum of the function f equipped with the set Q.

Let $0 \le \alpha < \infty, 0 < \beta < \infty$ and $0 < q, r \le \infty$. In what follows G_k denotes the set of all segments Q from $\mathbb{N}$ for which the number of elements is greater than $k \in \mathbb{N}$. We say that a function f belongs to $n^{\alpha,\beta,r}(L_q)$, if $f \in L_1$ and

$$\|f\|_{n^{\alpha,\beta,r}(L_q)} := \left(\sum_{k=1}^{\infty}\left(k^{\alpha}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q}\right)^r\frac{1}{k}\right)^{\frac{1}{r}} < \infty$$

for $0 < r < \infty$, and

$$\|f\|_{n^{\alpha,\beta,r}(L_q)} = \sup_k k^{\alpha}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q} < \infty$$

for $r = \infty$.

Moreover, for the proof of Theorem 6 we need some embedding and interpolation results, also of independent interest. The first one reads:

Proposition 7. *Let* $0 < \alpha < 1,\ 0 < \beta < 1$ *and* $0 < q \le \infty$.
(a) If $0 < r \le r_1 \le \infty$, *then*

$$n^{\alpha,\beta,r}(L_q) \hookrightarrow n^{\alpha,\beta,r_1}(L_q).$$

(b) If $0 < \sigma < \min\{1-\alpha, 1-\beta\}$, $0 < r \le \infty$, *then*

$$n^{\alpha,\beta,r}(L_q) \hookrightarrow n^{\alpha+\sigma,\beta+\sigma,r}(L_q). \tag{11.4}$$

Let (A_0, A_1) be a compatible pair of Banach spaces (see e.g. [2]), and let

$$K(\tau, a; A_0, A_1) := \inf_{a=a_0+a_1}(\|a_0\|_{A_0} + \tau\|a_1\|_{A_1}),\quad a \in A_0 + A_1,\quad \tau > 0,$$

be the Peetre K-functional.

Moreover, for $0 < q < \infty$, $0 < \theta < 1$, let

$$(A_0, A_1)_{\theta,q} := \Bigg\{a \in A_0 + A_1 : \|a\|_{(A_0,A_1)_{\theta,q}} := \left(\int_0^{\infty}\left(\tau^{-\theta}K(\tau, a; A_0, A_1)\right)^q\frac{d\tau}{\tau}\right)^{\frac{1}{q}} < \infty\Bigg\},$$

and for $q = \infty$

$$(A_0, A_1)_{\theta,\infty} := \left\{a \in A_0 + A_1 : \|a\|_{(A_0,A_1)_{\theta,\infty}} := \sup_{0<\tau<\infty}\tau^{-\theta}K(\tau, a; A_0, A_1) < \infty\right\}.$$

The second auxiliary result reads:

Proposition 8. *Let* $0 < \alpha_1 < 1,\ 0 < \beta < 1$ *and* $0 < r, q \le \infty$. *Then*

$$(n^{\alpha_1,\beta,\infty}(L_q), n^{0,\beta,\infty}(L_q))_{\theta,r} \hookrightarrow n^{\alpha,\beta,r}(L_q),$$

where $0 < \theta < 1, \alpha = (1-\theta)\alpha_1$.

The third auxiliary result reads:

Proposition 9. *Let* Q *from* $\mathbb{N}$. *If* $1 \le p \le q \le \infty$, *then there exists* $c > 0$, *which depends only on* p *and* q, *such that*

$$\left\| S_Q(f) \right\|_{L_q} \le c|Q|^{\frac{1}{p}-\frac{1}{q}} \| f \|_{L_p}, \tag{11.5}$$

for every $f \in L_p$.

11.3 Proofs

We present the proofs in the order the results are used in later proofs.

Proof. (Proposition 7.) Let us first prove that $n^{\alpha,\beta,r}(L_q) \hookrightarrow n^{\alpha,\beta,\infty}(L_q)$.

Indeed,

$$\| f \|_{n^{\alpha,\beta,\infty}(L_q)} = \sup_k k^\alpha \sup_{Q \in G_k} \frac{1}{|Q|^\beta} \| S_Q(f) \|_{L_q} \le$$

$$\le c(\alpha, r) \sup_k \left(\sum_{i=1}^{k} i^{\alpha r - 1} \right)^{\frac{1}{r}} \sup_{Q \in G_k} \frac{1}{|Q|^\beta} \| S_Q(f) \|_{L_q} \le$$

$$\le c(\alpha, r) \left(\sum_{i=1}^{\infty} \left(i^\alpha \sup_{Q \in G_i} \frac{1}{|Q|^\beta} \| S_Q(f) \|_{L_q} \right)^r \frac{1}{i} \right)^{\frac{1}{r}}$$

$$= c(\alpha, r) \| f \|_{n^{\alpha,\beta,r}(L_q)}.$$

Then, by using this fact and the multiplicative inequality for the spaces L_{r_1}, we have that

$$\| f \|_{n^{\alpha,\beta,r_1}(L_q)} \le \| f \|^{\frac{r}{r_1}}_{n^{\alpha,\beta,r}(L_q)} \| f \|^{1-\frac{r}{r_1}}_{n^{\alpha,\beta,\infty}(L_q)} \le c(\alpha, r) \| f \|_{n^{\alpha,\beta,r}(L_q)}.$$

Moreover,

$$\|f\|_{n^{\alpha+\sigma,\beta+\sigma,r}(L_q)} = \left(\sum_{k=1}^{\infty}\left(k^{\alpha+\sigma}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta+\sigma}}\|S_Q(f)\|_{L_q}\right)^r\frac{1}{k}\right)^{\frac{1}{r}} \leq$$

$$\leq \left(\sum_{k=1}^{\infty}\left(k^{\alpha}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q}\right)^r\frac{1}{k}\right)^{\frac{1}{r}} = \|f\|_{n^{\alpha,\beta,r}(L_q)}.$$

The proof is complete. □

Proof. (Proposition 8.) Let $f = f_0 + f_1$, where $f_0 \in n^{0,\beta,\infty}(L_q)$ and $f_1 \in n^{\alpha_1,\beta,\infty}(L_q)$, be an arbitrary representation of f. Then

$$\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q} \leq 2^{\left(\frac{1}{2}-1\right)}\left(\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_1)\|_{L_q} + \sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_0)\|_{L_q}\right),$$

where $x_+ = x$, if $x > 0$ and $x_+ = 0$, if $x \leq 0$. If we denote $v(\tau) = \tau^{\frac{1}{\alpha_1}}$, $\tau > 0$, then

$$\sup_{1\leq k\leq v(\tau)} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q}$$

$$\leq c_1\left(\sup_{1\leq k\leq v(\tau)} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_1)\|_{L_q} + \sup_{1\leq k\leq v(\tau)} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_0)\|_{L_q}\right)$$

$$\leq c_1\left(\sup_{k\geq 1} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_1)\|_{L_q} + \tau\sup_{k\geq 1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f_0)\|_{L_q}\right).$$

Taking into account that the representation $f = f_0 + f_1$ is arbitrary, we have that

$$\sup_{1\leq k\leq v(\tau)} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q} \leq c_1 K(\tau, f; n^{\alpha_1,\beta,\infty}, n^{0,\beta,\infty}).$$

Thus, for $0 < r \leq \infty$, we obtain that

$$\left(\int_0^{\infty}\left(\tau^{-\theta}K(\tau, f; n^{\alpha_1,\beta,\infty}, n^{0,\beta,\infty})\right)^r\frac{d\tau}{\tau}\right)^{\frac{1}{r}} \geq$$

$$\geq c\left(\int_0^{\infty}\left(\tau^{-\theta}\sup_{k\leq v(\tau)} k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^{\beta}}\|S_Q(f)\|_{L_q}\right)^r\frac{d\tau}{\tau}\right)^{\frac{1}{r}} =$$

$$= c\left(\alpha_1\int_0^\infty\left(u^{-\theta\alpha_1}\sup_{k\le u}k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^\beta}\|S_Q(f)\|_{L_q}\right)^r\frac{du}{u}\right)^{\frac{1}{r}}\ge$$

$$\ge c\alpha_1^{\frac{1}{r}}\left(\sum_{i=1}^\infty\left(2^{-\theta i\alpha_1}\sup_{k\le 2^i}k^{\alpha_1}\sup_{Q\in G_k}\frac{1}{|Q|^\beta}\|S_Q(f)\|_{L_q}\right)^r\right)^{\frac{1}{r}}\ge$$

$$\ge c\left(\sum_{i=1}^\infty\left(2^{(1-\theta)\alpha_1 i}\sup_{Q\in G_{2^i}}\frac{1}{|Q|^\beta}\|S_Q(f)\|_{L_q}\right)^r\right)^{\frac{1}{r}}\sim$$

$$\sim c\|f\|_{n^{\alpha,\beta,r}(L_q)},$$

i.e.

$$(n^{\alpha_1,\beta,\infty}(L_q),n^{0,\beta,\infty}(L_q))_{\theta,r}\hookrightarrow n^{\alpha,\beta,r}(L_q),$$

where $\alpha=(1-\theta)\alpha_1$. The proof is complete. □

Proof. (Proposition 9.) Let us consider the following well-known integral representation of the partial sum of the Fourier series in a strong regular system:

$$S_Q(x)=\int_0^1 f(t)D_Q(x,t)dt,$$

where $D_Q(x,t)=\sum_{k\in Q}\varphi_k(x)\varphi_k(t)$. According to (11.1) and the proof of the Young inequality we obtain that

$$\|S_Q\|_{L_q}\le B\|f\|_{L_p}\left\|\min\left(Q,\frac{1}{t}\right)\right\|_{L_r},$$

where the parameters $1\le p,q,r\le\infty$ satisfy the equality $1+\frac{1}{q}=\frac{1}{p}+\frac{1}{r}$.

Since

$$\left(\int_0^1\left(\min\left(|Q|,\frac{1}{t}\right)\right)^r dt\right)^{\frac{1}{r}}=$$

$$=\left(\int_0^{1/|Q|}|Q|^r\,dt+\int_{1/|Q|}^1\left(\frac{1}{t}\right)^r dt\right)^{\frac{1}{r}}\le c(p,q)\,|Q|^{\frac{1}{p}-\frac{1}{q}},$$

we have

$$\|S_Q\|_{L_q}\le Bc(p,q)\,|Q|^{\frac{1}{p}-\frac{1}{q}}\|f\|_{L_p}.$$

This completes the proof. □

Proof. (Theorem 6.) Let $1\le r<q\le\infty$. By Proposition 9 there exists $c>0$, depending on r and q, such that

$$\left\| S_Q(f) \right\|_{L_q} \le c(r,q) |Q|^{\frac{1}{r}-\frac{1}{q}} \|f\|_{L_r}, \tag{11.6}$$

for every $f \in L_r$.

Let $0 \le \alpha \le 1 - \frac{1}{p}$. Then $\frac{1}{p} - \frac{1}{q} < \beta \le 1 - \frac{1}{q}$ and there exists p_0, such that $1 < p_0 < p$ and $\beta = \frac{1}{p_0} - \frac{1}{q}$. According to (11.6), applied with $r = p_0$, we have that

$$\|f\|_{n^{0,\beta,\infty}(L_q)} = \sup_k \sup_{Q \in G_k} \frac{1}{|Q|^\beta} \left\| S_Q(f) \right\|_{L_q} \le c(p_0, q) \|f\|_{L_{p_0}} \tag{11.7}$$

for every $f \in L_{p_0}$.

Further, let $p < p_1 < q$ and $\alpha_1 = \frac{1}{p_0} - \frac{1}{p_1}$. By using (11.6) with $r = p_1$ we obtain that

$$\sup_k \sup_{Q \in G_k} \frac{1}{|Q|^{\frac{1}{p_1}-\frac{1}{q}}} \left\| S_Q(f) \right\|_{L_q} \le c(p_1, q) \|f\|_{L_{p_1}},$$

for every $f \in L_{p_1}$.

Since

$$\begin{aligned} \sup_k \sup_{Q \in G_k} \frac{1}{|Q|^{\frac{1}{p_1}-\frac{1}{q}}} \left\| S_Q(f) \right\|_{L_q} &= \sup_k \sup_{Q \in G_k} \frac{|Q|^{\alpha_1}}{|Q|^\beta} \left\| S_Q(f) \right\|_{L_q} \\ &\ge \sup_k k^{\alpha_1} \sup_{Q \in G_k} \frac{1}{|Q|^\beta} \left\| S_Q(f) \right\|_{L_q} = \|f\|_{n^{\alpha_1,\beta,\infty}(L_q)}, \end{aligned}$$

it yields that

$$\|f\|_{n^{\alpha_1,\beta,\infty}(L_q)} \le c(p_1, q) \|f\|_{L_{p_1}} \tag{11.8}$$

for every $f \in L_{p_1}$. The inequalities (11.7) and (11.8) mean that

$$L_{p_0} \hookrightarrow n^{0,\beta,\infty}(L_q)$$

and

$$L_{p_1} \hookrightarrow n^{\alpha_1,\beta,\infty}(L_q).$$

Hence,

$$L_{p_0} + L_{p_1} \hookrightarrow n^{0,\beta,\infty}(L_q) + n^{\alpha_1,\beta,\infty}(L_q).$$

In what follows by I we mean the corresponding embedding operator. According to (11.7) and (11.8) we have that

$$I : L_{p_0} \to n^{0,\beta,\infty}(L_q),$$

and

$$I : L_{p_1} \to n^{\alpha_1,\beta,\infty}(L_q).$$

Moreover, in both cases the operator I is bounded.

Let $\theta \in (0,1)$ be such that $\frac{1}{p} = \frac{1-\theta}{p_1} + \frac{\theta}{p_0}$. Since

$$\alpha = \beta - \frac{1}{p} + \frac{1}{q} = \frac{1}{p_0} - \frac{1}{p} = (1-\theta)\alpha_1,$$

by the interpolation properties of the spaces L_p (see e.g. [2]), we have that

$$I : L_{p,\tau} = \left(L_{p_1}, L_{p_0}\right)_{\theta\tau} \to \left(n^{\alpha_1,\beta,\infty}(L_q), n^{0,\beta,\infty}(L_q)\right)_{\theta\tau},$$

where the operator I is bounded. Thus,

$$L_{p,\tau} \hookrightarrow \left(n^{\alpha_1,\beta,\infty}(L_q), n^{0,\beta,\infty}(L_q)\right)_{\theta\tau},$$

and the statement of the theorem follows by using Proposition 8.

When $1 - \frac{1}{p} \le \alpha < 1 - \frac{1}{p} + \frac{1}{q}$, the statement of the theorem follows from Proposition 7. Indeed, let $0 < \tilde{\alpha} \le 1 - \frac{1}{p}$. Then, by using what is proved above and by (11.4), we have that

$$L_{p,\tau} \hookrightarrow n^{\tilde{\alpha},\tilde{\alpha}+\frac{1}{p}-\frac{1}{q},\tau}(L_q) \hookrightarrow n^{\alpha,\beta,\tau}(L_q).$$

The proof is complete. □

Proof. (Theorem 3.) Let $n \in \mathbb{N}$. By using an Abel transformation and the Minkowski inequality we find that

$$\|S_n(f_\lambda)\|_{L_q} = \left\| \sum_{k=1}^{n} \lambda_k \hat{f}_k \varphi_k(x) \right\|_{L_q} =$$

$$= \left\| \sum_{k=1}^{n} (\lambda_k - \lambda_{k+1}) \sum_{m=1}^{k-1} \hat{f}_m \varphi_m(x) - \lambda_n \sum_{k=1}^{n} \hat{f}_k \varphi_k(x) \right\|_{L_q} \le$$

$$\le \sum_{k=1}^{n} |\lambda_k - \lambda_{k+1}| \, \|S_k(f)\|_{L_q} + |\lambda_n| \, \|S_n(f)\|_{L_q} := I_1 + I_2.$$

Moreover, according to (11.5) and taking into account that $L_{p,1} \hookrightarrow L_p$, we obtain that

$$I_2 = n^{\frac{1}{p}-\frac{1}{q}} |\lambda_n| \frac{1}{n^{\frac{1}{p}-\frac{1}{q}}} \|S_n(f)\|_{L_q} \le Ac(p,q) \|f\|_{L_p} \le Ac(p,q) \|f\|_{L_{p,1}}.$$

Furthermore,

$$I_1 = \sum_{k=1}^{n} k^{\beta} \left|\lambda_k - \lambda_{k+1}\right| \frac{1}{k^{\beta}} \left\|S_k(f)\right\|_{L_q} \leq$$

$$\leq \sum_{k=1}^{n} k^{\beta} \left|\lambda_k - \lambda_{k+1}\right| \left(\sup_{Q \in G_k} \frac{1}{|Q|^{\beta}} \left\|S_Q(f)\right\|_{L_q} \right).$$

Since $\sup_{Q \in G_k} \frac{1}{|Q|^{\beta}} \left\|S_Q(f)\right\|_{L_q}$ is a non-increasing sequence, then, by using the well-known inequality $\sum_{k=1}^{\infty} F_k G_k \leq \sum_{m=1}^{\infty} F_m^* G_m^*$, we have that

$$I_1 \leq \sum_{k=1}^{\infty} \left(m^{\beta} \left(\lambda_m - \lambda_{m+1}\right)\right)^* (k) \left(\sup_{Q \in G_k} \frac{1}{|Q|^{\beta}} \left\|S_Q(f)\right\|_{L_q} \right) =$$

$$= \sum_{k=1}^{\infty} k^{1-\alpha} \left(m^{\beta} \left(\lambda_m - \lambda_{m+1}\right)\right)^* (k) k^{\alpha} \sup_{Q \in G_k} \frac{1}{|Q|^{\beta}} \left\|S_Q(f)\right\|_{L_q} \frac{1}{k} \leq$$

$$\leq \sup_{k \in \mathbb{N}} k^{1-\alpha} \left(m^{\beta} \left(\lambda_m - \lambda_{m+1}\right)\right)^* (k) \sum_{k=1}^{\infty} k^{\alpha} \sup_{Q \in G_k} \frac{1}{|Q|^{\beta}} \left\|S_Q(f)\right\|_{L_q} \frac{1}{k}$$

$$\leq A \left\|f\right\|_{n^{\alpha,\beta,1}(L_q)}.$$

Hence, by Theorem 6 there exists $c = c(p, q, \alpha) > 0$ such that

$$I_1 \leq Ac \left\|f\right\|_{L_{p,1}}.$$

Consequently, there exists $c_1 > 0$, depending only on p, q and α, such that

$$\left\|S_n(f_\lambda)\right\|_{L_q} \leq Ac_1 \left\|f\right\|_{L_{p,1}}.$$

Let the couples of numbers (p_0, q_0) and (p_1, q_1) be such that $1 < p_0 < p < p_1 < \infty$, $1 < q_0 < q < q_1 < \infty$ and

$$\frac{1}{p_0} - \frac{1}{q_0} = \frac{1}{p_1} - \frac{1}{q_1} = \frac{1}{p} - \frac{1}{q}. \tag{11.9}$$

Similarly, it follows that for some $c_2 > 0$, depending only on p, q and α such that

$$\left\|S_n(f_\lambda)\right\|_{L_{q_0}} \leq c_2 A \left\|f\right\|_{L_{p_0,1}}$$

and

$$\left\|S_n(f_\lambda)\right\|_{L_{q_1}} \leq c_2 A \left\|f\right\|_{L_{p_1,1}}.$$

Let $\theta \in (0, 1)$ be such that $\frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}$. Then by (11.9) we obtain also that

$$\frac{1}{q}=\frac{1}{p}-\frac{1}{p_0}+\frac{1}{q_0}=\frac{1-\theta}{p_0}+\frac{\theta}{p_1}-\frac{1}{p_0}+\frac{1}{q_0}=$$
$$=\theta\left(\frac{1}{p_1}-\frac{1}{p_0}\right)+\frac{1}{q_0}=\theta\left(\frac{1}{q_1}-\frac{1}{q_0}\right)+\frac{1}{q_0}=\frac{1-\theta}{q_0}+\frac{\theta}{q_1}.$$

By the Marcinkiewicz interpolation theorem (see e.g. [2]) we have that for every $0<r\le\infty$

$$\|S_n(f_\lambda)\|_{L_{q,r}}\le c_3 A\,\|f\|_{L_{p,r}},$$

where $c_3>0$ depends on p,q,α and r.

If, specifically, $r=p$, then for some $c_4>0$, which depends on p,q and α, it follows that

$$\|S_n(f_\lambda)\|_{L_q}\le c_4 A\,\|f\|_{L_p},$$

since $L_{q,p}\hookrightarrow L_q$ for $p<q$.

Consequently,

$$\sup_{n\in\mathbb{N}}\sup_{f\neq 0}\frac{\|S_n(f,\lambda)\|_{L_q}}{\|f\|_{L_p}}\le c_{13}A.$$

The proof is complete. □

Finally we include also a

Proof. (Statement in Remark 5.) Let

$$d_m=\begin{cases}2^{-n\beta}n^{-(1-\alpha)}, & \text{if } m=2^n,\\ 0, & \text{in other cases,}\end{cases}\qquad m\in\mathbb{N},$$

$$\lambda_k=\sum_{m=k}^{\infty}d_m.$$

We remind that $\beta-\alpha=\frac{1}{p}-\frac{1}{q}$.

Let us prove the first condition. Let $k\in[2^{n-1},2^n]$, then

$$k^{\frac{1}{p}-\frac{1}{q}}\,|\lambda_k|=k^{\frac{1}{p}-\frac{1}{q}}\left|\sum_{m=k}^{\infty}d_m\right|\le c2^{n\left(\frac{1}{p}-\frac{1}{q}\right)}d_{2^n}$$
$$=c2^{n(\beta-\alpha)}2^{-n\beta}n^{-(1-\alpha)}=c\frac{1}{2^{n\alpha}n^{1-\alpha}}<\infty.$$

We note that

$$k^{\beta}\,|\lambda_k-\lambda_{k+1}|=k^{\beta}d_k=\begin{cases}n^{-(1-\alpha)}, & k=2^n,\\ 0, & \text{in other cases,}\end{cases}\qquad k\in\mathbb{N}.$$

Hence

$$\sup_{i\geq 1} i^{1-\alpha}\cdot\left(k^{\beta}(\lambda_k-\lambda_{k+1})\right)^{*}(i)=\sup_{i=1} i^{1-\alpha}i^{-(1-\alpha)}=1,$$

i.e. the second condition holds.

On the other hand,

$$\sup_{m\in\mathbb{N}} m^{1+\frac{1}{p}-\frac{1}{q}}\cdot|\lambda_m-\lambda_{m+1}|=\sup_{m\in\mathbb{N}} m^{1+\frac{1}{p}-\frac{1}{q}}\cdot d_m\leq\sup_{n\in\mathbb{N}} 2^{n\left(1+\frac{1}{p}-\frac{1}{q}\right)}2^{-n\beta}n^{-(1-\alpha)}$$

$$=\sup_{n\in\mathbb{N}} 2^{n(1+\beta-\alpha)}2^{-n\beta}n^{-(1-\alpha)}=\sup_{n\in\mathbb{N}}\left(\frac{2^n}{n}\right)^{1-\alpha}=+\infty.$$

This completes the proof. □

Acknowledgements This research has been done within the frame of the general agreement between Eurasian National University in Astana, Kazakhstan and Luleå University of Technology in Sweden concerning research and PhD education in mathematics. We thank both these universities for financial support, which made this cooperation possible.

References

1. Aubakirov, T.U., Nursultanov, E.D.: Interpolation theorem for stochastic processes. Eurasian Math. J. **1:1**, 8–16 (2010)
2. Bergh, J., Löfström, J.: Interpolation spaces. An introduction. Grundlehren der Mathematischen Wissenschaften, vol. 223. Springer, Berlin (1976)
3. Marcinkiewicz, J.: Sur les multiplicateurs des series de Fourier. Studia Math. **8**, 78–91 (1939)
4. Mihlin, S.G.: On the multipliers of Fourier integrals. Dokl. Akad. Nauk SSSR (N.S.) **109** 701–703 (1956) (in Russian)
5. Lizorkin, P.I.: Multipliers of Fourier integrals in the spaces $L_{p,\theta}$. Trudy Mat. Inst. Steklov (Russian) **89** 231–248 (1967)
6. Nursultanov, E.D.: On multipliers of Fourier series in a trigonometric system. Mat. Zametki (Russian) **63**(2) 235–247 (1998); Translation in Math. Notes **63**(1-2) 205–214 (1998)
7. Nursultanov, E.D.: S. M. Nikol'skii's inequality for different metrics, and properties of the sequence of norms of Fourier sums of a function in the Lorentz space. Trudy Mat. Inst. Steklov (Russian) **255** (2006); Funkts. Prostran., Teor. Priblizh., Nelinein. Anal. 197–215
8. Nursultanov, E.D.: Network spaces and inequalities of Hardy-Littlewood type. Sb. Math. (Russian) **189**(3-4) 83–102 (1998); Translation in Sb. Math. **189**(3-4) 399–419 (1998)

Chapter 12
Note on the Structure of the Spaces of Matrix Monotone Functions

Hiroyuki Osaka and Jun Tomiyama

Abstract Let $n \in \mathbf{N}$ and M_n be the algebra of $n \times n$ matrices. We call a function f matrix monotone of order n or n-monotone in short whenever the inequality $f(a) \leq f(b)$ holds for every pair of selfadjoint matrices $a, b \in M_n$ such that $a \leq b$ and all eigenvalues of a and b are contained in I. The spaces for n-monotone functions is written as $P_n(I)$.

For each $n \in \mathbf{N}$ and a finite interval I we define the class $C_n(I)$ by the set of all positive real-valued continuous functions f over I such that $f(I^\circ) \subset (0, \infty)$ and for any subset $S \subset I^\circ$ there exists a positive Pick function h on $(0, \infty)$ interpolating f on S. Then we characterize $C_n([0, 1))$ by an operator inequality. Moreover we show that for each n $C_{2n}([0, \infty)) \subsetneq P_n^+([0, \infty))$.

12.1 Introduction

Let I be nontrivial interval of the real line $\mathbf{R}$ (open, closed, half-open etc.). A real valued continuous function f on I is said to be operator monotone if for every selfadjoint operators a, b on a Hilbert space H ($\dim H = +\infty$) such that $a \leq b$ and $\sigma(a), \sigma(b) \subseteq I$ we have $f(a) \leq f(b)$.

Let $n \in \mathbf{N}$ and M_n be the algebra of $n \times n$ matrices. We call a function f matrix monotone of order n or n-monotone in short whenever the inequality $f(a) \leq f(b)$ holds for every pair of selfadjoint matrices $a, b \in M_n$ such that $a \leq b$ and all eigenvalues of a and b are contained in I. We denote the spaces of

H. Osaka (✉)
Department of Mathematical Sciences, Ritsumeikan University, Kusatsu, Shiga 525-8577, Japan
e-mail: osaka@se.ritsumei.ac.jp

J. Tomiyama
Prof. Emeritus of Tokyo Metropolitan University, 201 11-10 Nakane 1-chome, Meguro-ku, Tokyo, Japan
e-mail: juntomi@med.email.ne.jp

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_12,

operator monotone functions by $P_\infty(I)$ and the spaces for n-monotone functions is written as $P_n(I)$. We also denote that $P_n^+(I) = \{f \in P_n(I): f(I^\circ) \subset (0,\infty)\}$, where I° means the set of inner points in I. We note that $P_{n+1}(I) \subseteq P_n(I)$ and $\cap_{n=1}^\infty P_n(I) = P_\infty(I)$.

The first question is whether $P_{n+1}(I)$ is strictly contained in $P_n(I)$ for every n. Although most of literatures assert the existence of such gaps, no explicit example was given in case $n \geq 3$ in spite of the longtime since the paper [10] of Loewner in 1934. In [7] Hansen, Ji and Tomiyama presented an explicit example with the gap between $P_{n+1}(I)$ and $P_n(I)$ for every n and an interval I. More general discussions are treated in [12] by Osaka, Silvestrov, and Tomiyama about gaps of $\{P_n(I)\}_{n\in\mathbf{N}}$ and we have now abundant examples of polynomials in $P_n(I)\backslash P_{n+1}(I)$ using the truncated moment problems for Hankel matrices in [3] of Curto and Fialkow,

On the contrary, in [1] Ameur, Kaijser and Silvestrov studied subclass $C_n(0,\infty)$ of interpolation functions of order n of $P_n^+(0,\infty)$ and showed by a theorem of Doughue [4] that $C_n(0,\infty)$ coincides with the class of functions such that for each n-subset $S = \{\lambda_i\}_{i=1}^n$ there exists a positive Pick function h on $(0,\infty)$ interpolating f on S, that is, $h(\lambda_i) = f(\lambda_i)$ for each $1 \leq i \leq n$. They also showed that $P_2^+(0,\infty) \subsetneq C_3(0,\infty)$ and $C_4(0,\infty) \subsetneq P_2^+(0,\infty)$. We recall that a complex analytic function h defined on $\{z \in \mathbf{C}: \Im(z) > 0\}$ is called a Pick function if their range is in the closed upper half plane $\{z \in \mathbf{C}: \Im(z) \geq 0\}$.

In this note we characterize n-monotone functions from the point of Jensen's type inequality for operators. For each $n \in \mathbf{N}$ and a finite interval I we define the class $C_n(I)$ by the set of all positive real-valued continuous functions f over I such that $f(I^\circ) \subset (0,\infty)$ and for any subset $S \subset I^\circ$ there exists a positive Pick function h on $(0,\infty)$ interpolating f on S. Then we characterize $C_n([0,1))$ by an operator inequality. Moreover we show that for each n $C_{2n}([0,\infty)) \subsetneq P_n^+([0,\infty))$. This is an answer to a question in [1].

The authors would like to thank Dr. Yacin Ameur for a fruitful discussion about interpolation class $C_n(0,\infty)$ and Professor Sergei Silvestrov for hearty hospitality when they stayed at Lund Univ. in May, 2006 and later visits to Lund.

12.2 The Class C_n

Definition 1. Let I be a finite interval (open, closed, or open-closed). For $n \in \mathbf{N}$ we denote $C_n(I)$ be the set of all positive real-valued continuous interpolation functions f over I such that for any $\{\lambda_i\}_{i=1}^n \subset I^\circ$ there is a Pick function $h: (0,1) \to \mathbf{R}$ such that $f(\lambda_i) = h(\lambda_i)$ for $1 \leq i \leq n$, where I° denotes the set of inner points in I.

For two finite intervals of the same type such as an open, half-open like $[\alpha,\beta)$ and $[\gamma,\delta)$ one can easily find an monotone increasing linear function $h: [\gamma,\delta) \to [\gamma,\delta)$ with the inverse function $h^{-1}: [\gamma,\delta) \to [\alpha,\beta)$ having the same property. As both functions h and h^{-1} are operator monotone and operator convex functions the set

$C_n([\alpha,\beta))$ and $C_n([\gamma,\delta))$ is easily transferred each other. So we consider the case that $I = [0,1)$.

The following is the characterization of a class $C_n([0,1))$

Theorem 2. *Let $f:[0,1] \to \mathbf{R}$ be a continuous function. The followings are equivalent.*

(1) $f \in C_n([0,1))$.
(2) For any $\{\lambda_i\}_{i=1}^n \subset (0,1)$ *if*

$$\sum_{i=1}^n a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i} \geq 0$$

for any $\{a_i\}_{i=1}^n \subset \mathbf{R}$ *we have*

$$\sum_{i=1}^n a_i f(\lambda_i) \geq 0.$$

(3) For any $A, T \in M_n(\mathbf{C})$ *with* $T^*T \leq 1$ *and* $\sigma(A) \subset (0,1)$

$$T^*AT \leq A \Longrightarrow T^* f(A) T \leq f(A).$$

Proof. (1) → (3):

Take $T, A \in M_n(\mathbf{C})$ satisfying $T^*T \leq 1$ and $\sigma(A) \subset (0,1)$. Set $\phi:(0,1) \to (0,\infty)$ by $\phi(t) = \frac{t}{1-t}$. Then ϕ is operator monotone. Hence $T^*\phi(A)T \leq \phi(A)$ by [5].

Since $f \circ \phi^{-1}:(0,\infty) \to \mathbf{R} \in C_n((0,\infty))$, by [1, Corollary 2.4] we have

$$T^*((f \circ \phi^{-1})(\phi(A))T \leq (f \circ \phi^{-1})(\phi(A)),$$

and $T^* f(A) T \leq f(A)$.

(3) → (1):

Take $A, T \in M_n(\mathbf{C})$ with $T^*T \leq 1$ and $\sigma(A) \subset (0,\infty)$. Since $\phi^{-1}:(0,\infty)$ is operator monotone and $\sigma(A) \subset (0,\infty)$, from [5] we have

$$T^*\phi^{-1}(A)T \leq \phi^{-1}(A).$$

Note that $\sigma(\phi^{-1}(A)) \subset (0,1)$. Then from the assumption for f we have

$$T^* f(\phi^{-1}(A))T \leq f(\phi^{-1})(A)$$
$$T^*(f \circ \phi^{-1})(A)T \leq (f \circ \phi^{-1})(A).$$

Hence $f \circ \phi^{-1} \in C_n((0,\infty))$ from the definition, and we know $f \in C_n([0,1))$ from [1, Corollary 2.4] and the definition.

(1) $\to$ (2):

Let h be a Pick funtion on $(0, 1)$. Then $h \circ \phi^{-1}$ is one on $(0, \infty)$. Then since there is a positive Radon measure on $[0, \infty]$ such that

$$h \circ \phi^{-1}(\lambda) = \int_{[0,\infty]} \frac{(1+t)\lambda}{1+t\lambda} d\rho,\ \lambda > 0,$$

we have

$$h(\lambda) = \int_{[0,\infty]} \frac{(1+t)\lambda}{1+(t-1)\lambda} d\rho,\ \lambda \in (0, 1).$$

For $\{\lambda_i\}_{i=1}^n \subset (0, 1)$ suppose that

$$\sum_{i=1}^n a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i} \geq 0$$

for any $\{a_i\}_{i=1}^n \subset \mathbf{R}$. Since there is a Pick function on $(0, 1)$ such that $f(\lambda_i) = h(\lambda_i)$ for $1 \leq i \leq n$,

$$\begin{aligned}\sum_{i=1}^n a_i f(\lambda_i) &= \sum_{i=1}^n \int_{[0,\infty]} a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i} d\rho \\ &= \int_{[0,\infty]} \sum_{i=1}^n a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i} d\rho \geq 0\end{aligned}$$

(2) $\to$ (1):

Take $\{\lambda_i\}_{i=1}^n$ in $(0, 1)$ and fix them. Set $A = C_{\mathbf{R}}[0, \infty]$ and

$$G = \{g: [0, \infty] \to \mathbf{R} \mid g(t) = \sum_{i=1}^n a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i}, \{a_i\}_{i=1}^n \subset \mathbf{R}\}.$$

Here A is a Banach space with respect to a norm $||k|| = \sup_{t\in[0,\infty]} |k(t)|$.

Then G is a linear subspace of A. Let $\ell: G \to \mathbf{R}$ be a linear functional defined by

$$\ell(\sum_{i=1}^n a_i \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i}) = \sum_{i=1}^n a_i f(\lambda_i).$$

Then ℓ is positive from the assumption. Note that for any $\lambda \in (0, 1)$ we have

$$\min_{t\in[0,\infty]} \frac{(1+t)\lambda}{1+(t-1)\lambda} > 0.$$

Take $c > 0$ such that $c\frac{(1+t)\lambda_1}{1+(t-1)\lambda_1} \geq 1$ and $t > 0$, and set $g_0(t) = c\frac{(1+t)\lambda_1}{1+(t-1)\lambda_1}$.

Define $m: G \to \mathbf{R}$ by $m(g) = \sup\{g(t) \mid t \in [0, \infty]\}$.

We will show that

$$\ell(g) \leq ||g + h||_{\ell(g_0)},\ \forall h \in C_{\mathbf{R}}[0, \infty]_+,$$

where $||k||_{\ell(g_0)} = ||k||\ell(g_0)$ and $C_{\mathbf{R}}[0, \infty]_+$ denotes a set of all positive functions in $C_{\mathbf{R}}[0, \infty]$.

For any $g \in G$ $m(g) < 0$ or $m(g) \geq 0$. If $m(g) < 0$, $g(t) < 0$ for any $t \in [0, \infty]$, and

$$\ell(g) < 0 \leq ||g + h||_{\ell(g_0)},\ \forall h \in C_{\mathbf{R}}[0, \infty]_+.$$

If $m(g) \geq 0$, we have

$$\begin{aligned} g(t) &\leq m(g) \leq m(g)1 \\ &\leq m(g)g_0 \\ \ell(g) &\leq m(g + h)\ell(g_0) \\ &\leq ||g + h||\ell(g_0) = ||g + h||_{\ell(g_0)},\ \forall h \in C_{\mathbf{R}}[0, \infty]_+. \end{aligned}$$

By Sparr's theorem [13, Lemma 2] there is a positive linear functional $L: C_{\mathbf{R}}[0, \infty] \to \mathbf{R}$ such that

$$\begin{aligned} L(k) &\geq 0,\ \forall k \in C_{\mathbf{R}}[0, \infty]_+ \\ L(h) &\leq ||h||_{\ell(g_0)},\ \forall h \in C_{\mathbf{R}}[0, \infty]. \end{aligned}$$

From the Riesz representation theorem there is a positive Radon measure ρ on $[0, \infty]$ such that

$$L(k) = \int_{[0,\infty]} k(t)d\rho(t),\ k \in C_R[0, \infty].$$

Set $g_i(t) = \frac{(1+t)\lambda_i}{1+(t-1)\lambda_i}$ for $1 \leq i \leq n$. Then we have

$$\begin{aligned} f(\lambda_i) &= \ell(g_i) \\ &= L(g_i) \\ &= \int_{[0,\infty]} \frac{(1 + t)\lambda_i}{1 + (t - 1)\lambda_i} d\rho(t) \\ &= h(\lambda_i) \end{aligned}$$

for $1 \leq i \leq n$ and a Pick function

$$h(\lambda) = \int_{[0,\infty]} \frac{(1 + t)\lambda}{1 + (t - 1)\lambda} d\rho(t).$$

□

The following a partial answer to [1, conjecture].

Proposition 3. *For each* $n \in \mathbf{N}$ $C_{2n}([0,\infty)) \subsetneq P_n^+([0,\infty))$.

Proof. Take $n \in \mathbf{N}$ and consider a gap function $g_n \in P_n^+([0,\alpha_n])$ for some $\alpha_n > 0$:

$$g_n(x) = x + \frac{1}{3}x^3 + \cdots + \frac{1}{2n-1}x^{2n-1}$$

Suppose that $g_n \in C_{2n}([0,\alpha_n])$. Take a set $S \subset (0,\alpha_n)$ of $2n$ numbers and take a subset $S' \subset S$ with $|S'| = 2n-1$. Since $g_n \in C_{2n}$, there is a Pick function of ϕ which are equal at points of S. Then ϕ and g_n are equal at points of S'.

Then in [4, XIV Theorem 3] since g_n does not satisfy condition $(i), (ii)$ (See [7].), ϕ and g_n are equal only at points of S'. But this is a contradiction to the fact that ϕ and g_n are equal at $S \supsetneq S'$.

Hence $g_n \notin C_{2n}([0,\alpha_n])$. Using an operator monotone function $h(t) = \frac{t}{\alpha_n - t} : [0,\alpha_n) \to [0,\infty)$. We know that $g_n \circ h^{-1} \in P_n^+([0,\infty))$, but $g_n \circ h^{-1} \notin C_{2n}(0,\infty))$. □

Acknowledgements Research partially supported by Ritsumeikan Rsearch Proposal Grant, Ritsumeikan University 2007–2008. The authors also are grateful to the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) and to Lund University for support and hospitality during their visits to Lund University where parts of this research have been performed.

References

1. Ameur, Y., Kaijser, S., Silvestrov, S.: Interpolation classes and matrix monotone functions, J. Operator Theory **57**(2), 409–427 (2007)
2. Bhatia, R.: Matrix Analysis, Graduate Texts in Mathematics. Springer, New York (1997)
3. Curto, R., Fialkow, L.A.: Recursiveness, positivity, and truncated moment problems, Houston J. Math. **17**(4), 603–635 (1991)
4. Donoghue, W.F.: Monotone Matrix Function and Analytic Continuation. Springer, Berlin (1874)
5. Hansen, F.: An operator inequality. Math. Ann. **246**, 249–250 (1980)
6. Hansen, F., Pedersen, G.K.: Jensen's inequality for operators and Loewner's theorem. Math. Ann. **258**, 229–241 (1982)
7. Hansen, F., Ji, G., Tomiyama, J.: Gaps between classes of matrix monotone functions, Bull. Lond. Math. Soc. **36**, 53–58 (2004)
8. Hiai, F., Yanagi, K.: Hilbert spaces and Linear Operators, Makino Publications (1995) (Japanense)
9. Horn, R.A., Johnson, C.R.: Topics in Matrix Analysis. Cambridge University Press, Cambridge (1991)
10. Loewner, K.: *Über* monotone Matrixfunktionen, Math. Z. **38**, 177–216 (1934)
11. Mathias, R.: Concavity of monotone matrix functions of finite order, Linear Multilinear Algebra **27**, 129–138 (1990)
12. Osaka, H., Silvestrov, S., Tomiyama, J.: Monotone operator functions, gaps and power moment problem. Math. Scand. **100**(1), 161–183 (2007)
13. Sparr, G.: A new proof of Löwner's theorem on monotone matrix functions. Math. Scand. **47**(2), 266–274 (1980)

Chapter 13
Interpolation of Normed Abelian Groups and Non-Commutative Integration

Sergei Silvestrov

Abstract This is a concise review concerned first of all with the two pioneering far reaching and in many ways yet to be fully explored important papers by Gunnar Sparr and Jaak Peetre on interpolation of normed abelian groups and on non-commutative integration. These papers introduced a general framework unifying many previously known interpolation results and methods in the ways applicable for non-commutative integration and non-commutative extensions of the function spaces, the directions of importance for example in non-commutative geometry and applications in quantum physics. Whence some notions and methods from these papers have been applied in various contexts, many other methods and ideas are yet to be discovered and developed further. In addition to the concise review of these important works by Jaak Peetre and Gunnar Sparr, a brief review is presented also of some related works on non-commutative spaces and non-commutative integration in the contexts of the theory of operator algebras and non-commutative geometry.

13.1 Introduction

Interpolation theory has strong tradition in Lund University with many pivotal contributions to the subject and its applications by several generations of mathematicians since 1960th which have profoundly influenced the development of the area both nationally and internationally and made it one of the pearls of modern functional analysis (see [5–7, 41, 42, 48] and references their).

S. Silvestrov (✉)
Centre for Mathematical Sciences, Lund University, Box 118, 22100 Lund, Sweden

Division of Applied Mathematics, The School of Education, Culture and Communication, Mälardalen University, Box 883, 72123 Västerås, Sweden
e-mail: sergei.silvestrov@math.lth.se; sergei.silvestrov@mdh.se

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0_13,

The development of a broad interplay of interpolation theory with the theory of operator algebras with emphasize on the non-commutative integration, non-commutative function spaces and their applications in non-commutative geometry is an excellent strategic long term project with many open research directions extending the state of the art in these areas of mathematics and in related developments in quantum physics, scientific computing and engineering. This line of research fits timewise with maturity of the subjects and aims at significantly strengthening conceptual and analytic base for unified treatment of broader classes of quantum and engineering systems, improvement and unification of approximation methods such as splines and rational approximation and new classes of inequalities on the levels of functions, matrices and the more general level of non-commutative operator algebras and operator spaces.

Among the potential applications of this research are the better error correction and computation algorithms and entropy estimates in quantum information processing and quantum computing based on norm inequalities for matrix and operator factorizations, in-depth unified analysis of symmetries in dynamical spin and lattice models and further development of fundamentals of quantum field theory based on non-commutative operator algebras and operator theory, perturbation analysis and dynamics, and more efficient approximation and transformation technics in geometric and operator modeling of linear and nonlinear input-output complex control systems with further links and perspectives for applications in for example signal and image processing, numerical analysis and optimization.

The fundamentally important part of this research is the development of the unified interpolation methods applicable to general non-commutative quasinormed, topologial and algebraic spaces by extending interpolation theory of quasinormed abelian groups developed in [41, 42] where such unified framework and methods were built using I. E. Segal gage spaces consisting of von Neumann algebras equipped with gage functionals such as traces, or general weights, states and dimension-like functionals. These functionals on non-commutative von Neumann algebras serve as replacement in non-commutative context of measures and integration for commutative spaces and functions. The methods involved are based to a large extent on the order structure and projections in von-Neumann algebras. As in case of commutative spaces, the main problems involve constructions of non-commutative generalizations of important function spaces, such as L_p, Lorentz $L_{p,q}$, Besov, Sobolev, Orlich, bounded variations and measure spaces, and on continuous side spaces of continuous, smooth and holomorphic functions with various important extra properties related to convexity or monotonicity. In order to be able to apply these non-commutative spaces to models of quantum physics it is desirable to define them in such way that allows to extend to non-commutative context main duality tools such as Fourier, Weyl and related transforms so that the main properties and methodological essence of these transforms is extended too.

For integration of functions on commutative measure spaces the most fundamental parametric family of function spaces extending the L_2 Hilbert space are L_p spaces, that is the spaces of power integrable functions. The L_p spaces are important for applications of Fourier and related integral transforms, since such transforms

move functions between such spaces (or their subspaces and dual spaces). Establishing when this happens in a continuous way with explicit inequalities estimations for the norms is of central importance for efficient convergence and stability of numerical iteration algorithms involving these transforms and for harmonic and smoothness analysis of solutions and discretization methods for differential and integral equations of mathematical physics. The most general and efficient methods for establishing and proving such mapping and continuity results between families of spaces and corresponding norm estimates are based on the ideas and methods of interpolation theory combined with convexity and monotonicity results for functions or functionals on such function spaces and their duals.

In [41], it was shown that approximation and interpolation spaces are very well suited for describing construction of L_p and Lorentz spaces, L_p spaces with weights and Besov spaces, and have interesting applications to approximation with rational functions and with spline functions. For example, it was observed in [41] that using the approximation space (quasinormed abelian group) between the spaces (quasinormed abelian groups) L_0 and L_∞ of measurable functions on a domain of $\mathbb{R}^d$ with L_p quasi-norms in the limits $p \to 0$ and $p \to \infty$, one gets the Lorentz space (see e.g. [7] for definition) $(L_0, L_\infty)_{\frac{1}{p},q;E} = L_{p,q}$. Therefore, for $p > 0, q > 0$, the Lorentz spaces $L_{p,q}$ can be also obtained by K method of interpolation for interpolation couple $\{L_0, L_\infty\}$ ([5, theorems 7.2.2, 7.3.3] and [41, Theorems 6.1, 6.2]).

Similarly, using the general framework of E-functionals and approximation spaces, the approximation and interpolation constructions have been obtained in [41] for weighted L_p spaces, Besov spaces and it was indicated how the methods can be applied to Sobolev spaces, and other related spaces. Moreover, in [41], it was shown that approximation and interpolation of quasinormed abelian groups using E method can be used as a framework for approximation of functions. In [41], the application of E method approximation (and interpolation) spaces have been especially detailed for pairs (A_0, C) where A_0 is a quasinormed abelian group of functions and C is the space of continuous functions on a compact interval $D = [a, b] \subset \mathbb{R}$. The applications to the cases when A_0 are rational functions and A_0 are spline functions, having important applications in numerical analysis and engineering, in [41] were emphasized in connection to the approximation and interpolation between spaces of functions of bounded variation, Sobolev spaces, Lipschitz functions and continuous functions.

Furthermore, in [41, Sect. 7] it was demonstrated that the framework of E-functional approximation spaces incorporates also spaces of linear and non-linear operators providing important insights in their structure and interplay. For example for quasinormed abelian couple $\{\mathfrak{S}_0, \mathfrak{S}_\infty\}$ where $\mathfrak{S}_0 = \mathfrak{S}(A, B)$ is the space of all linear operators of finite rank between Banach spaces A and B with the norm $\|T\|_{\mathfrak{S}_0} = \operatorname{rank} T = \dim_B T(A)$, and $\mathfrak{S}_\infty$ is the space of all bounded linear operators from A to B with the usual operator norm $\|T\|_{\mathfrak{S}_\infty} = \sup_{\|a\|_A \le 1} \|Ta\|_B$, the E-functional

$$E = E(t, T; \{\mathfrak{S}_0, \mathfrak{S}_\infty\}) = \inf_{\|S\|_{\mathfrak{S}_0} \le t} \|T - S\|_{\mathfrak{S}_\infty}$$

is intimately related to the approximation (singular, Schmidt) numbers s_n of operator T (in the notation of Gohberg-Krein book [11] for Hilbert space case):

$$s_{n+1}(T) = E(t, T; \{\mathfrak{S}_0, \mathfrak{S}_\infty\}), \text{ for } \quad t = n \text{ integer}.$$

One of the typical basic results often used in this context for instance asserts that restriction of Fourier transform to functions from $L_1(G) \cap L_2(G)$, $1 < p < 2$ on a compact (or locally compact) commutative group G with Haar measure can be continued to linear operator with norm ≤ 1 mapping $L_p(G)$ to $L_q(\hat{G})$ where $q = \frac{p}{p-1}$. In particular $q = 2$ when $p = 2$ recovering classical Fourier transform on the Hilbert space of square integrable functions L_2. The proof of this and similar more general duality statements is based on the beautiful idea of interpolation and convexity arguments involving L_q-norms as functions of $\frac{1}{p}$.

J. Peetre and G. Sparr in their fundamental article [41] on interpolation of normed abelian groups among other things raised a natural question of whether interpolation of the usual L_p spaces over a measure space and interpolation of the trace classes $\mathfrak{S}_p$ of compact operators in a Hilbert space can be treated within the same framework.

In the followup article J. Peetre and G. Sparr [42] have answered this question in affirmative by showing that both cases can be treated within the same framework if one combines interpolation with the theory of non-commutative integration and the corresponding theory of non-commutative L_p spaces over a (regular) gage space developed by I. Segal [47], R. A. Kunze [28] and W. Stinespring [50].

These general methods discovered in [41], [42] have appeared well ahead of time and have strong potential for further expansion and applications. The general interpolation and approximation approach developed by Sparr and Peetre in [41] and [42] is very fruitful and can be applied in many at first seemingly unrelated contexts. For example, it was shown in [42], that the interpolation theory of [41] can serve as a unified framework for non-commutative L_p spaces, various transforms as for example the Fourier (-Segal) transform on unimodular groups improving results of Kunze [28], the Weyl transform for finite and infinite number of dimensions using formalism of (boson or fermion) Fock spaces of interest from the point of view of quantum field theory, and the spinor transform generalizing results by Lavine [30] and R. F. Streater [51] involved deeply with Lie algebras and Lie groups, Clifford algebras, Gelfand-Naimark-Segal (GNS) construction and central states in operator algebras (von Neumann algebras, C^*-algebras). In [42], it was also remarked that in all these cases the gage spaces involved are of the two simplest types and one can thus use [41] directly, thus emphasizing that new important advances might be achieved when the general approach of [42] is applied to other gage or gage-like spaces combined with developing interpolation methods for other algebraic and topological structures than normed linear spaces following in footsteps of [41].

The main ingredients of the general approach in [41] and [42], in addition to the use of general gage spaces of Segal, include general K, J and E-functionals and generalized quasi-norms on groups, decreasing rearrangements, interpolation functors and retracts in categories, spectral decompositions and duality for spaces, operators and functors.

13.2 Interpolation of Normed and Quasinormed Abelian Groups

Following [41], a (c,ρ)-normed abelian group, where $1 \le c < \infty$, $0 < \rho \le \infty$, is an abelian group A in which there is defined a (c,ρ)-norm, i.e. a function

$$A \ni a \mapsto ||a|| = ||a||_A \in \mathbb{R}_+ = [0,\infty)$$

obeying the axioms:

1. $||a|| = 0 \Leftrightarrow a = 0$

2. $||-a|| = ||a|| \Leftrightarrow a = 0$

3. $||a+b|| \le c(||a||^\rho + ||b||^\rho)^{\frac{1}{\rho}}$ ((c,ρ)-triangle inequality).

When $c = 1$, one speaks of ρ-norms and ρ-normed abelian groups, and when $\rho = 1$ one speaks of c-quasi norms and c-quasinormed abelian groups, or simply quasinorms and quasinormed abelian groups when there is no need to accentuate the value of c. A $(\overrightarrow{c},\overrightarrow{\rho})$-normed abelian couple, with $\overrightarrow{c} = \{c_0,c_1\}$, $\overrightarrow{\rho} = \{\rho_0,\rho_1\}$, is defined as a pair $\mathbf{A} = \{A_0,A_1\}$ where A_0 is (c_0,ρ_0)-normed and A_1 is (c_1,ρ_1)-normed abelian group and where both A_0 and A_1 are subgroups of some abelian group, satisfying the following axiom:

$$||a_n - a^{(0)}||_{A_0} \to 0, \quad ||a_n - a^{(1)}||_{A_1} \to 0 \quad \Rightarrow \quad a^{(0)} = a^{(1)}$$

If A_0 is ρ_0-normed and A_1 is ρ_1-normed (i.e. $c_0 = c_1 = 1$) one speaks of a $\overrightarrow{\rho}$-normed abelian couple, and if moreover $\rho_0 = \rho_1 = \rho$ of a ρ-normed abelian couple. If A_0 is c_0-quasinormed and A_1 is c_2-quasinormed (i.e. $\rho_0 = \rho_1 = 1$), then one speaks of $\overrightarrow{c}$-quasinormed abelian couple or, less precisely, of a quasinormed abelian couple.

In [41], Peetre and Sparr defined K-, J and E-functionals on every quasinormed abelian couple $\mathbf{A} = \{A_0,A_1\}$ (of quasi normed abelian groups) by

$$K_p(t,a:\mathbf{A}) = \inf_{a=a_0+a_1} (\|a_0\|_{A_0}^p + t^p\|a_1\|_{A_1}^p)^{\frac{1}{p}}, \text{ for } 0<t<\infty, a\in\Sigma(\mathbf{A}), 0<p\le\infty$$

$$J_p(t,a:\mathbf{A}) = (\|a\|_{A_0}^p + t^p\|a\|_{A_1}^p)^{\frac{1}{p}}, 0<t<\infty, \text{ for } a\in\Delta(\mathbf{A}), 0<p\le\infty$$

$$E(t,a:\mathbf{A}) = \inf_{\|a_0\|_{A_0}\le t} \|a-a_0\|_{A_1}, \text{ for } 0<t<\infty, a\in\Sigma(\mathbf{A})$$

where $\Delta(\mathbf{A}) = A_0 \cap A_1$ with the norm

$$\|a\|_{\Delta(\mathbf{A})} = J_\infty(1,a\,;\,\mathbf{A}) = \max(\|a\|_{A_0},\|a\|_{A_1})$$

and $\Sigma(\mathbf{A}) = A_0 + A_1$ is the hull of A_0 and A_1 with the norm

$$\|a\|_{\Sigma(\mathbf{A})} = K_1(1, a\ ;\ \mathbf{A}) = \inf_{a=a_0+a_1}(\|a_0\|_{A_0} + \|a_1\|_{A_1}).$$

If $\mathbf{A}$ is a ρ-normed abelian couple, then

$$K_p(t, a+b\ ;\ \mathbf{A}) \leq ((K_p(t, a\ ;\ \mathbf{A})^\rho + (K_p(t, b\ ;\ \mathbf{A})^\rho)^{\frac{1}{\rho}}$$

and the same for J-functional by Jessen's and ρ-triangle inequalities. This together with separability axiom implies that $K_p(t, a\ ;\ \mathbf{A})$ is a ρ-norm in $\Sigma(\mathbf{A})$ and $J_p(t, a\ ;\ \mathbf{A})$ is a ρ-norm in $\Delta(\mathbf{A})$ provided $p \geq \rho$.

If $\mathbf{A}$ is a $\overrightarrow{c}$-quasinormed abelian couple ($\overrightarrow{c} = \{c_0, c_1\}$), then $K_1(t, a\ ;\ \mathbf{A})$ is a c-quasi-norm in $\Sigma(\mathbf{A})$, and $J_\infty(t, a\ ;\ \mathbf{A})$ is a c-quasi-norm in $\Delta(\mathbf{A})$, with $c = \max(c_0, c_1)$. The c-quasi triangle inequality

$$K_1(t, a+b\ ;\ \mathbf{A}) \leq c(K_1(t, a\ ;\ \mathbf{A}) + K_1(t, b\ ;\ \mathbf{A}))$$

(and similarly for J-functional) follows from more exact inequalities

$$K_1(t, a+b\ ;\ \mathbf{A}) \leq c_0(K_1(\frac{c_1}{c_0}t, a\ ;\ \mathbf{A}) + K_1(\frac{c_1}{c_0}t, b\ ;\ \mathbf{A}))$$

$$J_\infty(t, a+b\ ;\ \mathbf{A}) \leq c_0(J_\infty(\frac{c_1}{c_0}t, a\ ;\ \mathbf{A}) + J_\infty(\frac{c_1}{c_0}t, b\ ;\ \mathbf{A}))$$

The E-functional, instead of norm properties, has following subadditivity properties. If $\mathbf{A}$ is a $\overrightarrow{c}$-quasinormed abelian couple, $\overrightarrow{c} = \{c_0, c_1\}$, then

$$E(s+t, a+b\ ;\ \mathbf{A}) \leq c_1(E(\frac{s}{c_0}, a\ ;\ \mathbf{A}) + E(\frac{t}{c_0}, b\ ;\ \mathbf{A})),$$

and if $\mathbf{A}$ is a $\overrightarrow{\rho}$-normed abelian couple, $\overrightarrow{\rho} = \{\rho_0, \rho_1\}$, then

$$E((s^{\rho_0} + t^{\rho_0})^{\frac{1}{\rho_0}}, a+b\ ;\ \mathbf{A}) \leq (E(s, a\ ;\ \mathbf{A})^{\rho_1} + E(t, b\ ;\ \mathbf{A})^{\rho_1})^{\frac{1}{\rho_1}}.$$

For $\rho_0 = \infty$ this inequality and the fact that E is a decreasing function of t yield

$$E(t, a+b\ ;\ \mathbf{A}) \leq (E(s, a\ ;\ \mathbf{A})^{\rho_1} + E(t, b\ ;\ \mathbf{A})^{\rho_1})^{\frac{1}{\rho_1}}$$

since E is a decreasing function of t, meaning that $E(t, a\ ;\ \mathbf{A})$ satisfies the ρ_1-triangle inequality in this case.

In the most of the works up to date, the broad use has been made of K and J-functionals as a powerful tool for construction of interpolation functors and interpolation spaces and for proving various old and new inequalities in a unified way.

In [41, 42] important methodological point was made that the E-functional has fundamental significance and can be used as a general unifying tool for interpolation, inequalities and approximation. For example, in quasinormed abelian couples the K-functional can be always expressed through E-functional:

$$K_p(t) = \inf_s (s^p + t^p (E(s))^p)^{\frac{1}{p}}.$$

For $p = \infty$, if $E(t)$ is continuous then $K_\infty(t)$ is the inverse of the function $t/E(t)$. For $p = 1$, $K_1(t) = \inf_s (s + tE(s))$ and it is possible at least partially to express E via K:

$$E^*(t) = \sup_s (\frac{K_1(s)}{s} - \frac{t}{s}) \leq E(t) \leq \frac{1}{1-\alpha} E^*(\alpha t), \quad (0 < \alpha < 1)$$

where E^* is the greatest convex minorant of E, and $E^* = E$ if E is convex. In [41], further relations between K-functional and E-functional and related references are described. Relations between E, K, and J-functionals and interpolation methods is one of the important themes and tools in interpolation theory addressed for instance in books by J. Bergh, J. Löfström [5] and Yu. A. Brudnyi, N. A. Krugljak [6]

The distance-like way E-functional is defined suggests its importance for investigation of the relative geometry and topology of the involved spaces, mutual position for these spaces and their intersections and subspaces within the interpolation pairs and triples, and approximation and interpolation of functions. An interpolation functor or interpolation method $\mathfrak{G}$ from category of quasinormed abelian couples to category of quasinormed abelian groups associates to each couple $\mathbf{A} = \{A_0, A_1\}$ a quasinormed abelian group $\mathfrak{G}(\mathbf{A})$ called interpolation group (or interpolation space) so that

$$\Delta(\mathbf{A}) := A_0 \cap A_1 \subseteq \mathfrak{G}(\mathbf{A}) \subseteq \Sigma(\mathbf{A}) := A_1 + A_0$$

where the inclusions are natural. For any two quasinormed abelian couples $\mathbf{A}$ and $\mathbf{B}$ and quasinormed abelian groups A and B such that $A \subseteq \mathfrak{G}(\mathbf{A})$ and $\mathfrak{G}(\mathbf{B}) \subseteq B$, the interpolation property holds:

$$T : \mathbf{A} \to \mathbf{B} \quad \Rightarrow \quad T : A \to B$$

where notation $T : \mathbf{A} \to \mathbf{B}$ means that T is a homomorphism of quasinormed abelian couples and $T : A \to B$ means that T is a homomorphism of quasinormed abelian groups [41]. By the direct extension of Aronszajn-Gagliardo theorem to this general context, the converse statement also holds, that is in brief, the interpolation property implies the existence of the interpolation group [2, 41].

Let $\mathfrak{L}(X, Y)$ denote the quasinormed abelian group of all bounded homomorphisms between quasinormed abelian groups X and Y with quasi-norm

$$\|T\|_{\mathfrak{L}(A,B)} = \inf\{C \geq 0 \mid \|Ta\|_B \leq C\|a\|_A\}.$$

An interpolation functor is said to be bounded if there exists a constant C such that

$$\|T\|_{\mathfrak{L}(A,B)} \leq C\,\|T\|_{\mathfrak{L}(\mathbf{A},\mathbf{B})} := max(\|T\|_{\mathfrak{L}(A_0,B_0)}, \|T\|_{\mathfrak{L}(A_1,B_1)})$$

provided the inclusions in $A \subseteq \mathfrak{G}(\mathbf{A})$ and $\mathfrak{G}(\mathbf{B}) \subseteq B$ have norm ≤ 1. If moreover $C = 1$, then the interpolation functor is called exact.

The references [6, 25, 40, 59] are recommended for further reading on categories and functors in interpolation. (see also [9, 17–20])

In [41], a family of interpolation spaces has been introduced using the family of functionals $\Phi_{\theta,q}, -\infty < \theta < \infty, -\infty < q \leq \infty$ with values in $\overline{\mathbb{R}_+} = [0,\infty]$ defined for all positive measurable functions on $(0,\infty)$ by

$$\Phi_{\theta,q}[\phi] = (\int_0^\infty (t^{-\theta}\phi(t))^q \frac{dt}{t})^{\frac{1}{q}}.$$

The set $\mathbf{A}_{\theta,q;K}$, where $0 < \theta < 1, 0 < q \leq \infty$ (or $\theta = 0, q = \infty$ or $\theta = 1, q = \infty$), consists of all $a \in \Sigma(\mathbf{A})$ for which

$$\|a\|_{\mathbf{A}_{\theta,q;K}} = \Phi_{\theta,q}[(K_1(t;a)] < \infty.$$

In [41] it was shown that $\mathbf{A}_{\theta,q;K}$ is a (c,q^*)-normed abelian group with $q^* = \min(1,q)$, $c = c_0^{1-\theta}c_1^\theta$ and the (c,q^*)-norm $\|\cdot\|_{\mathbf{A}_{\theta,q;K}}$ (and consequently also a quasinormed abelian group); and the correspondence $\mathbf{A} \mapsto \mathbf{A}_{\theta,q;K}$ is an exact interpolation functor. More precisely, $T : \mathbf{A} \to \mathbf{B} \Rightarrow T : A \to B$ holds for any quasinormed abelian couples $\mathbf{A} = \{A_0, A_1\}$ and $\mathbf{B} = \{B_0, B_1\}$ with $A = \mathbf{A}_{\theta,q;K}$ and $B = \mathbf{B}_{\theta,q;K}$, and the following convexity inequality holds

$$\|T\|_{\mathfrak{L}(A,B)} \leq \|T\|^{1-\theta}_{\mathfrak{L}(A_0,B_0)}\|T\|^{\theta}_{\mathfrak{L}(A_1,B_1)}.$$

implying exactness. Also, in [41], interpolation functor (method) and interpolation spaces $\mathbf{A}_{\theta,q;J}$ associated with J-functional where defined, and the equivalence theorem, $\mathbf{A}_{\theta,q;J} = \mathbf{A}_{\theta,q;K}$ if $0 < \theta < 1, 0 < q \leq \infty$, was proved via equivalence of norms, thus allowing to drop K and J from the notations $\mathbf{A}_{\theta,q}$.

For the E-functional, $\mathbf{A}_{\alpha,q;E}$ consists of all $a \in \Sigma(\mathbf{A})$ for which

$$\|a\|_{\mathbf{A}_{\alpha,q;E}} = \Phi_{-\alpha,q}[(E(t;a)] < \infty, \quad 0 < \alpha < \infty, 0 < r \leq \infty \quad (\text{or } \alpha = 0, r = \infty).$$

In [41] it was proved that if $\mathbf{A}$ is a $\overrightarrow{c}$-normed abelian couple with $\overrightarrow{c} = \{c_0, c_1\}$, then $a \mapsto \|a\|_{\mathbf{A}_{\alpha,q;E}}$ defines a (c,q^*)-norm in $\mathbf{A}_{\alpha,q;E}$ with $q^* = \min(1,q)$ and $c = c_1c_0^\alpha$. The interpolation spaces obtained by K and hence also J methods can be expressed via approximation spaces $\mathbf{A}_{\theta,q;K} = \mathbf{A}^{[\theta]}_{\frac{1}{\theta}-1,\theta q;E}$ where $A^{[\theta]}$ is the original abelian group A with the new quasi-norm equal to the θ-th power of the original quasi-norm [41, Theorem 5.10]. Moreover, $(\mathbf{A}_{\alpha_0,r_0;E}, \mathbf{A}_{\alpha_1,r_1;E})_{\theta,q;K} = \mathbf{A}_{(1-\theta)\alpha_0+\theta\alpha_1,q;E}$ for $\alpha_0 \neq \alpha_1$ (see [41, Theorem 5.11] or [5, Theorem 7.1.8]). The spaces $\mathbf{A}_{\alpha,q;E}$

are called approximation spaces due to their important connection to approximation of functions (see [41] and [5]). In [41], it is explained how L_p spaces, L_p spaces with weights, Lorentz spaces $L_{p,q}$, Besov spaces can be expressed as interpolation and as approximation spaces and connection of interpolation and approximation spaces to approximation by rational functions and by spline functions and to spaces of functions of bounded variation and the spaces $Lip(\alpha, p)$ defined by Lipschitz condition of exponent α in the L_p metric.

The interpolation and approximation spaces defined in [41] can be naturally applied to spaces (abelian groups) of linear, or sometimes non-linear operators from one Banach space to another.

Let $\{\mathfrak{S}_0, \mathfrak{S}_\infty\}$ be the 1-normed abelian couple linear operators of finite rank $\mathfrak{S}_0 = \mathfrak{S}_0(A, B)$ and bounded linear operators $\mathfrak{S}_\infty = \mathfrak{S}_\infty(A, B) = \mathfrak{L}_\infty(A, B)$ with the norms defined by $\|T\|_{\mathfrak{S}_0} = \operatorname{rank} T = \dim_B T(A)$ and $\|T\|_{\mathfrak{S}_\infty} = \sup_{\|a\|_A \le 1} \|Ta\|_B$. The E-functional for this abelian couple is

$$E(t, T; \{\mathfrak{S}_0, \mathfrak{S}_\infty\}) = \inf_{\|S\|_{\mathfrak{S}_0} \le t} \|T - S\|_{\mathfrak{S}_\infty} = \inf_{\|S\|_{\mathfrak{S}_0} \le t} \sup_{\|a\|_A \le 1} \|Ta - Sa\|_B.$$

One of the important facts linking E-functional and thus interpolation and approximation spaces to operator theory and spectral theory is the fact that approximation (singular, Schmidt) numbers s_n of operator T (in the notation of Gohberg-Krein book [11] for Hilbert space case) can be expressed as values of the E-functional at integer points for this abelian couple

$$s_{n+1}(T) = E(t, T; (\mathfrak{S}_0, \mathfrak{S}_\infty)), \quad \text{for} \quad t = n \text{ integer}.$$

Within the framework of E method the p-th trace class $\mathfrak{S}_p = \mathfrak{S}_p(A, B)$ can be defined (see [41]) as the group corresponding to the norm

$$\|T\|_{\mathfrak{S}_p} = \left(\int_0^\infty E(t, T; (\mathfrak{S}_0, \mathfrak{S}_\infty))^p dt\right)^{\frac{1}{p}}$$

and viewed as the approximation space obtained by E method and as interpolation space obtained by K method (see [41, Theorems 7.1, 7.2], [5] and for Hilbert space case [33, 57]):

$$\mathfrak{S}_p = (\mathfrak{S}_0, \mathfrak{S}_\infty)_{\frac{1}{p},p;E} = (\mathfrak{S}_0, \mathfrak{S}_\infty)^{[\frac{1}{\theta}]}_{\theta,p;K}, \qquad \theta = \frac{p}{p+1}; \tag{13.1}$$

$$\mathfrak{S}_p = (\mathfrak{S}_{p_0}, \mathfrak{S}_{p_1})_{\theta,p;K}, \qquad \frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}. \tag{13.2}$$

A closely related class of p-nuclear operators $\mathfrak{N}_p = \mathfrak{N}_p(A, B)$ consists of linear operators which can be represented in the form $Ta = \sum_{i=1}^\infty \lambda_i \langle a, a' \rangle b_i$ where $(\sum_{i=1}^\infty |\lambda_i|^p)^{\frac{1}{p}} < \infty$, $\|a_i\|_{A'} \le 1$, $\|b_i\|_B \le 1$, A' is the dual of A and $\langle \cdot, \cdot \rangle$

is the duality between A and A'. The norm in $\mathfrak{N}_p$ is defined by $\|T\|_{\mathfrak{N}_p} = \inf(\sum_{i=1}^{\infty} |\lambda_i|^p)^{\frac{1}{p}}$. When A and B are Hilbert spaces, $\mathfrak{N}_p = \mathfrak{S}_p$ (see [11] and [32]). For Banach spaces the inclusions $\mathfrak{N}_p \subseteq \mathfrak{S}_q$ and $\mathfrak{S}_p \subseteq \mathfrak{N}_p$ if $p < 1$ and $\frac{1}{q} > \frac{1}{p} - 1$ were proven in [41, Proposition 7.1]. For the product of operators the following inequalities hold: $\|TS\|_{\mathfrak{S}_0} \leq \min\{\|T\|_{\mathfrak{S}_0}, \|S\|_{\mathfrak{S}_0}\}$ and $\|TS\|_{\mathfrak{S}_\infty} \leq \|T\|_{\mathfrak{S}_\infty}\|S\|_{\mathfrak{S}_\infty}$ similarly to the product of functions $\|fg\|_{\mathfrak{L}_0} \leq \min\{\|f\|_{\mathfrak{L}_0}, \|g\|_{\mathfrak{L}_0}\}$ and $\|fg\|_{\mathfrak{L}_\infty} \leq \|f\|_{\mathfrak{L}_\infty}\|g\|_{\mathfrak{L}_\infty}$. The classes $\mathfrak{S}_p(A, B)$ and $\mathfrak{N}_p(A, B)$ are quasinormed ideals in the sense of Triebel (see [58], and for the normed ideal case [43, 44]). This indicates that the interpolation couples $\{\mathfrak{S}_0, \mathfrak{S}_\infty\}$ and $\{L_0, L_\infty\}$ (and it's discrete analogue $\{l_0, l_\infty\}$) are closely related not only as spaces but also as rings (algebras) with respect operator composition and pointwise multiplication of functions respectively. In this direction, substantial progress is made in operator theory and operator algebras in investigation of ideals of operators on Hilbert spaces. Many important problems and important aspects of approximation and interpolation methods for operator ideals in Banach algebras and quasi-Banach algebras in the context of Banach spaces are to large extent open research directions with many interesting potential applications in other parts of mathematics and in important Physics and Engineering problems.

As pointed out in [41], the framework of interpolation and approximation spaces of quasinormed abelian groups can be applied to non-linear operators as well. The space of all bounded linear operators $\mathfrak{L} = \mathfrak{L}(A, B)$ can be embedded isometrically in $\mathfrak{G} = \mathfrak{G}(A, B)$, the abelian group of mappings from the unit ball of A into B with 1-norm $\|T\|_{\mathfrak{G}(A,B)} = \sup_{\|a\|\leq 1} \|Ta\|_B$. For any couple $\{\mathscr{G}, \mathfrak{G}\}$ where $\mathscr{G}$ is any normed abelian group of operators, the E-functional is

$$E(t, T; \{\mathscr{G}, \mathfrak{G}\}) = \inf_{\|S\|_{\mathscr{G}} \leq t} \sup_{\|a\|_A \leq 1} \|Ta - Sa\|_B.$$

and can be used to define the corresponding approximation and interpolation spaces (groups). For example, if $\mathscr{G}$ are either operators with range of finite dimension or operators with range of finite cardinality then $(\mathscr{G}, \mathfrak{G})_{\alpha,\infty;E}$ are the diameter and entropy quasi-norm ideals of Triebel [58]. Another special case when $\mathfrak{G}$ is taken to be $\mathfrak{L}(A_0, B)$ is of interest in connection with the problem of best numerical differentiation [37, 39, 41, 53].

13.3 Interpolation and Non-commutative Integration

In [42], J. Peetre and G. Sparr made important progress in the direction of unification of interpolation and approximation for commutative and non-commutative spaces by developing a general framework combining interpolation with the theory of non-commutative integration and non-commutative L_p spaces over a (regular) gage space as developed by I. Segal [47], R. A. Kunze [28] and W. Stinespring [50]. The

general approximation and interpolation approach developed by Sparr and Peetre in [41] and [42] is very fruitful and can be applied in many at first seemingly unrelated contexts and serve as a unified framework for non-commutative L_p spaces, various transforms as for example Fourier (-Segal) transform on unimodular groups [28], the Weyl transform for finite and infinite number of dimensions using formalism of (boson or fermion) Fock spaces of interest from the point of view of quantum field theory, and the spinor transform generalizing [30], [51] involved deeply with Lie algebras and Lie groups, Clifford algebras, Gelfand-Naimark-Segal (GNS) construction and central states in operator algebras (von Neumann algebras, C^*-algebras). Another interesting paper, called "Non-commutative interpolation", by Peetre [38], concerned with interpolation spaces between a given Banach space and definition domains of operators on this space representing elements (Lie algebra generators or linear basis) in a Lie algebra. This paper might provide ideas and motivation for investigation of similar interpolation properties for definition domains of unbounded operators arising in connection to other important non-commutative algebras then Lie algebras.

Segal's gage spaces are triples $\Gamma = (\mathscr{H}, \mathscr{A}, m)$ where A is a von Neumann algebra, that is some weakly/strongly closed $*$-algebra of the algebra of bounded operators $\mathscr{L}(\mathscr{H})$ on a Hilbert space $\mathscr{H}$ and m is a (regular) gage on $\mathscr{A}$, that is a mapping m : projections in $\mathscr{A} \Rightarrow \mathbb{R}_+$ such that $m(P) = 0$ only if $P = 0$, m is additive on orthogonal sums (hulls) of projections, invariant under action of the group of unitaries $m(UPU^{-1}) = m(P)$ if $U^{-1} = U^*$ unitary, and every projection in $\mathscr{A}$ is a hull of m-finite projections. One can define classes of measurable operators on a gage space Γ and using spectral resolution $T = \int_0^\infty \lambda dP(\lambda)$ extend m to positive measurable operators by $m(T) = \int_0^\infty \lambda dm(P(\lambda))$. If $m(T) \leq \infty$ then T is called integrable and m can be extended by linearity to general integrable operators which belong to hull of positive integrable ones. Important examples of such gage spaces are numerous:

(a) (commutative space case) triples $(L_2(M), L_\infty(M), m)$ with elements in $L_\infty(M)$ identified with the corresponding multiplication operator on $L_2(M)$ where $M = (X, \mathscr{B}, m)$ consists of a measure m on a Boolean ring of subsets $\mathscr{B}$ of a space X;

(b) the triple $(\mathscr{H}, \mathscr{L}(\mathscr{H}), tr)$ where tr is ordinary (von Neumann) trace, in particular on projections $m(P) = tr(P) = rank(P)$ and $m(\infty) < \infty$ if and only if $rank(P) < \infty$;

(c) more general class of gage spaces $\Gamma = (\mathscr{H}, \mathscr{A}, m)$ consist of $\mathscr{H}$ being sections $f = \{f_x\}$ of a Hilbert space bundle $\mathscr{K} = \{\mathscr{K}_x\}_{x \in X}$ over a measure space $(X, \mathscr{B}, m)$ with $\int_X \|f_x\|^2 dx < \infty$, $\mathscr{A}$ being sections $T = \{T_x\}$ of the operator bundle $\{\mathscr{L}(\mathscr{K}_x)\}_{x \in X}$ with $\sup_{x \in X} \|T_x\| < \infty$, and $m(T) = \int_X tr(T_x) dm$;

(d) Gage spaces associated to groups $\hat{G} = \Gamma = (L_2(G), \mathscr{A}, m)$ where $L_2(G)$ is a Hilbert space of square integrable functions on a unimodular group provided with Haar measure dg, $\mathscr{A}$ is the von Neumann algebra generated by all left regular representation operators $\{L_g f(\cdot) = f(g^{-1}\cdot)\}_{g \in G}$, and $m(P) = \|f\|^2$ is a projection of the form $P = L_f : \phi \mapsto f * \phi = \int L_g f \phi(g) dg$;

(e) for compact groups one can take more explicitly $\hat{G}$ to be the space of all equivalence classes of irreducible unitary representations of G provided with the discrete measure m which for a corresponding to a point $x \in X$ irreducible unitary representation U^x on a (finite-dimensional) Hilbert space V^x associates the mass equal to $\dim V^x$.

In [42], it was remarked that all these cases of the gage spaces are of the simplest type which can be treated using [41] directly, but also that there are many other interesting spaces and algebraic structures where the methods of [42] and [41] may be applied, thus motivating further investigations of applications of the general approach of [42] to other gage or gage-like spaces combined with developing interpolation methods for other non-commutative algebraic and topological structures than normed linear spaces following in footsteps of [41].

Many interesting important extensions of the described classes of gage spaces are possible and important for applications of interpolation theory. For instance, in the case of gage spaces of operators or in general operator bundles, $\mathscr{L}(\{\mathscr{K}\}_x)_{x\in X}$, it is a direction of great interest to develop further the interpolation approach for bundles of other von-Neumann algebras provided proper extension of trace-like functionals can be defined (like for example Dixmier trace, center valued traces etc ...) [8]; and in case of gage spaces associated to groups extend it to non-unimodular groups and to semigroups in various ways.

In [42], for a general gage space $\Gamma = (\mathscr{H}, \mathscr{A}, m)$, the non-commutative $L_p = L_p(\Gamma)$ spaces are defined as consisting of measurable operators T with

$$\|T\|_{L_p} = \|T\|_{L_p(\Gamma)} = m(|T|^p)^{\frac{1}{p}} < \infty$$

where $|T|$ is a positive measurable operator equivalent to T (e.g. $|T| = \sqrt{TT^*}$). For positive operators

$$\|T\|_{L_p} = (\int_0^\infty \lambda^p dm(P(\lambda)))^{\frac{1}{p}}.$$

If $1 \leq p < \infty$, then $\|T\|_{L_p}$ is norm, but if $0 < p < 1$ it is only a quasi-norm. It can be shown that L_p is complete and so for $1 \leq p < \infty$ it is a Banach space and for $0 < p < 1$ it is a quasi-Banach space. In [42], with this non-commutative L_p was associated another analogous family of spaces

$$\tilde{L}_p = \tilde{L}_p(\Gamma) = L_p^{[\tilde{p}]}$$

$$\|T\|_{\tilde{L}_p} = \|T\|_{L_p}^{\tilde{p}}, \quad \tilde{p} = \min(1, p).$$

For $1 \leq p < \infty$ nothing changed $\tilde{L}_p = L_p$. However if $0 < p < 1$ then $\tilde{L}_p$ is a quasinormed abelian group (p-normed vector space) thus requiring the framework of interpolation of quasinormed abelian groups developed in [41]. The spaces $\tilde{L}_0$

and $\tilde{L}_\infty$ are obtained by passing to the limits $p \to 0$ and $p \to \infty$, and so $\tilde{L}_0$ is the space corresponding to the 0-norm $\|T\|_{\tilde{L}_0} = m(\mathrm{supp}T)$ where $\mathrm{supp}T$ is the smallest projection $P \in \mathscr{A}$ such that $PT = T$; and $\tilde{L}_\infty = L_\infty$ is the space corresponding to $\|T\|_{\tilde{L}_\infty} = \|T\|_{L_\infty} = \|T\|_{\mathscr{L}(\mathscr{H})}$ the restriction to $\mathscr{A}$ of the norm in $\mathscr{L}(\mathscr{H})$. In special cases, these spaces give the usual commutative L_p space as well as trace classes $\mathfrak{S}_p(\mathscr{H})$ of operators and also $\tilde{\mathfrak{S}}_p$.

Interpolation of these non-commutative L_p and $\tilde{L}_p$ spaces for arbitrary gage spaces $\Gamma = (\mathscr{H}, \mathscr{A}, m)$ can be naturally developed within the framework of interpolation of quasinormed abelian groups from [41]. For a quasinormed abelian couple $\{\tilde{L}_0, \tilde{L}_\infty\}$ the E-functional of $T \in \mathscr{A}$

$$T^\star(t) = E(t, T; (\tilde{L}_0, \tilde{L}_\infty)) = \inf_{\|S\|_{\tilde{L}_0} \le t} \|T - S\|_{\tilde{L}_\infty},$$

the decreasing rearrangement of T, satisfies the very useful important inequality

$$(T_1 + T_2)^\star(t_1 + t_2) \le T_1^\star(t_1) + T_2^\star(t_2).$$

In [41], non-commutative Lorentz spaces are defined analogously with the commutative situation, as approximation spaces $L_{p,q}(\Gamma) = L_{p,q} = (\tilde{L}_0, \tilde{L}_\infty)_{\alpha,q;E}$ with $\alpha = \frac{1}{p}$ meaning that $T \in L_{p,q}$ if and only if

$$\|T\|_{L_{p,q}} = \Big(\int_0^\infty (t^{\frac{1}{p}} T^\star(t))^q \frac{dt}{t}\Big)^{\frac{1}{q}} < \infty.$$

For $p = q$ this becomes

$$\Big(\int_0^\infty (T^\star(t))^p \frac{dt}{t}\Big)^{\frac{1}{p}} < \infty.$$

Since the decreasing rearrangement $T^\star(t)$ as function of t is inverse of the function $\lambda \mapsto m(P(\lambda))$ where $\lambda \mapsto P(\lambda)$ is the spectral resolution of the operator $|T|$, the previous inequality simply is equivalent to $\int_0^\infty \lambda^p d\, m(P(\lambda)) < \infty$, which means that $L_{p,q} = L_p$ if $p = q$.

By [41], the Lorentz spaces can be also expressed as interpolation space by K method [42, Theorems 2.1, 2.2]:

$$L_{p,r}(\Gamma) = L_{p,r} = (\tilde{L}_0, \tilde{L}_\infty)^{[\frac{1}{\theta}]}_{\theta,q;K}, \quad \theta = \frac{p}{p+1}, r = \theta q.$$

and also by reiteration theorem for K method [41, Theorem 5.11],

$$L_{p,q} = (L_{p_0,r_0}, L_{p_1,r_1})_{\theta,q;K}, \qquad L_p = (L_{p_0}, L_{p_1})_{\theta,q;K}, \quad \frac{1}{p} = \frac{1-\theta}{p_0} + \frac{\theta}{p_1}.$$

Therefore the spaces $L_{p,q}$ and in particular L_p can be reconstructed from the couple $\{\tilde{L}_0, \tilde{L}_\infty\}$ via interpolation. For the Banach couple $\{L_1, L_\infty\}$, the following important general formula connecting K-functional and decreasing rearrangement holds:

$$K(t, T; \{L_1, L_\infty\}) = \int_0^t T^\star(s)ds.$$

Using this representation of K-functional one can show that

$$(L_1, L_\infty)_{\theta,q;K} = L_{p,q}, \quad (L_1, L_\infty)_{\theta,p;K} = L_p, \quad \theta = 1 - \frac{1}{p}$$

and moreover for $1 < p < \infty$, the noncommutative L_p is an interpolation K space for the Banach couple $\{L_1, L_\infty\}$, i.e. there exists a functional Φ such that $\|T\|_{L_p} = \Phi[K(t, T)]$.

These results on interpolation and non-commutative integration and their further improvements and applications have intimate connection to inequalities of Hardy, Hölder, Minkowski, Hausdorff-Young, Carson and other famous inequalities and open many interesting opportunities for their improvements and generalizations [4, 27, 29].

The expression relating E-functional and decreasing rearrangement of operators is important tool in interpolation theory, its application for non-commutative integration as well as in several other directions in analysis, in operator theory and in various applications. One of such directions is investigation of operator and matrix monotone or convex functions including also Löwner theory closely concerned with analytic continuation, Pick functions, generalizations of spline approximations and several other deep topics in analysis and in matrix and operator theory as well as in applications to quantum physics, especially in quantum information theory and quantum computing, and also in engineering control methods in signal processing (for example in MIMO systems). An interesting original contribution in this direction is a paper by Gunnar Sparr [49]. In addition to a new proof of Löwner theorem on integral representation of operator monotone functions, this paper contains several interesting open problems about inclusions and gaps between some new function classes arising in this context and used in the proof, and classes of matrix monotone functions. For matrix monotone functions the problem of description of the gaps between classes of matrix monotone or matrix convex functions on matrices of different dimensions, as well as classes of matrix and operator monotone and convex functions in C^*-algebras context, have been addressed in [34–36, 52]. In particular, in [15] it was proved by presenting explicit function that all the gaps between classes of these functions for different sizes of matrices are non-empty, which is a solution of a problem of proving this fact that remained open for several decades. In [35] it was shown that there are infinitely many functions in the gaps and that many such functions can be constructed using solutions of truncated moment problem. That abundance and the explicit way of construction of such functions in the gaps allows to use them in investigation of

the open problem about the gaps between classes of functions defined in [49]. The connections of operator monotone and convex functions with interpolation of functions, Pick functions and completely positive maps, both in context of operators and in context of C^*-algebras were considered in [3], where some open problems on inclusions between interpolation function classes and classes of matrix monotone functions have been solved using in substantial way techniques developed in [49].

Non-commutative integration and non-commutative interpolation is one of the important themes considered in the area of operator algebras, especially von-Neumann algebras, for example in connection to Tomita-Takesaki theory, classification problems and many other important topics in operator algebras and its applications in Non-commutative geometry. There is an increasing interest in operator algebras, Banach algebras, operator spaces and non-commutative geometry in building closer links of non-commmutative integration and non-commutative spaces with methods results and notions developed in interpolation theory. As a gateway to those closely related important developments the following references are highly recommended [1, 4, 8, 10, 12–14, 16, 21–24, 26, 31, 45, 46, 54–56]. In the context of operator algebras, operator spaces, Banach algebras and non-commutative geometry, the works of Peetre and Sparr on interpolation of normed and quasinormed abelian groups and on interpolation and non-commutative integration [41, 42] certainly stand out as important pioneering contributions, yet to be properly discovered and explored.

Acknowledgements This research was supported in part by The Swedish Foundation for International Cooperation in Research and Higher Education (STINT), The Swedish Research Council, The Swedish Royal Academy of Sciences and The Crafoord Foundation. The author also is grateful to Institut Mittag-Leffler for support of his participation in Quantum Information Theory program in Autumn, 2010, which has been beneficial for completion of this work.

References

1. Araki, H., Masuda, T.: Positive cones and L^p-spaces for von Neumann algebras. Publ. Res. Inst. Math. Sci. **18**, 339–411 (1982)
2. Aronszajn, N., Gagliardo, E.: Interpotation spaces and interpolation methods. Ann. Mat. Pura Appl. **68**, 51–118 (1965)
3. Ameur, Y., Kaijser, S., Silvestrov, S.: Interpolation classes and matrix monotone functions. J. Oper. Theory **57**(2), 409–427 (2007)
4. Boivin, P., Renault, J.: A Hausdorff-Young inequality for measured groupoids, Von Neumann algebras in Sibiu. Theta Ser. Adv. Math. **10**, Theta, Bucharest, 9–19 (2008)
5. Bergh, J., Löfstrom, J.: Interpolation Spaces. Springer, New York (1976)
6. Brudnyi, Yu.A., Krugljak, N.Ya.: Interpolation Functors and Interpolation Spaces. I, North-Holland, Amsterdam (1991)
7. Butzer, P.L., Berens, H.: Semi-groups of operators and approximation. Springer, Berlin (1967)
8. Connes, A.: Noncommutative geometry. xiv+661 pp., Academic, San Diego, CA (1994)
9. Cwikel, M., Janson, S.: Real and complex interpolation methods for finite and infinite families of Banach spaces. Adv. Math. **66**(3), 234–290 (1987)

10. Effros, E., Ruan, Z-J.: $\mathcal{OL}_p$ spaces, Operator algebras and operator theory (Shanghai, 1997). Contemp. Math. **228**, Amer. Math. Soc., Providence, RI, 51–77 (1998)
11. Gohberg, J.C., Krein, M.G.: Introduction to the theory of nonselfadjoint operators. Izd. Nauka, Moscow (1965) (Russian)
12. Haagerup, U.: L_p-spaces associated with an arbirary von Neumann algebra. In: Algebres d'operateurs et leurs applications en Physicue Mathematique, Edition CNRS, pp. 175–185 (1979)
13. Haagerup, U., Rosenthal, H.P., Sukochev, F.A.: Banach embedding properties of non-commutative L^p-spaces. Mem. Am. Math. Soc. **163**(776), vi+68 pp. (2003)
14. Haagerup, U., Junge, M., Xu, Q.: A reduction method for noncommutative L_p-spaces and applications. Trans. Am. Math. Soc. **362**(4), 2125–2165 (2010)
15. Hansen, F., Ji, G., Tomiyama, J.: Gaps between classes of matrix monotone functions. Bull. London Math. Soc. **36**(1), 53–58 (2004)
16. Izumi, H.: Constructions of non-commutative L^p-spaces with a complex parameter arising from modular actions. Int. J. Math. **8**(8), 1029–1066 (1997)
17. Janson, S.: Minimal and maximal methods of interpolation. J. Funct. Anal. **44**(1), 50–73 (1981)
18. Janson, S.: On interpolation of multilinear operators. In: Function spaces and applications (Lund, 1986). Lecture Notes in Math., 1302, pp. 290–302, Springer, Berlin (1988)
19. Janson, S.: Interpolation of subcouples and quotient couples. Ark. Mat. **31**(2), 307–338 (1993)
20. Janson, S., Nilsson, P., Peetre, J.: Notes on Wolff's note on interpolation spaces. With an appendix by Misha Zafran, Proc. London Math. Soc., (3) **48**(2), 283–299 (1984)
21. Junge, M., C. Le Merdy, Xu, Q.: H^∞-functional calculus and square functions on noncommutative L^p-spaces. Astérisque **305**, vi+138 pp. (2006)
22. Junge, M., Musat, M.: A noncommutative version of the John-Nirenberg theorem. Trans. Am. Math. Soc. **359**(1), 115–142 (2007)
23. Junge, M., Parcet, J.: Rosenthal's theorem for subspaces of noncommutative L_p. Duke Math. J. **141**(1), 75–122 (2008)
24. Junge, M., Parcet, J.: Mixed-norm inequalities and operator space L_p embedding theory. Mem. Am. Math. Soc. **203**(953), vi+155 pp. (2010)
25. Kaijser, S., Pelletier, J.W.: Interpolation functors and duality. Lecture Notes in Mathematics, 1208. iv+167 pp. Springer, Berlin (1986)
26. Kosaki, H.: Application of the complex interpolation method to a von Neumann algebra: Noncommutative L^p-spaces. J. Funct. Anal. **56**, 29–78 (1984)
27. Krugljak, N., Maligranda, L., Persson, L.-E.: The failure of the Hardy inequality and interpolation of intersections. Ark. Mat. **37**(2), 323–344 (1999)
28. Kunze, R.A.: L_p Fourier transforms on locally compact unimodular groups. Trans. Amer. Math. **89**(2), 519–540 (1958)
29. Larsson, L., Maligranda, L., Pecaric, J., Persson, L.-E.: Multiplicative inequalities of Carlson type and interpolation. xiv+201 pp., World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ (2006)
30. Lavine, R.: The Weyl transform. Fourier analysis of operators in L_2 spaces, thesis, M.I.T. (1965)
31. Musat, M.: Interpolation between non-commutative BMO and non-commutative L_p-spaces. J. Funct. Anal. **202**(1), 195–225 (2003)
32. Oloff, R.: p-Normideale von Operatoren in Banaehräumen. Wiss. Z. Friedrich-Schiller- Univ. Jena/Thüringen, **18**, 259–262 (1969)
33. Oloff, R.: Interpolation zwischen den Klassen $\mathfrak{S}_p$ von Operatoren in Hitberträumen. Math. Nachr. **46**, 209–218 (1970)
34. Osaka, H., Silvestrov, S., Tomiyama, J.: Matrix operator functions on C^*-algebras. Int. J. Math. **16**(2), 181–196 (2005)
35. Osaka, H., Silvestrov, S., Tomiyama, J.: Monotone operator functions, gaps and power moment problem. Math. Scand. **100**(1), 161–183 (2007)
36. Osaka, H., Tomiyama, J.: Double piling structure of matrix monotone functions and of matrix convex functions. Linear Algebra Appl. **431**(10), 1825–1832 (2009)

37. Peetre, J.: Approximation of linear operators. Proceedings of the International Conference on Constructive Function Theory, Varna, May 19–25, 1970, pp. 245–263. Publishing House of the Bulgarian Academy of Sciences, Sofia (1972)
38. Peetre, J.: Non-commutative interpolation, Matematiche (Catania), **25**, (1970), 159–173 (1971)
39. Peetre, J.: Zur Interpolation von Operatorenräumen (German). Arch. Math. (Basel) **21**, 601–608 (1970)
40. Peetre, J.: Interpolation functors and Banach couples, Actes du Congrès International des Mathématiciens (Nice, 1970), pp. 373–378. Tome 2, Gauthier-Villars, Paris (1971)
41. Peetre, J., Sparr, G.: Interpolation of normed Abelian groups. Ann. Mat. Pura Appl. **92**, 217–262 (1972)
42. Peetre, J., Sparr, G.: Interpolation and non-commutative integration. Ann. Mat. Pura Appl. **104**(4) 187–207 (1975)
43. Pietsch, A.: Ideale von Operatoren in Banachräumen. Mitt. Math. Gesselsch. D.D.R. **1**, 1–14 (1968)
44. Pietsch, A.: Ideale von S_p-Operatoren in Banaehräumen. Studia Math. **38**, 59–69 (1970)
45. Pisier, G.: Non-commutative vector valued L^p-spaces and completely p-summing maps, Asterisque, vol. 247, Societé Mathématique de France (1998)
46. Pisier, G., Xu, Q.: Non-commutative L^p-spaces. In: Handbook of the geometry of Banach spaces, Vol. 2, pp. 1459–1517. North-Holland, Amsterdam (2003)
47. Segal, I.E.: A non-commutative extension of abstract integration. Ann. Math. **57**(2), 401–457 (1953)
48. Sparr, G.: Interpolation of weighted L_p-spaces. Stud. Math. **62**(3), 229–271 (1978)
49. Sparr, G.: A new proof of Löwner's theorem on monotone matrix functions. Math. Scand. **47**(2), 266–274 (1980)
50. Stinespring, W.: Integration theorems for gages and duality for unimodular groups. Trans. Amer. Math. Soc. **90**, 15–56 (1959)
51. Streater, R.F.: Interpolating norms for orthogonal and spin Lie algebras. Symposia Mathematica, Vol. XIV (Convegno di Geometria Simplettica e Fisica Mathematica, INDAM, Rome, 1973), pp. 173–179. Academic, London (1974)
52. Silvestrov, S.D., Tomiyama, J.: Matrix monotone functions on C^*-algebras. Trends Math. **6**(2), 125–127 (2003)
53. Stečkin, S.B.: Best approximation of linear operators (Russian). Mat. Zametki **1**, 137–148 (1967)
54. Sukochev, F.A., Xu, Q.: Embedding of non-commutative L^p-spaces: $p < 1$. Arch. Math. **80**(2), 151–164 (2003)
55. Takesaki, M.: Duality and von Neumann algebras. Lectures on operator algebras; Tulane Univ. Ring and Operator Theory Year, 1970–1971, vol. 2; (dedicated to the memory of David M. Topping), pp. 665–786. Lecture Notes in Math., vol. 247. Springer, Berlin (1972)
56. Takesaki, M.: Theory of operator algebras, II, Encyclopaedia of Mathematical Sciences, 125, Operator Algebras and Non-commutative Geometry, 6. xxii+518 pp. Springer, Berlin (2003)
57. Triebel, H.: Über die Verteilung der Approximationszahlen kompakter Operatoren in Sobolev-Besov-Räumen. Inv. Math. **4**, 275–293 (1967)
58. Triebel, H.: Interpotationseigenschaften von Entropie- und Durchmesseridealen kompakter Operatoren. Stud. Math. **34**, 89–107 (1970)
59. Triebel, H.: Interpolation theory, function spaces, differential operators, Second edition. p. 532, Johann Ambrosius Barth, Heidelberg (1995)

Index

K. Åström et al. (eds.), *Analysis for Science, Engineering and Beyond*,
Springer Proceedings in Mathematics 6, DOI 10.1007/978-3-642-20236-0,